面向用户创造内容的大数据加工方法与技术

蔡淑琴　周　鹏　胡慕海　张　宇　马玉涛　石双元　著

MIANXIANG YONGHU CHUANGZAO NEIRONG DE

DASHUJU JIAGONG FANGFA YU JISHU

科学出版社

国家科学技术学术著作出版基金资助出版

面向用户创造内容的大数据加工方法与技术

蔡淑琴　周　鹏　胡慕海
张　宇　马玉涛　石双元　著

科学出版社
北京

内 容 简 介

用户创造内容（UGC）作为互联网中大数据中的重要组成部分，使得信息资源在获取、传播、效率上表现了前所未有的优势。但是其碎片化、非结构化、无序化和去中心化，加剧了组织的“信息过载”和“信息迷失”，由此产生的网络舆情、网络热点、网络危机等成为互联网虚拟世界中组织面临的新问题。本书以在线客户评论、微博两类主流 UGC 为对象，提出了 UGC 序化、中性化加工思想，研究了 UGC 的加工方法与技术，包括在线客户评论的序化方法、基于产品族设计的加工方法；研究了面向危机事件识别的微博加工方法、面向网络热点的发现模型、面向网络广告定向的中心化方法；面向动态情境的信息推荐方法及系统、个性化移动内容服务的模型和支持技术。

本书可作为管理科学与工程、计算机应用等专业的研究生教材，也可作为从事大数据研究与应用人员的参考资料。

图书在版编目（CIP）数据

面向用户创造内容的大数据加工方法与技术/蔡淑琴等著. —北京：科学出版社，2016

ISBN 978-7-03-048747-6

Ⅰ. ①面… Ⅱ. ①蔡… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 131831 号

责任编辑：陈晓萍 马琦杰 / 责任校对：王万红

责任印制：吕春珉 / 封面设计：卓然进

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

三河市骏杰印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2016 年 5 月第 一 版 开本：787×1092 1/16

2016 年 5 月第一次印刷 印张：20 1/4

字数：468 000

定价：64.00 元

（如有印装质量问题，我社负责调换〈骏杰〉）

销售部电话 010-62136230 编辑部电话 010-62138978-2010

前　言

随着以互联网技术为核心的信息技术在科学研究、商业领域的广泛应用，大数据引起了越来越多行业的关注，而互联网中以 Web 2.0 技术为基础的用户创造内容（User Generated Content，UGC）是大数据中的重要组成部分。用户创造内容作为一类社会化媒体信息在传播速度、广度、效率上取得了前所未有的优势，改变了信息产生、传播和产生影响的方式，并以惊人的速度影响个人、企业乃至社会。但是 UGC 的碎片化、非结构化、无序化和去中心化，加剧了组织（包括营利性组织和非营利性组织）的“信息过载”和“信息迷失”，由此产生的网络舆情、网络热点、网络危机等成为互联网虚拟世界中组织面临的新问题。因此以人工智能理论和大数据理论为指导，针对用户创造内容的信息特征、网络结构和传播机理，构建挖掘加工方法与技术，成为大数据中亟待解决的问题。

本书以大数据中在线客户评论、微博两类重要的用户创造内容为对象，重点研究用户创造内容的内容、结构特性及加工处理方法，包括在线客户评论的序化方法、在线客户评论的产品族设计和序化方法、微博的危机事件识别、热点发现和广告定向、面向动态情境的信息推荐、个性化移动内容服务的模型和支持技术等主要内容。

本书的研究工作得到国家自然科学基金项目“微内容生产加工模式及其支持平台的研究”（编号 71071066）、教育部人文社会科学项目“基于互联网信息的企业危机事件识别研究”（编号 11YJA630098）的支持。本书的出版得到了国家科学技术学术著作出版基金资助（2015）。

本书总结了作者和企业商务智能工程研究所历年来的研究成果，吸取了国内外同行的研究成果和有关文献的精华，在此谨向这些成果所有者和文献的作者表示感谢，他们的丰硕成果和贡献是本书学术思想的重要源泉。本书能够顺利撰写与出版还得益于企业商务智能工程研究所全体成员的贡献和工作成果，如张静、邱洁、段磊、邓运、谭婷婷、王旸、崔晓兰、王艺兴、王文龙等，在此一并表示感谢。

目　录

第1章 导 论

1.1 问题背景

1.1.1 大数据中的用户创造内容

随着以互联网技术、计算机技术及传感技术为核心的信息技术的广泛应用，人类第一次面临大数据的挑战和机遇，而以 Web 2.0 思想和技术为支撑的用户创造内容（User Generated Content，UGC）是大数据中的主要组成部分。

根据维基百科的定义，用户创造内容是指网络或其他开放性媒体上由网民生成的信息。从 2005 年开始，互联网上的许多网站使用 UGC 的方式提供服务，如图片、视频、博客、播客、论坛、评论、社交、Wiki、问答、新闻、研究类的网站。著名科幻作家 William Gibson 在 1996 年预言了职业博客的影响；《自然》杂志（2005）的研究表明由用户创建和编辑的在线百科全书“维基百科”在准确性上超过了《大英百科全书》，且在条目数量上也大大超出；在欧洲，超过一半的在线用户成为博客、播客、在线客户评论等 UGC 的生产者或需求者，或成为互联网中的社会网络参与者；我国个体网民也正成为我国互联网内容的主动传播者、作者和生产者。提供以 UGC 为源的 Web 内容生产加工与服务的互联网企业成为互联网行业中迅速发展的一类新型企业。

UGC 即网民将自己的内容通过互联网平台提供给其他网民。在这里，网民不再只是信息的需求者，他们作为互联网信息资源价值网的重要节点，将自己的喜好体验、经验教训、知识和信息通过互联网提供给他人，希望与他人分享，产生了一种新的网络交流方式。UGC 来自网民，每一个网民都可以生成自己的内容，互联网上的内容不再只由少数组织生成、控制，形成了一个多、广、专的局面。

与基于 Web 1.0 技术运营的互联网网站上的内容相比，基于 Web 2.0 技术的 UGC 是一种草根文化，它是众多网民智慧的反映。

互联网技术和 Web 1.0 技术的应用，只解决了人们在互联网上获取信息的需求。以博客（Blog）、社区网络服务（SNS）、聚合内容（RSS）等为代表的 Web 2.0 技术的应用，解决了众人自我实现的需求。他们不仅能够通过互联网获取信息，并且能够参与其中，产生和发布信息，并与他人共享信息。

随着信息技术的发展，大众可以利用 PC、手机、移动终端等随时随地接入互联网、上传 UGC，因此它具有体量大、产生速度快、非结构化等特征，是一类非常重要的大数

据，它丰富了互联网中的信息资源。

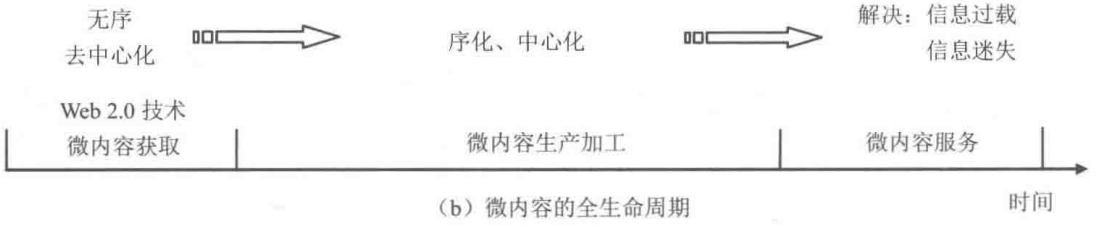
随着 Web 2.0 技术的深入应用，UGC 传播在时间、空间、效率上已经逐渐展现出非常显著的优势，改变着人类信息生产、加工、传播的方式，并以惊人的速度渗透到社会的各个方面。

1.1.2 UGC 价值实现的大数据加工问题

Web 2.0 技术解决了内容的原创、规模问题，但是产生的无序、去中心化 UGC 带来了内容的信息碎片，加剧了网民在互联网世界中的“信息过载、信息迷失”[图 1-1（a）]。以 UGC 为源的 Web 内容作为一类信息/知识型产品，其增值来自于对 UGC 的生产加工和使用。如何以合理成本对 UGC 进行生产加工，提供满足不同客户需求的 Web 内容、应对产品同质化竞争是其面临的一个严峻的问题。

互联网企业根据实际具体问题设计了一些解决方案，如内容聚合、广告位精细管理等，但是这些均属于个案。由于缺乏系统性理论的支持与指导，互联网企业仍被这些问题困扰着。随着互联网企业成为经济社会中越来越重要的组成部分，亟须与 UGC 以及其生产加工相适应的理论和方法的支持[图 1-1（b）]。这已成为业界关注的热点问题，也正成为学术界在互联网企业管理研究领域中的热点问题。

（a）互联网中的微内容现象/问题



（b）微内容的全生命周期

图 1-1 互联网中的 UGC 加工

信息碎片影响人们使用信息，并且制约信息价值的实现，因此一直是学术界关注的热点问题。例如，数据库技术的规范化解决了结构化数据的存储问题，但是带来的信息碎片影响决策用户使用，而数据仓库技术的反规范化解决了信息碎片问题，满足了决策用户的需求。Web 2.0 技术解决了信息资源的获取问题，但是 UGC 的信息碎片使网民、企业“信息过载、信息迷失”。到目前为止，技术性、单项应用的研究较多，UGC 无序和去中心化的理论研究、将 UGC 的信息处理与产品生产加工运作相结合的理论研究甚少，将管理和信息技术结合的系统性理论研究不多。以 UGC 为源的 Web 内容的非结构化使得其信息碎片的解决更难，是企业在互联网世界信息/知识管理面临

的难题，需要从信息/知识管理、企业生产与运作管理、系统科学相结合的综合视角进行深入研究（图 1-2）。

图 1-2 UGC 信息碎片、加工的研究现状

本书将以信息管理、知识管理、生产运作与管理、系统科学、超图理论等为基础，以提供 UGC 服务的互联网企业为背景，以解决 UGC 的碎片化为切入点，在序化、中心化的基本框架指导下，研究 UGC 生产加工模式及平台的相关基础理论与方法。

本书旨在为上述问题的解决建立相应的理论和方法，在互联网企业 UGC 生产加工与管理理论与方法上有实质性的突破，所研究的理论成果也为政府的互联网监管、传统出版行业的出版模式变革等提供可借鉴的理论与方法。

1.2 几类典型的 UGC

1.2.1 在线客户评论

1. 在线客户评论的定义

在线客户评论（Online Customer Reviews，OCR）是随着互联网的出现而产生的一种特殊口碑类型（Godes，2004），是从传统的线下口碑演化而来。要进行 OCR 加工，首先有必要深入了解 OCR 的本质，对其进行清晰的界定。已有不同学者给出过 OCR 的定义，但这些定义很少对 OCR 追根溯源进行系统的分析，此外，学者们大多站在各自特定的视角，而缺乏从信息加工的视角对 OCR 的界定。表 1-1 对从传统口碑演化到 OCR

过程中学者们提出的相关代表性定义进行了系统的对比分析。

表 1-1 前期口碑相关代表性定义的比较

文献	正式	评论主体	评论对象	评论极性
Arndt，1967	N	非商业传播者	产品，品牌，服务	
Richins，1983		消费者		负面
Westbrook，1987	N		产品，服务，卖家	
Haywood，1989	Y	企业		
File 等，1992		消费者	销售人员，经销商，制造商	正面，负面
Tax 等，1993	N	消费者	供应商，产品，服务	
Bone，1995		消费者		
Helm 和 Schlei，1998 Mangold 等，1999	N	卖家，专家，消费者		正面，负面
Emanual，2000			产品，服务，企业，品牌	正面，中立，负面
Hanson，2000			产品，服务	
Chatterjee，2001		消费者	产品，品牌	
Thurau 等，2004		消费者	产品	正面，负面
Litvin，2008		消费者	产品，服务	
Park 和 Kim，2008		消费者	产品	正面，负面

注：Hanson 和 Litvin 特指在线口碑，Chatterjee 特指网络口碑，Thurau 等特指电子口碑，Park 等特指 OCR，其他则都指口碑。

从表 1-1 中可以看出，Arndt 等学者强调口碑是非正式的信息交流或沟通，是一种未经正式组织的行为，而仅有 Haywood 认为口碑是企业正式的信息交流。虽然学者们对口碑相关的定义各不相同，但通过比较分析可以看出均包括如下三个基本的要素：

1）口碑传播者。Arndt 提出是非商业传播者，Haywood 特指企业，而 Helm 等则指出包括卖家、独立专家和消费者，除这些少数学者持有不同观点以外，绝大多数学者认为口碑传播者是消费者。

2）口碑客体，即传播者评价的对象。大多数学者认为包含产品和服务，Westbrook 等提到了企业，而 Arndt 等提出还应包括品牌。

3）评价极性。Richins 认为口碑中传播者对口碑客体评价的个人情感倾向是负面的，而 File 等提出评价极性包含正面和负面，Emanual 则指出不仅应包括正面、负面，还应包含中立。此外，Thurau 强调口碑传播中意见领袖的突出作用；Haywood 则指出企业的代表性行为能成为口碑点；Mangold 等认为口碑是被需求所激发，这表明用户对口碑的需求引导着口碑的形成；Emanual 提出口碑是给定时间内的评述，这说明口碑具有动态性和时效性。

从以上分析而知，不同定义中的三个基本要素并无根本变化，只是在内容上有所拓展。OCR 与传统口碑最大的区别是传播媒介的不同。对互联网企业而言，从生产运作的

视角出发，在线口碑、网络口碑、电子口碑等，这些本质上都是互联网上需要生产加工的零散的消费者评论信息，可以统一为“在线客户评论”。

因此，本书对 OCR 界定：OCR 是指客户根据自身消费体验或他人购物经历，自发地在网络上以评分或文本的形式发布的关于产品、服务、品牌或企业的正面、中立或负面评价。

2. OCR 的成因

区分 OCR 的成因有助于从起源的角度分析 OCR，其成因包括用户发起的 OCR 和企业引导发起的 OCR。

（1）客户发起

1）基于体验的 OCR。客户发布 OCR，通常是由对自身的购买或消费体验的满意或者不满意所引发，这和现实环境下的传统口碑类似。正面 OCR 的形成，通常是由于企业的品牌、产品或服务使客户感到满意所触发。客户可能会认为，一次愉悦的购物经历应该用正面的 OCR 作为对企业的回报，也可能是出于对品牌的某种忠诚（Cheung，2009）。而如果客户经历了一次不满意的消费，客户对购物经历的负面印象可能导致心理的不平衡，会有发泄不满情绪的欲望，而通过发布负面 OCR 的方式，与其他客户分享自身的负面购物经历，可以帮助缓解客户的负面情绪（Thurau，et al.，2004）。而对购物经历极度不满或者极度满意的客户，比起那些中等满意的客户，更容易发布和传播 OCR。

2）基于分享的 OCR。由于 OCR 的丰富性和易得性，当网民无意或有意看到 OCR 后，往往会将自己感兴趣的或者认为自己熟悉的他人会感兴趣的 OCR 转发和分享给其他人。这种行为在微博用户中经常发生。当看到有趣的 OCR 时，用户可以轻松地通过“@”分享给自己的好友，而用户本身可能并没有 OCR 中描述的消费体验。从 OCR 特征来看，权威性高或趣味性更强的 OCR 更容易被客户所接受，同时，客户与好友分享 OCR 和再传播的意愿就越强。从分享 OCR 的传播者角度来看，意见领袖能力越强的传播者，将 OCR 分享给他人的意愿也就越强，而 OCR 搜索倾向越强的传播者，将 OCR 再传播的意愿越高。而从社会网络的角度分析，程度中心性越高的成员，分享 OCR 的可能性越大（Ghose，2011）。

（2）企业引导

许多企业并不只是被动地依靠客户发布正面 OCR，以向周围的人群推广企业的产品或服务，而是采取主动措施，积极引导广大客户对自身产品、服务或品牌的讨论和关注。企业所引导的 OCR，能视为传统营销和客户口碑的综合体，对传统营销而言，通常是由企业主动发起的，而口碑则通常是由客户主动发起的，口碑营销（Word-of-Mouth Marketing）则是首先通过企业发起，然后由客户来积极推动，以获得广泛的关注（Liu，2006；Walsh，2012）。口碑营销又称为病毒式营销，通常是首先在知名的网站或者人气很旺的社区发布 OCR，即植入病毒，而后利用网络推手来做初期的回复和转载进行推动，达到所期望的营销效果。前期发布者的意见领袖能力越强，社区的互动性、人气和扩散性越高，口碑营销的成功概率就越高。

（3）OCR 的影响

OCR 的内容和极性会对消费者购买决策和企业营销效果都有直接且显著的影响，OCR 作为一类重要的 UGC 成为中外学者们研究的热点。

1）对消费者购买决策的影响。2007 年 4 月，全球知名的市场研究公司 AC·Nielsen 针对不同类型的广告信任度进行了一次“尼尔森在线全球消费者研究”的调查评估（Pepe，et al.，2011），结果如表 1-2 所示。由表 1-2 可以看出，消费者对 OCR 的信任度已经超过电视、杂志和广播等传统媒介。

表 1-2 全球互联网使用者对不同类型广告的信任程度

类型	信任度/%
消费者推荐（Recommendation from consumers）	78
报纸（Newspapers）	63
网络上贴出的消费者意见（Consumer opinions posted online）	61
品牌网站（Brand website）	60
电视（Television）	56
杂志（Magazines）	56
广播（Radio）	54
品牌赞助（Brand sponsorships）	49
我收到的电子邮件（E-mail I signed up for）	49
电影前的广告（Ads before movies）	38
搜索引擎广告（Search engine ads）	34
网络旗帜广告（Online banner ads）	26
收集短信广告（Text ads on mobile phones）	18

Koh 等（2010）通过引入行为理论，对中国、美国和新加坡的在线电影评论进行了对比分析，得出文化背景会左右 OCR 对消费者的影响力的结论。Zhang 等（2010）引入调节焦点理论对 OCR 的说服效果进行了度量，其研究表明消费者的消费目标对说服效果有调节作用。Mahony 等（2010）探讨了 OCR 可读性与 OCR 有用性之间的关系，提出了一种监督分类方法，结果显示，对 Amazon 网站而言，可读性对有用性有显著影响，而对于 TripAdvisor 而言，影响很小。Park 等（2009）研究了消费者特征对在线评论感知有用性的影响，结果表明，消费者怀疑对感知有用性的影响负相关，消费者越怀疑，在线评论对购买决策的影响越小，而团购经验对感知有用性的影响正相关，团购经验越多的人，在线评论对购买决策的影响越大。

Mudambi 等（2010）探讨了 OCR 哪些属性能辅助消费者购买决策，如图 1-3 所示，研究显示评论极性和评论深度能影响 OCR 的感知有用性，大多数情况下，负面评论比正面评论感知有用性更大，而评论字数越长，感知有用性也更大，而产品类型作为调节变量，不同产品类型，OCR 的感知有用性存在较明显的差异。郝媛媛等（2010）基于影评数据分析了 OCR 有用性影响因素，结果表明正面极性、句子长度和主客观混杂表达等有显著的正面影响。

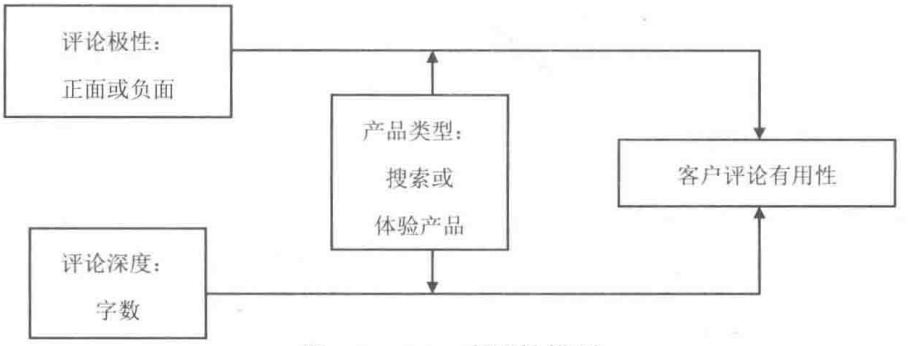


图 1-3 OCR 有用性模型

2）在线客户评论对企业经营绩效的影响。很多学者通过搜集 OCR 的数据，来对其影响企业经营绩效的程度进行深入的测量，总体而言，影响因素主要有三个，即 OCR 数量、OCR 极性和 OCR 离散度。

① OCR 数量。通常情况下，OCR 数量越多，表明有越多的客户投入企业产品或服务的评价之中，便越可能吸引到更多其他客户对该企业的注意（Dellarocas，2007；Gu，2012；Zhang，2012）。已有许多学者对 OCR 数量对企业经营绩效的正面效应做了实证性研究。Liu（2006）选取了雅虎网站（yahoo.com）上 20 部电影的 OCR 作为语料，研究结果表明无论是正面的还是负面的 OCR，OCR 数量的增加都会带来电影票房收入的增加。Duan 等（2008）的实证研究得出了类似的结论，即某电影的 OCR 数量与该电影的票房收入正相关。然而，也有学者得出不同的结论。卢向华等（2009）收集大众点评网的数据进行了实证分析，研究发现，OCR 数量对餐馆销售收入有着显著影响，不过受到价格的调节作用。对高价位的餐馆，OCR 数量太多，反而会对销售收入带来负面影响，而对于低价位的餐馆，OCR 数量的增加，会显著提升销售收入。

② OCR 极性。OCR 极性包含正面和负面两种，正面的 OCR 有助于提升企业经营绩效，而负面的 OCR 会给企业带来负面影响。Clemons（2006）以啤酒生产商为研究主体，收集两年的历史 OCR 数据，分析表明啤酒销量与正面 OCR 正相关，正面 OCR 越多，当期的销售量越高。Chevalier 等（2006）挑选亚马逊网站（amazon.com）上的书籍为研究对象，收集不同书籍的 OCR 进行分析，结果表明，正面 OCR 多于负面 OCR 的书籍比那些正面 OCR 少于负面 OCR 的书籍销量更好。

针对负面 OCR 对企业的影响，也有许多学者做了研究。DeCarlo（2007）引入归因理论，对零售企业的负面 OCR 影响进行了研究，结果显示，并不是所有的负面 OCR 均对企业收入带来负面影响，当合理组织负面 OCR，使其具有逻辑性时，能减弱负面 OCR 对企业的消极影响，甚至化危机为转机。这一点能在国内网站中得到印证，例如，携程网（ctrip.com）的口碑管理就允许酒店和旅游景点经营管理者针对客户的负面 OCR 进行在线的反馈，通过及时有效的沟通，不但能消除负面影响，还能吸引到更多的客户关注，扩大了影响力。Doh（2009）采用实验研究的方法，给每个参与者分配 10 条 OCR，并对其中的 5 组进行正面 OCR 比率的变换，结果发现一直获得正面评价和正面 OCR 占绝对比例的组织，更容易让人产生怀疑，同时会对该组织真实的信誉建立产生负面影响，

相反，有一定的负面 OCR，反倒让人对该组织更加信任。Sweeney（2012）研究了负面 OCR 对客户转换行为的影响，研究结果表明，负面 OCR 强度越强烈，越显著地影响客户的品牌转换行为，造成企业更多的客户流失。

③ 责制 OCR 离散度。OCR 离散度指的是 OCR 在不同群体间传播的程度，OCR 的离散度越高，说明 OCR 的影响范围越广（Moore，2012）。Clemons 等（2006）建立了一种基于 OCR 来评判产品市场影响力的方法，以啤酒的销售和相对应的 OCR 为分析数据，结果表明 OCR 的离散度对啤酒的市场接受度有着显著的影响。Berger 等（2011）通过收集 300 种产品在不同城市的 OCR 来分析 OCR 离散度对企业经营绩效的影响，结果表明越是趣味性强的产品越能快速获得更多的 OCR，但随着时间的推移，OCR 数量会迅速减少，反而，那些客户经常看见、方便购买的产品，OCR 的数量会随着时间的推移而逐渐增加。Berger 的研究结果启示企业管理者，应该尽可能持续地在当地投放经常能看得见的方便获得的赠品，这样能持续获得客户关注，扩大企业影响力，而不是靠短期的趣味产品来吸引眼球。

（4）OCR 实例

国内的 OCR 网站发展迅速，其中较有影响力的包括提供生活咨询类点评信息服务的大众点评网、口碑网等，以及提供图书电影类点评服务的豆瓣网、时光网等。

1）大众点评网。大众点评网（www.dianping.com）创建于 2003 年 4 月，是一家基于 Web 2.0 技术的第三方评论网站。网站的定位是为用户提供城市生活服务指南，主要致力于为中国消费者提供本地的餐饮、休闲、娱乐等生活服务发表评论、分享信息的平台。其用户定位是中国城市中具备一定消费能力的人士。其第三方点评模式吸引了千万网友的积极参与，由用户点评的包括餐饮、休闲、娱乐等生活服务商户已覆盖全国 300 多个城市 70 多万家，用户规模达到 1000 多万，且信息量和覆盖范围还在不断快速增长中。其中，餐饮点评是其发展最早也是目前消费者极为喜爱并聚集信息量极多的内容之一。大众点评网还提供优惠券下载、团购、手机服务等业务，为了提高会员黏性还建设社区，注册会员就可以申请大众点评网会员卡享受合作商户打折的优惠。在大众点评网，每一条点评信息都有唯一的编号和唯一的地址，因此，每一条点评信息都是一条微内容。用户在提交点评信息时需要遵守一定的规则，其中包括：

① 不发表抄袭的点评。

② 不发表灌水的点评（字数需超过 50 字）。

③ 不发表非亲身经历的点评。

这些规范措施在一定程度上能够保证评论的质量，网站也会采取技术或人工的方式对点评信息进行筛选，但是在评论内容不违反网站规定的条件下，网站会将评论内容全部保留，并且不会改变内容和结构，仅对点评信息进行排序分类等简单组织。每一条大众点评网的评论包含许多有价值的属性信息，其中包括：

① 评论地址：每条点评信息链接地址，该地址具有唯一性。

② 用户 ID：发表评论的用户编号，该编号具有唯一性。

③ 发表时间：用户发表评论的时间。

④ 修改时间：同一用户不能对同一家餐馆点评多次，如果已经点评，并且有新的体

验需要分享，可以修改以前的点评，点评可以修改多次，该属性记录的是最近一次修改时间。

⑤ 鲜花数：网站推出的用户间的互动方式之一，用户在阅读另一位用户发表的点评时，如果认为对自己有帮助，就可以为对方送上鲜花。鲜花数在一定意义上可以反映一条评论的价值。

⑥ 用户等级：用户等级是用户的属性，在反映用户资历的同时，在一定程度上也可以反映点评信息的价值。通常认为用户级别越高，给出的评价在可信性和价值方面都较高；相反则较低。

2）豆瓣网。豆瓣网创办于 2005 年 3 月，是一个典型的 Web 2.0 网站。在豆瓣网上，用户可以自由发布对书籍、电影或音乐的评论，也可以搜索别人的推荐。所有的内容、分类、筛选、排序都由用户产生和决定，甚至在豆瓣主页出现的内容上也取决于用户的选择。

在豆瓣网，用户可以选择提交简短评论或是长篇评论，其中简短评论规定不超过 140 字。用户在提交评论的过程中，可以执行的操作有：按照五分制为评论对象打分；按照自己的理解为评论对象添加标签，网站也为用户推荐使用量较多的标签供用户选择；书写点评信息。其他用户在阅读点评时可以执行的操作有：对评论给出回应；直接选择评论是否有用。网站对评论信息进行的处理包括：根据回应数、有用数将评论进行排序；为用户推荐最有用的点评信息；将点评信息按照倾向性分为好评和差评以便用户选择性查看。

1.2.2　微博

由于国外出现较早的微博平台是 Twitter，目前对微博的分析研究大多数以 Twitter 为对象。根据微博“内容+社会网络”的特征，可以将对微博的研究分为对微博结构的研究、对微博内容的研究和微博应用研究。Sagolla（2009）将微博（Twitter）描述为让人们用 140 个字来分享帖子的平台；Kaplan 等（2011）将微博定义为一种基于互联网的交换工具，允许用户之间交换短篇内容，如句子、图像和视频链接等。

1. 微博结构的相关研究

微博结构主要描述微博内容、微博用户之间的关系。由于微博用户按照社会网络关系组织，对微博结构的研究主要着眼于用户间的关系。日本学者 Yamaguchi 等（2010）用“用户-微博”图（图 1-4）来表示 Twitter 平台上的对象和关系。“用户-微博”图把用户和微博表示为图 $\mathrm{UTG}_S=(V_S,E_S)$，其中，V_S 表示用户、微博在内的节点集合，E_S 表示发布、关注、转发等关系，E_S 的权重反映关系的语义。

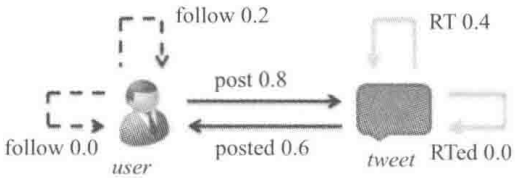


图 1-4　“用户-微博”图

Yamaguchi（2010）还根据用户间关系把微博用户分为三类：

① 信息来源：提供有用的信息，很少关注其他用户。

② 信息搜寻者：很少发布有用的信息，但是关注大量信息来源用户以获取有用信息，很少被其他信息搜寻者关注。

③ 朋友：现实世界中的朋友、亲人或同事。

Weng 等（2010）从微博用户间的关系视角，通过微博用户数据搜集，发现 Twitter 平台上超过 72.4%的用户“关注”（即 follow）了超过 80%关注自己的用户，而 80.5%的用户的 80%的“粉丝”是自己关注对方后对方关注自己的。Ghosh 等（2012）比较了 Twitter 和其他在线社会网络（OSN）的区别，提出一般的 OSN 的社会联系存在“硬界限”，而 Twitter（微博）间的社会联系是“软界限”，即界限内的社会联系的创建取决于用户的名气，并用复杂网络理论构建了可预测度分布的分析模型。

从微博信息间的关系视角，闫幸等（2011）将微博的传播机制分为：裂变式传播与聚合式传播，其中裂变式传播是指粉丝通过关注实现信息接收，通过转发和评论实现信息的再传播，从而一条消息可以迅速通过用户的社会网络实现一对多的病毒式传播；聚合式传播是指微博网站通过热帖热词的排名形成聚焦机制。

现有研究还对微博平台上用户的重要性进行排序，基于社交网络的“同质性”特征（Mcpherson, et al., 2001），Weng 等（2010）在改进 PageRank 算法的基础上提出 TwitterRank 算法，综合考虑用户话题的相似度和用户间关系结构来度量用户的影响力。类似的，日本学者 Yamaguchi（2010）用 TURank 算法度量用户的权威性。

从现有的微博结构研究界定了微博中对象及其关联，并分析了微博用户间的关系、微博信息间的关系，基于上述两种关系研究了微博用户的重要性排序。这些研究为微博及其应用的研究提供了基础，但其中对微博用户间的关联的研究较多，而缺乏对微博信息之间的传播路径和传播规律的研究，而这正是微博和一般社会关系网络的最重要区分。

2. 微博内容的相关研究

若微博结构的特殊性是微博区别于一般在线社会媒体的特征，那么微博内容的特殊性质是微博区别于一般社会关系网络的重要特性，微博内容和一般互联网媒体的文本内容具有显著区别，杨晓茹（2010）将微博内容的特点总结为：用户草根化、内容微小化、介质移动化、传播碎片化、交互多样化。

从微博内容的构成来看，Horn（2010）将微博内容分为三类：个人用户提供的信息、新闻和公司广告，其中超过 85%微博内容都是关于用户信息的。Cheong 等（2010）将微博内容的特征归纳为发布工具、信息长度、微博类型（转发、评论和加标签）、链接、图片附件；把微博用户的特征归纳为性别、所在地、Web 使用习惯、微博使用习惯、关注/被关注比率、账户年龄、发布频率、违规次数；并用案例研究的方法研究这些特征与用户情感倾向间的关联。

热点话题是微博内容区别于一般在线文本内容的另一重要特征。微博内容根据话题可以分为：常规内容和热门话题。常规内容的话题主要集中在娱乐、运动等一般性话题。

特殊事件发生时，微博沟通的内容主要是围绕事件展开的（闫幸等，2011）。Cheong（2011）通过对 Twitter 上 40 天的 667 个热门话题的研究发现，其中娱乐 27%、体育 20%、meme[①] 内容 12%、科技 12%，并发现热门话题具有下面四个特征：

① 排名在前 10 的热门话题会较长时间成为热门话题。

② 大部分热门话题会持续较短时间。

③ 某些热门话题会突然间到达关注的顶峰然后迅速失去关注。

④ 超过 5%的最热话题受到超过 50%的关注。

现有对微博内容的研究归纳了微博内容的特征与构成及热点话题的特征与构成，热门话题所指向的事件的产生机制、话题侦测和内容摘要都可能成为重要的研究问题。

3. 微博应用的相关研究

由于微博以内容为核心、以社会关系网络为传播基础的特征，研究者往往将其视为一个海量的信息来源池，通过信息采集、过滤和加工为具体应用问题服务，并且在这一过程中拓展微博的相关理论。按应用问题分类，大事件预测、恐怖袭击感知、重大疾病发现和品牌口碑评价成为微博解决的主要问题；按使用分析方法分类，情感分析（Sentiment Analysis）、话题侦测（Topic Aspect）和内容分析法等。

Tumasjan 等（2011）以大事件预测为应用问题，用情感分析技术分析了 Twitter 平台上 100 000 条有关 2009 年德国大选的微博，通过在线微博合理地验证了线下的政治版图和大选结果；Kim（2012）用三重螺旋指标体系分析了 5 位候选人之间的关注和被关注对象的重叠密度，发现非主流的、缺乏资源的候选人往往更倾向利用微博的资源。

Cheong 等（2011）以恐怖袭击感知为应用问题，将微博信息用于恐怖主义信息的侦测，构建了如图 1-5 所示的简洁的四阶段框架，包括检查突发事件、信息搜集和垃圾过滤、情感分析和人口统计学侦测、报告和数据挖掘；Onook（2011）在对孟买恐怖袭击事件微博内容分析的基础上用状态感知理论（SA）构建了一个包括知觉（Perception）、理解（Comprehension）和映射（Projection）的三阶段框架。

Paul（2011）以重大疾病发现为应用问题，采用话题侦测模型（Topic Aspect Model）在 Twitter 平台上的 50 万条微博发现了超过 12 个疾病话题信息。

Jansen 等（2009）以品牌口碑评价为应用问题，通过分析 Twitter 平台上 150 000 条包含品牌评价的微博得出结论：19%的微博包含品牌信息，其中超过 20%的品牌信息微博包含情感判断，其中约 50%是正面信息，33%对品牌有关键作用；Zhang（2011）用路径分析验证了 5 周内 96 725 个用户关于 9 个品牌的 164 478 条微博，验证了在 Twitter 平台上的企业运作和客户在线客户口碑交流直接相关，此外转发微博是客户表达对 twitter 平台上企业运作的明确反应。

此外还有一些在分析方法上值得借鉴的微博应用研究，Thelwall 等（2011）用情感分析结合时间序列分析研究证明了微博平台上的流行事件往往增强了公众的负面情绪，

① meme：这个词最初源自英国著名科学家理查德·道金斯（Richard Dawkins）所著的《自私的基因》（The Selfish Gene）一书，其含义是指“在诸如语言、观念、信仰、行为方式等的传递过程中与基因在生物进化过程中所起的作用相类似的那个东西”。