HZ BOOKS
华章教育

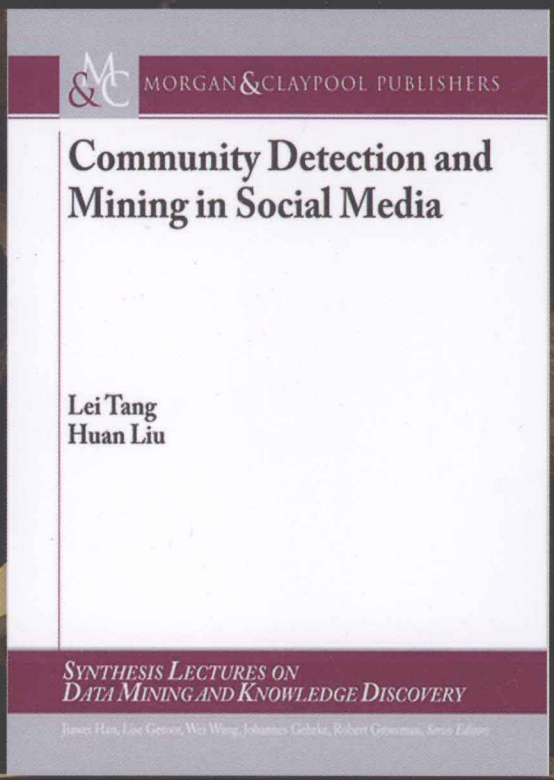
计 算 机 科 学 丛 书

# 社会计算

## 社区发现和社会媒体挖掘

Lei Tang Huan Liu 著 文益民 闭应洲 译

Community Detection and Mining in Social Media




机械工业出版社
China Machine Press

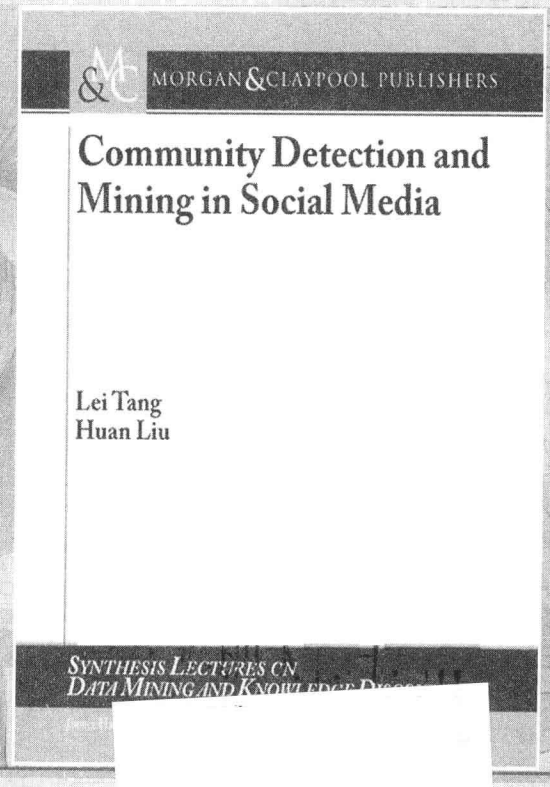
计 机 科 学 丛 书

# 社会计算

## 社区发现和社会媒体挖掘

Lei Tang Huan Liu 著 文益民 闭应洲 译

Community Detection and Mining in Social Media




机械工业出版社
China Machine Press

本书从数据挖掘角度介绍社会媒体的性质，评述社会媒体计算的代表性工作，并描述社会媒体带来的挑战。书中介绍了基本概念，使用浅显易懂的例子展示了最新的算法和有效的评价方法，阐述了混杂社会网络中的社区发现技术和社会媒体挖掘技术。

本书简明易懂，是研究社会媒体中的社区发现与挖掘技术的入门级读物，适合从事社会媒体数据挖掘研究与应用的学生、研究者和实践者阅读。



**本书版权登记号：图字：01-2012-7581**

**图书在版编目（CIP）数据**

社会计算：社区发现和社会媒体挖掘/（美）唐磊（Lei Tang）等著；文益民，闭应洲译．—北京：机械工业出版社，2012.11

（计算机科学丛书）

书名原文：Community Detection and Mining in Social Media

ISBN 978-7-111-40287-9

Ⅰ.社…　Ⅱ.①唐…　②文…　③闭…　Ⅲ.数据收集－计算机算法　Ⅳ.TP274

中国版本图书馆 CIP 数据核字（2012）第 258643 号

机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码　100037）

责任编辑：盛思源

北京瑞德印刷有限公司印刷

2013 年 1 月第 1 版第 1 次印刷

185mm×260mm · 9 印张

标准书号：ISBN 978-7-111-40287-9

定价：35.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage 等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出 Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不

仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

**华章网站**：www. hzbook. com
**电子邮件**：hzjsj@ hzbook. com
**联系电话**：（010）88379604
**联系地址**：北京市西城区百万庄南街1号
**邮政编码**：100037


华章科技图书出版中心

社会媒体（social media）是社会化媒体的简称，也称为社交媒体。在社会媒体中，人们既是信息的消费者也是信息的生产者。2004 年上线的 Facebook（脸谱）如今拥有的注册用户数超过 10 亿；2006 年上线的 Twitter（推特）拥有的注册用户数达到了 5 亿；2005 年创立的人人网拥有的注册用户数超过了 2 亿；每天人们在视频分享网站 YouTube 上观看的视频数超过 1 亿。奥巴马因在美国总统选举中使用 Facebook 而争取到了更多选民的支持；2009 年美国东部地震时，人们发现 Twitter 的传播速度比地震波还快。由于社会媒体在政治经济活动和人们日常生活中日益发挥着更加重要的作用，所以近年来针对社会媒体的研究成为热点。为了推进社会媒体的研究与应用，我们在多年从事社会计算和社会媒体研究的基础上出版了此书。

我们非常欣喜地看到针对社会计算和社会媒体的研究与应用在中国也得到了政府、产业界和学术界的高度重视，一批高级别的研究项目得到立项，同时高水平的研究成果也在不断产生。

拙著能够得到同行文益民教授和闭应洲教授的认可，我们倍感荣幸。他们于繁忙的教学工作中专门抽出大量时间和精力翻译本书，并且进行了精细的审查，指出和勘正了英文原稿中的一些错识和失漏。对于他们这种一丝不苟的学术研究态度，我们深表敬意和感谢。

学术无国界，希望本书能使有兴趣研究社会媒体和社会计算的读者缩短学习过程并拓展研究视野，同时还能推进美中学术交流。

Lei Tang 和 Huan Liu

2012 年 10 月

# 译 者 序

Community Detection and Mining in Social Media

Web 在过去十年当中得到了快速发展，涌现出许多用户参与的 Web 应用程序和社会信息网络，其中包括博客、论坛、共享媒体平台、微博、社会网络、社会新闻、社会书签和维基百科，学术界称其为社会媒体。与传统的 Web 应用和传统的媒体相比，上述社会媒体具有一个共同的特点：广大用户既是内容、信息和知识的消费者同时也是相应的生产者。由大量用户“贡献”的海量社会行为数据，为观测和研究社会创造了前所未有的条件。社会媒体的另一个显著特点在于它具有丰富的用户交互特性。通过互动，用户之间产生了相互关系。比如，微博中的关注关系、社会网络中的好友关系、在线商店中因共同购买或评论产品形成的共同兴趣关系等。这导致了各种用户关系网络的涌现。利用数百万甚至数以亿计的用户在线娱乐、在线工作和在线社交所产生的海量数据，可以进行前所未有的大规模社会网络分析，为研究人类的交互和集体行为提供了新的机会。

传统的社会学研究往往使用调查、问卷、面谈、参与者观察与统计的形式获取数据，所使用的数据规模较小，并且难以得到个人完整的信息行为记录。因此，传统研究的成果更多来源于直观认识，缺乏基于大规模真实数据的实验验证。社会媒体给人们提供了一个研究人类社会的新平台。计算社会学认为：网络上的大量信息，如博客、论坛、聊天、消费记录、电子邮件等，都是现实社会的人或组织的行为在网络空间的映射。这些网络数据可以用来分析个人或群体的行为模式，从而深化我们对社会的了解。就像大规模基因数据催生了生物信息学一样，海量的社会数据催生了社会计算，即以计算手段研究社会学中的定性问题并解决传统社会学中的实验问题。

本书作者在社会媒体挖掘和社会计算方面进行了深入的研究，不仅熟稔社会计算的理论研究，而且具有非常丰富的社会计算应用经验。本书介绍了社会计算的基

础知识，描述了社区发现的典型方法，并讨论了社区发现评价的问题，阐述了混杂社会网络中的社区发现问题和社会媒体挖掘技术。本书英文版深受读者欢迎，在Morgan &Claypool出版的数据挖掘和知识发现综合类电子书中，该书的下载量最高（http：//www. morganclaypool. com/action/showMostReadArticles? journalCode = dmk）。全书笔调清新，简明易懂。无论对社会计算感兴趣的学生还是专业人士，都非常值得一读。

感谢本书作者Lei Tang（唐磊）博士和Huan Liu（刘欢）教授。他们多次仔细阅读本书译稿，提出了许多宝贵意见。他们还专门为本书的中文版撰写了序言。

感谢机械工业出版社华章公司的编辑们，是他们对学术的敏感和细致的工作使得本书中文版能够尽快与读者见面。

最后，由于译者水平有限，译文中难免存在问题，敬请专家和读者指正。

文益民　闭应洲

2012年9月于凤凰城

## 译者简介

Community Detection and Mining in Social Media

**文益民**，桂林电子科技大学计算机科学与工程学院教授，上海交通大学计算机软件与理论专业博士，中国计算机学会高级会员。长期从事计算机软件与理论的教学和研究。主要讲授的课程包括：程序设计、数据库系统原理、机器学习、数据挖掘与数据仓库等。2007—2011 年在湖南大学和中南大学从事博士后研究，2012 年访问亚利桑那州立大学 Huan Liu（刘欢）教授领导的数据挖掘与机器学习实验室。研究兴趣包括：机器学习与数据挖掘、社会媒体挖掘、推荐系统、系统综合评价等。主持省部级科研项目 8 项；参与国家自然科学基金项目 2 项；发表学术论文 30 余篇；获得省部级教学、科研奖励 6 项；主编普通高等教育“十一五”国家级规划教材 1 部。多次应邀担任国际会议程序委员。

**闭应洲**，广西师范学院教授，武汉大学软件工程国家重点实验室计算机软件与理论专业博士，硕士生导师，广西高校优秀人才资助计划资助人选、广西师范学院软件研究所副所长。主要研究方向：智能计算、智能信息处理及社会计算。主持和参加了 10 多项科研项目的研究工作，发表了 30 多篇论文。2012 年 2 月至 2013 年 1 月在美国亚利桑那州立大学访学，师从 Huan Liu（刘欢）教授，重点研究从海量数据中获取知识所必需的理论和技术：1）高效的特征选择算法，试图通过特征选择和特征离散化来处理大量的高维数据；2）集成多数据源解决不确定性和模糊性问题；3）应用领域知识，在协同演化算法的框架下融合多种机器学习方法实现有效的挖掘和信息集成，使得计算机更加“聪明”，能够处理更复杂的问题。

我们由衷感谢为本书写作及出版提供实质性贡献的诸位同事。亚利桑那州立大学的社会计算社区、数据挖掘与机器学习实验室的成员使得本书的写作过程充满了乐趣。他们是：Ali Abbasi、Geoffrey Barbier、William Cole、Gabriel Fung、Huiji Gao、Shamanth Kumar、Xufei Wang 和 Reza Zafarani。特别要感谢 Reza Zafarani 和 Gabriel Fung，他们通读了本书手稿的早期版本，并且提供了非常好的、增强本书可读性的建议。

我们非常感谢 Sun-Ki Chai 教授、Michael Hechter 教授、John Salerno 博士和 Jianping Zhang 博士，他们提供了许多富有启发性的建议。本书的出版也部分得到了 AFOSR 和 ONR 的资助。

我们非常感谢 Morgan & Claypool 出版社。特别感谢执行编辑 Diane D. Cerra 给予的帮助和在整个出版过程中的无私与耐心。我们还要感谢阿肯色大学小石城校区的 Nitin Agarwal 教授，本书编辑过程中的问题他都在第一时间给予回答。

最后且最重要的，我们要感谢我们的家人在本书写作及出版过程中给予的支持。我们诚挚地将本书献给他们！

Lei Tang 和 Huan Liu

2010 年 8 月

# 目　录

Community Detection and Mining in Social Media

第1章
Community Detection and Mining in Social Media

# 社会媒体与社会计算

## 1.1 社会媒体

Web 和 Internet 在过去的十年中得到了快速的发展，同时也产生了巨大的变化。涌现出无数互动的 Web（participatory web）应用程序和社会网络站点（social networking site）。这些应用和站点拉近了人们之间的距离，使得他们拥有了新的合作与交流方式。大量的在线志愿者以协作的方式共同撰写百科全书，其所涉及的内容和范围都超出了人们的想象。通过利用对用户（user）的购物行为和评论进行分析所得到的集体智慧，在线商店（online marketplaces）可以更有效地推荐产品。而政治活动也可以从新的参与形式和新的集体行为中获益。

表 1-1 列举了不同的社会媒体（social media），包括博客、论坛、媒体共享平台、微博、社会网络、社会新闻、社会化书签和维基（wikis）。与传统的 Web 应用和传统的媒体相比，上述的各种应用和媒体尽管表面上不一样，但都具有一个共同的特点：内容、信息和知识的消费者（consumer）同时也是相应的生产者（producer）。

在电视、收音机、电影和报纸等传统媒体中，只有一小部分的权威和专家能够决定提供哪些信息和如何发布这些信息。大部分用户只是消费者，他们被分隔在信息的产生过程之外，不能参与其中。传统媒体的通信模式是单向传播，即从一个集中的生产者流向广泛的消费者。

然而，社会媒体的用户可以既是一个消费者又是一个生产者。对于在各种社会媒体

□1 网站（social media site）上活跃的亿万用户来说，其实每个人都可以是一个媒体出口（outlet）（Shirky，2008）。这种新的大众出版方式可以产生实时的新闻和大量来自底层的信息，从而出现海量的用户创建的内容（user-generated contents），并形成群体智慧（wisdom of crowds）。其中的一个例子就是发生在2005年的伦敦恐怖袭击（Thelwall，2006），一些当事者在博客上发布了他们的经历，提供了相关事件的第一手报道。另一个例子是发生在2009年伊朗总统选举过程中的流血冲突，许多人通过Twitter（一个微博平台），实时更新了相关信息。社会媒体也使得协作写作高质量的作品成为可能。比如，从2001年创办以来，维基百科（Wikipedia）已迅速发展成为最大的参考站点之一，在2009年每月吸引大约65 000万访问者。超过85 000名活跃的作者，使用260多种语言撰写了1400万篇文章[⊖]。它是一部基于互联网、内容开放的全球多语言百科全书，也是目前世界上最大的百科全书。

**表1-1 各种形式的社会媒体**

| | |
|---|---|
| 博客 | Wordpress、Blogspot、LiveJournal、BlogCatalog |
| 论坛 | Yahoo! answers、Epinions |
| 媒体共享平台 | Flickr、YouTube、Justin. tv、Ustream，Scribd |
| 微博 | Twitter、foursquare、Google buzz |
| 社会网络 | Facebook、MySpace、LinkedIn、Orkut，PatientsLikeMe |
| 社会新闻 | Digg、Reddit |
| 社会标记 | Del. icio. us、StumbleUpon、Diigo |
| 维基百科 | Wikipedia、Scholarpedia、ganfyd、AskDrWiki |

社会媒体的另一个显著特点在于它具有丰富的用户交互特性。社会媒体的成功依赖于用户的参与。越多的用户互动就会促使越多的用户参与，反之亦然。比如，Facebook声称在2010年8月拥有超过5亿活跃的（active）用户[⊖]。用户参与（user participation）是社会媒体成功的关键因素。根据Alexa在2010年8月3日发表的Internet流量数据，表1-2中列出了前20名网站。正是由于大量的用户参与，使得有8个社会媒体网站名列其中。通过互动，用户之间产生了相互联系，并导致了用户网络的涌现。这为研究大规

⊖ http：//en. wikipedia. org/wiki/Wikipedia：About.

⊖ http：//www. facebook. com/press/info. php? statistics.

模的人类交互和集体行为提供了新的机会，也产生了对新计算的挑战，促使高级计算技术和算法的进一步发展。

表 1-2　美国前 20 名网站

| 排名 | 网站 | 排名 | 网站 |
|---|---|---|---|
| 1 | google. com | 11 | blogger. com |
| 2 | facebook. com | 12 | msn. com |
| 3 | yahoo. com | 13 | myspace. com |
| 4 | youtube. com | 14 | go. com |
| 5 | amazon. com | 15 | bing. com |
| 6 | wikipedia. org | 16 | aol. com |
| 7 | craigslist. org | 17 | linkedin. com |
| 8 | twitter. com | 18 | cnn. com |
| 9 | ebay. com | 19 | espn. go. com |
| 10 | live. com | 20 | wordpress. com |

本书提出了与社区（community）发现相关的社会网络分析的基本概念，并利用一些简单的例子来说明目前用来分析社会媒体数据的最新算法。这是一本配套齐全的书，包含了在社会媒体中与社区发现相关的重要问题。本书从需要用到的概念和定义开始。 2

## 1.2　概念与定义

网络数据与属性 - 值类型的数据不同，它具有自己独特的性质。

### 1.2.1　网络与表示

一个社会网络就是一个由结点和边组成的社会结构，其中结点表示个人或组织，边用来连接结点表示各种关系，比如朋友关系、亲属关系等。表示一个网络一般有两种方法。一种是基于图的表示方法，这是一种直观的方法。图 1-1 给出了一个只有 9 个用户的社会网络。一个社会网络也可以表示为一个矩阵，称为社会矩阵（sociomatrix）或邻接矩阵（adjacency matrix）（Wasserman and Faust, 1994），参见表 1-3。一个社会网络通常是一个稀疏网络，对应的矩阵包含很多 0。这个稀疏性质可以用来加速网络分析的效率。在邻接矩阵中，没有指定对角元素的值。按照定义，对角元素表示自我连接，也就

是从一个结点到自身的连接。一般来说，对角元素常常设为0。但在某些场合下，对角元素应该设置为1。因此，除非特别指定，否则对角元素一般默认为0。

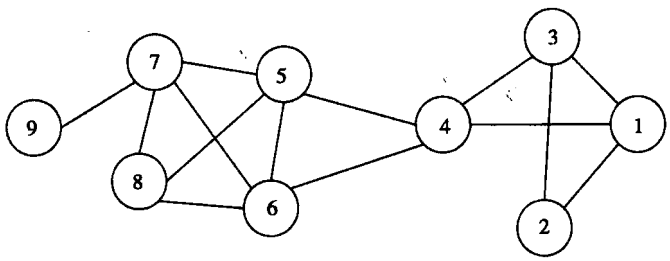



图1-1 一个包含9个用户和14个联系的社会网络。网络的直径是5。结点1～9的聚类系数分别是：$C_1=2/3$、$C_2=1$、$C_3=2/3$、$C_4=1/3$、$C_5=2/3$、$C_6=2/3$、$C_7=1/2$、$C_8=1$、$C_9=0$。平均聚类系数为0.61，而具有9个结点，14条边的随机网络的聚类系数期望值是$14/(9\times 8/2)=0.19$

**表1-3 邻接矩阵**

| 结点 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | — | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | — | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | — | 1 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 0 | 1 | — | 1 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | — | 1 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 | 1 | — | 1 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 1 | 1 | — | 1 | 1 |
| 8 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | — | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | — |

一个网络可以是加权的、有符号的和有方向的。在一个加权网络中，每一条边都关联着数值；在一个符号的网络中，一些边是正的联系，而另一些边则是负的联系。有向网络是指网络的边是有方向的。在图1-1的例子中，网络是无向的，与此相对应的邻接矩阵是对称的。然而，在一些社会媒体站点中，交互是有方向的。比如，在Twitter中，一个用户$x$关注（follow）$y$，但$y$不一定关注$x$。在这种情况下，关注者（follower）－被关注者（followee）网络是有向的，也是不对称的。除非特别说明，本书将只关注最简单的网络形式，也就是无向网络，而边的权值是布尔型，就像表1-3中的例子一样。

但本书中的许多技术通过扩展就可以处理加权、有符号的和有方向的网络。

表 1-4　名称

| 符号 | 含义 |
| --- | --- |
| $A$ | 一个网络的邻接矩阵 |
| $V$ | 一个网络的结点集合 |
| $E$ | 一个网络的边集合 |
| $n(n=\|V\|)$ | 结点的数量 |
| $m(m=\|E\|)$ | 边的数量 |
| $v_i$ | 结点 $v_i$ |
| $e(v_i, v_j)$ | 结点 $v_i$ 与结点 $v_j$ 之间的边 |
| $A_{ij}$ | 值为 1，表示结点 $v_i$ 与结点 $v_j$ 之间存在边；值为 0，表示结点 $v_i$ 与结点 $v_j$ 之间不存在边 |
| $N_i$ | 表示结点 $v_i$ 的邻接结点的集合 |
| $d_i(d_i=\|N_i\|)$ | 结点 $v_i$ 的度 |
| geodesic | 两个结点之间的最短路径 |
| geodesic distance | 最短路径长度 |
| $g(v_i, v_j)$ | 两个结点 $v_i$ 与 $v_j$ 之间的最短路径 |

书中分别使用 $V$ 和 $E$ 来表示结点的集合和边的集合，$n$ 表示结点的数量，$m$ 表示边的数量。矩阵 $A \in \{0,1\}^{n\times n}$ 表示网络的邻接矩阵，其中的某个元素 $A_{ij} \in \{0, 1\}$ 表示在结点 $v_i$ 与结点 $v_j$ 之间是否存在连接。结点 $v_i$ 与结点 $v_j$ 之间的边记为 $e(v_i, v_j)$。如果 $A_{ij}=1$，则表示结点 $v_i$ 与结点 $v_j$ 邻接（adjacent）。$N_i$ 表示结点 $v_i$ 的所有邻接结点集合。结点 $v_i$ 的所有邻接结点数量称为它的度（degree），记为 $d_i$，比如在图 1-1 的网络中，$d_1=3$、$d_4=4$。如果一个结点是一条边的终端结点，就说这条边邻接这个结点，比如边 $e(1, 4)$ 邻接结点 1 和 4。

结点之间的最短路径（比如 $v_i$ 与 $v_j$）称为 geodesic（测地线），最短路径中两个结点之间的跳数就是它们之间的最短路径距离（记为 $g(v_i, v_j)$），在图 1-1 的例子中 $g(2, 8)=4$，结点 2 和结点 8 的最短路径是（2-3-4-6-8）。表 1-4 对本文经常使用的记号进行了总结。

## 1.2.2　大规模网络的属性

社会媒体中的网络一般都是非常巨大的，通常包含数百万的用户与连接。这些大规