文章编号：1003-0077(2012)06-0027-11

微博及中文微博信息处理研究综述

文坤梅，徐 帅，李瑞轩，辜希武，李玉华

（华中科技大学 计算机科学与技术学院，武汉 430074）

摘 要：微博即微博客，是 Web2.0 时代下衍生出的一种新型社会网络，其简单快捷的操作方式和随时随地发布信息的互动形式成为互联网的一大亮点。自 2006 年美国 Obvious 公司推出全球首个微博服务 Twitter 后，微博以惊人的发展速度受到国内外研究人员的广泛关注。该文首先对以 Twitter 为代表的微博其研究现状进行综述，主要包括(1)微博社会网络的特性分析，如微博用户网络的结构特征、微博用户的影响力分析及消息网络的信息传播机制等；(2)微博内容的语义分析，对微博中的情感语义分析进行了重点阐述；(3)微博的相关应用，包括微博在事件监测与预警、安全隐私及实时检索中的应用。然后概述了中文微博的研究现状，包括中文微博的特性及知识发现，分析了中文微博与英文微博的主要区别。最后讨论目前微博研究中存在的问题及未来中文微博的研究方向。

关键词：Twitter；中文微博；信息处理

中图分类号：TP391　　**文献标识码**：A

Survey of Microblog and Chinese Microblog Information Processing

WEN Kunmei, XU Shuai, LI Ruixuan, GU Xiwu, LI Yuhua

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Microblog is a new social network developed in the Web2.0 era, with the simple and quick operation for a post anytime and anywhere through the interaction form. These features make Microblog boom with a highlight in the Internet since 2006, when the Obvious company of the United States launched the world's first Microblog service named Twitter. This paper firstly introduces the state-of-art research on Twitter, including 1) feature analysis on Microblog social network, e.g. the structure of Microblog users' network, the Microblog users' impact analysis and the data diffusion mechanics in the information network; 2) semantic analysis, i.e. emotional semantic analysis on Microblog; 3) related applications in Microblog, e.g. event monitoring and warning, security, privacy and real time search. Then we summarize the research on Chinese Microbolg, including the feature and knowledge discovery of Chinese Microblog, and the differences between English and Chinese Microblog. Finally, we discuss the problems in the future research on Chinese Microblog.

Key words: Twitter; Chinese microblog; information process

1 引言

微博(Microblog)即微型博客，是一种集成化、开放化的互联网社交服务，用户可通过 Web、即时通信、电子邮件和手机等方式用很短的文字描述其当前状态。为方便与移动终端的接入，其每一篇“微博”限定在 140 字左右，即一条手机短信的长度限制，同时也可上传音视频、图片。用户与用户之间信息的传递通过“关注—被关注”(Following-Follo-

收稿日期：2011-10-27　**定稿日期**：2012-03-29

基金项目：国家自然科学基金资助项目(61173170，60873225，70771043)；国家高技术研究发展计划(863 计划)资助项目(2007AA01Z403)；湖北省自然科学基金资助项目(2009CDB298)；中央高校基本科研业务费专项资金资助项目(华中科技大学自主创新研究基金 2011TS135，2010MS068)；CCF 中文信息技术开放基金

作者简介：文坤梅(1979—)，女，讲师，CCF 会员，通信作者，主要研究方向为信息检索、社会网络；徐帅(1984—)，男，硕士研究生，主要研究方向为情感挖掘；李瑞轩(1974—)，男，教授，主要研究方向为 Web 数据管理、信息安全。

wer)来实现,用户之间在微博平台上通过转发的方式对其他用户的微博进行传播。微博的出现以2006年推特(Twitter)[①]的创办为标志,从2008年开始Twitter得到了广泛的应用,至2011年底,Twitter拥有注册用户4.65亿。中文微博在近两年也得到了迅速的发展,以新浪微博为代表,包括腾讯、搜狐、网易、凤凰等其他门户纷纷加入微博阵营。中国互联网信息中心(CNNIC)数据显示,至2011年12月,国内微博用户人数已达2.5亿。

微博在国内外获得了广泛的应用,已成为一种具有强大影响力的新型媒体。2008年11月印度孟买的恐怖袭击事件、2008年5月汶川地震等事件都是通过Twitter首发。微博具备4A特性(任何时间、任何地点、任何方式、任何人),随时随地任何人都可以成为信息传播者。在对热点事件的报道中,微博可以占据信息发布的制高点,但与此同时也存在多方面的问题尚待解决。

近年来在KDD、WWW、SIGIR、WSDM以及其他刊物和会议上有越来越多的研究工作开始关注微博。目前,对微博的研究主要从两方面展开:一是分析微博的社会网络特性,微博是一种新兴的社会网络,因此也具有社会网络的各种特性,微博社会网络可以划分为两类,微博用户形成的社会网络和微博消息在传播过程中形成的社会网络,很多研究都是围绕这两种类型的社会网络展开特性分析;二是分析微博内容中蕴含的语义信息,微博信息呈现文本碎片化、海量等特点,信息利用成本高,无法直接获取微博中蕴含的语义信息,因此很多研究试图从微博内容中挖掘语义信息,特别是情感语义,用于发现用户对于热点事件的观点。除此之外,还有相关的微博应用研究等。大部分的研究工作是基于Twitter的,目前面向中文微博的研究工作还很少。

研究微博及中文微博信息处理技术具有重要的理论和应用价值,在管理领域,能够帮助领导者更快地了解群众对各类政策措施的反馈意见;在政策风险及舆情分析上,微博用户具有更高的自由度,其内容比博客更加难以监控,进行面向中文微博的语义分析和观点挖掘研究,也是实现内容监控、突发事件预警及舆情分析的基础;在商业领域中,能够为企业进行市场分析、市场调查、顾客反馈提供更多有价值的信息。该领域的研究成果在政府舆情分析、事件监控及企业商业智能系统等诸多领域有着广阔的应用空间和发展前景。

本文首先阐述以Twitter为代表的微博研究现状,主要包括微博社会网络(用户社会网络和消息传播网络)的特性分析、微博内容的语义分析以及微博在应用领域的研究等,然后概述中文微博的研究现状,最后对微博研究目前存在的问题进行总结,进一步探讨中文微博的研究前景。

2 微博社会网络的特性分析

一般在线社会网络中存在用户网络和消息网络两类,Twitter也不例外。不同之处在于,Twitter社会网络中用户间的关联是基于一种"关注—被关注"的特殊关系建立起来的,微博用户可以任意关注某个用户而不需得到对方同意,同时该用户也可被任意用户所关注,其结果是微博用户社会网络成为一有向图,而不同于一般社会网络中的用户关系无向图。另外,Twitter中任意用户发布的消息都会被该用户的跟随者收到,跟随者中部分用户会因兴趣将其转发,使更多用户看到这条消息,基于这种特殊的转发关系(Retweet),使得Twitter消息网络在传播力上有明显的优势。因此,在微博社会网络的特性分析方面,研究人员主要集中在这两种不同类型的社会网络上,通过分析用户网络拓扑结构,研究其基本社会网络特性,如小世界、6段分隔和幂律分布等;以及微博作为一种新兴社会网络,研究其在用户关联关系、消息传播机制等方面所具有的特性。

2.1 微博用户网络的特性分析

2.1.1 基本社会网络特性分析

自2006年Twitter获得广泛应用后,微博这一新兴社会网络逐渐引起学术领域的关注,研究人员对其社会网络特性进行了相关分析。Java等人[1]对Twitter的基本功能及特点进行了详细介绍,并对其社会网络特性进行了初步分析,数据集包括76 177个用户和1 348 543条微博信息,结果表明Twitter表现出一定的幂律分布和小世界等特性;同时还研究了Twitter用户社会网络的拓扑结构和地理位置等特征,并从个人和社区两个不同层次对用户使用Twitter的意图进行了分析,结果表明用户一般通过Twitter讨论日常事件或共享信息。Kwak等人[2]对整个Twitter进行了定量分析,数据集包括Twitter上的4 170万用户、14.7亿用户社会关系、4 262个热点话题和1.06亿微博等大量数

① http://www.twitter.com

据信息，通过分析 Twitter 用户间“关注—被关注”的拓扑结构，对 Twitter 用户社会网络统计特性进行了分析，统计结果表明 Twitter 在一定程度上表现出用户间的互惠性，但其社会网络特性较一般社会网络存在一定的偏差，例如，用户的 follow 数并不呈现幂律分布以及分割度更小等不同于一般社会网络的基本特性。然而，Wu 等人[3]则发现 Twitter 中存在明显的互惠性，通过将用户分为名人、媒体、博主和组织这四种类别，发现相同类别的用户间往往更可能存在关注关系。Gupte 等人[4]通过研究分析现实社会中社会阶层的概念，提出一种有效的探测和度量算法可以在有向用户社会网络图中发现社会阶层。

2.1.2　用户影响力探测

在用户网络中发现用户影响力不仅有助于用户推荐，对于微博网络中的商业运营模式也有着重要的意义，如利用用户影响力实现广告推送等。因此，用户影响力探测也是微博用户网络特性分析中的一个研究热点。微博用户影响力探测的方法可分为两类。一种方法是利用用户关系网络图的整体拓扑结构探测用户影响力；另外一种方法则是通过用户发布微博的网络传播影响力间接探测用户影响力。

基于用户关系网络图的方法从两个不同的角度去度量用户影响力。最简单直接的方法是利用用户的关注数大小，即网络图中节点度的大小，来评定用户影响力的大小。这种方法计算简单但效果不佳。另一种方法则是将用户这种“关注—被关注”关系看作是 Web 网页间的超链接关系，利用 Web 网页排名中常用的 PageRank 和 HIT 等算法进行用户影响力评定。例如，Java 等人[1]利用 HITS 算法对 Twitter 用户网络图中用户影响力进行探测；Kwak 等人[2]则利用 PageRank[5]算法对 Twitter 中的用户影响力进行探测，并通过用户的跟随者数和用户发布微博的转发数等不同方法与之进行对比；Weng 等人[6]提出了一种 TwitterRank 算法，在 PageRank 算法基础上，考虑用户所关注话题间的相似度和用户关系拓扑结构，从而发现 Twitter 中与话题相关且具有一定影响力的用户。

基于微博在整个用户网络中的传播覆盖度，即用户发布微博的被转发次数或其他用户在微博中提到该用户的次数，来度量用户的影响力大小。如 Cha 等人[7]对比分析了 3 种不同的用户影响力度量方法：用户的跟随者数、用户的微博转发数和用户在微博中通过“@”被关联的次数，认为用户的跟随者数越多，并不能真正说明该用户在用户群中的认可度越高，而用户的微博转发数以及用户在微博中通过“@”被关联的次数则能更准确地度量用户的实际影响力。前文提到的 Kwak 等人[2]在分析 Twitter 用户网络的基本特性时，提出了微博转发树的概念，但并未用于度量用户影响力，而 Bakshy 等人[8]则利用转发树的概念作为用户影响力的度量标准，认为在 Twitter 中用户的影响力是通过用户发布微博的转发规模所决定的，即消息传播的广度和深度。

2.1.3　用户特征分析与分类

通过微博用户社会网络分析用户特征，并根据这些特征进行用户分类也是重要的研究内容之一。例如，Krishnamurthy[9]等人通过分析 Twitter 用户关注和被关注数之间的关系分析了用户的特征，将用户分为三类：广播人（broadcaster）、一般人（acquaintance）和垃圾虫（miscreant）。有研究人员通过定量分析用户使用 Twitter 的行为模式，探测用户网络中的垃圾消息传播者，并分析用户使用 Twitter 的目的，如信息查询、信息共享以及维持自己的社会关系等[10]。Pal 等人[11]收集同一主题中的微博，然后提取该主题下所有微博发布者的特征，并根据其特征将用户聚成两类，将所聚类别中的作者进行排序，并找出最具权威的用户，实验结果对权威用户的发现提供了许多有用特征。

现实社会网络中，用户间各种不同的关联关系是不尽相同的。例如，用户 A 与用户 B 是基于朋友关系建立的关联关系，而用户 A 与用户 C 可能存在一种敌对的关联关系。微博作为现实社会网络在虚拟互联网中的具体展现，相应地，微博用户社会网络中不同类型的链接关系也必然存在差异，研究用户间不同的链接关系对于更深入的理解微博社会网络特性有着重要的作用。Welch 等人[12]认为在 Twitter 网络结构图中不同节点之间的边代表用户间不同的链接关系，分别针对用户间的 follow 关系和微博转发关系进行了相应的分析，并指出利用这种链接关系对用户排名算法有较好的改进。

2.2　微博消息网络的特性分析

与一般在线社会网络相同，Twitter 也允许用户在线、实时发布文本信息，然而，不同的是 Twitter 在信息长度上限制在 140 个字符之内，同时语法结构自由，支持手机等移动设备实时发布信息，这使得 Twitter 消息传播网络无论是在传播范围上还是速度上都具有更大的优势。因此，微博消息网络的特

性分析及消息在网络中的传播机制也是最近的研究热点。

Yang 等人[13]从用户贡献模式(即用户每月发布微博数目的分布情况)、Web 导航(即用户发布微博中含有超链接的目的指向)和用户社会网络整体结构模式等三个方面对比分析了 Twitter 与传统博客在信息传播结构上的区别。Kwak 等人[2]认为转发方式是 Twitter 消息传播中最有效的方式之一,基于微博转发关系,针对不同的热点话题,构建了微博转发树,并对微博转发机制进行了研究。通过对微博转发树的广度进行分析发现,Twitter 用户并非通过直接接收的方式获得信息,即大部分用户并不是该消息发布者的直接关注者,而是通过用户与用户间转发微博而间接收到消息,且微博一经转发,不管用户关注者有多少,该微博总会被传播到一定数量的用户。对微博转发树的深度进行分析发现,微博转发树中约占 97.6%其转发深度小于 6。这体现了 Twitter 消息网络中信息传播范围广且速度快的特点,即病毒式传播特点。在这种病毒式传播网络中,研究分析哪些微博被转发的可能性较大,从而预测出可能被转发的微博,然后在此基础上根据不同需求利用预测结果,其价值是相当可观的,如文献[14-16]等就针对该研究点进行了相关工作。在微博中对实时热点话题的广泛讨论是一大特色,然而不同类型的话题在传播机制上存在一定的差异。Romero 等人[17]基于 Twitter 中利用"#"符号来标示话题的特点,研究分析了 Twitter 消息网络中不同类型话题的传播特性。Sadikov 等人[18]还针对消息在传播的过程中导致信息丢失的问题做了相关研究。

笔者认为深入研究微博这一新兴社会网络的整体拓扑结构特性,无论是对于评估当前的微博本身,还是实现基于微博的应用都具有重要意义。然而,目前大部分的研究都是基于 Twitter,而针对中文微博的相关研究还很少,因此,在中国以新浪微博为代表的在线社会网路快速发展的同时,如果能够深入研究中文微博的拓扑结构及其基本特性,将为国内在线社会网络未来的良性发展提供重要的保证。

3 微博内容的语义分析

微博不仅具有社会网络的结构性特征,微博内容本身也包含了丰富的语义信息。基于微博内容的语义分析,其研究工作主要是从用户发布的微博内容中挖掘出有价值的信息,可分为面向事实(Fact-Oriented)的文本挖掘和面向观点(Opinion-Oriented)的文本挖掘两类。其中面向事实的文本挖掘主要包括热点话题探测[19]、主题抽取、垃圾信息处理、自动摘要等,在本文中归纳为微博内容的基本语义分析。而面向观点的文本挖掘即情感分析或观点挖掘是指从用户发布的信息中挖掘出其对讨论主题的潜在情感信息。因此,基于微博的情感语义分析研究工作主要是指对微博内容进行情感分析和观点挖掘。

3.1 微博内容的基本语义分析

微博为用户提供了更加便捷的日志工具,用户可通过微博发布大量的日常信息,而这些信息中通常隐含着用户的兴趣爱好,因此,与基于 Web 网页内容的自动标注[20]类似,可利用微博内容自动为用户生成标签,如 Wu 等人[21]利用 TF-IDF 与 TextRank[22]两种不同的算法来自动提取用户发布微博中的关键词,从而标注用户的兴趣爱好,其中 TextRank 算法的效果明显好于 TF-IDF 算法。

基于微博内容的文本自动摘要较传统文本摘要技术存在以下两方面的困难,一是微博消息内容短小,垃圾信息较多;二是和传统文本相比,微博中涉及的话题范围较广。Zhao 等人[23]针对微博的特点提出了基于上下文话题相关的 PageRank 算法,对微博进行关键词提取和排序,然后利用基于概率的得分函数计算关键词短语间的相关度和兴趣度,最后利用这些关键词对某话题特定时间段内的所有微博进行自动摘要生成。

在基于微博的话题探测方面,Zhao 等人[24]还提出了非监督 LDA 话题模型的改进形式 Twitter-LDA 模型,对 Twitter 与纽约时报在信息传播力(包括内容和速度两方面)进行了对比,并认为 Twitter 传播力更强。研究微博内容的价值也是值得关注的方向之一,如 Hong 等人[25]利用微博的转发次数作为度量微博流行程度的度量标准,并利用机器学习的方法,通过分析微博的内容、微博的时间特性、消息和用户的元数据以及用户社会网络图作为特征,预测新的微博发布后在多长时间内会被转发。

由于每个用户都可使用微博发布信息使得在微博网络中信息泛滥,最终导致信息的平均可靠度也随之下降。Castillo 等人[26]分析了微博的可信度,利用四个特征来度量微博的可信度:基于消息的特

征，如消息的长度、是否存在“#”符号、是否存在问号或感叹号以及情感词汇的数目等；基于用户的特征，如用户注册时间、关注人数、被关注人数、过去发布微博的数量等；基于话题的特征，如有多少微博包含 URL；基于消息传播的特征，如微博转发树的深度和广度等。结果表明：可信度高的微博被转发次数也较多；微博的原始发布者一般集中在少数用户中；转发微博的用户往往具有转发的习惯。曹鹏等人[27]提出了一种 Twitter 中近似重复消息的判定方法，统计字符种类和最短编辑距离两种字符串距离以判定 Twitter 中近似重复的消息。该方法可在一定程度上提高微博的信息利用率。

3.2　微博内容的情感语义分析

3.2.1　传统的情感分析和观点挖掘

情感分析（Sentiment Analysis）也可称为观点挖掘（Opinion Mining）[28-30]，随着 Web2.0 的发展，越来越多的用户在网络上发布具有不同情感趋向的信息，研究这些用户信息中潜在的情感信息，挖掘用户的潜在观点一直是研究的热点，但已有的观点挖掘研究主要集中于在线产品评论或传统博客上，较少有针对微博的观点挖掘研究。

文献[31]中提出了一种基于词汇的方法，该方法用简单的观点词汇来确定观点的情感语义倾向。观点词汇是指经常被用于表达正面或者负面情感的词，这种方法从根本上取决于出现在对象或对象特征附近的正面或负面观点词个数。如果正面观点词个数大于负面观点词个数，那么最终观点就是正面的，否则为负面的。观点词汇集合利用英文词网（WordNet）[32]通过引导过程得到，这种方法简单有效，能给出较合理的结果，但也存在较大的问题，观点词是依赖于内容的，在不同的语境中它所表达的语义倾向可能完全不同。Ding 等人[33]提出了一种基于全局词汇的方法，该方法充分利用了外部证据和自然语言表达中的语言约定。在中文领域，复旦大学朱嫣岚等人[34]利用类似于 WordNet 的中文知网（HowNet）进行了一些理论和试验研究；章剑锋等人[35]将同一句子中共现的评价词与评价对象作为候选集合，应用最大熵模型并结合词、词性、语义和位置等特征进行抽取评价词和目标对象之间的关联关系，具有一定的效果。哈尔滨工业大学杜伟夫[36]提出一个可扩展的词汇语义倾向计算框架，将词语语义倾向计算问题归结为优化问题，通过实验证明了方法的有效性。中国科学院计算所刘群等人[37]提出了一种基于知网的词汇语义相似度计算方法；廖祥文等人[38]提出一个基于概率推理模型的博客倾向性检索算法，该算法把主题相关性评分和倾向性评分合并到一个统一的概率推理理论模型，实验证明该算法针对传统博客是有效的。

综上所述，尽管有许多研究工作针对观点挖掘展开，但始终没有一个一般性的框架或者模型能清楚的描述观点挖掘中的各个方面及它们之间的联系。微博无论是在内容和形式上与传统 Web 信息都存在较大的差别，微博具有单一性、碎片化、开放性及实时性等特点，而传统 Web 信息具有多样性及完整性等特点，其更新及传播速度也相对较慢。已有的观点挖掘方法是针对传统 Web 信息，并不能完全适用于中文微博中的观点挖掘。

3.2.2　基于微博的情感语义分析

微博正日益成为一个普遍流行的实时性交流工具，大量网络用户每天都会发布并传播高达几千万的微博，在这些微博中包含着不同用户的日常生活记录，因此微博为情感分析与观点挖掘提供了丰富的数据来源，从中挖掘出相关用户对某个特定主题或事件的观点，如对使用过的产品或服务的满意程度以及用户的政治或宗教观点等。同时，由于微博内容简短、结构自由、实时性高且数据量大也为进行用户情感分析和观点挖掘提出了挑战。

微博用户和用户之间存在社会网络关系，用户通过微博所体现的观点集之间也存在语义上的关联关系，而事件特征之间同样存在隐含的关系。这些隐含的关联对情感分析会产生潜在的影响。如图 1 所示，在微博中，用户对事件或者事件特征表达某种观点，u 代表用户，f 代表事件特征，o 代表用户观点。在 u 空间形成了微博用户社会网络，在 f 空间形成特征关系网络，而在 o 空间形成观点语义网络，因此，这三者（用户、事件特征和观点）之间存在关联关系，需建立三维关联关系 R(u, f, o)。

Bermingham 等人[39]研究结果表明，针对微博进行情感分析相对传统博客的效果将会更好，微博已经成为情感分析与观点挖掘的有效文本领域。Go 等人[40]利用机器学习的方法对微博消息进行情感分类，即判断一条微博消息的情感倾向是正面还是负面。在训练集的选择上，利用微博中的表情作为类别标记，然后利用朴素贝叶斯和支持向量机等不同的分类算法训练分类器，从而实现微博的情感分类。Kim 等人[41]研究了 Michael Jackson 的死亡对 Twitter 用户产生的情感影响，结果表明在这段

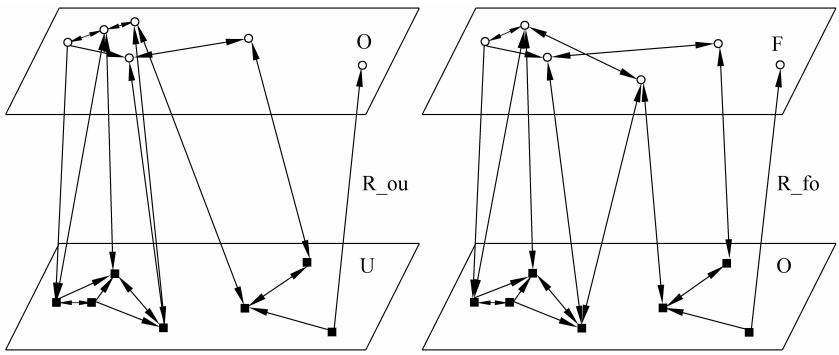

图1 用户、观点和事件特征的内外部关联关系

时间内用户的情感普遍表现出低落的趋势。Jansen等人[42]对微博进行随机抽样分析,结果表明大概19%的微博会涉及到针对某个产品品牌的评论,并利用自动分类的方法提取不同用户对相应产品的情感倾向性,指出Twitter可以作为在线市场营销的重要工具。另外,还有研究人员根据Twitter用户发布的微博探测目前股票市场的走势[43]以及预测总统选举的结果[44]等。在应用方面,已开发了用来测量微博褒贬倾向性的在线工具TweetFeel①。

笔者认为以上研究主要存在以下问题:(1)仅利用微博中包含的表情分类训练集中文本的褒贬存在一定偏差,并可能导致各类的训练文本数相差较大;(2)大部分的工作都是利用统计微博中词汇的出现频率来确定其情感倾向性,并没有考虑其上下文语境,从而影响最终结果的正确性;(3)大部分的分析方法将传统的情感分析方法移植到微博中,并没有深入考虑微博自身具有的特点。因此,如果能在传统情感分析方法的基础上,更多考虑微博自身的独特之处,将能取得更好的情感挖掘效果。

4 微博中的应用研究

4.1 微博事件检测与预测

Sakaki等人[45]通过实时监控微博用户的状态更新来进行地震探测,并实现了一个地震探测系统。该方法首先对目标事件进行分析,提取目标事件的特征属性,然后利用机器学习算法将监控到的所有微博用户的状态信息进行分类,最后对分类结果计算出目标事件信息,利用基于概率的时空模型定位地震源,将每个微博用户看作是一个传感器,每个用户发表的每一个微博状态信息被视为传感信息,利用普适计算中普遍使用的过滤方法定位地震源。文献[46]将利用图像中的像素概念来表示微博用户对某事件的兴趣,并结合微博包含的时间信息,将事件相关微博表示成类似视频文件的动态结构,用来监测事件发展的形势变化情况。

4.2 微博中的安全及隐私

由于共享信息的私有性,数据隐私在微博中具有独立的安全需求。Zhang等人[47]对在线社会网络中的隐私安全问题进行了探讨,提出了在线社会网络包括Twitter等在设计上存在的问题和挑战,给出了一个统一的框架来评价当前及下一代在线社会网络的安全性及隐私保护。Sun等人[48]提出了一种有效废止的方法提供在线社会网络的隐私保护,一旦联系人从社会组中被移除,将会遏制该联系人访问权限,同时该方法还具有高级特征,如有效搜索加密文件以及动态改变社会组成员。

4.3 微博实时检索

微博具有较高的实时性,关于提高搜索质量的研究有许多。这些研究大都集中在:(1)根据查询和文档的相似性进行排序。目前,这种技术包括产生锚文本、抽取元数据、分析链接关系和挖掘用户日志等。(2)根据链接关系计算出文档质量。而微博的实时性及其海量数据,决定了传统的搜索技术并不能完全应用于微博信息检索。Teevan等人[49]给出了关于微博搜索的相关分析工作。通过工具栏的内嵌方法采集了大量用户的微博搜索日志与其Web搜索日志,并对日志数据进行了对比分析。结果表明,微博搜索用户更愿意搜索时间性强的内容,包括突发新闻、实时报道和时势动态;微博搜索语句往往更短、更热门且常常被重复查询。

① http://www.tweetfeel.com/

4.4 微博中的其他应用

微博在不同领域已得到了广泛的应用，包括其在政府、教育、市场等方面的应用。Barau 等人[50]将 Twitter 应用到英语教学中；Ebner 等人[51]针对教育方向研究了 Twitter 的适用性，特别是在移动学习方面；文献[52]通过微博用户的兴趣和微博内容来定义领域描述特征，从而将微博短文字分为预先定义好的类。Pujol 等人[53]设计并实现了一个可以提高微博在线网络服务可扩展性的中间件 SPAR，该中间件通过平衡社会图的结构，以最小的复制成本在本地获得数据。Duan 等人[54]提出另一种新的排序策略，不仅利用了微博内容的相关性，还考虑了其权威性以及 URL 等特征，其结果表明微博是否包含 URL、微博长度及其权威性是排序的最佳组合。Sarma 等人[55]研究了微博的排序机制。Huang 等人[56]利用统计的方法研究了 Twitter 上的标注现象，包括标注本身和用户标注的意图，并对 Twitter 上出现的标注现象进行了解释。除了上述基于微博的学术研究，目前也开始出现基于微博的应用平台。例如，Tweettronics① 提供了对品牌与产品相关微博信息进行分析的平台，主要用于市场目的。该平台可以将用户微博分为正面和负面两种评价，同时可识别出有影响力的用户。

5 中文微博的信息处理

以 Twitter 为代表，基于英文微博的研究正不断取得新的进展，最近召开的 WWW2011、SIGIR2011 以及 WSDM2011 等世界著名计算机会议上微博相关的文章占了较大比例，说明针对微博的研究是目前的一大热点，然而在中国虽然有新浪，网易、腾讯等知名微博服务提供商的蓬勃发展，中国微博用户都已数以亿计，但针对中文微博的学术研究比较匮乏。分析其原因，除了中文微博属新兴服务之外，比较重要的因素在于，中文微博中的信息以汉语语言形式存在，在信息处理领域，中文信息处理一直要比英文信息处理更具挑战性。因此，笔者认为随着国内互联网的快速发展，无论从商业价值还是社会研究价值，分析研究以微博为代表的中国在线社会网络的意义重大，应用前景广阔。大致可以从以下几个方面进行研究。

5.1 中文微博的特性

针对微博社会网络的特性分析研究已经有了初步的研究成果，在前文中也提到，主要有微博用户社会网络特性和微博消息网络特性两方面，目前仍有研究人员在进行更深入的研究工作。与此同时，大部分人认为以新浪微博为代表的中文微博与以 Twitter 为代表的英文微博没有太大的区别，然而最近惠普实验室发表的一篇论文[57]中指出，新浪微博无论在实现模式上还是在微博内容上都与 Twitter 存在较大的差别，Sina 微博与 Twitter 的统计对比分析如表 1 所示。究其原因，笔者认为这是由于中国拥有世界上最大的网民数量，在线社会网络服务的崛起呈现出爆炸性增长，且大部分的用户都来自中国，大部分的内容信息以中文语言形式存在。同时，与西方国家（包括美国）的社会网络发展轨迹不同，国内社会网络呈现出不同的特性。这与中国宏观环境密切相关，包括经济的飞速发展、技术基础设施的快速扩张以及社会的转型等，因而中国在线社会网络服务（以新浪微博为例）呈现出不同的发展趋势和特征。

表 1　Sina 微博与 Twitter 的统计对比分析表

	Sina 微博	Twitter
热门话题的主要来源	娱乐类内容	新闻类内容
微博的转发频率（前 10 名用户）	48 622.53（次/人）	77.98（次/人）
微博涉及的话题数（前 10 名用户）	37.5（个/人）	67.4（个/人）
热门话题的转发比例	62%	31%
微博的内容	包含文本、图片、视频、链接等	仅包含文本、链接
影响力用户类别	大部分是非认证用户	大部分是认证用户

微博在中国正获得蓬勃发展，国外科学家和社会学家也逐渐开始从国际视角来对中国的在线社会网络进行研究。因此，国内研究工作者也应该把握住这样的机遇，中文微博是国内在线社会网络的研究热点，对中文微博社会网络的特性分析和中文微博内容的语义分析，无论对当前系统的优化还是开发新的应用系统都有着至关重要的作用。

① http://www.tweettronics.com

5.2 中文微博的知识发现

目前,我国互联网用户已达 4.52 亿,成为仅次于美国的互联网用户世界排名第二的国家,另外,微博在我国近两年发展迅速,受到广大网民的喜爱。究其原因,主要在于微博的方便、快捷、实时和高效等特点,同时微博内容更加自由且对用户的写作能力要求更低,微博用户可以较传统博客用户更方便地发表观点。由于微博客观真实地反映了由个体所组成社会的整体状态,因此笔者认为应该针对微博背后的社会价值进行有效深挖,微博应成为执政者了解民意、分析舆情以及制定对策的快捷通道。特别是当今社会处于转型期,社会问题和社会矛盾前所未有的激化和突出,及时了解社会动态意义重大。

中文微博是适合中文信息处理的一种新文本模式,已有的中文信息处理技术有部分可直接应用于中文微博中,而微博不完全同于已有的短文本,它自身具有的简短、实时性及社会性等特征,应在研究中充分考虑。谢丽星等人[58]使用新浪 API 获取数据,针对中文微博消息展开了情感分析方面的初步研究。对于三种情感分析的方法进行了深入研究,包括表情符号的规则方法、情感词典的规则方法、基于 SVM 的层次结构的多策略方法,实验表明基于 SVM 的层次结构多策略方法效果最好。

在国内,针对中文微博的研究还不多见。笔者认为,将中文信息处理技术与微博自身特性相结合,推出更为智能化、更为个性化、更易于操作以及更加有利于组织和利用中文微博信息的方法与技术,将是未来中文微博领域较为前沿的研究课题。特别是针对中文微博的内容,进行基于中文微博的语义分析,挖掘中文微博中用户隐含的情感信息,在此基础上进行相关的预警预测及舆情分析,是具有重要理论及使用价值的研究课题。

6 存在问题和未来研究方向

目前,微博开始呈现比较广泛的研究,但是由于难以对其语义进行管理和应用,微博及中文微博的研究成果还不能令人满意。目前,在微博信息处理研究领域依然存在以下问题和挑战。

(1) 微博内容松散、信息呈现碎片化

微博追求快速传递,很多信息在发送过程中未经加工,文字内容松散,不能清晰有效地向受众传达事件信息。简洁的信息发布方式,促使用户频繁上传信息,信息超载现象较为严重。据调查,在 Twitter 上有 40.55%的内容属于毫无价值的信息。有效信息很容易被淹没,信息提取成本高昂。笔者认为针对这一问题,微博未来的发展可以呈现多元化的特点,如针对不同领域,提供不同方向的专业微博平台,从而部分解决微博内容太过松散的问题。

(2) 微博可信度

微博信息的发布取决于用户的自律,可信度受到质疑。不完整的信息经过用户不断转发后,难以找到信息源。而微博可信度从根本上依赖于发布微博的用户。因此对微博用户的研究也是很重要的课题之一。如何对微博用户进行分类,如何识别具有重要影响力的用户,这些都是需要重点解决的问题。通过建立完整的微博用户社会网络,分析微博用户社会网络的基本特性及社会属性,以此为基础形成比较完善的微博可信度评价体系,是解决这一问题的根本途径。

(3) 微博语义挖掘

微博由大众产生,当微博汇集在一起时,由于缺乏规范和层次性,导致很难从大量的微博以及微博用户间建立起层次结构的语义关系。如对热点事件跟进时,无法获取大众对热点事件的整体观点,也无法从整体层面获取大众的舆论导向。获取海量微博中有价值的信息,需挖掘微博社会网络所隐含的语义信息及情感关联。解决这一问题,需要在已有的语义分析及情感挖掘技术基础之上,结合微博自身的特点,提出新的微博语义分析方法,这也是目前尚待解决的重点及热点问题之一。

(4) 中文微博中的观点挖掘

专门针对中文微博的研究还属于起步阶段,很多方面的问题亟待解决。由于文化差异导致语言表达方式不同、语言结构的差异以及中英文词汇语法的差别,因此,研究基于中文微博的语义分析和观点挖掘就凸显其必要性,如何从浩瀚的中文微博中特别是热点话题中获取有效的信息,发现用户对于热点事件的观点,用于事件监测和趋势预测,是目前亟待解决的重点研究课题。另外在政策风险及舆情分析上,微博比博客更加难以监控其内容,进行面向中文微博的语义分析和观点挖掘研究,也是实现内容监控及舆情分析的基础,具有重要的研究和应用价值。在企业商业智能系统、政府舆情分析等诸多领域有着广阔的应用空间和发展前景。

(5) 垃圾微博的处理

目前,在提供微博的在线社会网络服务中,存在

大量恶意且无用的垃圾微博信息，这些垃圾微博十分不利于对网络资源的共享、检索和定位。对于垃圾微博，目前主要依靠手工检查和删除，其他很多提供此服务的微博应用大多采用手工方式。由于微博不同于一般网页，目前已有的垃圾网页识别方法并不能直接用于垃圾微博的处理中。因此，能够自动检测垃圾微博是当前须解决的一个问题，这一问题的解决将大大提高海量微博信息的有效利用率。

(6) 微博实时信息检索

微博在信息检索中的应用研究尚处于探索阶段，考虑到微博的高实时性及微博数据的海量性，微博检索将成为未来的研究热点之一。已有的信息检索技术并不能完全解决微博的实时检索问题。问题的难点在于快速且实时地将搜索结果更新给查询用户，并对实时结果进行正确排序。如何将已有的信息检索方法有效地融入到微博实时检索中，提出有效的微博索引构建机制和微博实时检索结果排序方法，帮助用户快速发现所需要的信息，在此基础上尽可能保证微博信息的实时性，也是目前亟待解决的问题之一。

7　结束语

随着微博在线社会网络服务的普及和微博用户的急剧增加，对微博的研究成为目前关注的重点，研究者已经在这方面做了大量的工作，本文对近几年来国内外在该领域的主要成果进行了回顾与总结，综述了微博的研究现状，包括微博社会网络的特性、微博内容的语义分析及其在观点挖掘及信息检索中的应用等，同时指出了仍然存在的问题和将来进一步研究的方向。总的来说，对微博及中文微博的研究仍然处于探索阶段，离商业应用还有很长的路要走，仍然有大量关键问题还需做深入细致的研究。

随着中国微博服务的蓬勃发展，新浪、腾讯等公司相继对外开放了其微博 API 接口，这对于研究中文微博所需的实验数据提供了方便，为研究中文微博提供了一个良好的契机。随着中文微博用户的日益增长，对中文微博的研究日趋重要。如何在已有的微博相关研究成果和观点挖掘方法的基础上，结合中文自然语言处理技术和中文微博自身的特点，提出新的模型和方法，挖掘中文微博中蕴含的语义信息及用户观点，使之能有效应用于热点事件监测及趋势预测中，是需要重点解决的问题。

参考文献

[1] A. Java, X. Song, T. Finin, et al. Why we twitter: understanding microblogging usage and communities. [C]//Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, 2007: 56-65.

[2] H. Kwak, C. Lee, H. Park, et al. What is Twitter, a social network or a news media[C]//Proceedings of the International Conference on Word Wide Web (WWW), 2010: 591-600.

[3] S. Wu, J. M. Hofman, W. A. Mason, et al. Who says what to whom on Twitter[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 705-714.

[4] M. Gupte, P. Shankar, J. Li, et al. Finding hierarchy in directed online social networks[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 557-566.

[5] A. Arasu, J. Cho, H. Garcia-Molina, et al. Searching the web [J]. ACM Transactions on Internet Technology, 2001, 1(1): 2-43.

[6] J. Weng, E. Lim, J. Jiang, et al. TwitterRank: finding topic-sensitive influential twitterers[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2010: 261-270.

[7] M. Cha, H. Haddadi, F. Benevenuto, K. P. Gummad. Measuring user influence on twitter: the million follower fallacy[C]//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 2010.

[8] E. Bakshy, J. M. Hofman, W. A. Mason, et al. Everyone's an influencer: quantifying influence on Twitter[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2011: 65-74.

[9] B. Krishnamurthy, P. Gill, M. Arlitt. A few chirps about twitter[C]//Proceedings the 1st Workshop on Online Social Networks, 2008: 19-24.

[10] D. Zhao, M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work[C]//Proceedings of the International Conference on Supporting Group Work, 2009: 243-252.

[11] Aditya Pal, Scott Counts. Identifying topical authorities in microblogs[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2011: 45-54.

[12] M. Welch, U. Schonfeld, D. He, et al. Topical semantics of Twitter links [C]//Proceedings of the ACM Conference on Web Search and Data Mining

(WSDM), 2011: 327-336.

[13] J. Yang, S. Counts. Comparing information diffusion structure in weblogs and microblogs[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[14] J. Yang, S. Counts. Predicting the speed, scale, and range of information diffusion in Twitter[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[15] S. Petrovic, M. Osborne, V. Lavrenko. RT to win! predicting message propagation in Twitter[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[16] J. Leskovec. Social media analytics: Tracking, modeling and predicting the flow of information through networks[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 277-278.

[17] D. Romero, B. Meeder, J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 695-704.

[18] S. Sadikov, M. Medina, J. Leskovec, et al. Correcting for missing data in information cascades[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2011: 55-64.

[19] 杨亮，林原，林鸿飞. 基于情感分布的微博热点事件发现 [J]. 中文信息学报，2012，26(1)：84-90，109.

[20] 靳延安，李瑞轩，文坤梅，等. 社会标注及其在信息检索中的应用研究综述 [J]. 中文信息学报，2010，24(4)：52-62.

[21] W. Wu, B. Zhang, M. Ostendorf. Automatic Generation of Personalized Annotation Tags for Twitter Users[C]//Proceedings of the Annual Conference of the North American Chapter of Association for Computational Linguistics (ACL), 2010: 689-692.

[22] Mihalcea, P. Tarau. TextRank: bringing order into texts[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004: 404-411.

[23] X. Zhao, J. Jiang, J. He, et al. Topical keyphrase extraction from Twitter[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011: 379-388.

[24] W. Zhao, J. Jiang, J. Weng. Comparing Twitter and traditional media using topic models[C]//Proceedings of the European Conference on Information Retrieval (ECIR), 2011: 338-349.

[25] L. Hong, O. Dan, B. D. Davison. Predicting popular messages in twitter[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 57-58.

[26] C. Castillo, M. Mendoza, B. Poblete. Information credibility on twitter[C]//Proceedings of the International Conference on World Wide Web (WWW), 2011: 675-684.

[27] 曹鹏，李静远，满彤，等. Twitter 中近似重复消息的判定方法研究 [J]. 中文信息学报，2011，25(1)：20-27.

[28] M. Hu, B. Liu. Mining and summarizing customer reviews[C]//Proceedings of the Annual Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2004: 168-177.

[29] N. Kaji, M. Kitsuregawa. Automatic construction of polarity-tagged corpus from HTML documents[C]//Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ ACL), 2006: 452-459.

[30] L. Zhuang, F. Jing, X. Zhu, et al. Movie review mining and summarization[C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), 2006: 43-50.

[31] A. Andreevskaia, S. Bergler. Mining WordNet for fuzzy sentiment: sentiment tag extraction from WordNet glosses[C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006: 209-216.

[32] G. A. Miller. WordNet: a lexical database for English [J]. ACM Transactions on Communication, 1995, 38(11): 39-41.

[33] X. Ding, B. Liu, P. Yu. A holistic lexicon-based approach to opinion mining[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2008: 231-240.

[34] 朱嫣岚，闵锦，周雅倩，等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报，2006，1(20)：14-20.

[35] 章剑锋，张奇，吴立德，等. 中文观点挖掘中的主观性关系抽取 [J]. 中文信息学报，2008，22(2)：55-59，86.

[36] 杜伟夫，谭松波，云晓春. 一种新的情感词汇语义倾向计算方法 [J]. 计算机研究与发展，2009，46(10)：1713-1720.

[37] 刘群，李素建. 基于《知网》的词汇语义相似度的计算 [C]//第三届汉语词汇语义学研讨会，2002.

[38] 廖祥文，曹冬林，方滨兴，等. 基于概率推理模型的博客倾向性检索研究 [J]. 计算机研究与发展，2009，46(9)：1530-1536.

[39] A. Bermingham, A. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? [C]//Pro-

ceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), 2010: 1833-1836.

[40] A. Go, L. Huang, R. Bhayani. Twitter sentiment analysis [R]. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.

[41] E. Kim, S. Gilbert, M. Edwards, et al. Detecting sadness in 140 characters: sentiment analysis of mourning Michael Jackson on Twitter [R]. Web Ecology Project, Boston, MA, 2009.

[42] B. J. Jansen, M. Zhang, K. Sobel, et al. Microblogging as online word of mouth branding[C]//Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, 2009: 3859-3864.

[43] J. Bollen, H. Mao, X. Zeng. Twitter mood predicts the stock market [J]. Journal of Computational Science, 2011, 2(1): 1-8.

[44] A. Tumasjan, T. O. Sprenger, P. G. Sandner, et al. Predicting elections with Twitter: what 140 characters reveal about political sentiment[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[45] T. Sakaki, M. Okazaki, Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors[C]//Proceedings of the 19th International World Wide Web Conference (WWW), 2010: 851-860.

[46] V. K. Singh, M. Gao, R. Jain. Situation detection and control using spatio-temporal analysis of microblogs[C]//Proceedings of the 19th International World Wide Web Conference (WWW), 2010: 1181-1182.

[47] C. Zhang, J. Sun, X. Zhu, et al. Privacy and security for online social networks: challenges and opportunities [J]. IEEE Network, 2010, 24(4): 13-18

[48] J. Sun, X. Zhu, Y. Fang. A privacy-preserving scheme for online social networks with efficient revocation[C]//Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM), 2010: 1-9.

[49] J. Teevan, D. Ramage, M. Morris. Twittersearch: A comparison of microblog search and web search [C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2011: 35-44.

[50] K. Borau, C. Ullrich, J. Feng, et al. Microblogging for language learning: using twitter to train communicative and cultural competence[C]//Proceedings of International Conference on Web-based Learning (ICWL), 2009: 78-87.

[51] M. Ebner, M. Schiefner. In microblogging more than fun? [C]//Proceedings of IADIS International Conference on Mobile Learning, 2008: 155-159.

[52] B. Sriram, D. Fuhry, E. Demir, et al. Short text classification in Twitter to improve information filtering[C]//Proceedings of the 33rd Annual Conference of the ACM Special Interest Group on Information Retrieval (SIGIR), 2010: 841-842.

[53] J. Pujol, V. Erramilli, G. Siganos, et al. The little engine(s) that could: scaling online social networks [C]//Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), 2010: 375-386.

[54] Y. Duan, L. Jiang, T. Qin, et al. An empirical study on learning to rank of tweets[C]//Proceedings of the 23rd International Conference on Computational Linguistics (COLING), 2010: 295-303.

[55] A. D. Sarma, S. Gollapudi, R. Panigrahy. Ranking Mechanisms in Twitter-Like Forums[C]//Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2010: 21-30.

[56] J. Huang, K. M. Thornton, E. N. Efthimiadis. Conversational tagging in Twitter[C]//Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 2010: 173-178.

[57] L. Yu, S. Asur, B. A. Huberman. What trends in Chinese social media[C]//Proceedings of the ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD), 2011.

[58] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取 [J]. 中文信息学报，2012，26(1)：73-83.