个人数据银行——一种基于银行架构的个人大数据资产管理与增值服务的新模式

郭 兵[1)] 李 强[1)] 段旭良[1)] 申云成[1)] 董祥千[1)]
张 洪[1)] 沈 艳[2)] 张泽良[3)] 罗 键[4)]

[1)](四川大学计算机学院 成都 610065)
[2)](成都信息工程大学控制工程学院 成都 610054)
[3)](成都数银科技有限公司 成都 610041)
[4)](西南财经大学工商管理学院 成都 611130)

摘 要 随着移动互联网的不断深入发展,个人大数据呈现指数级增长,却面临着产权模糊、管理散乱和流通困难等问题,严重影响个人大数据市场的有序发展.文中基于银行个人货币资产管理的模式及架构,以保护个人数据产权、知情权、隐私权和收益权为核心,提出了一种个人大数据资产管理与增值服务系统——个人数据银行,包括数据确权、汇聚、管理、交易和增值服务等功能,以个人为主体对象组织数据,有效连接人与数据,使得个人数据可授权访问且有序流通.文中对个人大数据资产管理与增值服务平台面临的问题进行了讨论,明确了个人数据银行定义,探讨了个人数据银行平台的组成架构和关键技术,并研发了一种个人数据银行平台——数汇宝,从理论和实践两个方面分析了个人数据银行建设的可行性,为个人数据资产化管理奠定了基础.

关键词 个人大数据;个人数据管理;个人信息管理;个人数据资产管理;个人数据银行
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2017.00126

Personal Data Bank: A New Mode of Personal Big Data Asset Management and Value-Added Services Based on Bank Architecture

GUO Bing[1)] LI Qiang[1)] DUAN Xu-Liang[1)] SHEN Yun-Cheng[1)] DONG Xiang-Qian[1)]
ZHANG Hong[1)] SHEN Yan[2)] ZHANG Ze-Liang[3)] LUO Jian[4)]

[1)](*College of Computer Science, Sichuan University, Chengdu* 610065)
[2)](*Control Engineering College, Chengdu University of Information Technology, Chengdu* 610054)
[3)](*Chengdu Databank Technologies Co., Ltd., Chengdu* 610041)
[4)](*School of Business Administration, Southwestern University of Finance and Economics, Chengdu* 611130)

Abstract With the further development of mobile Internet, personal big data as a valuable asset is growing exponentially. But the vague of ownership, chaos of management and difficulty of circulation severely affect the regular development of personal big data market. Personal data bank is a new mode of personal big data asset management and value-added service based on the monetary asset management architecture of current banks, and its core principle is to protect the

收稿日期:2016-03-08;在线出版日期:2016-08-09.本课题得到国家自然科学基金(61332001,61272104,61472050)、四川省科技计划项目(2014JY0257,2015GZ0103,2014-HM01-00326-SF)资助.**郭 兵**,男,1970 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为绿色计算、个人大数据. E-mail: guobing@scu.edu.cn. **李 强**,男,1963 年生,博士,副教授,主要研究方向为移动云计算、个人大数据. **段旭良**,男,1982 年生,博士研究生,讲师,主要研究方向为个人大数据、大数据清洗和数据挖掘. **申云成**,男,1979 年生,博士研究生,讲师,主要研究方向为个人大数据、大数据定价. **董祥千**,男,1975 年生,博士研究生,讲师,主要研究方向为个人大数据、信息安全和个人隐私保护. **张 洪**,男,1980 年生,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究方向为移动互联网数据溯源、个人大数据. **沈 艳**(通信作者),女,1973 年生,博士,教授,主要研究领域为智能终端、智能仪器. E-mail: shenyan02@163.com. **张泽良**,男,1967 年生,高级工程师,主要研究方向为软件构架、个人大数据. **罗 键**,男,1969 年生,博士,副教授,主要研究方向为企业管理、人力资源管理.

ownership, right to know, privacy and usufruct of personal data. Personal data bank organizes all data centered with individuals, and provides some key functions of ownership verification, aggregation, management, trading and value-added services for personal big data, which can effectively guarantee all data access authorized and all data circulation orderly. This article mainly discusses the system architecture and key technologies of personal data bank, introduces an implement of personal data bank platform "ShuHuiBao", analyzes the feasibility of personal data bank from the perspective of theoretical and practical aspects, and lays a strong foundation for future development of personal data capitalization.

Keywords　personal big data; personal data management; personal information management; personal data asset management; personal data bank

1　引　言

从 2011 年提出大数据概念以来，大数据几乎应用到人们生活、工作的各个方面，无论是对社会、公司还是个人来说，都是一次世界观的改变，而且数据和土地、劳动力、资金一样成为国家的重要经济资源，作为一种商业资本和重要的经济投入，于信息社会就如燃料之于工业革命，可以创造出新的经济效益，激发新产品和新型服务，影响着国家和社会经济的安全、稳定与发展. 大数据产业本身具有巨大的商业价值和社会价值，国际数据公司(IDC)将大数据作为 4 个支柱研究领域之一，预计到 2018 年大数据技术和服务市场总额将达 415 亿美元①，大数据将成为全球下一个促发创新、角力竞争和提高生产力的前沿领域[1].

个人数据是移动互联网时代呈指数级增长的资产，我们每个人每天都会产生大量工作和生活数据，一个人一生产生的数据量上限可达到 PB 级，且类型繁多. 目前，国内外学术界和工业界许多研究关注的大数据主要是企/事业大数据、行业大数据、政府大数据，甚至是城市大数据和国家大数据，而与个人紧密相关、广泛存在而又经常使用的、以个人为数据对象主体的数据——个人大数据，往往被忽视，缺乏相应的研究成果、处理平台和分析工具[2].

1998 年英国颁布《数据保护法》，比较全面的界定了个人数据的内涵和外延，认为个人数据指"一个活着的自然人的数据集合，通过这些数据，或者这些数据和使用者占有的其他信息的组合可以辨识该人"，另外，有关该人观点的表述及涉及到该人时数据使用者或其他人的意图也属于个人数据范畴，去掉个人识别信息后个人数据将失去原有价值. 个人数据是个人与周围环境交互过程中产生的数据，是个人互联网足迹、个人电子历史(e-history)和个人习惯的表达，具有高分散、高多样、高相关、高隐私等特征，同时个人数据具有典型的大数据 4V(Volume, Variety, Velocity, Veracity)特征，对个人、社会都具有很重要的价值[3].

因此，个人大数据是个人生活和工作中产生的、个人可以拥有或控制的数据，主要指个人生活活动中产生的原始数据，其数据来源复杂，数据形式多样，包括个人基础信息以及个人收支、个人财产、个人朋友圈、个人健康、个人教育、个人日志、个人文档、个人观点、个人感知数据等信息，是一类重要的大数据.

2011 年，世界经济论坛(World Economic Forum)启动一项名为"重新思考个人数据(Rethinking Personal Data)"的项目，旨在汇集个人、公司、公共部门、隐私保护机构、学术机构等一批利益相关者，探索如何形成一个平衡、互相协作、自律的基于个人数据的生态系统，并发布《个人数据——一种新资产的崛起》报告，将个人数据作为"最新的经济资源"，列为"新的资产类别"[4].

从数据使用和价值角度看，尽管单个人的数据价值量相对有限，但大量个人的数据累加，量变就会引起质变，个人大数据的分析和使用将对许多行业的商业模式带来革命性的影响. 国际互联网行业巨头 Google、FaceBook、Twitter 以及国内三家巨头企业 BAT(百度、阿里和腾讯)的重要业务都是"To C (Consumer 消费者个人)"模式，以个人数据为核心开展相应的搜索、电子商务或社交网络服务. 2014 年 10 月，硅谷精神教父、科技商业预言家凯文·凯利在

① BigData. http://www.idc.com/prodserv/4Pillars/bigdata/, 2015-4-24

斯坦福大学演讲，预言未来 20 年科技潮流，明确指出“个人数据才是大未来”①.

当前，个人大数据管理中存在的主要问题包括：

(1) 数据产权模糊. 目前行业潜规则是“谁采集，谁拥有”，出售和利用个人数据获利，侵犯用户数据产权、知情权、隐私权和收益权的现象时有发生. 对个人大数据的采集、处理和使用笼罩在企业商业秘密之下[5]，由于数据产权模糊带来用户权益保护困难，这是许多现有大数据应用中的一个重要缺陷.

(2) 数据管理散乱. 众多线上/线下企业通过各种业务平台采集和管理个人数据，如电商平台对用户商品交易和支付信息的采集，银行、第三方支付平台等对用户金融交易信息的采集，社交通信软件、搜索引擎、安全杀毒、浏览器、输入法等对用户信息资料、个人习惯、操作信息的采集，导致个人数据分散其中，带来个人数据的存储碎片化和管理复杂化，形成个人信息孤岛，甚至一些业务平台成为“数据黑盒”，用户的个人隐私和收益更难以保障.

(3) 数据开放、流通困难. 对企业和政府部门而言，由于个人数据采集和管理的散乱现状，严重影响个人数据的流通、分享使用和依法监管，汇聚线上/线下等多维度的个人数据非常困难，导致个人征信、互联网金融服务、产品精准营销等新的增值服务难以实现，个人数据的经济价值和社会价值也难以发挥. 当前的大数据价值实现，基本是因循旧的数据价值增值模式，实现大数据价值的充分发挥，需要的不仅仅是更好的算法、更快的计算速度，还需要有创新的数据使用和流通模式[6].

目前，中国经济进入“新常态”，大数据作为经济增长的新动力，迫切需要探索一种新的个人大数据采集、管理和使用模式，规范个人数据的增值服务和交易行为，保护用户的隐私和数据安全，保障用户数据的合法权益，不断推动个人大数据领域相关技术和应用的发展.

本文以个人大数据资产管理与增值服务平台研究和设计为主线，描述了个人大数据的发展与现状、个人数据银行定义、系统构架、关键技术和系统实现，其中涉及到计算机科学和管理学交叉融合的一些研究成果.

2 个人大数据的发展与现状

2.1 产业界的发展与现状

在个人大数据产业链中，从数据采集、存储、管理、分析、交易到应用服务，较困难的地方是个人数据的收集、管理和流通环节，总体呈现数据产权模糊、数据管理散乱和数据开放、流通困难等现状.

国内外互联网服务公司大都已经意识到个人数据处理和应用带来的潜在价值，多着眼于个人信息管理(Personal Information Management，PIM)的某一个具体方面，如个人邮件、理财信息和笔记等. 在发达国家，个人管理信息化程度非常高，家庭和个人已经实现了手机、PC、服务器的个人数据同步处理，我国目前大多只局限于手机、PC、服务器的单点应用，数据记录、跟踪和管理功能单一化，数据集成度低，一体化个人管理信息平台很难实现. 为了管理个人数据，用户不得不使用多种硬件设备和不同功能的 APP 应用平台，操作复杂且管理多组密码，不仅增加使用成本，也大大降低了用户管理和使用个人信息的效率.

有关个人数据存储与管理的一些典型项目包括：

(1) 美国国会图书馆发起“国家数字信息基础结构保存项目(National Digital Information Infrastructure and Preservation Program)”，其中个人数字存档计划(Personal Digital Archiving Campaign)是重要的组成部分②. 个人数字存档注重收集、整理、组织、保存、检索和利用归属于个人的原生或再生电子文件，是近年来国外的研究热点；英国大英图书馆通过“数字生活探索工程(The Digital Lives Research Project)”，研究个人数据存档、收藏的缺陷.

(2) 中国的人事档案涉到组织部门、人事部门、劳动部门、教育部门和综合档案部门等多个组织，多头管理下的利益冲突不可避免. 现行人事档案缺乏正常合理的交流利用渠道，与就业、医保、社保等社会管理职能逐步脱离关系，具有的“政治管理性档案”特点也已经与时代脱节，解除一定的保密性，档案的信用化是人事制度改革的大趋势，可和个人信用库共享、共建，将知情权、监督权交还公众. 未来的发展趋势是“单位人”开始向“社会人”转变，建立公民档案中心，让档案管理回归社会，并建立个人电子档案库.

关于个人数据管理与交易的一些典型服务平台包括：

(1) 许多行业性的业务运营与服务公司，往往

① 凯文·凯利 6000 字未来商业预言. http://news.ittime.com.cn/news/news_2418.shtml，2015，11，9

② 国家数字信息基础结构保存项目. http://www.loc.gov 2015，11，9

涉及对某类个人数据的采集、分析与应用，如中国移动采集的个人通信消费数据、中国电力采集的个人用电消费数据和淘宝平台采集的个人电商消费数据等，2010 年 3 月 30 日，阿里巴巴开放淘宝平台所有的个人电商交易数据，将该计划命名为“数据魔方”①.

(2) 2013 年 1 月 30 日，日本富士通公司建立数据交易市场 Data plaza，目前在 Data plaza 市场上买卖的数据包括购物网站上的购物记录、出租车传感器获得的交通堵塞记录、智能手机的位置信息、社交网站(SNS)帖子等；美国 DataCoup 和 Personal 公司，对个人银行交易、社交等数据进行价值评估，按固定周期与用户分享数据交易的收益.

(3) 2014 年 2 月，国内组建首个面向数据交易产业组织——中关村大数据交易产业联盟，通过开放的 API 进行数据的录入、检索和调用，以期为政府机关、科研单位、企业组织和个人提供数据交易和使用场所，2014 年 6 月，中关村大数据交易产业联盟发布了我国首个大数据交易行业规范《中关村数海大数据交易平台规则》；2011 年成立的数据堂(北京)科技股份有限公司，提供个人、企业和行业等各类数据的交易平台.

(4) 华为网盘(数据银行)、百度文库平台提供文档等非结构化数据的存储和托管服务.

实际上，智能手机等终端上的许多 APP 软件，本质上都是个人数据的采集与管理软件，如挖财理财软件、个人健身健康数据采集平台 iHealth 等. 总体来看，当前个人数据的采集和管理主要由服务企业依托业务平台，在未明确用户个人数据产权等权益的情况下采集的，普遍存在数据“谁采集、谁拥有”的潜规则，并进一步依靠采集的个人数据，通过提供增值服务扩展企业的盈利途径.

2.2　学术界的发展与现状

在个人大数据研究中，个人大数据作为一种资源和资产，如何对个人数据进行有效的管理和挖掘其潜在价值是个人大数据领域一直研究的课题.

迅猛增长的数据量使管理个人信息的负担日益加重. IDC 2012 年在美国调查发现，从事信息相关工作的人员每个月大约花费 20 小时在 PIM 上. 随着移动互联网、智能设备、穿戴设备、大数据等技术的发展，大多数个人信息被保存于各类电子设备中，个人信息管理支持更好的对工作和生活信息的存储、组织和检索，以达到信息重用目的. 当前，用户信息空间(Personal Space of Information，PSI)主要研究如何将产生的个人数据及时采集、保存、分析、按需快速检索等，可视为 PIM 的理想方式.

目前公认 PIM 思想的最早提出者是美国科学家 Vannevar Bush，他在 1945 年发表《As we may think》一文，构想一种机器 Memex 能够帮助人们记忆和管理个人信息. 2001 年，美国微软公司旧金山研究院的首席计算机科学家 Gordon Bell 决定尝试启动 MyLifeBits 项目，制造世界上第一台 Memex，通过照相机和录音设备实现在一台机器中储存一个人一生中所有的信息，包括全文检索、文字及影音说明和超文本链接，该项目持续七年，可以精确到每一分钟，再现 Bell 过去七年中的每一天.

从根本上来说，PIM 是一系列操作行为的集合，在信息的生成、保存、利用和用户需求之间建立映射关系，主要功能包括多样化个人信息的统一存储管理；PSI 中共性的处理服务，如标签服务；建立在 PSI 之上的涉及个人信息管理的其它功能，如文件、日程管理等. 文献[7]介绍个人信息管理的发展历程、相关定义及其概念框架，分析当前个人信息管理过程中所存在的主要问题，探讨优化个人信息管理的若干对策；文献[8]介绍个人信息管理系统的协议基础——SyncML 协议，个人信息管理系统的设计和实现，并设计新的安全模型，使得协议更加安全可靠；文献[9]提出基于 WSRF 的个人信息空间管理系统的分层体系结构，探讨本地及网络数据资源统一封装技术，并基于封装数据为用户提供个人信息管理服务；文献[10]提出一种基于云计算的个人信息管理系统设计方法，利用云的特点和优势，结合数据库同步和网络信息的抓取，充分利用网络资源，解决信息的封闭性问题，增进信息的流通.

从内容上来看，PIM 倾向于对人们直观上认为有意义的“信息”进行管理，目的是采用各种新技术、新方法通过对个人数据更好的组织、存储实现高效的查找与利用，并没有对用户数据进行分析加工，没有产生新的信息. 随着数据处理效率的不断提升和存储成本的降低，更多类型的数据纳入到个人信息管理范畴，原来直观上认为价值不大的一些数据通过分析挖掘也可反映出用户的习惯或其它特征，得到越来越多的重视和应用，个人数据管理概念和方法逐步发展起来. 个人信息管理和个人数据管理并没有明确的界限，一般认为，个人数据管理是个人信息管理的进一步发展和拓展，本质上是管理大规模、异构、分布式复杂数据的理论与技术在个人数据管

① 淘宝网. https://www.taobao.com 2015,11,9

理领域的应用,涉及数据存储、索引、查询、分析、挖掘、安全与隐私保护等诸多问题,目的是建立用户的个人图谱,提高用户对个人数据的管理和利用效率.

个人数据管理的目标是针对个人数据,在Internet环境下构建一个分布式、集成化的个人数据空间,是一个包含主体、信息、工具的人机交互系统,在保护个人隐私的前提下,实现个人信息的分类、组织、存储、索引、检索、共享和使用,提高个人工作和生活的质量与效率.

麻省理工学院媒体实验室 Sandy Pentland 教授2010年提出人类动力学(Human Dynamics)①,通过收集和分析海量个人数据,揭示人性,甚至预测人们的行为,称为现实挖掘(Reality Mining)[11-12].同时,他鼓励建立个人数据商店(Personal Data Store)机制,鼓励人们贡献和分享数据,使自己、他人甚至整个社会从中受益.

数据空间(Data space)是与主体相关的数据及其关系的集合,是管理个人数据的一个新理念.数据空间中数据的主体特征将个人数据管理和传统企业数据管理区分开来,主体相关性和可控性是数据项的基本属性,所有数据对于主体来说都是可控的.一般意义上的数据空间是指主体数据空间,相对应的还有公共数据空间.主体数据空间是公共数据空间的子集,在主体需求的变化过程中,数据项也随之从公共数据空间转入到主体数据空间中.文献[13]从技术方面对个人数据管理与企业数据管理进行比较,提出以主体为中心的个人数据管理系统框架,在数据模型、数据存储、数据查询检索、安全隐私保护、评价技术、系统实现等方面,分析已有研究工作,提出基于图的个人数据空间模型、自适应的个人数据集成方法、多级的数据存储策略和基于用户记忆规律的数据查询处理策略;文献[14]在PIM及数据空间的基础之上,采用Web领域广泛应用的标签(Tag)技术组织个人数据,并采用了P2P(Peer-to-Peer)的方式使个人数据得以有效发布及共享.

2005年,Dong等人发表了一系列文章介绍数据管理最新研究成果,文献[15]从数据管理视角研究个人信息管理与信息集成,提出构建PIM关键是为用户提供个人信息的逻辑视图,表现出数据及其关联的直观的语义特征;文献[16]介绍了个人数据管理与集成平台SEMEX构架及功能结构,SEMEX通过在各类型数据中抽取有意义的对象和关联实现个人信息的单一逻辑视图.

个人数据从单纯的管理,逐渐发展到分析、挖掘和利用.文献[17]指出,大数据时代,数据不再是简单的处理对象,而是一种重要的基础资源,更好的管理和利用大数据是当前热点研究问题.大数据之"大",在数据存储、管理及分析应用上带来极大的挑战,数据管理理念、管理方式上正在酝酿发生着变革;文献[18]提出一种"以用户为中心"模型,个人数据被假定为数字世界的"能量"或"新货币",旨在通过使个人能控制与自己有关数据的收集、管理、共享和使用,特别值得关注的是,文中分析了一种新的个人数据生态系统,围绕"个人数据银行"(Bank of Individuals' Data,BID)进行搭建.

随着个人数据管理范围越来越广、数据量越来越大和组织管理手段多样化,人们对数据管理的关注点也从简单的数据存储、检索,逐步发展到数据的使用、交流和共享.在个人数据使用中,个人隐私保护往往是第一位的.个人隐私一般是指对个人敏感且不愿公开的信息,通常有信息隐私、通信隐私、空间隐私以及身体隐私等几类.对于隐私的定义主要有两类观点:一是基于价值,将隐私视为一种人权和商品;二是基于同源,认为隐私是个人的思想、认识、感知和状态.隐私会随着生活经验而改变,也依赖于特定场景,是动态的、多维的,很难给出通用概念.

Web2.0技术将个人有效转变为数据的生产者,通过博客、微博、社交网络等新兴服务和各类移动平台产生了类型繁多、体量巨大的个人大数据,数据关联分析与应用带来的隐私问题不容忽视.文献[19]从数据层、应用层、表示层阐述个人隐私保护相关技术,并从法律法规和行业规范层面探讨个人隐私的保护问题;文献[20]指出,信息化和网络化的高速发展使得大数据成为当前学术界和工业界的研究热点,在提高经济和社会效益的同时,也为个人和团体的隐私保护以及数据安全带来极大风险与挑战.当前,隐私成为个人大数据应用领域亟待突破的重要问题,该文描述个人大数据的分类、隐私特征与隐私类别,分析大数据管理中存在的隐私风险和隐私管理关键技术,提出大数据隐私主动式管理建议框架以及该框架下隐私管理技术的主要研究内容,并指出相应的技术挑战.

在具体司法实践中,各国对个人隐私的范围有明确的界定,如中国征信业管理条例规定中,禁止征信机构采集个人的收入、存款、商业保险、有价证券、纳税数额、不动产、宗教信仰、血型、指纹、基因、疾病

① Human Dynamics. http://hd.media.mit.edu 2015,11,9

以及病史信息.但是,个人同意的除外,以及依照法律、行政法规规定公开的不良信息除外.

在个人隐私保护和个人数据共享使用方面,世界各国都主张采取二者均衡的基本政策,即在一定个人隐私保护基础上,用户授权数据的正常开放、分享和使用,但不允许数据乱用、滥用.

综上,当前个人大数据的研究侧重在个人数据的管理、使用与隐私保护方面,而个人数据的资产化管理方法、价值评估、共享流通等理论和技术方面尚缺少相应研究成果.

3　个人数据银行定义

3.1　个人数据银行的定义及目标

文学作品、专利、软件、硬件 IP 核、电路版图等人类高智力成果都已实现资产化,包括所有权清晰、货币可计量价值和给所有者带来经济利益,并可以用于企业入股和清偿债务等.个人数据与专利等高智力成果相比,二者都和个人相关,个人大数据价值密度低、分布范围广、数量大,并与个人隐私关联度高,是人类生产、生活等活动的伴生物,部分非有意识创造,是一种个人的低智力成果.

从 IT 到 DT(Data Technology)时代,谈论最多的资产是数据,数据产权是数据经济的基石,数据经济从宏观上说是以数据为核心生产要素的产业模式和经济形态,在微观上则是数据驱动的企业业务流程,数据产权模糊将直接导致数据市场无序及数据产业模式的畸形发展.

中国工程院院士邬贺铨曾指出,目前我国一些部门和机构拥有大量数据,但宁愿自己不用也不愿提供给有关部门和机构间分享,导致信息不完整或重复投资①.究其原因,主要是数据产权模糊,用户权益不能得到有效保证,不愿将数据将进行开放共享.

个人大数据作为一种资产,无法回避使用和流通中所涉及的所有权、交易规则和定价问题,这些问题国内外都还没有一个合理的解决方案.其中所有权问题涉及到技术、商业伦理、法律等诸多方面的问题,关系到个人、服务企业和政府等关联方权益与责任的平衡,借用《著作权法》等法规,一般认为个人数据产权归属个人更为合适.

由于个人数据资产与货币资产本质上具有共同点,个人数据是个人财产的一部分,就像在银行里的存款一样,因此,简而言之,个人数据资产能够采用银行模式进行管理和运营,既可以实现个人数据的集中有效管理,又可以实现个人数据的增值和有序流通,给个人带来一定收益,即个人数据银行.

个人数据银行是基于银行个人货币资产的管理与运营模式,以保护用户个人数据的所有权、知情权、隐私权和收益权为核心,建立个人大数据资产的管理与运营综合服务系统,包括数据确权、汇聚、管理、交易与增值服务等功能.

将银行发展过程与个人数据银行发展过程对比如表 1 所指示,可进一步明确个人数据银行的定位.

表 1　银行与个人数据银行发展过程的比较

银行发展过程	个人数据银行发展过程
(1) 个人钱币自我保存.古代社会经济整体不发达,个人财富有限,个人钱币大多自我保存.	(1) 个人数据的自我保存.PC 时代或以前,将个人数据以纸张、磁带、硬盘等单一孤立的存储形式保存.
(2) 钱庄.银号、钱铺,明朝中期诞生,主要功能是汇兑和存放款.	(2) 个人数据存储与管理平台.云计算时代,主要功能是以云盘或网盘等形式,个人数据存储可靠、管理方便,部分平台有交易功能.
(3) 现代银行.近代诞生,主要功能是货币资产的集中托管和专业化的综合运营,实现安全、便捷、收益和普惠的金融服务,是货币资产的管理者与经营者.	(3) 个人数据银行.大数据时代,主要功能是个人数据资产的集中托管和专业化的综合运营,实现安全、方便、收益和普惠的数据服务,是个人数据资产的管理者与经营者.
(4) 互联网银行.依靠互联网技术,在线为客户提供全方位无缝、快捷、安全和高效服务的金融机构.	(4) 互联网银行.互联网金融不能被神话,关键是数据与资金的整合,数据银行未来是现代银行功能的外延性发展,二者将趋于融合.

综上,与现有的数据集市、数据交易所等模式相比,个人数据银行是尝试解决个人数据流通问题的一种新模式,具有更多的优势和可行性,实现的目标包括:

(1) 在个人数据产权和隐私保护的前提下,所有权和使用权可分离,个人让渡数据的使用权,即产权"换"使用权,为个人数据有序的社会化流通提供重要基础.

(2) 实现个人数据的产权化、资产化、商品化、集中化、服务化、专业化和收益化,使得个人数据量化有用、授权访问且可有序流通,降低个人数据的交易成本,使得数据资源配置更加优化,也利于加强个人隐私保护和国家网络空间信息安全.

3.2　个人数据银行的形式化描述

个人数据银行是一个 8 元组{$U,D,P,C,T,BS,$

① 邬贺铨. 需制定大数据国家战略. http://news.xinhuanet.com/info/2013-12/15/c_132969258.htm

$MS,VS\}$,其中:

(1) $U=\{u_0,u_1,\cdots,u_n\}$.

U 表示用户的非空有限集.

(2) $D=\{d_0,d_1,\cdots,d_n\}$.

D 表示用户数据集的非空有限集,集合中元素表示某个用户某类数据的集合,每一个数据集都唯一的隶属于某个用户,即对于 $\forall d_i\in D$, $\exists u_i\in U$,映射关系 $f:d_i\mapsto u_i$ 成立.

(3) $P=\{p_0,p_1,\cdots,p_n\}$.

P 表示数据产品的非空有限集,数据产品是经过治理之后的用户数据集合,用于交换或交易,是个人数据增值的主要形式.一个数据产品可包含多个用户多类数据,或者是多个用户多类数据融合之后的结果.

令 $PR=\{pr_0,pr_1,\cdots,pr_n\}$ 为产品构建规则(或构建需求)的集合, $P'\subseteq P$, $PR'\subseteq PR$, $p_i\in P$,则每一个数据产品 $p_i=(P',PR')$.

(4) $C=\{c_0,c_1,\cdots,c_n\}$.

C 表示客户集合,是数据交换的对方或者数据交易的买方.

(5) $T=\{t_0,t_1,\cdots,t_n\}$.

T 表示数据产品交易的有限集.集合中的元素表示一次数据交易,每次交易都有一定收益,交易可能涉及一个或者多个数据产品,每次交易的客户也可能有一个或多个.定义 $TT=\{tt_0,tt_1,\cdots,tt_n\}$ 为数据交易方式的非空有限集合,对于 $t_i\in T$, $P'\subseteq P$, $C'\subseteq C$, $tt_i\in TT$,令 $pf\in R^+$ 为此次交易收益,则每次交易可表示为 $t_i=(P',C',tt_i,pf)$.

(6) $BS=\{bs_0,bs_1,\cdots,bs_n\}$.

BS 表示个人数据银行基础服务的非空有限集,主要包括数据确权与溯源服务、隐私保护服务、数据安全服务、数据计量计价服务等.

(7) $MS=\{ms_0,ms_1,\cdots,ms_n\}$.

MS 表示数据管理服务的非空有限集,是数据存储、数据分析、数据表示等几大类服务的集合.

(8) $VS=\{vs_0,vs_1,\cdots,vs_n\}$.

VS 表示增值服务的非空有限集,主要包括数据查询与共享服务、数据使用与交易服务、数据加工与产品服务、数据收益与结算服务等.

4 个人数据银行关键技术

4.1 个人数据银行构架

个人数据银行主要由数据众筹、数据管理、增值业务以及基础服务 4 大模块组成[①],如图 1 所示.其中,数据基础服务是个人数据银行框架中的核心基础模块,通用大数据处理框架中通常没有数据确权、溯源、计量计价.与通用大数据处理框架不同的是,数据的存储、管理、分析、挖掘和交易等是建立在数据确权、溯源及计量计价基础服务之上的,这也是个人数据银行构架明确数据产权、保护数据主体利益的核心所在.

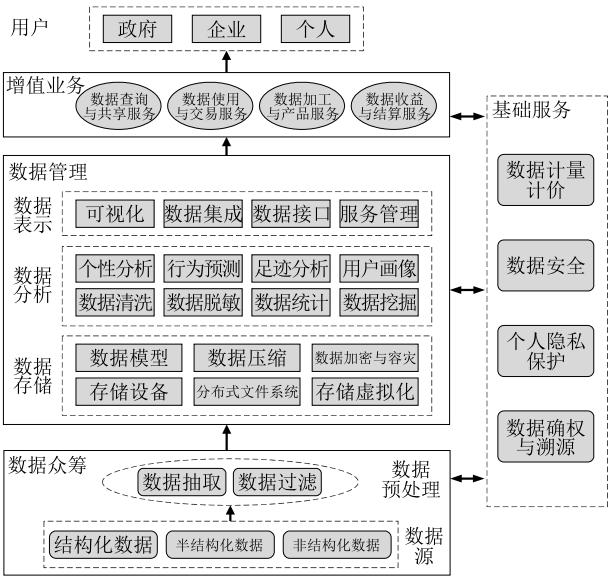


图 1 个人数据银行架构

其中:

(1)数据众筹.具有低门槛、多样性、依靠大众力量、注重创意等特征,但在应用领域、实现方式上与一般众筹模式有许多区别.个人数据银行数据众筹以互联网平台为基础,以多用户、多终端协同数据参与共享为核心,能够进行用户化、网络化、动态化的数据发现、分享与动态反馈.个人是数据的产生者、管理者、使用者和监督者,对个人结构化、半结构化和非结构化数据,采用线上/线下输入、传感器自动采集和第三方数据源交换等方法众筹个人数据,将个人汇聚起来,存放到个人数据银行中,减少数据的重复采集成本.

(2)数据管理.将个人数据分类、集中存储到个人数据银行平台中,对数据开展清洗、分析和挖掘等处理工作,通过数据集成、数据接口等机制,方便用户进行数据访问、管理和使用,为个人生活提供便利服务.

① Big Data Reference Architecture. http://thinkbig.teradata.com/leading_big_data_technologies/big-data-reference-architecture 2015,11,9

(3)增值业务. 对用户个人数据资产进行商业化运营,开展数据产权交易和数据服务交易,交易过程对数据所有者透明,支持数据查询与共享、数据使用与交易、数据加工与产品等增值业务,为个人提供数据收益与结算服务.

(4)基础服务. 为数据众筹、数据管理和增值服务提供与个人数据资产有关的一些共性基础服务,包括隐私保护、数据安全、数据计量与计价和数据确权与溯源等服务.

在个人数据银行中,数据计量与计价、数据收益模型、数据确权与溯源、数据分享与交换机制、个人隐私保护和个人数据的分析与挖掘是特色的关键技术,下面对这些技术做进一步的描述.

4.2　数据计量与计价

数据定价一直以来都是一个世界性难题,个人数据的计量与计价主要解决个人数据资产化的第二个问题.

数据计量的基本单位可以采用数据调用次数、数据包、查询、视图或数据元组,数据计价的方式通常包括两种:

(1)绝对计价. 数据所有者直接标价或者按行业内的规定定价.

(2)相对计价. 根据市场历史交易成功的类似数据价格定价.

数据计量与计价的难点在于如何体现出数据包含的真正价值,用一种相对公平合理又兼具效率的定价模式对数据进行计量与计价,让数据买卖双方都能从数据交易中得到实惠,形成一个良性的数据交易生态链. 文献[21]提出按使用付费定价模型(Pay-per-use Pricing Model)和预订定价模型(Subscription Pricing Model);文献[22]提出免费定价策略、按使用付费定价策略、数据包定价策略、固定费用关税策略、二部收费制策略和免费增值策略;文献[23-24]认为当前定价模式存在4个缺点,即当前定价模型允许套利、模型假设所有数据集是等价的、数据由客户进行缓存以及数据提供者没有如何设定价格的指导性建议,并提出一种细粒度的数据定价策略. 文献[25]提出了数据市场中的视图概念,一个数据实例的视图相当于实例的一个版本,视图也许只包含数据的一个子集或仅仅某些列或粗粒度级别的信息,并根据不同的价格进行销售. 文献[26]提出了一种根据噪音查询应答来进行定价的理论框架,划分数据拥有者之间的价格,根据他们隐私丢失情况获得应有补偿,并指出隐私估值也许与数据本身是强烈相关的.

本文提出一种以数据元组作为数据计量的基本单位,采用正向定级、反向定价的数据计价方式,对原始数据和派生数据进行计量与计价,即根据数据包最终的市场销售价格,结合数据所包含的信息量、价值权重、引用指数等正向价格影响因素,进行动态反向定价,以标定数据的市场价格,作为数据收益分成和未来数据交易的基础.

定义一个四元组集 $p=(q,r,w,s)$描述正向定级、反向定价模型,其中 q 表示数据的信息量,r 表示数据的引用指数,w 表示数据的价值权重,s 表示数据的销售价格. 假设以数据包为基本销售单位,数据包由 n 条数据元组构成,数据包中第 i 条数据元组的价格为 p_i,则计算公式为

$$p_i=s\times d\times\left(\frac{w_i}{w}\times\alpha+\frac{q_i}{q}\times\beta+\frac{r_j}{r}\times\gamma\right),\quad i=1,\cdots,n;\ j=1,\cdots,m \tag{1}$$

其中,s 表示数据包的销售价格,d 表示数据卖方的收益分成比例,n 表示数据包中的数据元组数,w_i 表示第 i 条数据元组的价值权重,w 表示数据包中所有数据元组的价值权重之和,q_i 表示第 i 条数据元组的信息量,q 表示数据包的信息量,r_j 表示第 j 个用户的 r 指数值,r 表示数据包中所有用户的 r 指数值之和,α 为价值权重因子、β 为信息量因子、γ 为 r 指数因子,且满足下面的约束条件:

$$\alpha+\beta+\gamma=1 \tag{2}$$

下面通过一个具体算例进行正确性和有效性验证. 为了简便性,假定有这样一个个人消费数据包,其中包含了2条消费记录,每条记录都属于同一人所有,此数据包的销售价格为10元,即 $s=10$,设定数据卖方的分成比例 $d=0.4$,价值权重因子 $\alpha=0.3$,信息量因子 $\beta=0.3$,r 指数因子,$r=0.4$,$w_1=1$,$w_2=2$,$q_1=1.5$,$q_2=2.5$,$r_1=1$,$r_2=3$,那么由式(1)得到

$$p_1=10\times0.4\times\left(\frac{1}{3}\times0.3+\frac{1.5}{4}\times0.3+\frac{1}{4}\times0.4\right)=1.25,$$

$$p_2=10\times0.4\times\left(\frac{2}{3}\times0.3+\frac{2.5}{4}\times0.3+\frac{3}{4}\times0.4\right)=2.75.$$

从得出的结果可以看出此数据包中第1条记录的价格为1.25,第2条记录的价格为2.75,说明第2条记录比第1条记录价值更大. 从设定的参数来看,第2条记录的价值权重、信息量及引用指数都大于第1条记录,第2条记录应该比第1条记录值钱,

与得出的结论相一致，证明提出的模型是正确有效的.同时，定价模型可动态调整，具体可调整数据包的 4 个参数，即数据卖方的收益分成比例 d、价值权重因子 α、信息量因子 β 和 r 指数因子 γ，计算出较合适的数据元组市场价格，体现公平性和效率.

与此相比，现有计价方式对数据元组的价格通常采用平均定价方法，即以数据包为销售单位，将数据包销售价格平均分配到每一条数据元组，而没有考虑数据元组本身所包含的价值量，没有体现出定价的公平性.仍以前面的数据包为例进行说明.假定有一个包含 2 条消费记录的数据包销售价格也为 10 元，分成比例为 0.4，数据卖方的分成收益为 4 元，按平均定价的方式每条数据元组的价格为 2 元.可见，正向定级、反向定价方法得到第 1 条数据元组为 1.25 元，第 2 条数据元组为 2.75 元.第 2 条数据元组的价值权重、信息量和引用指数都高于第 1 条数据元组，进一步说明第 2 条数据元组价值更大、体现价值的相应价格更高，既合理又公平.

因此，正向定级、反向定价方法，能够对数据元组进行精确的细粒度控制，体现出每一条数据元组应有的价值，激励更多的个人参与到数据交易市场中来，形成具有规模效应的数据交易环境，交易各方都能从中获得更大的实惠，形成数据普惠的生态系统.

4.3 数据收益模型

数据收益是基于用户有效的数据，通过数据分析和增值服务产生的数据收入，用户按比例享受收益分配权，主要解决个人数据资产化的第 3 个问题.

文献[27]结合静态定价和动态定价，通过盈亏平衡分析、双因素试验和回归方程等数学模型对云计算公司的销售数据进行分析，在此基础上提出合理的定价策略，以供决策者参考，使中国的云计算公司以较低廉价格占领云计算市场，又有丰厚的利润可持续发展.

目前实现用户数据收益的方法包括：

(1) P2P 数据收益.通过交易平台，用户按每条数据固定的价格直接卖给买方，以获得数据收益.

(2) 数据交易分成收益.通过交易平台提供数据产权交易或服务交易，对买方进行收费，然后与用户进行收益分成.

(3) 数据期货佣金收益.以数据为标的物形成数据期货，通过数据期货交易，获取期货交易的结算收益和利息收益.

(4) 数据融资佣金收益.允许用户抵押数据产权进行贷款，收取适当的利息收入，可利用数据进行众筹，支持大数据创业，然后换取成长收益或股权.

本文将用户数据收益分为数据利息收益和数据分成收益，包括：

(1) 数据利息收益的计算方法

① 通过清洗和审核的个人数据成为用户的有效数据资产，才能开始计算数据利息.

② 一条数据元组为个人数据的基本计量单位，并进行估值.

③ 个人基本信息的完整性和真实性(即基本信息完整度)，以及各类个人数据的商用价值率(即数据的资产化率)，是影响数据资产利息的主要因素.

假定一条有效的个人数据资产估值为变量 v，个人基本信息完整度为变量 u，数据的资产化率为变量 c，则计息资产 a 的计算公式为

$$a = v \times u \times c \tag{3}$$

④ 个人数据资产利息计算.

由式(3)得出个人数据资产的计息资产 a，假定日利率为变量 r，则由计息资产所产生的日利息 i 的计算公式为

$$i = a \times r \tag{4}$$

⑤ 针对每个用户、每天各类的数据资产，每天进行计息和汇总工作.

⑥ 数据利息计算完成后，进行记账处理.

(2) 数据分成收益的计算方法

数据分成收益是基于个人数据形成的数据产品和增值服务，获得的收益按比例与用户分成.

下面通过一个有关数据利息收益的算例来验证.假定数据资产计量单位为元，有一条有效的个人收支数据资产估值为 1，个人基本信息完整度为 70%，数据的资产化率为 20%，日利率为 0.5%，则由式(3)和(4)可得到这条数据每天的数据利息 $i = v \times u \times c \times r = 1 \times 0.7 \times 0.2 \times 0.005 = 0.0007$.也就是说这条估值为 1 元的数据资产，根据给定的参数可以获得 0.0007 元的日利息.

收益模型可动态调整，具体可调 4 个参数：数据资产估值变量 v、个人基本信息完整度变量 u、数据的资产化率变量 c 和日利率变量 r，计算合理的计息资产利息，体现公平和效率.

4.4 数据的确权与溯源

数据确权是指确定数据的权利人，即谁拥有数据的所有权、知情权、隐私权和收益权以及对个人隐私权附有保护责任等.数据确权的关键在确定数据的持有人，包括数据的原始产生者、数据交易后的拥

有者两方面，而这两方的确定与数据溯源具有直接的关系. 数据溯源是追溯数据的演进过程，包括数据来源、管理、使用、交易、更新维护、失效退出等数据全生命周期的变迁过程，以及引起这些变化的因素，溯源技术的难点在于异构数据的处理，随着时间的推移和应用的需要，将产生各种各样异构的数据，如传统数据库的结构化数据、互联网的文档等. 这类异构数据如何实现溯源，是困扰业界悬而未解的一个难点问题. 解决数据溯源问题，对数据权利人的权益确定具有十分重要的意义.

目前数据溯源计算可分为两类[28]：

(1) 计算时，分析查询或视图定义，构造逆查询，求逆结果即为视图起源，亦即查询反演(Query Inversion)方法，主要是在早期把数据起源用于视图维护和更新问题时提出来的，但该方法不能完全适用于复杂查询.

(2) 基于标注的溯源计算，即直接在数据上注明其来源，在标注中记录数据来源或生产流通过程，用标注存放数据起源信息. 这种方式需引入额外的辅助数据——标注，因此，有关标注的组织、管理、维护等也带来一系列复杂问题.

近年来，基于图论和生成树的方法，对数据溯源提出新的研究思路. Karvounarakis 等人[29]提出基于有向图的交换环算法，利用多项式环进行数据溯源，但该算法还是利用标注信息，对数据的管理都会增加额外的负担. 王梁等人[30]提出构建共享路径表，在构建过程中对原子析取式进行预计算，构建溯源二叉树的方法. 该文中定义元组的规模为 n，溯源表达式包含的属性表达式个数为 m，其转换的时间复杂度为 $o(mn)$. 由于属性级溯源信息比元组级溯源信息的粒度更细，因此该方法存储开销必然大于 $o(ns)$，其中 s 为每条元组的溯源表达式存储代价.

本文提出一种基于网络流理论的数据溯源计算方法——数据源流向图 DRDG(Data Resource Direction Graph)，定义一种新型数据源有向图，描述如下：

有向图中节点 V 表示数据表中的数据项或者标注，有向边 E 表示节点之间的数据源关系，即有向图 $D=(V,E)$，且具有下列定义在集合 $V\times V$ 上的 3 个函数：

(1) 对于每一个 $(i,j)\in(V\times V)$，有一个下界(Lower Bound)函数 $l_{ij}\geq 0$；

(2) 一个容量函数(Capacity)函数 $u_{ij}\geq l_{ij}$；

(3) 一个费用(Cost)函数 c_{ij}.

本文用记号 $N=(V,E,l,u,c)$ 来表示一个有向数据源流向图，它对应着一个有向图 $D=(V,E)$ 以及函数 l,u,c.

对于 N，设其顶点集上的平衡向量函数 b_x，对任意的顶点 $v\in V$，有

$$b_x(v)=\sum_{vw\in A}x_{vw}-\sum_{uv\in A}x_{uv} \tag{5}$$

其中，x_{vw} 为以 v 为头的弧流，x_{uv} 为以 v 为尾的弧流，$b_x(v)$ 是以 v 为头的弧流之和与以之和的差. 按照顶点平衡值 $b_x(v)$，我们能够将有向图 N 的全体顶点分为 3 大类：若有 $b_x(v)>0$，顶点 v 是数据流的源(Source)；如果有 $b_x(v)<0$，称顶点 v 是数据流的收点(Sink)；当 $b_x(v)=0$ 时，称 v 是数据流的一个平衡点. 因此，可以通过计算 $b_x(v)$，求出数据起源. 在本方法中，某顶点的出度大于入度，即该顶点有新数据的输出，此数据必定来源于该顶点，因此，该顶点即为此数据的源，由此该方法的有效性得以验证. 与文献[29]相比，本算法中有向图中节点数据由邻接矩阵给出，算法时间复杂度为 $o(n^2)$；空间复杂度方面，由于基于平衡向量函数 b_x 和概率元组的计算大大减少了标注信息，增加的存储开销不超过 $o(ns)$，可有效节约存储空间.

4.5　数据共享与交换机制

数据作为一种资产，只有流通才能发挥其价值. 目前，数据流通存在诸多障碍，其中数据开放共享与交易交换是主要因素之一. 数据共享指政府、企业以及个人间共享数据的使用权(包括产权)及知情权，主要解决数据的获取问题. 数据交换泛指数据格式转换及数据作为商品的交易行为，主要解决数据的流通问题. 数据共享与交易可形式化描述为

$$\begin{cases} C\Leftrightarrow \Pi \\ \Pi\subseteq f(D_1,D_2,\cdots,D_n)\mid D_i\subset D, \\ \qquad u_i\in U,\forall d_j\in D_i\Rightarrow d_j\mapsto u_i \end{cases}$$

其中数据集 D_i 等分别表示属于用户 u_i 的数据. 该描述在客户 C 与来自各个用户的数据 Π 之间建立起联系.

建立数据共享与交换机制的目的是在数据需求方与数据提供方之间形成高效的信息交互渠道，从而避免信息孤岛，同时建立数据溯源、确权、定价等机制. 数据共享的难点在于建立数据共享机制(如数据共享激励机制以及共享收益机制)；而数据交换的难点则在于建立统一的数据交换格式，以及数据产权、数据在商品化过程中的定价、流通、交易机制.

激励个人共享私有数据的关键在于保障个人

的知情权以及个人对数据的控制权.技术层面上,数据空间或个人数据空间(Personal DataSpace,PDS)[31-32]强调了主体的相关性和可控性,为共享个人数据创造了条件.基于该思想,麻省理工学院的Living lab项目组在对个人数字生态系统(personal data ecosystem)的研究中,创造性地开发了Open-PDS(Open Personal Data Store),进一步发展了个人数据空间的概念,解决了个人数据采集、存储、数据产权、细粒度访问控制、权限管理(访问控制)、数据监控与审计等问题[33];诺丁汉大学数字经济研究组研发的Dataware,进一步强调数据所有者对其私有数据的控制,为激励数据所有者共享其数据创造了条件①.

在市场激励方面,系统实时采用随机奖励(random rewards)、社会影响、个性化定制[34]以及竞价机制等措施促使个人共享数据.

本系统采用图2所示的数据共享与交换机制.其中核心部分包括数据获取、ETL、数据索引、数据交易、云存储、认证管理以及数据服务.

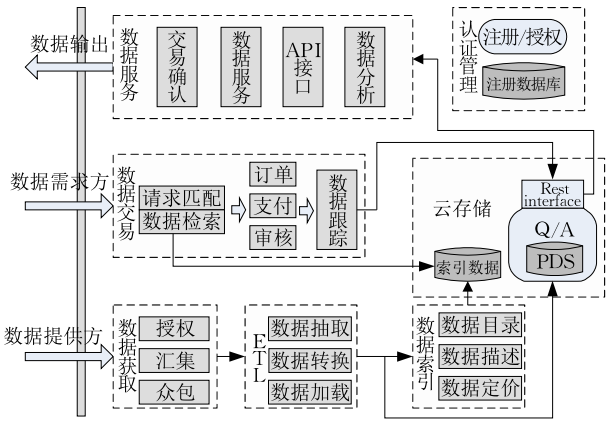


图2 数据分享与交换逻辑图

其工作流程如下:

(1)共享数据的获取.依据数据开放的形式,包括3种典型数据获取方式:授权访问(一般针对第三方数据,如线上/线下企业通过业务系统采集的个人数据,这类数据需提供方的认证和授权(OAuth)方可访问);汇集(本系统收集的由用户直接创建和管理的数据以及政府公开的数据,如个人文档信息以及人口统计数据等)及众包(由发布特定悬赏任务而获取的数据).

(2)数据格式规范化.负责数据的抽取、转换和加载.在本系统中主要完成将分散的、异构的数据源转换成JSON格式的数据,形成统一数据格式,以实现个人数据的交换.

(3)数据定价.根据元数据类型(存储在数据索引结构中),采用4.2节数据计量与计价和4.3节数据收益模型完成数据的静态定价.

(4)数据交易交换管理.包括订单管理、支付管理以及交易审核等.同时,为确保个人的知情权以及保护数据提供方的隐私.系统采用了PDS来分别管理各个数据提供者的数据,即最终的数据访问请求需要经过PDS的授权以及记录处理才能完成.

4.6 个人隐私保护

个人隐私保护形式化描述为

$$\begin{cases} C \Leftrightarrow \Pi \,\&\, C,\ \text{not know } U \\ \Pi \subseteq f(D_1, D_2, \cdots, D_n) \mid D_i \subset D, \\ \quad u_i \in U, \forall d_j \in D_i \Rightarrow d_j \mapsto u_i \end{cases}$$

目的是在客户C与Π之间虽然建立了联系,但客户C不能确定数据D来自哪个用户U.

个人隐私保护的难点在于,在保护个人隐私的前提下,实现个人数据价值的最大化,或者在隐私保护与数据价值之间寻找平衡点.实现个人数据的隐私保护首先应对数据中的标识符及准标识符信息进行匿名化处理,常用的方法有k-anonymity、l-diversity、(a,k)-anonymity等[35].其次,应防止攻击者利用背景知识推断某些敏感信息.在统计数据查询中,差分隐私[36]被证明是隐私保护最有效的模型.

目前通用的隐私保护是"聚合型"的,如求和、平均等,针对个体数据(即个人数据)隐私保护的研究相对较少.为解决个人数据的隐私保护问题,本文将个人标识信息与个人敏感数据分别处理:对个人标识信息采用传统的数据加密技术;对个人敏感信息,创新性的将安全问答(Safe Answer)机制、元数据发布机制及目录服务机制相结合,以实现数据的发现与保护、数据保护与共享之间的平衡以及个人对数据的完全控制(解决数据产权问题).修改后的隐私保护策略将式(2)改为

$$\begin{cases} C \Leftrightarrow \Pi \,\&\, C,\ \text{not know } U \\ \Pi \subseteq f(f_1(D_1), f_2(D_2), \cdots, f_n(D_n)) \mid D_i \subset D, \\ \quad u_i \in U, \forall d_j \in D_i \Rightarrow d_j \mapsto u_i \end{cases}$$

个人隐私保护的核心是安全问答机制,本节主要介绍该机制的工作原理.

安全问答(Safe Answer)[30]被认为是保护个人数据隐私的有效机制之一.在该机制下,用户提出问题,只有(安全的、合理的)答案才返回给用户.也是

① DEMO: Dataware, A Personal Data Architecture. http://mor1.github.io/publications/pdf/de13-dataware-demo.pdf 2015.10.5

本系统采用的主要机制.

系统采用的安全问答机制(SA)的组成及原理如下：

系统由元 PDS 及组 PDS 组成，其中元 PDS 特指逻辑上的个人数据集(物理上，元 PDS 可采用分布式存储)，基本组成如图 3 中的(a)所示.

工作流程如下：

(1) 请求被传到 SA 模块；

(2) SA 模块访问内部数据库，并执行相关请求计算；

(3) SA 模块返回计算结果.

该机制中，特定的应用(即请求)只能与该应用对应的 SA 模块交互信息，SA 模块实现对用户访问的审计，访问控制等功能，并且只有 SA 模块具有访问 PDS 原始数据的权限. PDS 用户主可以控制 SA 权限的管理以及对隐私的设置.

在元 PDS 中，所有的数据访问请求都是通过 SA 模块完成的. 不同应用对应不同的 SA，设计时，需设计专用的 SA，安装时也应安装对应的 SA 模块.

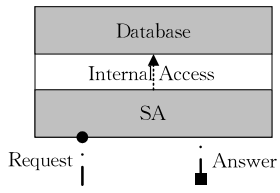
组 PDS 由多个用户的元 PDS 组成，采用安全多方计算的方式保护个人数据隐私，组成如图 3 中的(b)所示.

工作流程如下：

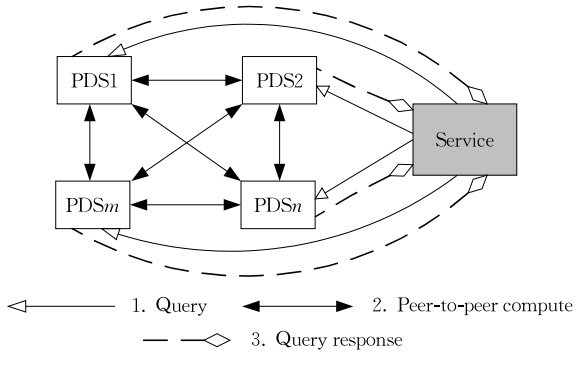
(1) Service 发布数据需求；

(2) 各个 PDS 交换信息，协作计算；

(3) 返回查询结果.



(a) 元PDS的组成



(b) 组PDS的组成

图 3　基于 PDS 的个人数据隐私保护机制

在该结构中，PDS 之间以加密的方式交互信息，服务请求者不知道信息的具体来源(出自某个具体的 PDS)，因而可有效的保护用户的隐私.

安全问答机制的难点在于如何确定对特定问题的回答是安全的，这与差分隐私中的隐私预算(ε)有相似性；同时，安全多方计算的效率问题也是制约系统性能的主要因素之一. 这两方面的问题是本系统在安全方面应着重解决的问题.

另外，文献[37]提出的空间数据众包结构以及文献[38]提出的空间数据差分隐私分解策略，也是本系统中下一步个人隐私保护的重要参考策略.

4.7　个人数据的分析与挖掘

传统数据统计和挖掘方法虽然已应用于许多领域的大数据分析处理，但在处理大规模、高速增长、类型复杂、结构复杂、模式复杂的大数据时，通常会遇到数据高度稀疏与维度灾难等问题，处理效率难以满足用户需求[39]，而且，由于数据大而全，挖掘中更要注意对数据隐私的保护. 同时，数据规模较大时，传统数据分析方法往往采用抽样技术对样本进行分析，根据样本推断整体，基于对原始数据的抽样或过滤，在数据样本空间中提取特征和规律，发现尽可能多的知识和信息.

大数据分析方法发展的一个重要特点是不再片面追求算法的精确性，而是力求对全体数据集整体的处理分析，在这种情况下，出现一些改进的专门针对大数据集成分析与处理的方法，如局部学习(Local Learning)、布隆过滤器(Bloom Filter)、散列法(Hashing)、索引技术、字典树(Trie)、并行计算、基于稀疏表达的学习方法等[40-41].

发掘个人大数据蕴含的巨大价值，本文针对具体类型数据进行分析处理，包括：

(1) 收支财产类数据. 通过分析和挖掘收支财产类数据，可对个人消费者、消费者群体的消费习惯、消费行为进行刻画和预测，支持各类消费服务提供商进行个性化服务，改进和提升服务质量；通过对消费者行为模式的提取和分类，可为征信服务、金融服务、营销服务提供稳定可靠的基础信息.

(2) 位置和时间数据. 智能手机、平板电脑、智能穿戴等设备接入互联网，几乎全天候记录个人的位置和时间，这些数据精确的反映人口驻流情况，结合其他类型数据可以分析区域内网格功能划分，为生活、商业、治安等提供精确选址服务；通过汇总和预测不同的地点、特定的时间、大量的人流的信息，可为政府部门和交通运输行业提供各类预警和预测.

(3)社交网络信息. 社交网络日趋发展成为获取信息、表达观点和交流信息不可缺少的网络传播媒介,其影响力逐渐覆盖人们生活的方方面面. 通过对社交网络中人际关系及其特征进行提取可划分圈子,支持精准的社会化营销;基于内容的分析包括观点挖掘、情感挖掘,情绪挖掘可为商业层面和社会层面上的危机预警、舆情监测提供服务.

(4)基于多源异构个人数据的用户画像. 对多源异构个人数据整体的高效处理,通过个人个性分析、行为分析和足迹分析,构造用户画像模型、用户价值模型、用户流失预警模型等.

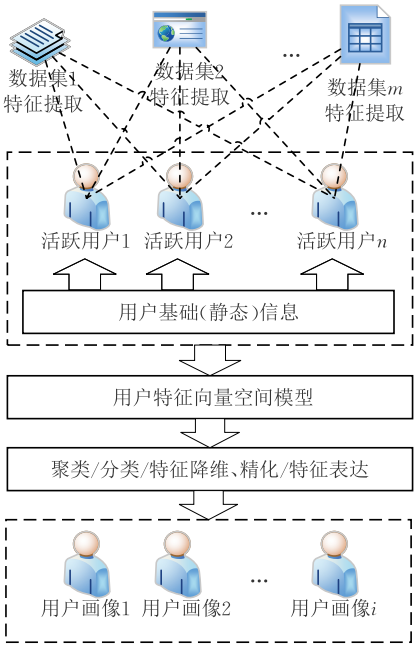


图 4 多源数据用户画像技术框架

用户画像即一系列拥有特定特征的人物模型,建立在一系列真实数据基础之上,是真实用户的虚拟代表,是需求建模、改善用户体验、产品设计开发、实现精准营销和各类增值服务的重要方法和手段. 多源异构数据用户画像框架如图 4 所示,画像过程以用户为中心,首先在各个异构数据集提取活跃用户特征,构建用户特征的向量空间模型(Vector Space Model,VSM),在此基础上进行分析挖掘,采用聚类分类算法依据用户特征进行归类. 为了增强用户画像的可辨识性和可理解性,在生成最终用户画像之前需要对用户特征进行降维,并对特征按应用需求进行表达和精化.

总之,个人大数据价值密度相对较小,时间敏感性较高,不同时期、不同年龄阶段人们会表现出不同的行为模式. 远期的个人历史数据很可能会干扰数据分析精度,因此,对个人大数据的分析应注重对数据阶段的划分和对人行为规律的特征提取,引入人类行为动力学研究机制,选择适宜的数据时间窗口进行模式提取、挖掘及预测预警.

5 个人数据银行的实现——数汇宝平台

四川大学嵌入式系统研发与测试实验室和成都数银科技有限公司基于个人数银行模式与架构,合作开发数汇宝平台,作为一种个人大数据资产的管理与增值服务平台解决方案,以保护用户个人数据的所有权、知情权、隐私权和收益权为核心,结合数据众筹/众包模式,为商品精准营销、个人征信和互联网金融等提供数据支撑.

数汇宝服务平台采用多点分布式构架搭建运行系统,可满足大数据存储处理需求,应对高访问、大并发等常见问题. 系统技术构架如图 5 所示,使用 HBase 分布式数据库存储非结构化数据,数据分析挖掘基于 Hadoop 构架实现.

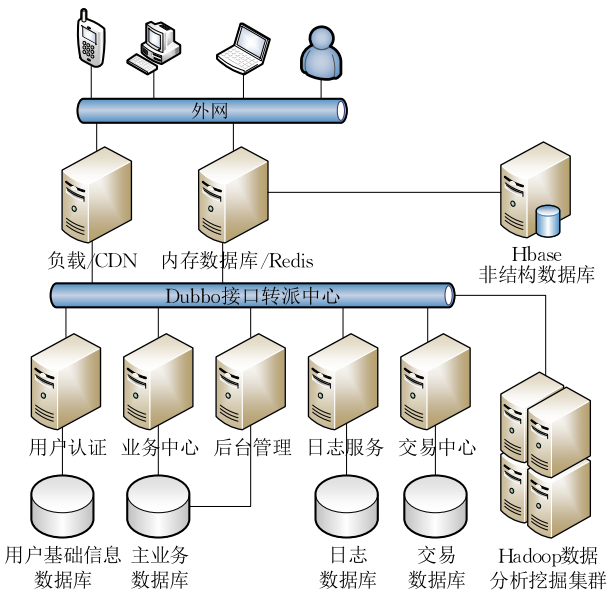


图 5 数汇宝平台技术构架

数汇宝通过个人数据资产的集中管理与协同服务,重组数据资源,实现跨界数据融合、流通,成为开放的个人基础数据平台,提升数据集成的价值,扩大数据分享、使用,推动个人数据银行新模式、新业态的发展. 通过数汇宝 API,个人数据有偿合理地提供给第三方使用和开发行业应用服务,形成数据到业务的闭环,为个人、企业和政府提供共赢合作的价

值，构建一个完善的个人大数据生态系统.

数汇宝是个人数据银行的一个实践性平台，基于英国标准协会(BSI)BS 10012:2009 PIMS 标准，全面集中管理个人分类生活数据，包括个人财产信息、个人收支信息、个人健康信息、个人教育信息、个人朋友圈信息、个人日志信息、个人数据保险箱等七大类数据，解决了数据的计量计价、确权与溯源、收益分成模型等难点.

作为一个个人数据汇聚和集中管理的服务平台，数汇宝的数据来源主要包括：

(1)个人提交. 包括手工提交和智能终端自动读取两种方式，其中手工提交效率低，对用户忠诚度要求较高，数据可靠性较差且验证成本高，若采用此种形式，需制定相应激励机制以保证数据真实有效.

(2)机构间数据交换. 包括企业、政府等机构之间的数据共享与交换，获得个人授权后，通过这些平台的开放 API 汇聚个人信息，如消费信息、社交网络信息和社保信息等，数据一致性和质量较好.

(3)传感器读取. 如从手环等可穿戴设备中读取个人健康信息，这类数据具有良好的结构性，但需要开放相应设备的共享接口.

数汇宝初步打通了个人数据资产化的流程，集数据获取、数据治理、增值服务和收益分成于一体，其平台组成如图 6 所示.

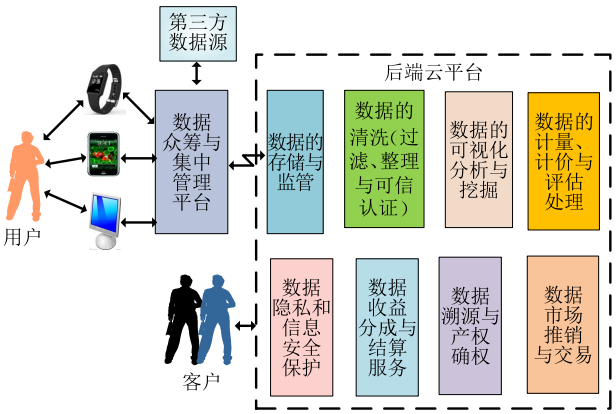


图 6　数汇宝平台组成

数汇宝平台主要功能包括：

(1)个人数据众筹与集中管理平台. 采用线上/线下输入、传感器自动采集和第三方数据源交换等方法汇聚个人数据，将个人数据存放到数汇宝中，方便用户进行数据管理、增值服务和收益核算等工作.

(2)数据的存储与监管. 将个人数据存放到后端云平台中，一方面进行数据日常运营的监管，防止异常、有害和违法数据的存出现；另一方面方便用户访问，形成个人的数字化足迹.

(3)数据的清洗. 包括过滤、整理与可信认证等，主要检查个人数据的真实性、可信性和一致性，对残缺数据、错误数据、冗余重复数据进行处理.

(4)数据的可视化分析与挖掘. 根据客户需求，对个人数据进行多源融合，构建可视化分析与多维数据挖掘模型.

(5)数据的计量、计价与评估处理. 采用“正向定级，反向定价”策略，首先，根据用户数据属性评估数据的等级，不同重要程度的数据具有不同的权重，然后，根据与客户完成的实际交易情况，基于用户数据权重，与用户进行收益分成，实现数据的定价.

(6)数据的市场推广与交易. 通过多种推销手段，调高数据交易额，改善数汇宝的数据产品和服务收益.

(7)数据确权与溯源. 以个人数据全生命周期为基础，建立溯源技术体系，确定用户数据的产权、分享、使用与服务等流动过程，完成数据的审计和治理，为用户数据权益保障提供支撑.

(8)数据收益分成与结算服务. 对用户数据进行增值服务，采用数据利息收益和数据分成收益等，交易过程对用户透明，并通过第三方支付接口与用户完成结算.

(9)数据隐私与信息安全保护. 通过多种管理和技术手段，保护个人数据的产权、隐私权与信息安全.

数汇宝平台创新点主要包括：

(1)基于个人数据产权的托管模式. 在明确个人数据产权等权益基础上，作为第三方中介平台开展个人数据资产的信托管理与增值服务工作.

(2)个人数据的计量与计价. 采用“正向定级，反向定价”策略，以数据记录和数据包形式，实现数据的定价.

(3)基于多源异构个人数据的用户画像. 对多源异构个人数据整体的高效处理，通过个人个性分析、行为分析和足迹分析，构造用户画像模型、用户价值模型、用户流失预警模型等.

数汇宝平台开通试运行 5 个月注册用户突破 100 万，有效数据 200 万条以上，数据交易成交量达到 50 万条，单次数据服务交易的平均响应时间 50 ms，包括数据计量计价、交易、确权与溯源等过程. 快速增长的用户数量和数据量反映了人们对数

据银行模式的认可，同时也体现出用户数据产权意识的增强.随着产业生态圈的逐渐成熟和技术的进一步完善，数汇宝及将进一步推动个人数据产业健康有序发展.

6 结论与未来展望

6.1 结 论

大数据时代，个人数据不但是个人的数据资产，也是社会经济发展的重要资源.个人数据银行作为一种个人大数据管理与运营的新模式，将从前以企业为中心组织、管理和控制个人数据，转变为以个人为中心汇聚、管理和控制个人数据，资源配置更加优化，使得个人数据量化有用、授权访问且有序流通，降低个人数据的使用成本，促进个人大数据市场的全面发展.

通过与银行架构比较研究，探讨个人数据银行平台的组成架构、关键技术和主要研究内容，并研发一种个人数据银行平台——数汇宝，在数据汇聚、确权、定价、隐私保护、共享流通和使用等方面提出并实现了一系列技术方案，从理论和实践两个方面分析个人数据银行建设的可行性与挑战，为个人数据有效流通奠定基础.

6.2 未来展望

下一步值得研究的一些问题主要包括：

(1)数据流通形式及技术实现.数据流通是促进数据市场发展的首要条件，包括开放与共享、交易与交换、租用与流转等多种形式.这些形式从不同角度反映了数据在流通过程中的特点，如数据的开放与共享描述数据发布方式，前者是自由、无限制的，后者受使用方式、访问权限、期限等共享协议的限制；数据的交易与交换从技术及市场的角度描述数据流通方式，前者指将数据视为商品，能够进入市场交易，后者主要从技术角度强调数据格式转换及统一；数据的租用与流转从数据权属的角度描述数据使用方式，前者将数据作为一种按需提供的云计算服务(即 DaaS 服务)，可以租赁使用，后者指数据使用权的转让.数据流通既涉及流通规则设计，又涉及相关计算机技术实现，是一门交叉学科问题.个人数据的有序使用与流通，需要在理清这些概念与技术的基础上，组合成有效的数据流通形式.

(2)基于个人数据的现实挖掘.数据深度分析与挖掘是实现个人大数据增值、服务社会管理的重要手段，汇聚的个人数据直接体现群体的各类倾向，可表现出社会运行规律，人类行为动力学诸多理念、方法和技术对个人大数据的分析挖掘具有非常重要的指导意义.

(3)数字化生存.理论上来讲，数字信息可以无限期的保存，人的生命终结后，他依然可以生活在数字化环境中，继续在虚拟空间中存在.依靠飞速发展的虚拟现实技术对个人大数据进行建模或重构，让普通个人以虚拟形式永远存在于数字空间中，形成家族的有效传承，也将是一个新的研究方向.

(4)个人数据产权的衍生权益保护.自主权、名誉权、人格权、遗忘权、可携带权等衍生权利，如当个人数据资产化后，人去世时，这个资产也能继承吗？离婚时，这个资产如何切分呢？这是社会问题，也涉及到计算机技术实现问题.

(5)数据遗忘技术.和其他事物一样，数据也是有生命周期的，互联网时代，采集数据容易，但数据消除和遗忘绝非易事，过期的数据占用了大量存储资源，也给数据的分析使用带来很多困难，集中化的个人数据存储与管理为数据遗忘提供了可能.在个人数据银行生态圈中，要实现数据遗忘权，就要保证用户选择遗忘的数据、或者超出数据生命周期的完全不被使用，需要探讨一种集成数据溯源、隐私保护、数据状态确认等数据遗忘技术.

(6)其他大数据的开放共享机制.如企业大数据、行业大数据和政府大数据的产权界定及开放共享机制，其中一些数据是否可以作为公共数据，免费向社会开放？如美国政府数据 2009 年开始进行开放与共享①；2015 年 9 月 5 日中国国务院发布的《促进大数据发展行动纲要》，规定从 2017～2020 年逐步将政府数据向社会开放，如何界定政府数据开放的边界和方式，是加速推动我国大数据产业发展的重要环节.

参 考 文 献

[1] CCF Task Force on Big Data. China's big data technology and industrial development white paper (2013). China Computer Federation, Beijing: Technical Report, 2013(in Chinese)
(CCF 大数据专家委员会. 中国大数据技术与产业发展白皮书(2013). 中国计算机学会，北京：技术报告，2013)

① 美国政府开放数据主页. http://www.data.gov 2015,10,5

[2] Wang Jia-Qiu, Wang Zhong-Jie. A survey on personal data cloud. The Scientific World Journal, 2014, 2014(2014): 1-13

[3] Gurrin C, Smeaton A F, Doherty A R. LifeLogging: Personal big data. Foundations and Trends in Information Retrieval, 2014, 8(1): 1-125

[4] Schwab K, Marcus A, Oyola J O, et al. Personal data: The emergence of a new asset class//Proceedings of the Initiative of the World Economic Forum. Geneva, Switzerland, 2011: 5-39

[5] Richards N M, King J H. Three paradoxes of big data. Social Science Electronic Publishing, 2013, 41(3): 41-46

[6] Huberty M. Awaiting the second big data revolution: From digital noise to value creation. Journal of Industry, Competition and Trade, 2015, 15(1): 35-47

[7] Xie Xiao. Pondering on personal information management. Library and Information Service, 2011, 55(24): 21-26(in Chinese)
(谢笑. 个人信息管理研究探析. 图书情报工作, 2011, 55(24): 21-26)

[8] Xiao Wei-Qing. The Research and Implement of Personal Information Management System Based on Data Synchronization [M. S. dissertation]. Beijing University of Posts and Telecommunications, Beijing, 2010(in Chinese)
(肖卫青. 基于数据同步的个人信息管理的研究与实现[硕士学位论文]. 北京邮电大学, 北京, 2010)

[9] Huang Lan-Hui. The Research and Implementation of Personal Space of Information Management System Based on WSRF [M. S. dissertation]. National University of Defense Technology, Changsha, 2008(in Chinese)
(黄蓝会. 基于 WSRF 的个人信息空间管理系统的研究与实现[硕士学位论文]. 国防科学技术大学, 长沙, 2008)

[10] Zhong Cheng. Design and Implementation of Personal Information Management System Based on Cloud Platform [M. S. dissertation]. Huazhong University of Science & Technology, Wuhan, 2014(in Chinese)
(钟承. 基于云平台的个人信息管理系统的设计与实现[硕士学位论文]. 华中科技大学, 武汉, 2014)

[11] Buchanan M. Secret signals: Does a primitive, non-linguistic type of communication drive people's interactions. Nature, 2009, 457(7229): 528-530

[12] Pentland A. Reality mining of mobile communications: Toward a new deal on data. The Global Information Technology Report 2008-2009, 2009: 1981

[13] Li Yu-Kun, Ren Biao, Zhao Xi-Yan, et al. Research on personal data management. Journal of Frontiers of Computer Science and Technology, 2014, 8(11): 1281-1295(in Chinese)
(李玉坤, 任标, 赵喜燕等. 个人数据管理技术研究. 计算机科学与探索, 2014, 8(11): 1281-1295)

[14] Zhong Shi-Lun. Personal Data Space Management System Based on Tag Technology [M. S. dissertation]. National University of Defense Technology, Changsha, 2008(in Chinese)
(钟世伦. 基于标签技术的个人数据空间管理系统研究与实现[硕士学位论文]. 国防科学技术大学, 长沙, 2008)

[15] Dong X L, Halevy A. A platform for personal information management and integration//Proceedings of the VLDB 2005 PhD Workshop. Trondheim, Norway, 2005: 26

[16] Cai Y, Dong X L, Halevy A, et al. Personal information management with SEMEX//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, 2005: 921-923

[17] Meng Xiao-Feng, Ci Xiang. Big data management: Concepts, techniques and challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169(in Chinese)
(孟小峰, 慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展, 2013, 50(1): 146-169)

[18] Moiso C, Minerva R. Towards a user-centric personal data ecosystem the role of the bank of individuals' data//Proceedings of the 16th Intelligence in Next Generation Networks (ICIN). Berlin, Germany, 2012: 202-209

[19] Liu Ya-Hui, Zhang Tie-Ying, Jin Xiao-Long, et al. Personal privacy protection in the era of big data. Journal of Computer Research and Development, 2015, 52(1): 229-247(in Chinese)
(刘雅辉, 张铁赢, 靳小龙等. 大数据时代的个人隐私保护. 计算机研究与发展, 2015, 52(1): 229-247)

[20] Meng Xiao-Feng, Zhang Xiao-Jian. Big data privacy management. Journal of Computer Research and Development, 2015, 52(2): 265-281(in Chinese)
(孟小峰, 张啸剑. 大数据隐私管理. 计算机研究与发展, 2015, 52(2): 265-281)

[21] Li Chun-Lan, Deng Zhong-Hua, Zhang Wen-Ping. Pricing the cloud services. Library and Information, 2013(1): 36-41(in Chinese)
(黎春兰, 邓仲华, 张文萍. 云服务的定价策略分析. 图书与情报, 2013(1): 36-41)

[22] Muschalle A, Stahl F, Löser A, et al. Pricing approaches for data markets//Proceedings of the Workshop Business Intelligence for the Real Time Enterprise. Istanbul, Turkey, 2012: 129-144

[23] Balazinska M, Howe B, Suciu D. Data markets in the cloud: An opportunity for the database community. Proceedings of the VLDB Endowment, 2011, 4(12): 1482-1485

[24] Mankiw N G. Principles of Economics. Ohio, USA: South-Western Cengage Learning, 2012

[25] Balazinska M, Howe B, Koutris P, et al. A discussion on pricing relational data. Lecture Notes in Computer Science, 2013: 167-173

[26] Li C, Li D Y, Miklau G, et al. A theory of pricing private data//Proceedings of the 16th International Conference on Database Theory. Genoa, Italy, 2013: 33-44

[27] Sun Hong, Tu Qian-Wei, Wang Xiao-Wan, et al. Pricing of SaaS in cloud computing based on mathematical models. Journal of Shanghai University of Science and Technology, 2014, 36(2): 199-204(in Chinese)
(孙红, 屠金炜, 王晓婉等. 基于数学模型的云计算 SaaS 定价的研究与实现. 上海理工大学学报, 2014, 36(2): 199-204)

[28] Galhardas H, Florescu D, Shasha D, et al. Declarative data cleaning: language, model and algorithms//Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, USA, 2001: 371-380

[29] Karvounarakis G, Green T J. Semiring-annotated data: Queries and provenance. ACM Sigmod Record, 2012, 41(3): 5-14

[30] Wang Liang, Zhou Guang-Yan, Wang Li-Wei, Peng Zhi-Yong. Attribute level lineage and probabilistic computation of uncertain data. Journal of Software, 2014, 25(4): 863-879 (in Chinese)
(王梁, 周光焱, 王黎维, 彭智勇. 不确定关系数据属性级溯源表示与概率计算. 软件学报, 2014, 25(4): 863-879)

[31] Halevy A Y, Franklin M J, Maier D. Dataspaces: A new abstraction for information management. Lecture Notes in Computer Science, 2006, 34(4): 1-2

[32] Jiang Shuo. Research on Some Key Problems of Data Integration in Dataspace [Ph. D. dissertation]. Donghua University, Shanghai, 2014(in Chinese)
(姜朔. 数据空间中数据集成若干关键问题研究[博士学位论文]. 东华大学, 上海, 2014)

[33] Yves-Alexandre D M, Erez S, Wang S S, et al. openPDS: Protecting the privacy of metadata through safe answers. Plos One, 2014, 9(7): e98790

[34] Pluntke C, Prabhakar B. INSINC: A platform for managing peak demand in public transit. Journeys, Land Transport Authority Academy of Singapore, 2013,(9): 31-39

[35] Feng Deng-Guo, Zhang Min, Li Hao. Big data security and privacy protection. Chinese Journal of Computers, 2014, 37(1): 246-258(in Chinese)
(冯登国, 张敏, 李昊. 大数据安全与隐私保护. 计算机学报, 2014, 37(1): 246-258)

[36] Dwork C, Lei J. Differential privacy and robust statistics// Proceedings of the 41st ACM Symposium on Theory of Computing. Washington, USA, 2009: 371-380

[37] Cormode G, Procopiuc C, Srivastava D, et al. Differentially private spatial decompositions//Proceedings of the 2014 IEEE 30th International Conference on Data Engineering. Chicago, USA, 2011: 20-31

[38] To H, Ghinita G, Shahabi C. A framework for protecting worker location privacy in spatial crowdsourcing. Proceedings of the VLDB Endowment, 2014, 7(10): 919-930

[39] Wang Yuan-Zhuo, Jin Xiao-Long, Cheng Xue-Qi. Network big data: Present and future. Chinese Journal of Computers, 2013, 36(6): 1125-1138(in Chinese)
(王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望. 计算机学报, 2013, 36(6): 1125-1138)

[40] Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics, 2008, 37(1): 246-270

[41] Wu X, Zhu X, Wu G Q, et al. Data mining with big data. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(1): 97-107

GUO Bing, born in 1970, Ph. D., professor, Ph. D. supervisor. His current research interests include green computing and personal big data.

LI Qiang, born in 1963, Ph. D., associate professor. His research interests include mobile cloud computing and personal big data.

DUAN Xu-Liang, born in 1982. Ph. D. candidate, lecturer. His current research interests include personal big data, big data cleaning and data mining.

SHEN Yun-Cheng, born in 1979, Ph. D. candidate, lecturer. His current research interests include personal big data and big data pricing.

DONG Xiang-Qian, born in 1975, Ph. D. candidate, lecturer. His current research interests include personal big data, information security and personal privacy protection.

ZHANG Hong, born in 1980, Ph. D. candidate, lecturer. His current research interests include mobile internet data provenance and personal big data.

SHEN Yan, born in 1973, Ph. D., professor. Her research interests include smart terminal and instruments.

ZHANG Ze-Liang, born in 1967, senior engineer. His research interests include software architecture and personal big data.

LUO Jian, born in 1969, Ph. D., associate professor. His current research interests include business management and human resource management.

Background

Big Data is one of the research hotspots whether in academic or industrial circle, it is being increasingly recognized that big data contains big values. Personal big data, as an important types of big data, is becoming a new economic "asset class", and a valuable resource for the 21st century that will touch all aspects of society.

With the further development of mobile Internet, personal big data as a valuable asset is growing exponentially. While, around the world, the application status of personal big data is not optimistic, the vague of ownership, chaos of management and difficulty of circulation severely affect the regular development of personal big data market. To release the full potentials of personal data, a balanced ecosystem with the increased trust between individuals, government and private sectors is necessary to be built. Current research topics about personal big data focus mostly on data management, data using, privacy protection and so on, while the research of personal data ownership, personal data asset management and personal data value-added services are few. In this situation, exploring a new mode for collection, management and usage patterns of personal big data will contribute to regulate the data trading patterns, and protect individual privacy and data security.

Therefore, we first propose the concept of Personal Data Bank, which is a new mode of personal big data asset management and value-added services based on the bank monetary asset management architecture. By developing a personal data bank prototype product "ShuHuiBao", we attempt to implement the system architecture and key technologies of personal data bank, which can systematically analyzes the feasibility of personal data bank from the perspective of theoretical and practical aspects, and lays a strong foundation for future development of personal data capitalization.