从语言计算到社会计算

刘 挺 徐志明 秦 兵 赵世奇 李 生
哈尔滨工业大学

关键词：社会计算 社会网络 自然语言处理

从大规模文本到社会媒体

目前，很多自然语言处理和信息检索的研究者都在谈论微博，研究微博，在自然语言处理和信息检索的顶级国际会议上也迅速出现了大量关于社会媒体分析的论文。这种情况让人不由得回忆起1990年8月在芬兰赫尔辛基举行的第十三届国际计算语言学大会（Coling′90），主题是“处理大规模真实文本的理论、方法和工具”。我国学者当时敏锐地意识到了这一变化，开始了大规模文本语料库的研究。在过去的20年中，自然语言处理的对象主要是大规模真实文本，新闻语料则一直是研究的重点。而如今，人们研究的对象正从孤立的中规中矩的文本开始转向以复杂社会关系网络为背景的“流动着的”短文本。

什么是社会媒体？社会媒体上的数据与传统的文本语料、网页有哪些区别？

社会媒体的出现与发展

社会媒体（social media）是一种在线交互媒体，具有广泛的用户参与性，允许用户在线发布和传播信息，相互沟通与协作，组成虚拟网络社区。

社会媒体兼具媒体属性和社交功能，其基础是社会网络（social network）。目前最受关注的社会媒体是社交网站（social networking services，SNS）和微博（MicroBlog），国外以脸谱（Facebook）和推特（Twitter）为代表，国内以人人网、新浪微博、腾讯微博为代表，典型的社会媒体还包括论坛、博客、维基百科和视频网站等。

从通信形式的演化进程看，社会媒体实现了人们通信的最复杂形式。最初的通信方式都是近距离的，包括一对一面谈（口口相传）、一对多演讲和多对多的会议。但人们不满足于当面交流，希望能够突破时空，因此从近距离通信到远距离通信实现了对空间的突破，从同步通信到异步通信实现了对时间的超越。电话/即时通信、广播/电视、视频会议分别实现了远距离的一对一、一对多和多对多的同步通信，而信件（无论是传统的信件还是电子邮件）和报纸/网站实现了远距离一对一和一对多的异步通信。但是，一直没有一种能够实现**多对多远距离异步通信**的形式，原因在于这是一种形式最复杂的通信。Web 2.0社会媒体填补了这一空白，产生了空前的信息传播效果。如表1所示。

表1 人类通信形式的类别与演化

	近距离（同步）	远距离	
		同步	异步
一对一	面谈	电话/即时通信	信件
一对多	演讲	广播/电视	报纸/网站
多对多	现场会议	视频会议	Web 2.0社会媒体

在各种社会媒体中，与语言处理关系最密切的是微博。微博以简短的文字记录用户的所思所见。用户与用户之间通过“关注”关系建立起人与人之间的关注网络，一个用户发出的微博可以在瞬

间显示在他的全体“粉丝”（关注他的用户）的屏幕上。微博内容的短小使得微博发布的门槛变得非常低，微博的网状传播结构也使得微博信息能够以极快的速度传播。这些特点使微博迅速成为一种新的社会媒体，并且其用户群不断得到扩大。根据中国互联网络信息中心最新的报告，截至2011年6月底，我国的微博用户已达1.95亿，在过去一年中微博用户数量的增幅高达208.9%。

社会媒体与Web 1.0传统媒体的区别

通常称传统的万维网为Web 1.0，而社会媒体是Web 2.0。下面以微博为例，分析一下Web 2.0与Web 1.0的区别。

表2 Web 1.0与Web 2.0的区别

	Web 1.0	Web 2.0（以微博为例）
信息来源	网站的编辑人员	广大用户
节点	网页（信息）相对静态	人（用户背景、微博）动态，内容更新频繁
节点分析	文本分析	用户兴趣分析
边	超链接（指向式）	关注关系（订阅式）边是信息转发通道
边的分析	链接分析	人物关系分析、社交圈分析
节点重要性计算	网页重要性计算	意见领袖发现
节点间的互动	转载	转发、评论
网络结构	变动慢	变动较快

从表2中可以看出，最根本的变化是网络的节点从“网页”变成了“人”，这个根本性的变化导致网络关系、网络结构以及信息内容等各方面发生变化。与以往网络媒体不同，社会媒体背后的社会网络以及借助社会网络而产生的社会话题传播的速度很快。要在社会媒体环境下开展语言计算的工作，必须充分利用社会网络的信息和话题传播的模式，在此基础上运用自然语言处理技术实现对社会媒体中人物和信息的更深层次的理解。

基于社会媒体的社会计算

社会计算研究框架

计算社会科学是社会计算研究的核心。2009年2月，戴维·莱兹（David Laze）（美国哈佛大学）等15位美国学者在《科学》（Science）上联合发表论文，题目是“Computational Social Science”[1]，宣告“计算社会科学”这门学科的诞生。大规模数据的出现推动了新学科的诞生和发展。20世纪50年代，大规模语料库催生了计算语言学；1989年，大规模基因图谱数据催生了计算生物学；今天大规模社会数据又催生了计算社会科学。

计算社会科学在各学科门类中处于什么位置呢？物质运动有五种形式：机械运动、物理运动、化学运动、生物运动和社会运动，此外，意识有其独特的运动形式。机械、物理和化学早已被计算技术所渗透，生物运动与计算技术的结合是计算生物学，意识运动与计算技术的结合是人工智能，只有社会运动目前还没有充分利用计算科学。如今，海量社会数据的出现为社会科学的研究提供了具有革命性的计算工具。

社会计算是一门理工管文相结合的交叉学科，研究范围非常广，在很短的时间内吸引了来自数学、物理学、计算机科学、管理学、社会学、传播学、心理学等诸多学科的学者。在计算机科学方面，部分过去从事网络科学、机器学习、数据库、多媒体、自然语言处理等方面研究的学者进入社会计算领域，从不同角度开展工作，形成了百家争鸣，异彩纷呈的局面。在此，笔者仅从自然语言处理研究者的角度尝试提出一个基于社会媒体的社会计算研究框架，供相关学者参考。

如图1所示，基于社会媒体的社会计算研究从社会媒体中采集社会网络（个人信息和社会关系信息）和媒体内容。以语言分析、数据挖掘和可视化为工具，重点研究社会媒体中人和信息的网络结构、传播模型，研究人和信息的各种属性，比如倾向性、可信度和影响力等。而群体智慧则是一个相

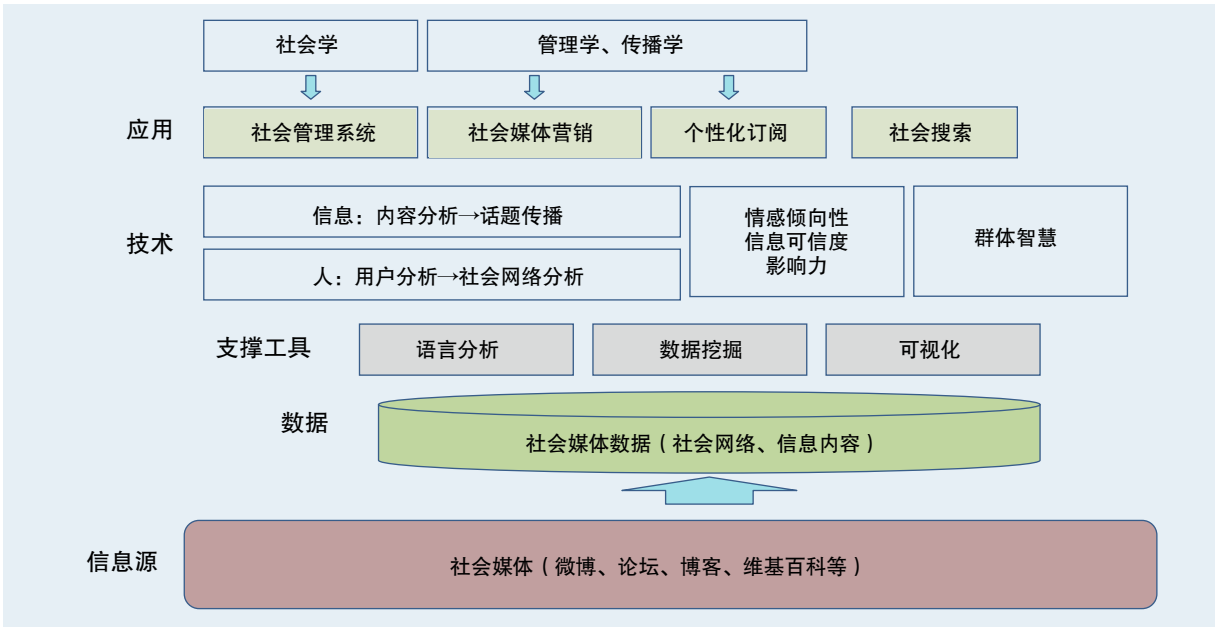

图1 基于社会媒体的社会计算研究框架

对独立的研究内容。在基础研究之上，有社会管理系统、社会媒体营销等主要的应用系统，这些系统受到社会学、管理学、传播学等社会科学的指导。

社会计算有三条研究主线，一是研究人，从点到面，从对个人兴趣的分析到社会关系分析，直至社会网络分析；二是研究信息，从对信息内容的解析到话题发现、话题传播；三是研究人和信息的属性，属性包括倾向性、可信度和影响力。

换个角度看，与信息检索类似，社会计算需要考虑的是内容、结构和用户行为三方面因素。内容是文字、图片和视频带来的信息内容，跟自然语言处理和信息检索密切相关；结构包括社会的结构和信息传播时的传播网络结构，要向网络科学、复杂系统学习；用户行为分析与心理学和管理学相关。

计算社会科学

根据上面的分析，计算社会科学包括社会网络分析、社会网络中的话题传播、倾向性分析、可信度计算和影响力分析等。

社会网络分析

对社会网络的分析（social network analysis）包括对人物节点的分析、社会关系的分析、社会群体的分析和社会网络拓扑结构分析等。

人物节点及社会关系分析 对人物节点的分析可以从内容、结构和行为三方面展开。一个人发布的信息内容能泄露出这个人的兴趣爱好和观点倾向，但仅从内容进行分析无法全面地认识一个人。人是社会关系的总和，通过分析一个人的社会关系能够更准确地发现一个人的特征。此外，一个人的发布、转发、评论、收藏等行为也泄露出其意图和倾向。按照人在社会网络中与他人之间的关系可以把人分为意见领袖、桥节点等。对社会关系的分析包括关系类型和关系强度等多个方面。

社区挖掘技术 社会系统具有层次性。在一个大的社会网络上分布着各种社区（community），这些社区相当于是小的社会系统。社区挖掘算法可以用于发现、识别这些关系紧密的社团、组织。如何对用户、社团进行分类以及分类体系的建设是一个重要的研究问题。社区挖掘包括发现真实社会组织，发现群体的层次结构，发现兴趣相同的用户群等多个方面。

社会网络拓扑结构分析 社会学家们最早开

展了社会网络的研究。他们研究了现实社会的小规模社群的人际关系、群体行为和社会结构等问题。对于人际关系，重点分析了强关系和弱关系对信息扩散的不同作用；对于群体行为，研究了群体构成的社会空间的结构特征、群体与环境的关系以及群体的凝聚力、合作、权力和领导等模型；对于社会结构，采用了图论、拓扑学和集合论等方法，研究了社会网络的拓扑结构特性，提出了基于网络指标（密度、中心度、模块度等）的团体分析等方法。复杂网络的研究者也较早开展了社会网络的研究，研究了社会网络的小世界特性和无标度特性以及网络信息传播的动力学特性；通常采用一些网络指标（密度、度分布、聚类系数、平均路径长度、度相关系数、介数、互惠指数和模块度）来测量网络的拓扑特性。

哈尔滨工业大学（简称哈工大）对用户相似度网络进行了研究，以自然语言处理（natural language processing，NLP）领域的5位研究者为种子节点，根据关注关系扩展为600个人物节点；然后对任意两个节点之间的相似度采用多种信息进行计算；最后用团体分析算法将600个节点自动划分为多个人物群体（如图2所示）。从图2中可以发现一个有趣的现象，图中最左边的一个矩形中绝大多数都是自然语言处理领域中的企业界人士，左数第2列的上图是自然语言处理领域内高校教师，左数第2列下图是自然语言处理领域内的学生。可以看出，教师和学生之间的关系比较密切，教师和企业界人士的关系比较密切，而企业界人士和学生的关联很少。这和现实中的情况也是相符的。

话题传播

对微博内容的分析包括对微博主题的分类（政治、经济、体育等）和微博意图的分类（叙事、抒情、查询、闲聊）等。由于微博内容短小，在分类方面与整篇文本的分类相比具有特殊性，但一条微博并不是孤立存在的，在它前后发表的微博以及博主的社会关系等都可对微博的分类起到支撑作用。

话题传播是社会计算中非常重要的研究内容。一条微博在社会网络上的连续转发构成一棵传播树，与一个社会事件相关的若干传播树组成传播森林。对于传播的理论模型，已经有大量的研究成果，如独立瀑布模型（independent cascade model）、线性阈值模型（linear threshold model）、流

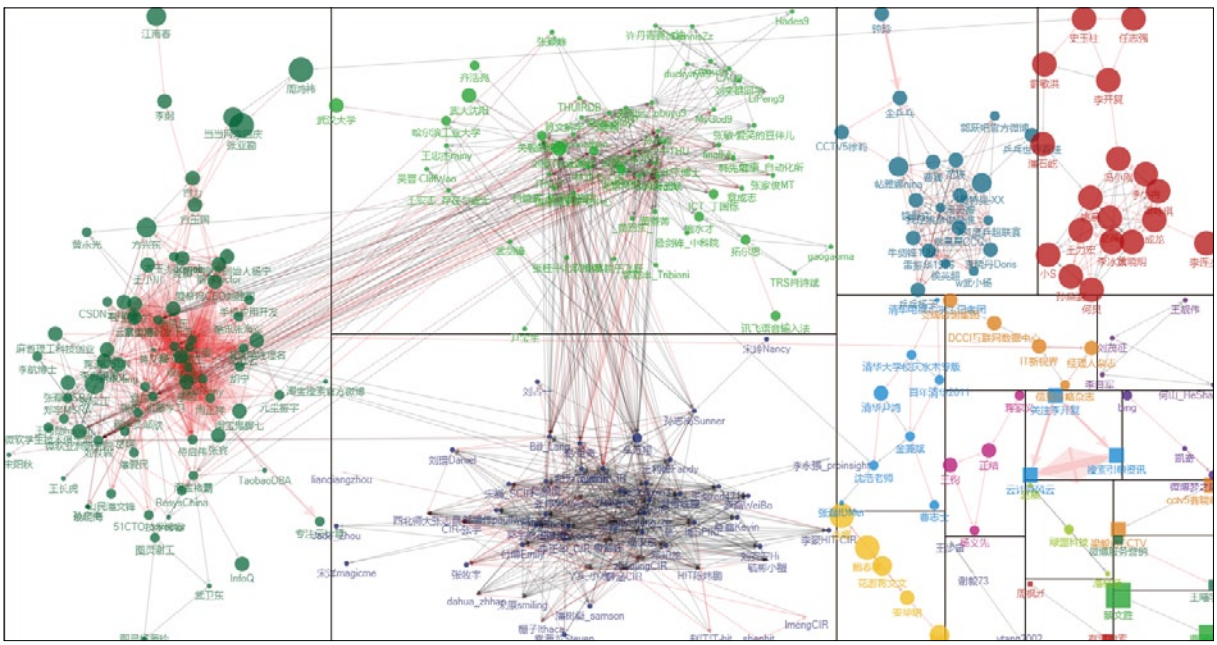

图2“自然语言处理”领域人物聚类

行病模型（epidemics model）和博弈论模型（game-theoretic model）等。在信息传播中，信息传播最大化问题、信息传播特性（比如传播速度等）以及信息扩散概率预测等都是值得研究的课题。

图3是一棵微博的传播树，由哈工大计算机学院和媒体系的老师联合绘制，其中包含了5000个参与该话题传播的人物节点，在传播过程中明显可以看到“二次爆炸”的传播形态，即一些意见领袖节点的转发引起了该话题的二次放大式传播。

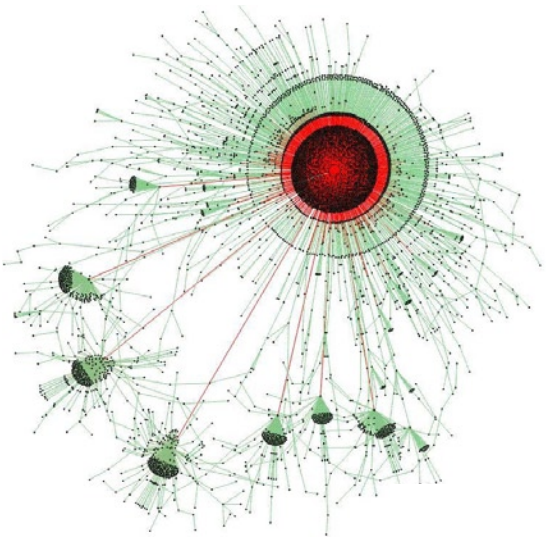

图3 微博的“二次爆炸”传播图

社会情感倾向性分析

虚拟空间中的社会媒体不只包含对客观事实的报道，还有很多主观情感的表达。如果用户的情感以“支持、反对或中立”这种简单三元分类方式来表达则称为“倾向性分析”，如果将网络情绪分为“喜悦、愤怒、悲哀、恐惧、惊慌”等，则称为“情感分析”，合称为“情感倾向性分析”。

目前针对论坛和博客的情感倾向性的研究已经有很多，由于微博的推动，这方面的研究有进一步加强的趋势。有的研究产品的倾向性，有的试图利用微博反映出的用户情绪来预测股票走势，有的研究社会舆论的走势以便辅助社会管理，有的研究大众幸福指数。

以往对产品评论研究较多，相关评测也往往基于产品评论组织数据，在微博上，针对社会事件进行情感倾向性分析的工作是一个很有价值的课题，包括社会事件的评论收集，针对事件的倾向性分析、情绪状态识别和情绪程度识别，特别是可以随着时间的推移，观测一个社会事件在传播过程中网民情感倾向的变化过程。此外，如何借助社会媒体网民的力量自动构建大规模情感词典和情感语料库也是一个值得探索的课题。

信息可信度

社会媒体极大地降低了在网络上发布信息的门槛，不可信信息伴随着“用户贡献的内容”（user generated content，UGC）大量涌现。不可信信息主要包括垃圾信息、虚假信息、错误信息和过时信息。不可信的信息会带来各种严重后果，如用户受骗，浪费用户时间，影响社会稳定等。

信息可信度（information credibility）是指信息或信息源被信任的程度，通常利用计算技术从互联网上挖掘佐证，对信息可信度进行评估。

信息可信度的判别是一项具有挑战性的课题，目前可以利用的技术手段包括：信息内容的不可信特征识别，包含广告语等；信息在网络中与其他信息的链接关系，被较多可信网页链指的网页可信度也较高；信息发出者的权威度，权威度越高，其发布信息的可信度也越高；用户意见，挖掘现有用户评论，或激发用户主动对该信息进行评论、打分、纠错等；事实检验，建立可信常识、事实数据库，并在此基础上对新事实进行推理验证；时效性，提取信息的发布时间、生效时间，判断该信息是否满足用户的时效性需求；人机协同判别，计算机收集事实，让网民自己判断等。

影响力

社会媒体中一个人的社会影响力包括静态和动态两方面的因素。静态方面是指个人在社会网络结构中是否处于核心位置；动态方面包括其发布信息的频度和信息被传播的广度。

影响力最大化问题（influence maximization）是社会影响力分析中涉及到的重要问题。该问题的目标是要通过找到社会网络中若干“有影响力的”的

用户使得最终影响可以在整个网络中最大化。已经有学者提出了定量描述用户之间存在的影响力的方法、影响力强度衡量的方法以及根据社会网络和行为日志建立用户的影响力模型等。

近年来，研究者开始关注社会影响力和社区分析之间的关系。一个人的影响力是有范围的，这个范围就是社区。在某一个社区内影响力很大的人，换了一个社区可能就毫无影响力。

社会计算应用

基于社会媒体可以设计出各种各样的应用，帮助人们充分沟通，协同工作或者获取信息。社会计算最重要的应用包括虚拟社会管理、社会媒体营销以及个人用户应用等方面**虚拟社会管理**。

虚拟社会的形成和发展改变了人类生存和活动的空间，影响了社会结构，使社会分化为现实社会和虚拟社会。它具有开放性、互动性等特征。虚拟社会首先是对现实社会的延伸，现实社会中诸多层面的矛盾也延伸到其中，以乘数效应反作用于现实社会。因此，其自身的调节能力是不够的，需要以政府为实施主体的社会管理介入。因此，在当前社会矛盾多发的时期，虚拟社会的管理已经引起政府的高度关注。

政府在对虚拟社会进行管理时，也会面临诸多问题和挑战。对社会媒体的监管，如果处置方法不当，会影响社会媒体在我国的良好发展势头。建立完善的基于社会媒体的社会管理平台，对社会舆论的管理方式变“堵”为“疏”，可借助网民的力量发现事实，鼓励参与事件的调查，缓解网络舆情，增强政府公信力，对社会媒体成为社会和谐的润滑剂有着重大的现实意义。

基于社会媒体的虚拟社会管理平台的主要功能包括：提高政府对于以社会媒体为代表的虚拟社会中个人、群体之间社会网络结构的了解，对突发群体事件的掌控，对人群情感倾向状态及演变的分析，预测群体事件的走势，并实施舆论引导。

社会媒体营销

社会媒体由于传播速度快和便于获取个人信息的特点，其广告价值将比传统网络媒体高出很多。谷歌和百度取得了巨大的成功，他们的主要盈利模式都是广告。社会媒体具有如此大的广告价值，这自然引起了业界高度重视。但如果广告泛滥会严重破坏社会媒体环境，所以如何实施社会媒体广告投放需要认真研究。

社会媒体广告投放一般按如下步骤进行：面对当前需求，参考以往案例，拟定广告信息；选择投放节点；广告投放；广告效果评估：利用传播树计算广告传播规模（定量），利用情感倾向分析技术计算广告投放效果（定性），综合定量和定性的指标可以设计出新的广告计费模式。

在广告投放过程中，运用了多项关键技术，包括人物分析（个性化）、影响力分析、传播分析和情感倾向性分析等，计算社会科学的基础研究方面的实质性进展将对社会媒体营销给予理论上的指导。

个人用户应用

与苹果公司的App Store一样，微博系统作为一个平台，积极鼓励第三方开发者研制面向个人用户的各类应用软件。

哈工大开发了一种基于新浪微博的微博应用，称为“围脖庞统”（“庞统”取“庞大统计”之意）。“围脖庞统”能够分析出一个用户每天不同时段使用微博的频度，能够找出与这个用户交互最频繁的若干其他用户。目前该应用已有1万多用户了。

图4 “围脖庞统”新浪微博应用

挑战与机遇

语言计算和社会计算有以下的区别和联系。

区别 在研究问题上，前者研究人类语言的分析技术，后者研究人类社会的分析技术；在数据资源上，前者研究文本数据库，后者研究社会媒体中的人和信息共同组成的网络数据；在内容分析上，前者分析文本的语义，后者除了分析话题的内容外，还分析人和群体的特点；在结构分析上，前者分析文本的结构，后者分析社会网络的结构。

联系 在模型表示上，由于社会媒体中的文本是用户表达信息的最重要的方式，因此，语言计算的文本表示方法均可用于用户的模型表示；在关系分析上，语言计算的文本分析技术可帮助计算社会网络的用户之间的相似度，发现他们之间潜在的关系。

挑战

语言是社会媒体最重要的表现形式，语言计算是社会计算的重要支撑技术，社会计算的实现离不开语言计算技术的发展和成熟。从另一方面讲，社会计算这一课题的出现也会带动传统的语言计算研究内容与应用形式发生改变，至少在以下几方面已见端倪。

语言底层分析技术 语言底层分析包括分词、词性标注、命名实体识别、句法分析等，目的是对线性无结构的自然语言表达转换成结构化表达，以便于计算机更好地对其进行处理。在社会计算的背景下，语言底层分析技术依然是必不可少的，然而所要处理的对象却发生了变化，即从标准的书面语文本转化为灵活多样的网络语言。这至少从以下几方面提出了挑战。首先，网络语言使用灵活，不遵循严格的用词和语法规则，因此自动分析的难度更大；其次，网络语言中夹杂大量的书写错误，这使得纠错成为必要的预处理环节；再次，网络语言演变迅速，新词、新词义和新概念层出不穷，有的却又转瞬即逝，这导致了未登录词的比例和影响增大。近年来，随着社会计算在研究界得到愈来愈多的关注，已经有学者开始尝试将传统的语言底层分析技术移植到社会计算的土壤中，其中包括对Twitter、微博等数据的分词、词性标注、拼写纠错等等。初步的研究表明，虽然传统的技术路线仍可沿用，但在面向社会计算的语言底层分析技术在问题定义、特征设计、数据选择方面都要做不小的调整。

网络搜索技术 网络搜索是一个宽泛的概念。从广义上讲，用户从互联网上通过搜索的方式获取信息都可称作网络搜索。而这里则特指以谷歌、百度等为代表的搜索引擎技术。在社会计算时代，传统的网络搜索技术不可避免地受到影响并产生变化。搜索引擎将针对网民对微博等即时信息的搜索需求，在爬取网页、建立索引、提供搜索服务等各个环节全面“提速”，以应对数据的快速更新和用户对即时信息的需求。此外，在社会计算的背景下，用户希望获得个性化的服务，这也将使各大搜索引擎不久就会推出酝酿已久的个性化搜索功能，即不同用户在搜索同样的查询时，将看到不同的搜索结果。而且，网络搜索将从以搜索信息为主演变为既搜信息又可搜人的情况。未来，在充分记录用户个性化信息并进行个性化建模的基础上，网络搜索不仅可以帮助我们搜找朋友，还能搜助手、搜生意伙伴等等。

信息抽取与知识挖掘技术 在社会计算时代，全民均可通过互联网发布信息、表达观点，这使得本已严重的信息过载问题更为凸显，从而也对信息抽取与知识挖掘技术提出了更迫切的需求。为此，信息抽取技术一方面必须提高信息处理的即时性，另一方面也要考虑信息的主观性。换言之，对主观性信息和知识的抽取、统计以及挖掘将成为研究的热点。近年来，已经有很多这方面的研究，包括对Twitter、微博等数据的情感倾向性分析、舆情分析等。此外，传统的自动文摘技术在这样的应用场景下也将发挥出更大的作用，只不过对象不是新闻、报告等传统的文本内容，而是数量庞大的精短微博。在微博的信息抽取和知识挖掘的过程中，我们要充分认识到微博数据的特点，即微博长度很短，必须结合更大范围的上下文（如转发/评论的其他微博）来实现信息和观点的准确挖掘。

自动问答技术 自动问答一直都是自然语言处理研究中的重要方向，研究者们希望借自动问答的实现来集成式地展现自然语言处理各项技术的成熟。多年来，互联网的自动问答（Web-based QA）一直是主要的研究对象，直到“百度知道”等为代表的社区式问答（community-based QA，cQA）走进人们的视野。人们发现，这种由人提出问题，再由其他人回答的形式，所提供的答案更加准确可信。因此，对社区式问答的研究越来越多，包括对社区式问答资源的检索与推荐以及社区式问答资源的质量自动评估等。随着社会计算的兴起，我们相信这种形式也将得以进化，即从现在的提出问题等待他人回答变成提出问题主动寻找他人回答，这个人是与当前问题最“匹配”的人。当然，这也要得益于用户个性化建模的实现。倘若答案的提供是有偿的，那么知识将成为一种特殊的商品成为交易的对象。

机遇

社会媒体中出现了大量的虚拟社区，并组成巨大的虚拟社会。虚拟社会中的人物和群体与现实社会中的人物和群体存在着一定的映射关系，两个社会的人群活动行为存在着相互作用关系。虚拟社会的人际交流有着超时空的及时性和距离无关性。与现实社会相比，微博构建的虚拟社会的可观测、可计算为社会科学超越问卷调查等传统手段，转而以计算的方式认知社会创造了条件。

群体智慧的运用引发知识获取方式的变革

知识获取是一切智能系统的关键，让我们从这个角度回顾一下语言技术的发展。

第一代语言技术的知识获取方式是专家系统式的，其知识获取形式是语言学家的语言学专业知识加上计算机专家的算法知识。这些知识是封闭的、有限的，以封闭而有限的知识面对开放而无限的语言现象，是难以应对的。

第二代语言技术的知识来自于大规模原始语料和小规模经过标注人员手工标注的深加工语料。方法上主要采用统计模型，语言学家插不上手。真实文本蕴含着用户提供的知识，但不经标注的原始纯文本只能提供很浅的不精确的知识，而组织一些学生对语料进行人工标注的工作每每令研究者备尝艰苦，以句法分析研究为例，几万句的标注树库搞搞研究尚可，面向不同领域的实际应用就捉襟见肘了。

第三代语言技术的知识获取方式是群体智慧，或者叫人本计算，是有大规模用户直接参与的计算。一个典型案例是reCAPTCHA系统，美国卡耐基梅隆大学的路易斯·冯·安（Luis Von Ahn）于2008年9月在《科学》（*Science*）上发表论文“re-CAPTCHA: Human-Based Character Recognition via Web Security Measures”[2]，他引导用户输入“双验证码”，其中一个是已知答案的，另一个是从《纽约时报》的文本图像中截取出来的一个单词。海量用户在输入验证码时不自觉地将单词的扫描图像识别为单词的字母序列，一年之内以此方式获得了《纽约时报》30亿字的文本。其他案例有图片的社会化标签和亚马逊购书时的社会协同推荐等。

群众一旦被发动起来，其力量是无比巨大的，但要发动他们参加高科技的研究又是非常困难的，需要找到带有趣味性的简单的参与模式。如果我们能够把复杂的语言学标注（如依存句法、语义角色）转换为普通用户能够理解的简单形式，并以文字游戏等形式吸引用户参与，实现“傻瓜式”标注，则有可能在社会媒体环境下以前所未有的高速度和低成本获取大量的语言学知识。美国2006年的MINDS（MT, IR, NLP, Data resources, Speech）报告[3]已经把语料标注问题上升为标注科学（annotation science），其中也谈到激励非专业人士参与标注的观点。

便于获取个性化信息

以往，我们从事一些与人有关的研究，比如个

表3 知识获取方式的变革

	知识获取方式	知识来源
第一代	专家系统	算法+专家知识 无用户参与
第二代	大规模真实数据	算法+真实数据驱动， 少数人进行标注 用户间接参与
第三代	群体智慧	大规模用户参与计算

性化信息检索与推荐等，总是苦于数据的匮乏，如今微博在很大程度上解决了这个问题。我们浏览一个人的微博，可以看到这个人的兴趣爱好、观点倾向，甚至能够猜出他的作息时间，知道他使用什么样的移动设备上网等。微博提供了一个记录个性信息的开放平台。

便于应用推广

微博还提供了一个应用推广平台。如果你有一个有价值的应用，很快会得到用户的使用和传播。在社会网络中，应用与信息一样，也在快速地传播，这对研究者来说是非常有价值的。研究者可以通过所掌握的应用进一步采集用户的使用信息，持续、递进式地研究人和社会群体。

结语

社会计算是一个方兴未艾的交叉学科，网络科学、复杂系统、数据挖掘、社会学、管理科学、语言处理和信息检索等不同背景的学者从不同的角度对社会计算进行了研究。社会计算的研究横跨文理，为社会科学提供了一条革命性的计算之路，其研究成果对于社会管理、社会生活都将产生重大影响。随着学术界、产业界和政府对社会计算的认识不断加深，关注度不断提高，社会计算正逐步进入蓬勃发展的阶段。■

刘 挺

CCF高级会员、2011～2012年度YOCSEF副主席。哈尔滨工业大学计算机学院教授。主要研究方向为社会计算、自然语言处理、信息检索。tliu72@163.com

徐志明

哈尔滨工业大学计算机学院教授。主要研究方向为社会计算、信息检索、自然语言处理。xuzm@hit.edu.cn

秦 兵

CCF高级会员。哈尔滨工业大学计算机学院教授。主要研究方向为文本挖掘、信息抽取。qinb@ir.hit.edu.cn

赵世奇

CCF会员。百度高级研发工程师。主要研究方向为语义计算、语义搜索等。zhaoshiqi@baidu.com

李 生

哈尔滨工业大学计算机学院教授。主要研究方向为中文信息处理、信息检索、社会计算。lisheng@hit.edu.cn

参考文献

[1] David Lazer, etc., Computational Social Science, Science 6 February 2009: Vol. 323 no. 5915, 721~723

[2] Luis von Ahn, etc., reCAPTCHA: Human-Based Character Recognition via Web Security Measures, Science 12 September 2008: Vol. 321 no. 5895, 1465~1468

[3] Liz Liddy, etc., MINDS Workshops Natural Language Processing Working Group Final Report, http://www-nlpir.nist.gov/MINDS/FINAL/NLP.web.pdf, 2006

CCF章程等五部规章修订获得通过

在2011年11月25日举行的CCF第十次会员代表大会上，学会秘书长杜子德代表会员代表大会工作委员会作了关于《CCF章程》、《CCF监事会条例》、《CCF会员条例》、《CCF理事会条例》和《CCF理事会选举条例》修订的说明。会员代表以举手表决的形式通过了五部新修订的规章。