社会智能与复杂数据处理

舆情计算方法与技术

◎饶　元　编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

社会智能与复杂数据处理

舆情计算方法与技术

饶　元　编著

阜阳师范学院图书馆
藏书

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

内 容 简 介

社会计算是近十年来快速发展起来的一个新兴的研究领域，它一方面依托于近年社交网络技术及应用的快速发展，使得越来越多的网络用户之间产生自联接、自媒体、自选择的内容传播新方式，并且带来了关于针对文本内容深入挖掘与分析研究的强大的动力基础；另一方面传统社会学对于社会活动领域中的分析方法，特别是基于网络化的社会化分析方法，使得人们发现在庞大的网络数据中可以充分地利用其中的一些指标与算法来进行有效的度量与分析，从而使得社会网络分析方法从其他的角度上再次获得了新的生命力。本书则是在这两个研究背景下，结合西安交通大学软件学院社会智能与复杂数据处理实验室近三年来的研究成果，对社会智能与复杂数据处理领域中有关社会舆情分析过程中所采用的核心技术、方法和机制进行了系统的梳理，以期望从技术角度为读者提供一个深入学习和了解该领域前沿动态与关键技术的一个视角。

本书不仅可以作为计算机领域、数据挖掘领域、自然语言处理领域及信息管理领域的高年级本科生、研究生的课程教学书与辅导参考书，还可以作为专业领域软件工程师与设计师进行深入技术研究与算法优化的工具参考书。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

舆情计算方法与技术/饶元编著.—北京：电子工业出版社，2016.8
（社会智能与复杂数据处理）
ISBN 978-7-121-29626-0
Ⅰ. ①舆… Ⅱ. ①饶… Ⅲ. ①数据收集－技术 Ⅳ. ①TP274
中国版本图书馆 CIP 数据核字(2016)第 185756 号

策划编辑：甄文全
责任编辑：王凌燕
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1 000　1/16　印张：22.25　字数：460 千字
版　　次：2016 年 8 月第 1 版
印　　次：2016 年 8 月第 1 次印刷
定　　价：88.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254760；QQ：112670423。

在繁杂中寻找简单，在喧闹中寻找和谐，机会就在困难的中央！

——爱因斯坦

前　言

在 CMU 访学的日子是这样的漫长和宁静，远处在匹兹堡城南一角，每天在考虑着自己的生存模式的同时，还时时被来自国内的信息所包围着、簇拥着，深深地感受到“世界上最遥远的距离就是没有网络”。但是，让自己印象最深刻的场景不是新鲜的空气和美丽的天空，也不是让我心生恐惧的 Pizza 大餐，而是当我在实验室内部进行的工作汇报中，刚说到自己希望完成这样一本书时，看到合作导师 Alex 教授一脸茫然和不确信。能完成吗？我自己同样也在问自己，这是一个巨大的、辛苦的工作，因为在每一个章节中都存在着巨大的挑战与工作量，尽管在每一章节中我们实验室都有一些相关的工作基础，但是其中大量的背景、算法及算法应用，如何在短短的篇幅中能够尽可能地介绍清楚，这对于我而言绝对是一次重大的挑战。

社会计算是近十年来快速发展起来的一个新兴的研究领域，它一方面依托于近年社交网络技术及应用的快速发展，使得越来越多的网络用户之间产生自连接、自媒体、自选择的内容传播新方式，并且带来了关于针对文本内容深入挖掘与分析研究的强大的动力基础；另一方面，传统社会学对于社会活动领域中的分析方法，特别是基于网络化的社会化分析方法，使得人们发现在庞大的网络数据中，可以充分地利用其中的一些指标与算法来进行有效的度量与分析，从而使得社会网络分析方法从其他的角度上再次获得了新的生命力。在这两股力量的共同作用与影响下，大量社会计算领域中的研究成果也在不断地涌现出来。特别是我们自从 2012 年起依托西安交通大学软件学院，建立了一个社会智能与复杂数据处理实验室，我们将研究和关注的目光聚焦到了这个让人兴奋的领域，同时，为了更好地研究和处理相关的数据，我们选择了几个特定的应用领域，一是针对社会网络中的舆情与涌现现象的分析；二是针对利用网络中的舆情与情感来分析金融证券，特别是股票与基金的风险预测；三是利用这些网络中存在的信息资源，如何能够更好地通过内容、实体及关系的挖掘，寻找出一个基于知识内容聚合的新的知识服务机制，如企业与专家之间知识服务合作与匹配的新机制与新模式。这些问题都让我们渴望从社会网络计算的过程中去探索并寻找出一些关键的技术与分析问题的方法；同时，也希望在针对这些非结构化的复杂数据处理过程中，为探索社会智能应用场景下的相关技术、方法和机制奠定研究基础。

但是，每次看到我们新加入的研究生们对这一个研究领域与工作方向的茫然和低水平的重复，也促使我努力来做这样一个尝试，即从社会计算领域中的关键问题

此为试读，需要完整PDF请访问：www.ertongbook.com

出发，特别是在社会舆情分析过程中所遇到的有关文本分析、分词、特征抽取、主题识别与追踪、内容的分类与聚类、摘要的生成、社会情感，以及所涉及的文本挖掘、自然语言理解和机器学习中相关的算法进行一个梳理，形成进入这个研究领域的敲门砖，帮助他们用最快的时间来建立一个领域的研究体系与技术框架，为他们未来的研究与技术生涯奠定坚实的基础。因此，这就需要我必须静下心、坐下来开始了这样一个痛苦的、煎熬的研究和写作过程。

在此过程中我阅读了大量的研究生论文、专著文献及视频资源，其中特别是斯坦福大学、CMU 大学及台湾大学的在线机器学习课程，曾经为了深入理解 CRF 算法的机制，在 YouTube 中浸泡了两天；为了进一步了解 Deep Learning 在文本挖掘中的应用，也曾详细研读了 CMU 许多教授的内部视频。同时，年过 40 还不得不重新学习高等统计学与随机过程，这个过程看似艰辛，但个人内心却是充满了一些收获的喜悦。

但是由于时间关系及个人水平与能力有限，本书内容还存在许多不太如意的地方，特别是对目前本领域最新的前沿研究的处理相对较弱，有许多好的算法因内容篇幅的原因也没有展开，仅仅是提供了一个算法的初步实现框架；同时，也可能在文中存在一些这样或那样的不足、缺陷、甚至是错误，但是希望能够通过本书来抛砖引玉，获得更多领域专家与从业同行们的批评与指正，从而也可以不断地更新与完善本书的内容体系，最终为读者呈现更具有价值的内容，这也是我们一直以来不断努力的奋斗目标。

感谢 Alex 教授给我提供的一个宽松的环境与氛围，使我不仅有机会听到图灵大师们精彩的报告与演讲，同时也有机会看看 CMU LTI 中的 Pioneer 们的创新与思考；还可以通过多场博士答辩会（defense）了解到一流博士工作的问题来源与解决问题的工作方法；更重要的是让我还有一个重新回到教室上课听讲的机会。这一切工作汇总成了我的访学生活及学习研究的动力。特别感谢 LTI 的 Infomedia 实验室的一群年轻优秀而又勤奋有加的博士和博士后们，与各位一起参加会议讨论、农场干活劳动和节日聚餐的日子都是一次次难得的也是难忘的人生体验。

还需要感谢的是来自交大的几位年轻的校友，特别是刘佳鹏和周抒睿两位博士，也正是因为有了你们的协助，使我少了许多初来美国时的各种不适，特别是在外与你们聚会和共处的日子，都像是我的节日，幸福其实很近也很简单；也特别感谢雷雪辰博士的帮助，感谢他的睿智与幽默及对创新的理解与思考。感谢邱晨博士专程驱车来看我，并和我沟通了许多博士求学过程中存在的问题、思考与建议。同时，也感谢戴涛博士，默默地、努力地协助我管理了整个实验室，保证了它的基本正常运行。另外，要感谢交大社会智能与复杂数据处理实验室里的那一群可爱、有为的青年学俊们，无论与钟旭辉、冯妮、王智民、毛伟宁、张伟奇，还是闫瑞涛、宋明

爽、刘笑天、赵亚丽……与你们每一位同学共同奋斗的岁月，让我虽早生华发但心里还仍然觉得自己还年轻。

此外，需要感谢软件学院的多位老师和领导，特别是曾明老师，也是他把我引入到了这一个充满挑战和创造性的研究领域之中，在各位老师与领导对我们工作的鼎力支持下，才有了今年这一点点初级的“成果”，相信假以时日，我们应该能够通过更多的人才培养与深入的研究工作，来回报学院所提供的各种基础条件，真心地希望我们的学院在未来的国际与国内竞争中越来越有影响力，也希望我们实验室在将来的某一天能够真正成为一个国际化的研究平台。同时，也需要感谢本书的责任编辑，正是他们专业与细致的工作，才使得本书这么快便完成样稿审订与出版发行。此外，本书的工作得到了 2015 年度陕西省科技厅陕西省协同创新计划（2015XT-21）——基于网络化感知计算的智慧社区管理关键技术应用示范研究项目；2015 年西安市科技局产学研协同创新计划（CXY1514（5））——基于内容挖掘的专利地图可视化服务的关键技术研究项目；国家社科基金重大项目“基于多学科理解的社会网络分析模型研究”第 5 子课题：虚拟 Web 网络空间中的社会网络模型与个体行为机制研究”等项目的联合资助，同时也感谢国家留学基金委所给予的出国资助。

感谢我的房东朱焕老师和张青老师，两位旅美多年的音乐家热心地给予了多方面的关照，让我也尽可能快速地融入到这个社区，了解到与已往完全不同的生活模式。同时也感谢罗敏女士等诸多旅美华人朋友对我们所提供的大量热心的帮助。

最后，感谢与我同在彼岸但居住两地的妻子，以及天天与我做伴、每天都有新故事的儿子，没有他们的支持很难在这么短的时间内完成这一个充满挑战的任务。同时，也将此书作为送给年近八旬的老父老母的一份小小礼物，祝愿亲人们健康幸福。

海内知己，天涯毗邻。

于美国卡内基梅隆大学－语言技术研究所
信息媒体实验室（Lab of Infomedia）
2015 年 9 月

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第 1 章　社会计算与社会舆情分析概述

伴随着 Web2.0 技术的迅速发展和广泛应用，因特网对人类社会的交往产生了深远的影响，人们利用网络不仅突破了对传统社会关系的理解，并且利用网络工具来构建一个更为广泛的社会生态环境，从而将物理世界中的现实社会和网络中的虚拟社会相互融合，促进了社会计算（Social Computing）领域相关技术与应用的快速发展。同时，作为计算机网络技术、社会科学及心理学等多领域之间的新兴交叉学科，社会计算正在深刻地改变着人与人、人与社会之间的交互模式，特别是通过去中心化、社会化、开放性、创造性、自下而上的新模式将互联网的主导权归还给普通的网络用户，并在各种社会化软件的协助下，一方面每一个用户创造的内容通过社会关系网络来实现信息的快速传播，改变了人们对信息获取、发布、分析和利用的传统渠道与模式；另一方面通过整合每个人在互联网上的各种社会资源来形成集体智慧与知识共享，从而促进了基于网络的社会智能的诞生。因此，本章在社会计算相关技术介绍的基础上，分析社会计算与社会智能领域的相关技术及社会舆情分析方法和应用之间的相互关系与影响，从而对复杂数据处理环境下的舆情分析与技术提供一个总体的概述，并形成舆情分析的总体技术框架与视图。

1.1　社会计算的概念与关键技术

1.1.1　Web2.0 与社会化软件的特征

O' REILLY 公司 CEO——Tim O' Reilly 在其公司的个人栏目里发表了名为《什么是 Web2.0——下一代软件设计模式与商业模式》一文中首次提出了 Web2.0 概念，他认为下一代网络的生命力主要来自于用户的积极参与，且这种源于用户自主贡献的网络效应才是 Web2.0 时代中统治市场的关键。这一点与基于传统 Web 1.0 的网络应用所强调的信息门户及对信息内容的集中式管理和控制相反，Web 2.0 从一开始就希望去中心化，打破信息的封闭与功能大而全，但无特色的、被动的应用管理模式，而是提倡以用户个人需求为核心，围绕用户个性化需求来提供集成、开放和针对性的信息服务，使得网络软件在支撑社会化应用的同时，也越来越多地具有了社会化的特征。

这种社会化的特征主要体现在：Web2.0 系统彻底改变了传统自上而下的由少数信息资源控制者集中控制和主导信息的网络管理体系。并采用自下而上的方式，通过用户参与、共享及集体智慧，实现了对传统互联网的管理理念和思想体系的变革与升级，极大地促进了普通用户对网络内容服务的贡献与创造力，以及对新技术的迫切需求。例如，传统门户中的信息发布，用户主要是对信息内容进行浏览；但是在 Web2.0 的条件下，用户不仅是信息内容的消费者，同时也是信息内容的创造者。在博客、播客、微博、微信、云存储等 Web2.0 应用（参见附录 A 所示的信息分类示意图）的支持下，越来越多的功能与应用使得用户对信息的使用方式与需求也发生了巨大的变化，这些变化直接推动了目前移动计算、云计算及大数据等技术和应用的发展与升级。表 1-1 从信息的控制、通信方式、信息发现、信息获取、内容的控制及技术等 10 个不同的角度对 Web 1.0 和 Web 2.0 进行了对比，从而可以清晰地发现两者之间的差异。

表 1-1　Web1.0 和 Web2.0 之间的差异比较

项　　目	Web1.0	Web2.0
网络管理方式	自顶向下（Top Down）	自底向上（Bottom Up）
交流方式	人—机交互（P2M）	机器—机器（M2M） 人—人（P2P）
信息的发现	浏览与搜索	发布与订阅
信息的获取	交易	关系共享
信息聚合	商业的聚集者，门户	微聚合（Micro-Aggregation）
市场推动模式	“推送”（Push）内容	“拉取”内容（Pull）Conversational, Personal
内容的控制者	新闻机构	内容作者本身
内容结构	网页和文本	标签对象
应用	私有的封闭的体系	基于标准的开放体系
技术	HTML, Solaris, Oracle	XML, AJAX, RSS

从表 1-1 中可以看出，Web2.0 在网络技术的基础上实现了人与人之间的广泛通信与交互，并在 BBS 论坛、Blog、微博、微信、Wiki 等社会软件（Social Software）的支撑下，使得人们可以更方便地建立起个人与朋友之间沟通与协助的虚拟网络空间。同时，在用户的广泛参与下，极大地促进了用户对网络内容的创作和贡献的潜力，直接推动了社会计算与社会软件技术与应用的发展，以及基于个性化与社会化的大数据时代的到来。其中，与传统的软件相比，社会化网络软件所具有的特征如下：

（1）互联网已从一个信息管理工具演化成为了一个公共平台，利用互联网已不再是为了单纯的统治和控制，而是为了更有效地促进交流和分享。

（2）平台的社会化，使人们更加充分地重视并利用网络集体的力量和群体智慧，促进了社会计算与社会智能的诞生。

（3）将数据变成“Intel Inside”，促进了大数据分析与应用时代的到来。

（4）广泛采用分享和参与的架构，促进了互联网的数据开放与传播，同时也驱动了互联网的社会化群体效应——众包的产生。

（5）通过接口的开放性与标准化，带动分散的、独立的开发者将不同的软件应用汇集并形成一个软件的生态环境。

（6）通过内容和服务的融合模式，促进了轻量级业务之间的分享机制的形成。

（7）注重用户体验的持续性服务（“永久的 Beta 版”）。

（8）有价值的服务成为了应用的本质，并且通过网络使得服务无处不在，改变了传统软件应用的单机版或单一平台版本概念与使用方式。

（9）不仅仅关注少数的重要用户，同时关注大量的普通用户，以及所形成的长尾效应。

因此，Web2.0 的特征归纳起来即为“主动”和“互动”的互联网。“主动”是指“以个人为中心”，开启了网络个性化的个人时代，个人深度参与到互联网中，并彼此相连；“互动”是指“以自组织为中心”，即个人与个人之间、个人创造的内容与内容之间及个人汇聚的群体与群体之间，越来越多的采用自组织的方式来架构，并通过自组织的方式让人、内容和应用等资源充分“流动”起来，并以这些网络资源价值最大化的方式来体现出应用的价值。

1.1.2　社会计算的概念

在各类社会化软件的应用中，越来越多的人利用网络工具建立起了个人与真实社会环境中的朋友及亲人之间进行沟通与交流的在线联系方式，随着用户数量的不断增加，带有真实物理社会特征的一个社会化虚拟网络也逐渐显示出了一种蓬勃的生命力。在这种虚拟网络中，不同人与人之间的交流内容、行为及情感特征，以文字、图片、视频及相应操作的形式记录下来，并成为人们重新审视与分析社会网络结构、演化、传播、情感等特征的一把新钥匙。2009 年 2 月，美国哈佛大学大卫·拉泽（David Lazer）等 15 位美国学者在 *Science* 上联合发表了一篇具有里程碑意义的文章“Computational Social Science”标志了“计算社会科学”这一研究领域正式兴起，这也使得人们在前所未有的深度和广度上通过网络来自动收集和利用数据，为社会科学的研究提供深入的数据分析服务。

社会计算是将网络技术、复杂系统、数据挖掘、社会学、管理科学、自然语言处理、信息检索及心理学等多个学科之间进行相互融合并形成的一个新兴的交叉学科（王飞跃，2004，2005），它主要研究在利用互联网与计算机系统协助人们进行沟

此为试读，需要完整PDF请访问：www.ertongbook.com