清华大学学术专著

社会计算：用户在线行为分析与挖掘

刘红岩 著

清华大学出版社

清华大学学术专著

社会计算：用户在线行为分析与挖掘

刘红岩 著

清华大学出版社
北京

内 容 提 要

近年来，随着 Web 技术的发展和应用的普及，大量用户将线下行为转移到线上进行，并且通过各种社会媒体随时随地进行社会交互和情感表达。这些海量的社会行为形成的大数据，催生了社会计算这个新的跨学科的研究和应用领域。本书在大数据的时代背景和社会计算的框架下，介绍从大量用户在线行为数据中发现其中隐含的用户行为模式和兴趣偏好的方法和技术。全书主要内容分为 7 个部分，分别介绍用户在线搜索行为、网上购物行为、浏览行为、社会标注行为、评论行为以及社交行为等方面的数据分析技术和方法，涉及搜索意图的分析、购物模式的发现、周期行为的挖掘、标签的有效聚类、评论意见的挖掘、用户偏好的发现、个性化推荐方法、链接分析以及社会网络的分析方法等最新研究内容。

本书内容新颖、丰富、易于理解，反映社会计算和商务智能的最新研究和应用趋势。本书主要面向高等院校和科研单位的研究生、博士生和相关研究领域的学者，对业界管理人员和信息技术人员也有一定的参考价值。

图书在版编目(CIP)数据

社会计算：用户在线行为分析与挖掘/刘红岩著. --北京：清华大学出版社，2014
清华大学学术专著
ISBN 978-7-302-35648-6

Ⅰ. ①社… Ⅱ. ①刘… Ⅲ. ①数据收集－技术 Ⅳ. ①TP311.13

中国版本图书馆 CIP 数据核字(2014)第 050779 号

责任编辑：索　梅　李　晔
封面设计：傅瑞学
责任校对：李建庄
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn,010-62795954

印 装 者：三河市中晟雅豪印务有限公司
经　　销：全国新华书店
开　　本：153mm×235mm　**印　　张**：13.5　**字　　数**：243 千字
版　　次：2014 年 7 月第 1 版　**印　　次**：2014 年 7 月第1 次印刷
印　　数：1～2000
定　　价：65.00 元

产品编号：056654-01

前　言

随着互联网、计算机等信息技术和应用的发展，人们的学习、生活和工作的诸多行为都从线下转移到线上进行，人们的在线活动行为和轨迹被记录下来。同时，Web 2.0 技术以及社会媒体的兴起和广泛使用，使得人们的社会交互和意见情感在网上表露无遗，这些都使得利用计算机技术分析人类的社会活动，发现其中蕴含的规律和模式成为可能，也使社会计算这门学科成为近年来的研究和应用热点，国内外知名高校和研究机构也纷纷设立社会计算专业学科。另外，消费者的众多网上在线活动中隐含着有意义的行为规律和兴趣偏好，挖掘这些规律和偏好，可以为电子商务实现个性化服务、精准营销和开发新型业务模式提供技术和理论支持。同时，大数据、稀疏数据、富媒体、非结构化等极端数据为传统的分析技术和方法带来了巨大的挑战。

本书主要介绍著者近几年在商务智能和社会计算领域的最新研究成果，重点探讨如何根据用户的在线行为特点，针对电子商务过程中出现的管理需求，提炼和定义新型用户行为模式，刻画用户的兴趣偏好；在极端数据情况下如何高效发现各类知识，构建用户兴趣偏好模型；最后通过严谨的理论分析和广泛的实验，对所提方法和发现的结果进行验证和评价，旨在为推动数据挖掘和社会计算的理论发展做出贡献，也为业界的实际应用提供参考。本书的内容根据用户在线行为的类别进行组织，分别介绍新型的分析和挖掘方法。在线行为包括搜索、网购、标注、浏览、评论以及社交等，分析方法涉及各种概率统计、数据挖掘、社会网络分析等。研究内容包括搜索意图的发现、热点话题的侦测、在线购物模式的挖掘、周期模式的定义和分析、标签相似度的衡量、高效的聚类，以及针对中文表达的意见挖掘、个性化推荐和社会网络中的相似度、影响度的有效度量和计算等。本书主要面向高校和科研单位的硕士生、博士生和相关研究领域的学者，对企业营销、运营及信息技术等方面的管理人员也有一定的参考价值。

本书由国家自然科学基金项目“基于数据挖掘的用户网上行为模式的

发现技术与应用研究”(项目编号 70871068)和“通过社会化媒体挖掘用户兴趣的方法及应用研究”(项目编号 71272029)的研究成果凝练而成。在项目的研究过程中,著者及其相关团队紧密配合,深入研究,在许多国内外热点和前沿研究问题上,勇于面对挑战,攻克了许多难点,取得了一系列创新型研究成果。为此,衷心感谢研究团队和合作者在研究项目中所做出的学术贡献。

作为新兴领域的一部专著,难免有疏漏之处,敬请读者指正。

著 者

2014 年 4 月于清华园

目　录

第1章
绪论

本章首先介绍大数据、社会媒体和社会计算等概念及其产生背景，概要介绍大数据分析、社会计算相关研究的意义和研究内容；然后介绍作者在社会计算的大框架下近年来所从事的用户在线行为模式的挖掘和分析工作，并对本书的主要内容进行概括介绍。

1.1 大数据分析与社会计算

近年来，随着互联网、移动设备等信息技术的迅猛发展，除了企业业务运营过程中不断积累的交易等业务数据之外，遍布全球的传感器无时无刻不在探测和收集物理世界的各种信息，移动互联网则在不断收集用户的地理位置信息，各种社会媒体中的数以亿计的用户也在随时随地产生交互信息。这些数据不仅数量巨大(以 TB 甚至 PB 为单位)，而且形式繁多，除了企业业务运营信息系统中的结构化数据之外，各种文本信息、声音、图片、视频、地理位置等各种不同类型的数据决定了数据的多样性。同时，这些时刻变化的来自各种数据源的数据又充满噪音，对这些数据的管理和分析已经超出了传统的数据管理技术的能力，因此，人们将其称为大数据(big data)。信息技术咨询与研究机构 Gartner 将大数据定义为“需要经济有效的新型信息处理技术才能获得更强的决策力和洞察力的海量、高速变化和多样化的信息资产”。IBM 对大数据的四方面的特性——数据量(volume)、时效性(velocity)、多样性(variety)和可疑性(veracity)通过图文结合的方式进行解释，如图 1.1 所示。从各种类型的大量数据中，快速获得有价值信息的能力，就是大数据技术。

进入大数据时代，无论是工业界、学术界还是政府都面临着新的机遇和挑战，促使各领域的管理改革，推进社会发展。全球管理资讯公司麦肯锡

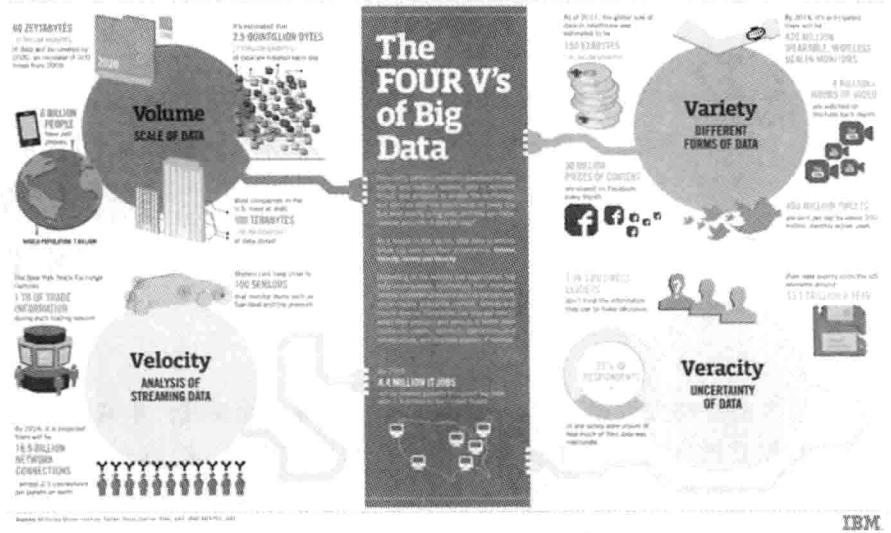

图 1.1 大数据的特点(资料来源：IBM 公司)

2011 年发布了一个题目为“大数据：创新、竞争和生产力的下一个前沿领域”的研究报告(Manyika, *et al.*, 2011)。报告中指出，大数据可以为世界经济创造重要价值，提高企业和公共部门的生产率和竞争力，并为消费者创造大量的经济剩余。报告调查并总结了大数据在 5 个领域可以创造的价值，如图 1.2 所示。

2011 年 2 月，《Science》推出名为 Dealing with data 的专题[①]，广泛讨论了与大数据研究相关的各方面问题，强调了数据的泛滥可能给人们带来的挑战，以及如何通过对数据的更好的组织和处理来抓住不可多得的机遇。

美国政府非常重视大数据，将大数据定义为“未来的新石油”，将大数据作为国家战略。2012 年 3 月，美国白宫科学与技术政策办公室(OSTP)宣布了“大数据研究和开发计划”，投资 2 亿美元，旨在推进用于收集、存储、管理、分析以及共享大量数据的先进核心技术，推动大数据的相关研究和应用。

2009 年联合国开始发起一个名为“全球脉动(Global pulse)”的行动计划，旨在利用大数据对全球经济发展起到推动作用。2012 年 5 月，全球脉动发布题为“大数据促发展：机遇与挑战”的报告，详细阐述了大数据给世

① http://www.sciencemag.org/site/special/data/

图 1.2　大数据在 5 个领域可以创造的价值图示(资料来源：麦肯锡报告“大数据：创新、竞争和生产力的下一个前沿领域”)

界带来的机遇、挑战和应用。2013 年 5 月发布的报告中，描述了根据推特(Twitter)中的微博信息预测大米价格以及利用谷歌趋势(Google trends)侦测登革热疫情等案例。

最近几年，作为大数据的一种主要来源，越来越多的企业和学者开始重视社会媒体这种新型的媒体形式。随着 Web 技术的发展，用户参与的 Web 应用大量涌现，如博客、论坛、社交网站、微博、社会标签、在线评论、多媒体信息分享以及基于位置的社会网络服务等。在这些新型的应用中，数量众多的普通百姓不仅是信息的消费者，更是信息的生产者，相应的信息称为用户生成内容 (User Generated Content，UGC)，产生这些信息的媒介称为社会媒体(social media)。这些社会媒体中的用户数量巨大。例如，截至 2013 年 3 月底，新浪微博用户数已达 5.36 亿。截至 2012 年 11 月底，中国移动电话用户数达到 11.04 亿户，其中 3G 用户数 2.2 亿户。如此大量的用户通过各种社会媒体每时每刻、随时随地地表达着自己的各种感受和情绪变化，并且与线上的朋友间产生着频繁的互动，分享信息，交换意见和观点。用户在线上的各种关系，例如关注关系、好友关系、信任关系等，构成了一个巨大的虚拟社会的社会网络。在这个社会网络中，大量的信息在快速传播，网络结构在动态演变。这些海量的社会行为形成的大数据，为进行社会分析和研究提供了便利条件，使得通过计算手段分析社会问题成为可能，

这催生了社会计算这个新的跨学科的研究和应用领域。

社会计算，又称计算社会科学（Lazer，*et al.*，2009），是通过对大量数据的收集和分析，揭示个人和群体行为规律的新兴领域，也可以将其看作是基于计算系统对社会行为和规律进行研究的一门学科。六度分割理论的创立者——著名学者 Duncan J. Watts 于 2007 年在《Natures》上发表的一篇题为“二十一世纪的科学”的文章中指出：如果进行适当的处理，人们在网上的交流和互动会彻底变革人们对人类群体行为的理解（Watts，2007）。文章分析了社会学研究在过去 50 年中遇到的问题。尽管社会学家十分重视个体及组织间的互动在决定群体社会行为中的重要性，但是长时间对分析对象的观察确实非常困难，他们通常借助于少量的某时间点的数据进行研究，或根据调查问卷形式的参与者的自报告得出结论，这往往会受到认知偏见、感知错误以及表达的模糊性等方面的影响。但是近年来基于 Internet 的交流和互动数据的急速增长提供了新的机会，使人们可以从个体的层面观察大量用户之间的实时交互，同时，计算机技术的发展也提供了处理大数据的能力，能够模拟社会交互的巨大网络（Watts，2007）。

社会计算的研究范围很广，常见的研究包括社会网络分析、社区分析、意见挖掘、个性化推荐、信息传播机制研究等。

1.2 用户在线行为的分析与挖掘

网络、通信和计算机技术的发展，使得越来越多的人的日常生活、工作以在线的形式进行。搜索引擎、电子商务网站、微博、微信、社交网站、在线评论网站、在线约会网站等 Web 应用以很细的粒度、很高的频度不断记录着人们的行为轨迹。这些数据背后隐藏着人们的生活习惯、兴趣偏好以及随时随地的情绪变化模式，同时也隐含着社会运行的规律和发展趋势。发现这些深藏于大数据中的知识，可以针对个体需求进行个性化服务，可以变革传统的管理和业务运营模式，可以促进人们对社会发展规律的认识，增强人与人的沟通和合作。然而如何从这些海量的、富含噪音的、动态变化的、结构形式多样的大数据中发现有意义的模式、规律和趋势，是学术界和工业界普遍关注的热点研究问题。通过分析大规模群体行为数据的应用案例已有很多，例如，根据在线评论预测商品价格，依据微博中表达的情绪预测股市变动，根据搜索日志预测和跟踪疾病的流行，利用微博辅助政治运作，组织大规模群众运动，等等。

我们的研究团队近年来一直从事用户在线行为数据的分析和挖掘方面

的研究工作，提出了许多有创新性的分析方法，发现了许多有意义的行为模式，解决了一些典型的大规模数据分析中存在的问题，在一流国际期刊和国际会议上了发表了许多学术文章。本书是对部分研究工作的汇总和总结，按照用户行为特点的不同，下面概括性地介绍相关的研究工作。

1.2.1　在线搜索行为分析

搜索行为是用户最频繁的网上行为之一，搜索日志中隐含了大量有价值的信息。通过搜索日志分析用户的搜索意图及潜在需求、探测用户关心的热门话题对于广告投放、个性化推荐具有重要意义，因此，用户搜索意图的挖掘和热门话题的探测是近几年的研究和应用热点。

通用搜索引擎，如 Google、百度、必应等记录了用户提交的每个查询以及点击行为，这种数据称为查询日志(search log)或点击数据(click-through data)。从查询日志发现搜索意图的研究旨在从大量用户的搜索行为中发现每个查询背后的真实目的，从而简化查询过程，提高广告投放的准确性。但是查询表达简单、模糊、存在的二义性对此方面的研究提出了很大的挑战。已有的相关研究或者无法充分利用各方面的信息，或者分析的时间复杂度过高。为此我们提出了多视图随机游走(multi-view random walk)模型，可以结合多方面信息，如查询语义、URL 点击以及会话(session)等，提高了分析的准确度；另外，可以将一个领域(domain)的查询同时进行处理，提高了分析的效率。通过真实的查询日志对多个领域的查询进行处理，挖掘出了很多有代表性的搜索意图。

对热门话题——探测方面，提出了一种新型的话题和事件挖掘的系统框架，设计并实现了高效率的话题探测方法。该方法基于分治策略，采用局部扩展的方法从重要结点出发发现话题包含的查询和 URL。以此为基础设计并实现了两种算法，分别利用随机游走模型和马尔科夫随机场模型，可以从大规模查询日志中快速发现热点话题，克服数据稀疏及富含噪音的问题。另外还提出了从话题中分解事件，推断事件之间演化关系等方法。根据所提出的方法，开发了可视化演示系统。

这一部分工作的相关论文发表在英文国际期刊 *ACM Transactions on Information Systems*(Liu, *et al.*, 2012)以及知识管理领域国际会议 ACM CIKM (Gu, *et al.*, 2011)上。所开发的话题探测以及事件演化演示系统被知识发现国际会议 SIGKDD 2010 录取为演示系统(Cui, *et al.*, 2010)。本书第 2 章将详细介绍搜索意图的挖掘以及话题探测方面的研究工作。

1.2.2 在线购物行为分析

随着电子商务的发展，越来越多的人选择在网上购物。已有的有关用户购买行为的关联分析方面的研究是针对实体店进行的，主要进行购物篮分析，即消费者一次购买物品之间的关联性分析，这方面的研究已有很多，然而已经提出的方法用在网上购物行为分析时只能发现用户在一个网站上的关联模式。而电子商务的便捷性使得用户可以很方便地在多个网站进行购物，因此发现跨网站的购物模式具有理论和实用价值。为此，通过分析跨网站购物的特点，提出了消费者网上购买行为的新型跨网站购物模式，在涵盖已有单网站关联模式以及层次关联模式的基础上，集成了多种新型关联模式，如不同网站购物类别之间的关联，购物类别与网站之间的关联等。利用这些新型模式，可以辅助商家更好地进行促销策略的制定、进行有效的推荐服务以及开发新型业务等。针对这种新型的知识类型，原有方法无法发现，为此我们提出了发现这种模式的两种高效方法，在处理不同特点数据集时各有特点。为了提高挖掘效率，我们提出了一种新颖的数据结构——OSP-tree，从而大大加快了挖掘发现过程。利用该方法分析用户实际网上游览数据，得到了许多有意义的模式。

在研究用户购买行为时，使用实际数据集进行实验存在一定的局限性，无法从各个方面对所提出的方法进行测试和评估。因此模拟用户购物行为，生成用户购物对应的交易数据流是一种辅助研究的方法(Liu，*et al.*，2009；Liu，*et al.*，2010；Liu，*et al.*，2011)。已有的数据流生成器不能构造数据项的概念层次结构以及改变数据流格式，存在一定的局限性。为此我们提出一种新的人工层次数据流生成器，用于解决这些问题，它可以按照用户要求随机构造出贴近现实的概念层次结构，既可以生成固定维数的数据流，也可以生成维数可变的数据流，同时兼容了一般数据流生成器的功能。

这两部分工作发表在国际期刊 *Informs Journal on Computing*(Yang，Liu 和 Cai，2013)和 *ACM SIGMIS Database*(Wang，Liu 和 Er，2009)上。本书第 3 章将详细介绍这两部分工作。

1.2.3 在线浏览行为分析

现实世界中很多时序数据中存在周期性频繁模式，如用户的周期性消费行为、周期性网上浏览行为、股票的周期性变化等，从大量时序数据中发现周期模式对于了解用户需求、为用户提供个性化服务、设计新型业务模式具有重要意义。已有方法通过检验每个周期是否符合给定阈值来判断。此

种方法有两个缺陷：一是从短的序列中发现的模式可能缺乏周期性；二是可能漏掉有意义的周期模式。为了解决这些问题，我们给出了衡量模式周期性的新的度量，该度量基于统计上的方差对模式的周期性进行量化，并将周期模式分为 4 种类型，其中包含前人尚未涉及的具有更高发现难度的类型，设计了高效的算法，可以发现更全面、更合理的周期模式。通过合成数据以及实际的用户网上浏览时序数据上的实验结果表明，所提方法比已有方法更合理、有效。

这部分工作的成果发表在国际期刊 *Informs Journal on Computing* 上(Yang, *et al.*, 2012)。本书第 4 章将详细介绍该部分工作。

1.2.4　在线评论意见挖掘

随着 Web 2.0 的发展，用户网上创建的信息急剧增长，产品评论就是其中的一类信息。产品评论反映了消费者对产品的意见和态度，具有很高的利用价值。一方面，评论中表达的观点和情感可以对其他消费者的购买意向产生影响；另一方面，便于商家对产品的质量或服务进行改进，提高客户满意度，也是开发和销售新产品的很好依据。比起传统的市场调查方法，通过网上信息的收集和分析能够获取更广泛用户的意见，而且也便于跟踪意见随时间的演变。辅以对客户群的分析也便于商家进行目标销售。因而，意见挖掘是近些年来学术界和工业界的研究和应用热点。

已有在线评论的分析方法大多针对英文较规范的文本，用于分析中文网上产品评论时具有较低的准确度和召回率。为此，我们提出了针对中文产品评论信息的意见挖掘方法。针对中文表达以及口语表达的特点提出了两种新颖的抽取产品特征和意见词的方法。该方法依据少量的可以用于跨领域的常用副词的集合，通过相互迭代的自助方法发现意见词和特征词，经过实际数据集的实验结果表明，新提出的方法具有更高的准确度，并且适用于各种不同的领域。同时提出的情感分析方法可以解决已有方法无法确定所有意见词极性的问题。

这一部分工作的相关论文发表在知识发现国际会议 ACM SIGKDD (Liu, *et al.*, 2008)和国际期刊 *Electronic Commerce Research and Applications* (Liu, *et al.*, 2013)上。本书第 5 章将详细介绍意见挖掘相关工作。

1.2.5　基于在线行为的推荐

随着电子商务的发展，网上可购买的产品和服务种类越来越多，用户欲购买的产品或服务淹没在海量的信息之中，需要付出大量的时间和精力进

行搜寻。为了解决这种信息过载问题，推荐系统一方面可以辅助商家为客户推荐个性化的商品，另一方面也可以节省用户搜索所需商品的时间。因此，有效的推荐方法一直是学术界和工业界关注的研究和应用热点。经典的推荐方法，协同过滤推荐方法是使用非常广泛而有效的一类方法，然而这种方法需要用户对商品或服务进行打分，打分数据的稀疏性影响了此类方法的有效性。为了发现用户的偏好，我们研究了如何从用户的产品或服务评论中提取用户偏好的方法，基于用户对已购商品或服务的细节评价，通过与其他用户的比较，发现用户对商品或服务的偏好特征，更详细地描述用户的偏好。进一步地，根据其他用户对商品的评论，提出了进行商品推荐的方法，有效地解决了数据稀疏性带来的无法推荐的问题，提高了推荐的准确率。

冷启动问题一直是推荐系统需要解决的另一个难题。我们针对在线社交系统中如何为新用户推荐合适的配偶问题展开研究，提出了基于群体智慧的行为预测模型，提出了一个三阶段推荐方法。首先，根据老用户之间的交互链接信息将用户进行分群，使得在择偶方面有相似偏好的用户聚集成群。其次，对于没有历史交友信息的新用户，计算其属于每个用户群的概率。最后，提出一种新颖的度量方法，估计一个用户群的用户对某推荐用户采取各种回应行动的概率，这样可以避免数据稀疏带来的问题。基于一个新用户隶属每个用户群的概率和用户群对推荐用户的概率，就可以预测新用户对某用户的反应行为并进而进行推荐。该方法有效地解决了冷启动问题和数据稀疏问题，提高了推荐的准确率。

这两部分工作的相关文章发表在国际期刊 *Electronic Commerce Research and Applications*（Liu，*et al.*，2013）以及国际会议 ADMA（Wang，*et al.*，2011）上。本书第 6 章将详细介绍这两部分工作。

1.2.6 在线标注行为分析

互联网上的资源，结构化数据和非结构化数据大量共存，文本与图像、影像以及声音等多媒体数据大量共存，如何有效管理和共享这些数据资源，从而为用户提供个性化服务是一个值得研究的问题。数据标签（tag）是解决该问题的方法之一。然而，不同的用户对相同或相似的资源采用不同的标签进行标注，这给资源的分析和共享带来了困难。为了解决此问题，鉴于数据的异构性，我们提出了利用资源与标签之间的标注关系衡量标签之间关系的方法，提出了通过二部图描述两者之间联系，并进而基于随机游走理论计算标签之间相关度的方法。实验结果表明，该方法比传统方法更加有