Research on public credit information collection technology and its application

公共信用信息采集技术及其应用研究

杨胜刚　吴志明　著

中国金融出版社

公共信用信息采集技术及其应用研究

杨胜刚　吴志明　著

中国金融出版社

责任编辑：肖丽敏
责任校对：李俊英
责任印制：陈晓川

图书在版编目（CIP）数据

公共信用信息采集技术及其应用研究（Gonggong Xinyong Xinxi Caiji Jishu Jiqi Yingyong Yanjiu）/杨胜刚，吴志明著. —北京：中国金融出版社，2018.6
ISBN 978-7-5049-9578-0

Ⅰ.①公… Ⅱ.①杨…②吴… Ⅲ.①信用—信息获取—研究—中国
Ⅳ.①F832

中国版本图书馆 CIP 数据核字（2018）第 102721 号

出版发行 中国金融出版社
社址 北京市丰台区益泽路 2 号
市场开发部 （010）63266347，63805472，63439533（传真）
网上书店 http://www.chinafph.com
（010）63286832，63365686（传真）
读者服务部 （010）66070833，62568380
邮编 100071
经销 新华书店
印刷 北京市松源印刷有限公司
尺寸 185 毫米×260 毫米
印张 23
字数 490 千
版次 2018 年 6 月第 1 版
印次 2018 年 6 月第 1 次印刷
定价 68.00 元
ISBN 978-7-5049-9578-0
如出现印装错误本社负责调换 联系电话（010）63263947

前　言

社会主义市场经济本质上就是信用经济，信用作为特定的经济交易行为，是商品经济发展到一定阶段的产物，而征信在促进信用经济发展和社会信用体系建设中发挥着极其重要的基础作用。国际金融危机后，世界各国普遍意识到信用信息服务对经济金融运行的深刻影响和重要作用，从防范非系统性风险上升为防范系统性风险，利用新技术推动征信业健康发展成为各国面临的共同任务。本著作研究目的就是通过对我国现有公共信用信息采集工作的全面分析，在借鉴国际先进经验教训的基础上，探索建立基于公共信用信息采集标准和信息归集、清洗、匹配、加载等技术规范的可持续机制。

而大数据时代的到来和互联网技术的发展，为我国征信机构实现公共信用信息的可持续采集创造了前所未有的发展机遇和开拓空间，需要做出客观的总结归纳并明确未来的路径选择。征信业意欲把握住大数据时代的发展机遇，须在采集庞大的数据信息、提高数据的加工与挖掘能力、实现数据的增值、培育多元化发展的征信机构等方面努力。而全面的信息采集、大数据库的建设、大数据挖掘技术的研究与应用等，这些仅仅依靠公营性的公共征信系统是难以实现的，更需要商业性征信机构的大力参与。因为有潜在的巨大盈利激励，商业性征信机构愿意投入大量财力、物力、人力到数据库建设、大数据技术研究等方面。所以，应以更加市场化的方式推动我国征信业的发展，实行政府与市场运作“双轮驱动”，紧紧把握大数据的战略时机。

公共信用信息是征信信息的重要组成部分，它的采集是一项庞大而复杂的系统工程，与一国消费者获得授信的渠道和方式有关。本著作深入研究我国地区社会信用体系建设中公共信用信息采集的实践经验与教训，重点探讨了完善我国公共信用信息采集的运行机制，系统提出了强化公共信用信息采集机制的顶层设计与安排；明确公共信用信息采集的原则与方法；推进公共信用信息采集与共享行为的规范以及公共信用信息共享平台建设；完善公共信用信息采集法律法规体系的公共信用信息可持续采集的运作机制。在此基础上，科学构建了我国公共信用信息采集指标体系。通过公共信用信息采集指标体系的科学设计，可以有效记录参与经济社会活动的个人和机构的信用信息，并逐步推动信用信息在全国范围的互联互通，实现信用信息采集的规范化。

本著作的主要学术创新体现在：

一是明确提出大数据时代为我国征信行业发展带来新契机。征信行业作为大数据产业链上拥有数据资产，甚至可能掌握大数据技术和思维的行业，必将迎来前所未有的发展机遇和空间。大数据时代“数据开放”的倡议使得信用信息的采集范围更加广泛，内容更加丰富，更多的公共信用信息纳入征信范围。征信业须在采集庞大的数据信息、提高数据的加工与挖掘能力、实现数据的增值、培育多元化发展的征信机构等方面努力。而全面的信息采集、大数据库的建设、大数据挖掘技术的研究与应用等，这些仅仅依靠公营性的公共征信系统是难以实现的，更需要商业性征信机构的大力参与。

二是明确指出欧美国家公共信用信息采集经验可以作为中国加强公共信用信息采集工作的有益借鉴。具体而言体现在：广泛采集和隐私保护是公共信用信息可持续采集的核心与关键；多样化渠道采集和统一化渠道管理是公共信用信息可持续采集的有力支撑；多层次和全覆盖的法律体系是公共信用信息可持续采集的根本条件；规范化信息共享方式是公共信用信息可持续采集的重要保障；创新性和多品种授信应用是公共信用信息采集的最终落脚点；公共信用信息采集是发达国家征信建设过程中至关重要的因素，也是发展中国家信用体系建立和完善的必要条件。

三是客观评价我国地区公共信用信息采集的经验与教训。本著作研究指出，从公共信用信息采集的主体来看，主要有以发展改革委、工商局为代表的政府机关、人民银行征信中心以及社会征信机构；从公共信用信息采集的对象上看，主要包括行政及司法部门、行业协会等民间组织、中介机构、具有社会公共管理职能的事业单位和公共团体以及其他非银行金融机构；从公共信用信息采集的内容上看，征信机构采集的企业与个人公共信用信息力求涵盖评价企业和个人信用状况的各个方面，前者指企业的身份信息、良好信息、提示信息、警示信息，后者指个人的基本信息、社会交易行为信息、公共记录信息以及收入与资产信息；从公共信用信息采集的方式来看，主要包括以强制数据源单位报送数据和协议征集相关信用数据两种方式。但从目前的实际情况来看，也存在着信用法律缺失；部门间协调困难；信用信息采集难以持续，信息质量得不到保障和对信息主体权益保护不充分等现实问题。

四是明确提出科学的信息采集原则和方法。在采集原则方面，基于我国公共信用信息采集现状与目标，我们需要在坚持“以服务为主线、以应用为中心”“惠及全民、安全可控”的基本要求下，遵循“统一管理、先易后难、标准兼容、适度采集、法律保障”五大原则，促使公共信用信息的采集工作有序、规范进行；在采集方法方面，针对各数据主体以及信息种类的不同，可以把主要的采集方法总结为协调采集法、协议采集法和有偿采集法三种，上述三种采集方法分别适宜于信息采集发展的不同阶段，

且在各阶段均可以同时存在，但在同一阶段的重要程度不同，所以在特定的阶段应以某一种采集方法为主，同时辅助以其他采集方法，以更有效、全面地采集公共信用信息。

五是提出了科学构建我国公共信用信息采集指标体系。通过构建科学的公共信用信息采集指标体系，可以有效记录参与经济社会活动的个人和机构的信用信息，并逐步推动信用信息在全国范围的互联互通，实现信用信息采集的规范化。本著作在总结吸收中国人民银行征信中心和部分省（市）公共信用信息指标体系研究成果的基础上，提出了具有可操作性的《企业公共信用信息采集规范》和《个人公共信用信息采集规范》两个规范性文本，构建起公共信用信息采集的基础工程。

六是明确提出加快推进公共信用信息可持续采集工作的五点政策建议。具体包括：以国情为基础彰显公共信用信息采集的中国特色；以需求为导向抢占公共信用信息采集战略制高点；以技术为依托保障公共信用信息采集的高效安全；以现实为考量推进公共信用信息采集的可持续性；以制度为保障优化公共信用信息采集外部环境。

公共信用信息广泛分布于各政府相关部门和社会组织机构之间，数据分散且数量繁多，要使今后信息采集工作规范化、标准化，顺利实现高质量的公共信用信息可持续采集与共享，就必须科学合理地设计公共信用信息的采集协议，中国人民银行应该同拥有公共信用信息的政府部门与社会组织充分协商，签署满足各自工作需求的双边合作协议，以建立权责更为明晰的合作机制，明确各部门的权利与义务。同时，由于各数据源单位数据库存在所采用的平台不同、数据库开发软件不同、数字化程度不一致等问题，征信机构与各数据源单位数据库无法进行直接对接。因此，在征信机构和各数据源单位数据库之间建设一个能兼容多种数据类型并顺利进行数据交换的信息采集与共享平台，实现异构数据源系统之间无缝对接，这对于加快数据交换效率、顺利实现公共信用信息的交换与共享是至关重要的因素。尽管在课题研究过程中我们看到了这些问题，但并未提出有效的解决办法，这是本研究报告最大的不足，只能留待以后的进一步研究中加以完善。

但客观而言，瑕不掩瑜，经过长达5年时间的努力耕耘，终于能够有今天的成果，非常感谢所有参与本研究成果调研、论证、写作等环节的团队成员辛勤的付出，他们是张学陶、张磊、陈佐、成程、孙杨斌、阳旸、邹子昂、朱桑之、刘姝雯、方敏、王芍、熊军凯等；衷心感谢国家发展改革委原副主任徐宪平同志，湖南大学原党委书记刘克利同志，中国人民银行征信中心主任王煜同志，中国人民银行征信中心原主任王晓明同志，中共湖南省委宣传部副部长肖君华同志，湖南省人民政府原副秘书长黄卫东同志，中国人民银行长沙中心支行行长马天禄同志，中国人民银行长沙中心支行副行长徐涌同志、副行长王地宁同志，湖南省发展改革委副主任汤兹同志、蒋天海同志，

湖南省社会科学规划办公室骆辉主任、陈湘文副主任，中国人民银行征信中心李连三同志，中国人民银行长沙中心支行内审处李海林处长、征信处刘敏处长等同志在著作研究过程中给予的大力支持；衷心感谢中国人民银行征信管理局、中国人民银行征信中心、湖南省社会信用体系建设领导小组、湖南省发展改革委、湖南省社会科学规划办公室、中国人民银行长沙中心支行、湖南省人民政府金融办公室、湖南大学金融与统计学院、计划财务处、社会科学处、审计处等单位在本成果调研、论证、研究与结题过程中给予的鼎力支持；此外，还要特别感谢中国金融出版社刘小平主编和肖丽敏编辑在本书出版过程中所给予的无私帮助。

热诚期待广大关心和关注中国社会信用体系建设与发展的读者和各界朋友们对本书提出宝贵的意见和建议，以便我们在今后的进一步研究中加以修改完善。

杨胜刚

2018 年 3 月 20 日

目　录

第一章　公共信用信息采集与中国的现状 …… 1

一、公共信用信息 …… 1

（一）公共信息 …… 1

（二）公共信用信息 …… 3

（三）公共信用信息采集 …… 6

二、国内公共信用信息采集现状 …… 10

（一）中国人民银行征信中心公共信用信息采集现状 …… 10

（二）社会征信机构的公共信用信息采集现状 …… 13

（三）中国人民银行征信中心与社会征信机构对比分析 …… 15

（四）对中国人民银行征信中心公共信用信息采集的影响 …… 16

第二章　征信机构的发展趋势与信用信息采集 …… 17

一、大数据时代行业发展前沿 …… 18

（一）大数据时代及其发展 …… 18

（二）大数据时代金融业的发展前沿 …… 19

（三）大数据时代征信业的发展前沿 …… 21

二、大数据时代征信机构的发展趋势 …… 23

（一）大数据时代征信机构发展的基本分析 …… 23

（二）大数据时代我国征信机构发展取向 …… 28

三、大数据时代公共信用信息采集的机遇与挑战 …… 32

（一）我国公共信用信息采集的现实判断 …… 32

（二）大数据时代公共信用信息采集面临的机遇 …… 33

（三）大数据时代公共信用信息采集面临的挑战 …… 35

第三章　公共信用信息采集的国际经验：国别比较 …… 39

一、美国经验 …… 39

（一）公共信用信息的内容与来源 …… 39

（二）公共信用信息采集的法律保障 …… 43

（三）公共信用信息采集方式 …… 45

（四）小结 …… 46
二、欧洲经验 …… 47
（一）公共信用信息的内容与来源 …… 47
（二）公共信用信息采集的法律保障 …… 51
（三）公共信用信息的采集方式 …… 54
（四）小结 …… 57
三、日本经验 …… 57
（一）公共信用信息的内容与来源 …… 58
（二）公共信用信息采集的法律保障 …… 60
（三）小结 …… 62
四、美国、欧洲、日本公共信用信息采集比较 …… 62
（一）相同点 …… 63
（二）不同点 …… 64
五、国际经验的启示 …… 66
（一）建立公共信用法律体系 …… 66
（二）加快公共信用环境建设 …… 68
（三）明确公共信用主体和范围 …… 69

第四章　公共信用信息采集的国际经验：技术分析 …… 70
一、国外公共信用信息内涵 …… 70
（一）征信内涵及起源 …… 70
（二）公共信用信息内涵 …… 71
二、国外个人征信系统公共信用信息采集经验借鉴 …… 72
（一）信息来源 …… 72
（二）信息分类 …… 73
（三）采集渠道 …… 75
（四）采集依据 …… 78
（五）共享方式 …… 80
（六）授信应用及效果 …… 82
三、国外企业征信系统公共信用信息采集经验借鉴 …… 89
（一）信息来源 …… 89
（二）信息分类 …… 89
（三）采集渠道 …… 90
（四）采集依据 …… 91
（五）共享方式 …… 93
（六）授信应用及效果 …… 93

四、国外公共信用信息采集的启示和建议 …… 95
（一）公共信用信息采集体系建设框架 …… 95
（二）采集的诊断式程序设计阶段 …… 97
（三）公共信用信息采集的实施阶段 …… 97
（四）公共信用信息采集的发展阶段 …… 98
（五）发达国家征信局公共信用信息采集的经验和启示 …… 99

第五章　我国公共信用信息采集的实践经验 …… 103
一、地区社会信用体系建设及其公共信用信息采集概况 …… 103
（一）地区社会信用体系建设概况 …… 103
（二）公共信用信息采集概况 …… 105
二、公共信用信息采集的主体、对象、内容与方式 …… 105
（一）个人公共信用信息的采集 …… 105
（二）企业公共信用信息采集 …… 109
三、公共信用信息采集的经验与教训 …… 114
（一）公共信用信息采集的经验 …… 114
（二）公共信用信息采集的教训 …… 116
四、典型省（市）公共信用信息采集实践 …… 117
（一）浙江省 …… 117
（二）陕西省 …… 119
（三）上海市 …… 121
（四）深圳市 …… 123
（五）江苏省 …… 125
（六）辽宁省 …… 127
（七）黑龙江省 …… 131
（八）湖南省 …… 134

第六章　公共信用信息采集与处理技术 …… 137
一、信息采集技术 …… 141
（一）网络爬虫 …… 142
（二）网页信息抽取技术 …… 147
（三）Web 信息检索、搜索引擎 …… 149
（四）应用软件 …… 152
二、信息清洗技术 …… 155
（一）征信信息清洗研究现状 …… 155
（二）征信信息清洗对象 …… 156

（三）征信信息常用的清洗原理 …… 157
（四）征信信息常用的清洗算法和工具 …… 159
（五）征信信息清洗评估 …… 161
（六）征信信息清洗应用展望 …… 161
三、信息匹配技术 …… 162
（一）征信信息匹配的概念和原理 …… 162
（二）征信信息匹配的意义 …… 166
（三）征信信息常用的匹配流程 …… 167
（四）征信信息清洗匹配技术 …… 169
（五）征信信息属性值相似度匹配介绍以及常见问题 …… 171
（六）征信信息有标识字段的实体匹配方法 …… 175
（七）无标识字段的自动实体匹配方法 …… 178
（八）征信信息匹配应用前景 …… 180

第七章　公共信用信息可持续采集机制研究 …… 182
一、完善多层次的征信机构体系和建设信息透明的征信环境 …… 182
（一）以市场化为目标的多层次征信机构体系的完善 …… 182
（二）以促进信息透明为重点的征信环境建设 …… 184
二、明确信息采集的相关主体及采集范围 …… 186
（一）信息采集的相关主体 …… 187
（二）采集范围 …… 189
三、统一公共信用信息采集与共享平台 …… 191
（一）公共信用信息采集协议 …… 191
（二）建立公共信用信息交换与共享平台 …… 193
四、规范公共信用信息行为 …… 197
（一）信息提供者的行为规范 …… 197
（二）征信机构的行为规范 …… 199
（三）信息使用者的行为规范 …… 200
五、完善公共信用信息采集的法律法规体系 …… 201
（一）国内征信法律法规现状 …… 201
（二）国外征信法律法规制度的借鉴 …… 204
（三）构建中国特色征信法律制度 …… 207

第八章　公共信用信息采集路径的选择 …… 211
一、公共信用信息采集原则 …… 211
（一）统一管理原则 …… 211

（二）先易后难原则 …… 211
（三）标准兼容原则 …… 212
（四）适度采集原则 …… 212
（五）法律保障原则 …… 212
二、公共信用信息采集方法 …… 213
（一）公共信用信息采集的发展趋势 …… 213
（二）公共信用信息主要采集方式 …… 214
三、公共信用信息采集范围 …… 223
（一）信息采集范围确定 …… 223
（二）企业公共信息具体采集范围及频率 …… 224
（三）个人公共信息具体采集范围及频率 …… 225
四、公共信用信息采集路径及指标体系 …… 226
（一）个人公共信用信息采集路径及指标体系 …… 226
（二）企业公共信用信息采集路径及指标体系 …… 242
五、公共信用信息采集难点 …… 258
（一）部门间协调难 …… 258
（二）信息采集标准不统一 …… 259
（三）信息质量得不到保障 …… 259
（四）信息采集不可持续 …… 259

第九章　信息采集技术在公共信用信息采集中的应用 …… 260
一、应用背景 …… 260
二、主要问题 …… 261
（一）征信信息主动搜索模块的主要问题 …… 261
（二）征信信息被动采集模块的主要问题 …… 262
三、关键技术 …… 263
（一）网络爬虫技术 …… 263
（二）开放式数据库互联技术 …… 264
（三）XML 中间件技术 …… 265
（四）Web Services …… 265
四、采集方案 …… 266
（一）征信信息主动搜索采集方案 …… 266
（二）征信信息被动采集方案 …… 280
五、可行性分析 …… 281
（一）经济可行性 …… 281
（二）技术可行性 …… 281

（三）网络可行性 …… 281
（四）使用可行性 …… 282

第十章　我国公共信用信息采集指标与规范 …… 283
一、公共信用信息指标体系设计理论及实证基础 …… 283
（一）公共信用信息指标体系设计理论基础 …… 283
（二）公共信用信息指标体系设计实证基础 …… 284
二、个人公共信用信息的分类思路 …… 287
三、个人公共信用信息采集内容 …… 288
四、个人公共信用信息采集指标设计及说明 …… 288
（一）基本信息说明 …… 289
（二）公共信息说明 …… 289
（三）提示信息说明 …… 289
（四）其他信息说明 …… 290
五、企业公共信用信息采集指标体系及说明 …… 291
（一）企业公共信用信息采集内容 …… 291
（二）企业公共信用信息采集指标设计及说明 …… 291

附录一　个人公共信用信息指标体系及规范 …… 298

附录二　企业公共信用信息指标体系及规范 …… 318

参考文献 …… 345

插图索引

图2-1：大数据技术工具 …… 23
图4-1：征信体系的主要参与者 …… 71
图4-2：按目标申请接受率分类的严重逾期率 …… 83
图4-3：按目标申请接受率分类的逾期率（使用电信支付信息计算出的结果）（Vantage Score 评分模型） …… 83
图4-4：征信体系建设：关键性因素和设计领域 …… 96
图4-5：诊断式程序设计——关键活动及时间段 …… 98
图4-6：数据提取的传统方法 …… 99
图4-7：数据提取的创新方法 …… 100
图4-8：公共信用信息采集发展阶段指导方案 …… 100
图6-1：JDL 由五个不同的功能部分组成 …… 140
图6-2：JDL 五个不同功能实现的逻辑流程 …… 141
图6-3：通用爬虫框架描述 …… 143
图6-4：Heritrix 系统框架 …… 145
图6-5：Heritrix 处理一个 URL 的流程 …… 146
图6-6：整体深层搜索的架构及流程 …… 147
图6-7：基于概念的信息检索模型 …… 151
图6-8：Goonie 信息采集系统 …… 153
图6-9：数据清洗大致框架 …… 158
图6-10：实体信息主题转变 …… 164
图6-11：复杂结构实体 …… 165
图6-12：征信信息匹配图形 …… 168
图6-13：字符串相似度算法总体框架 …… 175
图6-14：征信信息的匹配过程 …… 180
图7-1：基于联邦数据库的解决方案 …… 194
图7-2：基于数据集成中间件的解决方案 …… 194
图7-3：基于网关式数据交换控制器的解决方案 …… 195
图7-4：公共信用信息共享平台技术架构 …… 196
图7-5：信用系统网络结构 …… 196

图 7 -6：信用系统应用流程 …… 197
图 9 -1：分布式爬虫单个节点工作原理 …… 266
图 9 -2：爬行线程 …… 267
图 9 -3：全分布式非结构化拓扑结构 …… 268
图 9 -4：爬行过程中各个模块之间的关系 …… 268
图 9 -5：总体流程 …… 271
图 9 -6：中间件在三层结构中的位置 …… 272
图 9 -7：整体架构 …… 273
图 9 -8：ETL 过程和元数据的关系 …… 277
图 10 -1：按目标申请接受率分类的严重逾期率 …… 286
图 10 -2：按目标申请接受率分类的逾期率 …… 286
图 10 -3：所有消费者全样本信用评分分布情况 …… 287

附表索引

表1-1：公共信用信息分类 …… 3
表1-2：中国人民银行征信中心与社会征信机构对比分析 …… 15
表2-1：公共征信机构运作模式和私营征信机构运作模式的比较 …… 26
表3-1：来源于美国政府部门的可采集公共信用信息 …… 40
表3-2：按照分类查询 …… 42
表3-3：美国国内的公共记录查询 …… 43
表3-4：美国邓白氏公司的信息来源 …… 45
表3-5：英国政府机构可获取公共信用信息 …… 48
表3-6：法国政府部门及其公共信用信息 …… 50
表3-7：欧洲各国征信监督及消费者保护框架 …… 53
表3-8：日本行业协会会员报送的信用信息 …… 59
表3-9：来源于日本政府部门的可采集公共信用信息 …… 59
表3-10：三个国家（地区）公共信用信息采集比较 …… 66
表4-1：Equifax 公共记录代码 …… 75
表4-2：Experian 和 Trans Union KOB …… 75
表4-3：个人公共信用信息采集渠道 …… 77
表4-4：美国征信法规概览 …… 79
表4-5：基于不同数据报告行为分类的可获取资源 …… 81
表4-6：益百利主要征信产品及服务 …… 84
表4-7：环联主要征信产品及服务 …… 86
表4-8：艾可飞主要征信产品及服务 …… 88
表4-9：企业公共信用信息采集渠道 …… 91
表4-10：监督和执法机构介绍 …… 92
表4-11：邓白氏主要征信产品及服务 …… 94
表5-1：中国人民银行个人征信系统公共信用信息采集对象 …… 107
表5-2：部分省份企业公共信用信息对象 …… 112
表6-1：四种主要采集方式的对比 …… 138
表6-2：匹配精确度统计 …… 173
表6-3：输入同音字错误占漏配实体对比例统计 …… 174

表7-1：中国人民银行征信规章汇总 …… 202
表7-2：美国征信法律法规汇总 …… 205
表7-3：欧盟征信领域隐私保护法律 …… 206
表8-1：个人公开信用信息 …… 216
表8-2：企业公开信用信息 …… 217
表8-3：企业各类公共信息采集汇总 …… 224
表8-4：个人各类公共信息采集汇总 …… 225
表8-5：通过政府行政机关和司法机关采集的个人公共信用信息 …… 226
表8-6：通过事业单位采集的个人公共信用信息 …… 240
表8-7：通过企业单位采集的个人公共信用信息 …… 240
表8-8：通过非营利性组织采集的个人公共信用信息 …… 242
表8-9：通过政府行政机关和司法机关采集的企业公共信用信息 …… 242
表8-10：通过事业单位采集的企业公共信用信息 …… 256
表8-11：通过其他企业单位采集的企业公共信用信息 …… 257
表8-12：通过非营利性组织采集的企业公共信用信息 …… 257
表10-1：特征变量的最终分组结果 …… 284
表10-2：测试结果 …… 285