社会计算报告

周美廷 | 76066002

1. 对社会计算的基础概念

一开始我也不太了解社会计算这门课是会学一些什么样的内容。看英文有 'networking' 这个词还以为是跟网络有关。结果确实是跟网络有关，但不是我想象中的电脑网络虚拟世界的那个网络，而是学人与人之间的关系网络。是什么样的一个网络把每一个人连起来。

论基础，社会计算完全是靠逻思维和事实的，不是那种一直计算数学的那种。我印象最深刻的有几个东西：

- 三元闭包
- 人与人关系的投影
- 纳衡均值

1.1. 三元闭包

自然还有一些我不知道名字叫什么的。这些知识其实跟我们作为社会上的一部分，社会上的一个人，一个大世界小世界的一部分都很有关也很明显地可以看得出。比如说像三元闭包，三元闭包是说两个不认识的人因为认识同一个人就可以有可能互相认识。这个例子就很简单。比如说像我们年轻人，先说是单身独立的。然后找到了对象。那一开始我们跟那对象的家人都不认识，他们也不认识我们，可是因为我们认识同一个人，也就是我们的那对象，所以我们可以跟他的家人认识。

1.2. 人与人关系的投影

人与人关系的投影主要是解释人与人之间因为参与一样的活动就可以互相认识。我觉得这个跟三元闭包是挺像的，只是中介不一样，刚才是一个人，这个是一个活动。比如说像 4 月时我参加了微软的 hackathon 比赛活动。我是一个很宅的女生，完全不可能认识当时也参与活动的人。可是因为参与与那些人一样的活动，我就可以认识那些同样参与 hackathon 的人。

1.3. 纳衡均值

纳衡均值是我在学社会计算印象最深刻的。我作为外国人从未听过这个词，第一次听也同时需要了解其中的意思。来来回回它的道理其实就是：社会中的纳衡均值就是社会中最佳选择。比如说，有两个小偷然后这小偷被抓。他们可以选择要合作还是要背叛彼此。这会生成一个表，里面放着选项的值。选项表例子：

	合作	背叛
合作	{A, B}	{C, D}
背叛	{D, C}	{D, D}

比如说第一选项（两个都合作）值就是最高。这个值就是收益值，就是如果那些人选这个选项，他们会收到最大的收益。既然这是两个人最大的收益，彼此都可以保证和肯定，另一个人也会选这个选项，所以没有选项不一样的危险，毕竟每个人肯定都会选对自己有最大收益的选项。这就是纳衡均值。

2. 语言分析技术在社会计算中的应用

传统社会科学研究中的数据主要通过调查问卷或口头采访等方式获取，既耗时耗力，数据规模也很受限。进入互联网时代后，人类社会越来越多的信息以在线形式出现，为社会学研究提供了丰富的数据支持。特别是进入 Web 2.0 时代后，以用户为中心的服务（如微博、社交网站等）积累了大量的用户产生内容，包括用户个人档案（如性别、年龄、职业等信息）、用户社交关系网络（如关注关系、好友关系等）和文本信息（如微博、个人状态、博客等）等，成为社会学研究绝佳的数据来源。

顺应该趋势，2009 年由哈佛大学学者 David Lazer 牵头的来自信息科学、社会学和物理学的 15 位学者在 Science 杂志上联名发表文章，提出了“计算社会学”（Computational Social Science 或 Computational Sociology）（Lazer, et al. 2009），阐述了利用计算手段从大数据中揭示社会学规律的学术思想和趋势，标志着社会学进入到数据计算时代。短短几年内，计算社会学已成为人文社科领域近年来最重要的研究范式。Science、Nature 和美国国家科学院院刊等国际顶级学术期刊上大量涌现计算社会学的研究成果（Schich, et al. 2014, Lieberman, etal. 2007, Michel, et al. 2011, Bond, et al. 2012），众多学术期刊出版专刊介绍计算社会学研究进展。美国还成立了计算社会学学会，George Mason 大学甚至成立了计算社会学系，并成为世界上第一个正式授予计算社会学博士学位的单位。计算社会学无论对于揭示人类与社会规律，还是对于用户个性化服务，均具重要意义，因此基于社会媒体大数据的计算社会学研究，在学术界和产业界均引起广泛关注。

自然语言是社会媒体海量数据的重要组成部分，蕴藏了与用户及其复杂关系有关的丰富信息，是社会语言学、社会心理学等社会学分支的重要研究对象和研究角度，但是这些社会学分支所需的信息都隐藏在复杂的语言背后，需要利用自然语言处理和理解技术挖掘出来，才能被计算社会学研究进一步加以利用。 随着机器学习和自然语言处理技术的发展，如何更好地分析社会媒体大数据中的自然语言已经成为计算社会学中的研究热点，近年来吸引了众多学者的研究兴趣，并已初具规模。

本文将综述最近在这方面的典型工作，并试图总结未来的研究趋势，希望对我国学术界和产业界在计算社会学的研发能够有所助益。

1 面向社会媒体的自然语言使用分析

传统的自然语言处理主要面向正式文本，例如新闻、论文等。这些文本遣词造句比较规范，行文符合逻辑，因此比较容易处理。自然语言处理技术按照处理目标分为几个层次：（1）词汇层。主要是在词汇级别的处理任务，如中文分词、词性标注、命名实体识别等。（2）句法层。主要是在句法级别的处理任务，如针对句子的句法分析、依存分析等。（3）语义层。主要是在语义空间的处理任务，例如语义分析、语义消歧、复述等。（4）篇章层。主要是在篇章级别的处理任务，如指代消解、共指消解等。（5）应用层。主要是指利用自然语言处理分析技术完成的应用任务，如文本分类、信息抽取、问答系统、文档摘要、机器翻译，等等。关于自然语言处理技术的详细介绍可以参考（Jurafsky, et al. 2000，Mannin & Schütze 1999）。

进入社会媒体时代，用户产生的大量文本内容无论从词汇到造句都更加非正式，不仅存在大量拼写错误，还有很多网络产生的新用法，甚至出现专门的术语“网络用语”来命名这种现象。那么，自然语言处理技术如何分析社会媒体文本呢？研究者提出了文本正规化（text normalization）的任务，通过拼写纠错、词汇替换等方式，将非正式的网络文本转换为正式文本，然后再利用传统自然语言处理技术进行分析。当然这样还不够，研究者们还开始研究专门面向社会媒体文本特点的自然语言处理技术。

这里介绍的重点并不是面向社会媒体的自然语言处理技术，而是利用这些处理技术对社会媒体中的语言使用开展的分析工作。接下来，我们将介绍人们已经从社会媒体语言使用方面得到的主要成果。

• 1.1 词汇的时空传播与演化

词汇是自然语言的基本表意单位，也是自然语言处理的基础。利用词汇在时空中的变化开展社会学研究在国内外都不鲜见。金观涛和刘青峰通过分析近代文献中的特定词汇使用情况，探讨了中国现代重要政治术语的形成（金观涛&刘青峰 2009）。最近，哈佛大学研究团队利用 Google Books 收集并扫描识别的 1800 年到 2000 年之间的 500 万种出版物（占人类所有出版物的 4%），通过不同关键词使用频度随时间的变化，分

析了人类文化演进特点，做出了很多惊人的或有意思的发现。例如，他们发现在过去几百年里英语中越来越多的不规则变化动词演化成了规则变化动词（Lieberman, etal. 2007）。再如图 8.1 所示，通过 Google Books 中历年来使用“The United States is”和“The UnitedStates are”的统计趋势图，可以定量分析美国作为一个统一国家的概念是如何慢慢形成的（Aiden&Michel2013）。他们甚至为此提出“文化组学”（culturomics，仿照“基因组学”发明的新术语）的概念（Aiden&Michel2013，Michel, et al. 2011）。正如文献（Aiden&Michel 2013）的副标题“Big Data as a Lens onHuman Culture”所暗示的，基于大数据的定量分析为社会科学研究提供了一个全新的视角。

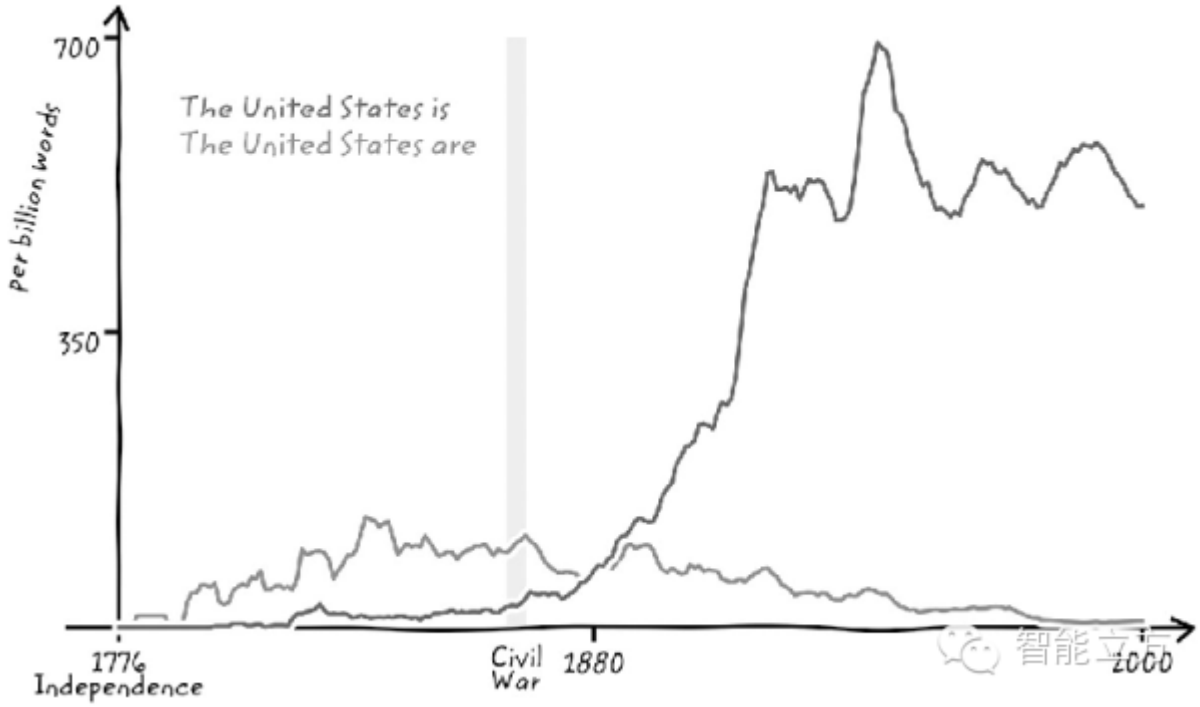

图 1　通过 Google Books 中历年来使用“TheUnited States is”和“The United States are”的统计趋势图，可以定量分析美国作为一个统一国家的概念是如何慢慢形成的。来自文献（Aiden & Michel 2013）

在社会媒体中，新的词汇产生后，就会随着信息流动而进行传播和演化。一方面，新词汇的流行程度和形式会随着时间而演化，出现爆发（burst）和变形（variance）。不同新词汇的爆发程度和变形情况可能会受到不同因素的影响。另一方面，社会媒体中的用户分布在世界各地，其社交圈子往往会受到地理位置的限制，因此新词汇在社会媒体中用户间的传播，也会反映在地理位置的扩散上。一个词汇可能会首先在某个地域流行，然后逐渐扩散到全国，甚至全世界。

探索词汇的时空传播与演化，研究意义重大，相关技术也比较容易做到。目前已有关于英语词汇在社会媒体中的时空传播的研究。斯坦福大学 Leskovec 等人（Leskover,

et al. 2009）从不同来源收集了约 9 千万篇新闻文章，利用引号从新闻中自动抽取流行语句，命名为模因（meme）。通过跟踪这些模因的使用频率随时间而变化的情况，能够及时、有效地把握美国政治、经济和文化生活，如图 8.2 所示。例如作者提到的典型模因“you can put lipstick on a pig”（为猪涂上口红）即是 2008 年美国总统大选中奥巴马讽刺竞选对手时引用的一句谚语，全句是“你就算给猪涂上口红，它也还是只猪”，当时引起了选民的广泛争议，也让最早出现于上世纪 20 年代的谚语“lipstick on a pig”重新流行起来，一时间成了美国人民很爱用的一个短语。通过文献（Leskover, et al. 2009），我们可以看到作者巧妙地使用了流行语作为社会热点问题的指标。

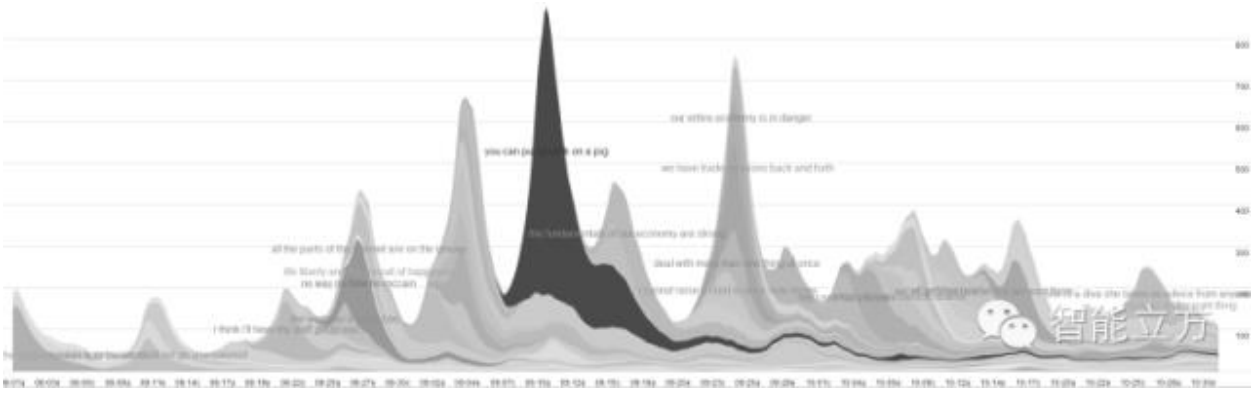

图 2　MemeTracker 提供的模因时序变化趋势，其中大红色代表“you can put lipstick on a pig”。
来自网站 http://www.memetracker.org/

此外，值得注意的是，文献（Leskover, et al. 2009）作者巧妙地借助引号这种“显式标注”从海量文本中自动发现长度可变的流行语，有效地降低了识别流行语的计算难度。近年来，清华大学计算机系孙茂松教授系统地总结了这类研究思路，提出了“基于互联网自然标注资源的自然语言处理”的研究范式（孙茂松 2011），这对于如何有效利用大规模互联网数据具有极大的启发意义。Leskovec 研究团队还更进一步，通过聚类算法研究信息扩散的时序特征，分析 Twitter 和博客中模因使用的时序信息，共总结出 6 种时序曲线的主要形状（Yang & Leskover 2011）。

上述研究主要对流行语使用频率的时序变化进行了分析，也有学者考察了社会媒体中词汇与地域的关系。Eisenstein 等学者（Eisenstein, et al. 2010）发现同样的话题在不同地域会以不同的方式提出和讨论，为了探究 Twitter 中文本与使用者所处地域的关系，他们建立了一个瀑布模型（cascading model），用来分析词汇变化如何同时受到话题和地域的双重影响，并把地理空间按照语言学上的群体进行分割，试图通过文本本身去预测那些没有标注的用户所处的地域。词汇在地域上的差异和演化，与许多因素有关，如不同地域的文化风俗、地标建筑、方言俗语，等等。

词汇是文本中负载信息的基本单位，考察社会媒体中词汇的时空传播与演化，无论对语言演化研究，还是对社会管理，均具重要意义。

- ### 1.2 语言使用与个体差异

人格心理学和社会语言学的相关研究认为，人们的个体差异会反映在他们语言使用的特点上。因此，如何定量建立起语言使用与个体差异之间的关联，是学者关心的重要话题。这方面最具代表性的工作，是 20 世纪 90 年代 Pennebaker 和 King 提出的 Linguistic Inquiry and Word Count（LIWC）方法（Pennebaker & King 1999）。其基本思想是以词汇作为语言使用定量分析的基本单位，首先通过人工收集、标注的方式建立不同类别的词典（如代词、数词、情感词等），然后在给定的个体或群体对应的文本中进行词频统计，从而建立起个体差异（即不同人格）与词类比例（即语言使用特点）之间的关联关系。经过数次修订后，LIWC 已经形成了 70 余种分类词典，相关软件可以通过官方网站 http://www.liwc.net/购买，而台湾地区学者黄金兰等人也在 Pennebaker 教授的授权下建立了中文版 LIWC 词典（Huang, et al. 2012），可以通过 http://cliwc.weebly.com/访问。
目前，从语言使用的角度探索个体差异的研究，大部分采用了类似于 LIWC 的研究范式。Pennebaker 教授的研究团队就在这方面做了大量有影响力的工作。他们发现，抑郁与自杀者往往会在文本中发出可侦测的求救信号（Chung & Pennebaker 2007）；初次约会的时候对象之间几分钟的对话就可以预测彼此的好感，而情侣间的对话也可以预测几个月后持续交往的概率（Ireland, et al. 2011）；团队的凝聚力和合作倾向也可以通过内部对话做出预测（Gonzales, et al. 2010）；谎言的相关语言特性也有助于分辨真假（Newman, et al. 2003）；语言使用分析还将有助于结识新朋友（Pennebaker & King 1999）；语言使用还与年龄有千丝万缕联系（Pennebaker & Stone 2003）等等。

然而，以上研究仍然未脱离传统社会学研究的藩篱，大部分是在受限的小规模数据上开展的。而在大规模在线社会媒体背景下，通过语言使用分析个体差异更凸显其重要性，一方面，很多在小规模数据上建立的社会理论需要在大规模真实数据进一步验证或再发现；而另一方面，利用社会媒体用户产生的文本数据推测用户的人格或心理特点，在个性化推荐服务中发挥重要作用。因此近年来，在社会计算领域提出了用户建档（user profiling）的研究任务，旨在利用用户产生内容预测用户的各种属性，既包括用户的各种简单属性，如性别（Burger, et al. 2011，Fink, et al. 2012）、年龄（Goswami, et al. 2009）和地理位置（Rao, et al. 2010，Li, et al. 2012）等，也包括用户的复杂属性，如兴趣（Yang, et al. 2011）、政治倾向（Rao, et al. 2010）、性格特点（Mairesse, et al. 2007，Schwartz, et al. 2013）和主观幸福感（Frank, et al. 2013，Mitchell, et al. 2013，Dodds, et al. 2011）等。

前述基于 LIWC 的研究与用户建档研究的主要不同在于：（1）前者侧重于人格差异与语言使用之间的关联关系的发现，而后者侧重于将语言使用作为特征来建立预测用户

属性的模型。（2）前者更纯粹地考察语言使用与个体差异的关联，而后者则会将语言使用与用户的其他方面的特征（如用户的社会网络结构、在线行为模式等）综合起来进行属性预测。（3）前者对语言使用的分析还基本停留在词频统计的层面，而后者则充分利用了机器学习和自然语言处理领域的最新研究成果，如向量空间模型（Manning et al. 2008）、隐含主题模型（Steyvers & Griffiths 2007）、时间序列分析（Hamilton 1994）等，其定量分析的广度和精度均为前者所不及。

目前面向大规模在线社会媒体的语言使用与个体差异的关系研究尚处于起步阶段，一方面在线社会媒体为研究提供了更丰富的分析素材和角度，而另一方面机器学习和自然语言处理的发展也为语言使用分析提供了更丰富的维度。可以预期，未来将能看到关于语言使用与个体差异的更多、更深层次的分析和发现。

- ### 1.3 语言使用与社会地位

语言是人类相互交流的工具，而社会中的人存在着地位差异。那么语言使用方式与人的地位差异有什么关系呢？这是一个社会语言学经典问题。

社会语言学理论提出，地位越低的发言者需要从语言上去适应地位越高的听者，而相反，地位越高的人则不需要调整自己的语言方式去适应别人（Gonzales, et al. 2010）。在过去由于缺少相关大规模数据，有关理论一直缺少定量分析的支持。美国康奈尔大学 Danescu-Niculescu-Mizil（以下简称 Mizil）等学者对这个问题进行了深入探讨，做出了一系列开创性的研究成果。

Mizil 等人（Danescu-Niculescu-Mizil, et al. 2012）选取线上和线下两个场景验证了交流行为是如何体现权力关系的。两个场景分别是维基百科中编辑的在线讨论，以及法庭庭审现场的辩护对话。值得注意的是，这里所谓的语言使用方式，并不是实词的使用，而是虚词的使用，甚至可能连发言者都没有注意自己这种发言方式的变化。该研究定量验证了参与讨论的人之间权力的差异会在两人如何回应对方的语言方式上有所体现。

该理论也在 Twitter 平台上得到了验证（Danescu-Niculescu-Mizil, et al. 2011）。首先，作者同样利用介词等虚词的使用情况，考察了交流双方的语言风格是如何彼此适应的。然后，作者考察了交流双方之间影响的不对称性，以及这种不对称性与社会地位的关系，即地位高的人不会去适应地位低的人，而地位低的人要付出更多去适应地位高的。研究结果表明，虽然 Twitter 对交流增加了一些限制（非面对面，非实时，而且只能说 140 个字），但交流中仍然有比较明显的语言适应行为。

礼貌用语的使用与社会地位之间也有密切关系（Danescu-Niculescu-Mizil, et al.2013A）。作者分别对维基百科编辑和 Stack Exchange 论坛的讨论者进行研究，把用户对他人提出请求时的对话摘录出来，其中一句是真正的请求，而另一句是客套话，然后由标注者为其礼貌程度进行评价。研究结果表明，维基百科编辑在选举中试

图获得更高地位时会更加礼貌，而一旦选上后，礼貌程度就会下降。这种情况也同样出现在 Stack Exchange 上，人们的礼貌程度与地位呈反比关系。

该理论还被用来定量分析社区用户的语言使用变化情况（Danescu-Niculescu-Mizil,et al. 2013B）。作者以两个大型啤酒讨论社区作为研究对象，发现用户在社区中一般会经历两个阶段，在第一个阶段他们刚进入社区，会积极学习适应社区的语言使用规则，而接下来他们逐渐不再做出改变，任由规则变化，最后逐渐退出社区主流群体。该研究工作的学术意义在于，定量探索了在社区与个人的相互作用下，语言使用规则变化的复杂性。

Mizil 等人开创性地在社会媒体大数据上定量验证了社会语言学中的重要理论，并进一步利用该理论展开社会学研究。社会语言学乃至社会心理学中仍有大量的理论，有待于在大规模社会媒体中得到验证和利用，而语言使用是不可忽视的重要角度。

- ### 1.4 语言使用与群体分析

作为广大互联网用户在线交流信息和观点的平台，社会媒体汇集了成千上百万用户的产生内容，这些内容从整体上反映了人们关注的社会焦点和主要立场。从语言使用的角度，可以通过两个方面对这些用户进行群体分析：（1）作为文本内容的客观部分，分析用户群体关注的话题及其趋势；（2）作为文本内容的主观部分，分析用户群体的情绪、观点及其演化过程。

作为文本内容的客观部分，文本的话题检测与跟踪（Topic Detection and Tracking，简写为 TDT）（Allan 2002）是自然语言处理和信息检索领域的传统研究问题。最初是面向新闻媒体流提出的这个研究问题，旨在发现与跟踪新闻媒体流中的热点话题的趋势。在该任务中，一个话题是由一个种子事件及与其直接相关的事件组成的。在话题检测中有很多子任务，例如话题检测、话题跟踪、首次报道检测、关联检测，等等。面向社会媒体的话题检测与跟踪已经成为 TDT 的最新研究趋势，如图 8.3 是利用隐含主题模型分析 Twitter 话题并做可视化的样例，图 8.4 则是对 Twitter 话题变化趋势的分析与可视化。当然我们可以用单词或短语来表示话题，这样就可以利用 8.2.1 节“词汇的时空传播与演化”中的技术。但是，从实用角度，为了增强话题检测与跟踪的表达和概括能力，我们往往需要借助于隐含主题模型等技术，同时使用隐含主题和词汇一起来展示社会媒体的话题及其演化趋势，这是近年来的最新发展趋势。

图 3　利用主题模型分析 Twitter 话题并用标签云进行可视化。来自文献（Ramage, et al. 2010）

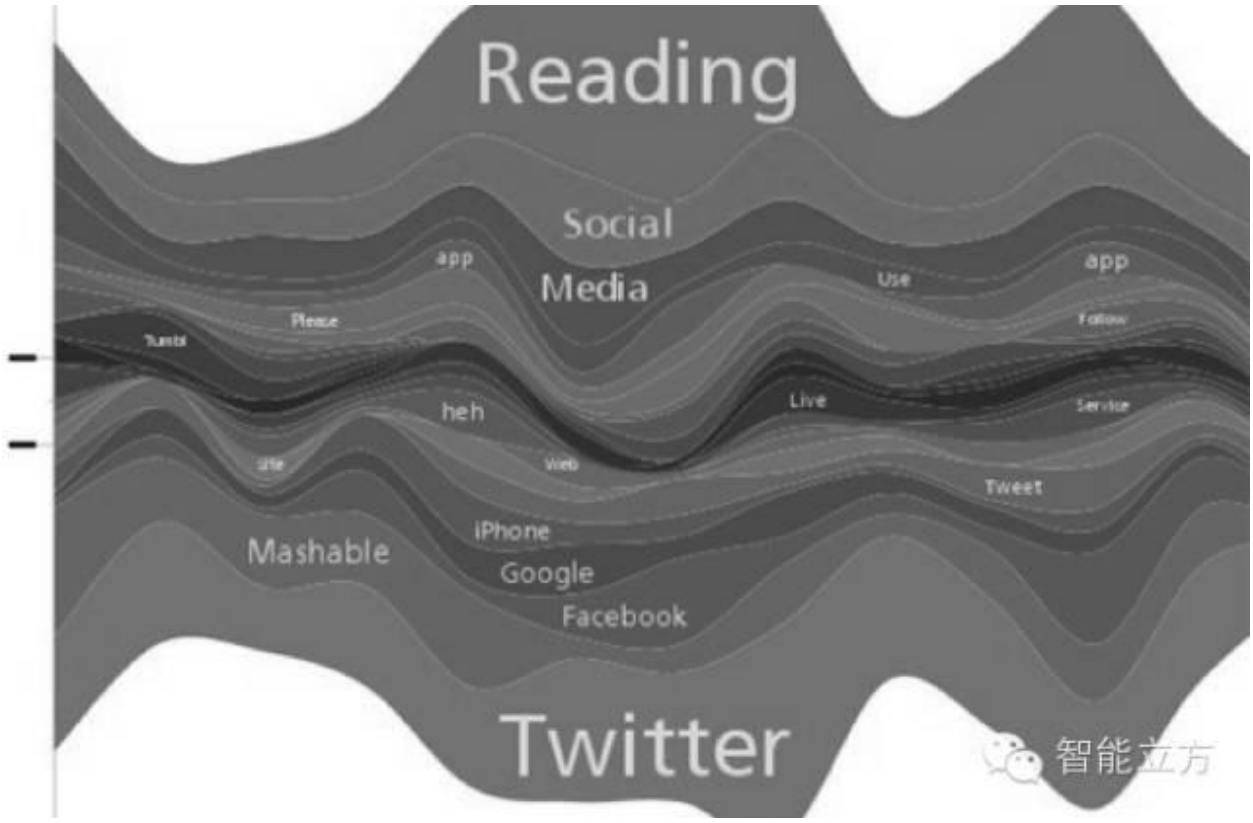

图 4　Twitter Stream Graphs 分析 Twitter 话题变化趋势并进行可视化

作为文本内容的主观部分，用户也会在社会媒体中表达他们的情绪、倾向和观点等主观情感。而社会媒体文本与传统媒体文本（如新闻）的最大不同也在于此，因此有大量研究聚焦于社会媒体的用户情绪和情感分析。如图 8.5 所示，作者通过分析 3 亿条 Twitter 数据中的情感词汇的使用情况，探索美国人的情绪随时间和地域的变化趋势，可以看到美国全国各地、一周七天以及每天 24 小时的情绪变化，得到很多有意思的结论。例如，美国人在下午的时候会变得烦躁，而在晚上开始好转；居住在美国西部的人普遍比东部沿海的人快乐，而位于美国南部的佛罗里达州几乎是最快乐的地方，等等。另外一个颇有影响力的工作是“We Feel Fine”项目，作者仅通过“We Feel X”的模板（其中 X 是待统计的情感词汇），在互联网博客等社会媒体中统计用户的情感分布，并用各种用户友好的可视化方案呈现给读者，可以很方便地查看不同类型用户（如男女、年龄）的主要情绪分布，如图 8.6 是该项目的搜索界面。可以说该工作也是充分利用互联网的海量、冗余的特点成功运用“基于互联网自然标注资源的自然语言处理”学术思想的典型代表。

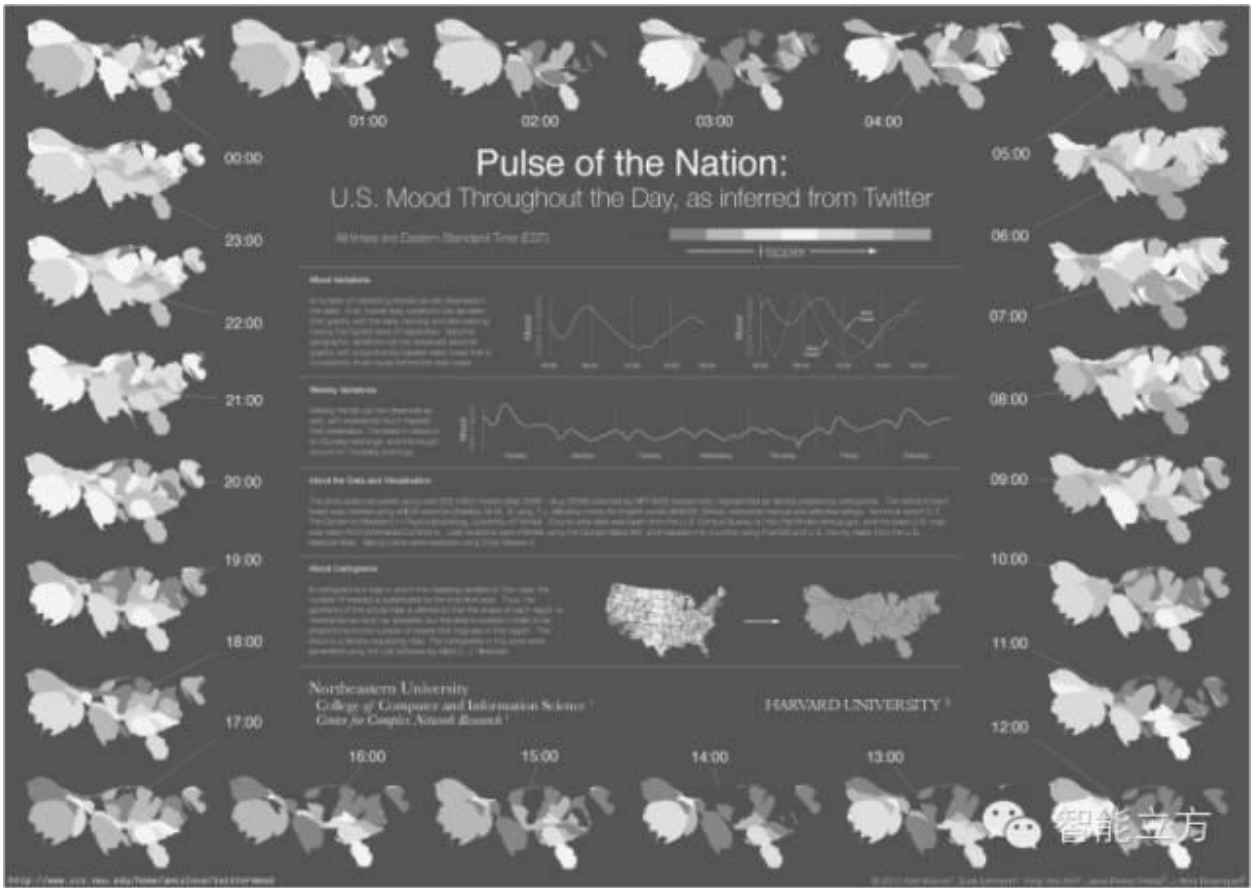

图 5　利用 Twitter 数据分析美国人情绪的时序变化。来自文献（Mislove, et al. 2010）

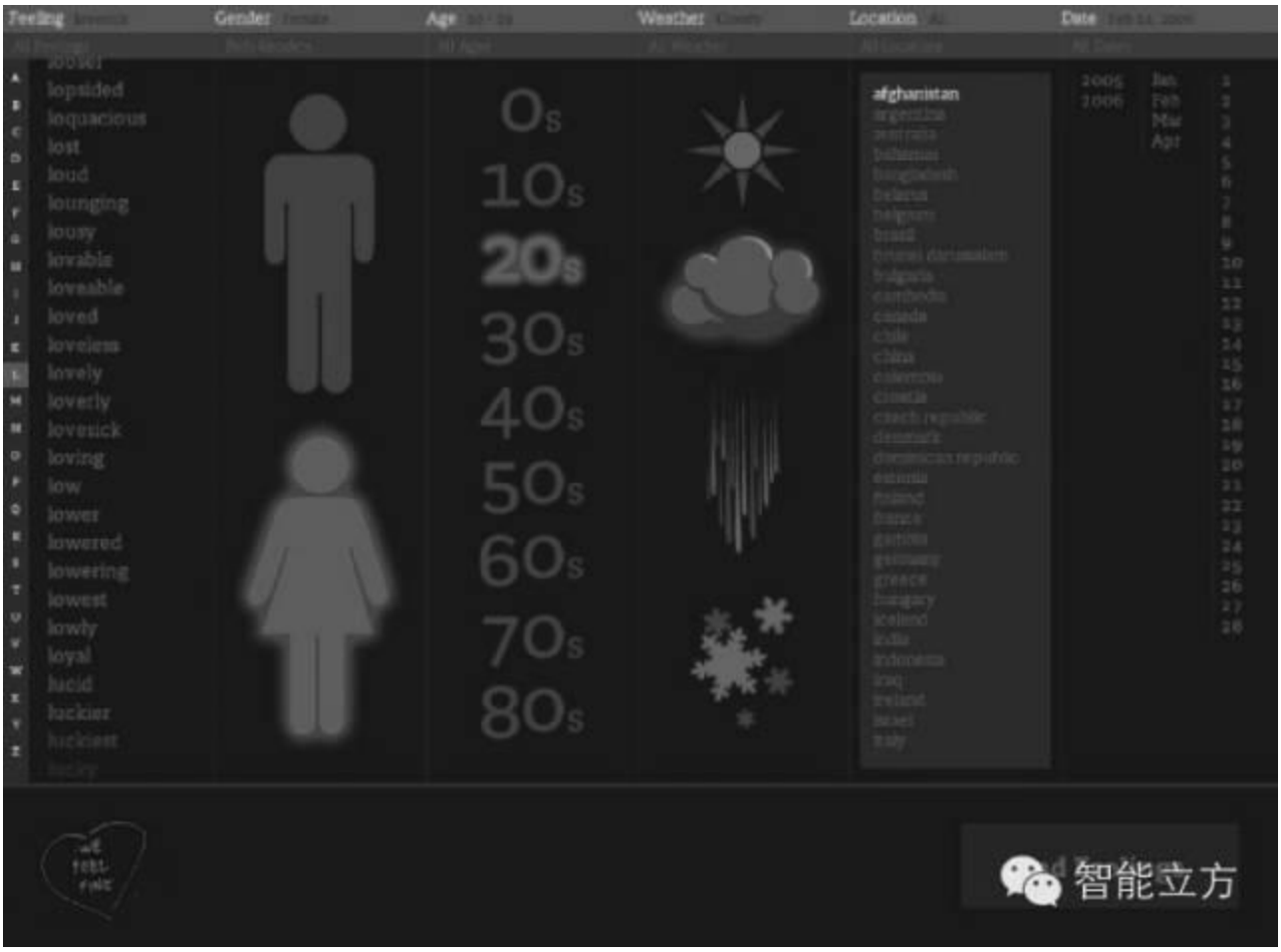

图 6　We Feel Fine 网站搜索界面。来自文献（Kamvar & Harris 2010）

2 面向社会媒体的自然语言分析应用

面向社会媒体中的自然语言分析技术有很多方面的应用，这里着重介绍几个有代表性的工作成果，相信在未来，会有更丰富而深入自然语言分析应用涌现出来。

2.1 社会预测

社会媒体用户产生内容在很大程度上反映了人们在社会生活方方面的关注和立场，因此，最近被广泛用来进行各种社会事件的预测，包括产品销量（如电影票房收入）（Joshi, et al. 2010）、体育比赛结果（Sinha, et al. 2013）、股市走势（Bollen, et al. 2011，Zhang, et al. 2011）、政治选举结果（如美国总统大选）（Gross, et al. 2013，Yano, et al. 2013，Chung & Mustafaraj 2011，Williams & Gulati ，Tumasjan, et al. 2010，O'Connor, etal. 2010）、自然

灾害传播趋势（如流行病传播）（St Louis &Zorlu 2012，Ritterman, et al. 2009），等等。
仅以政治选举为例，很多工作发现社会媒体中关于候选人的提及率就是很好的预测指标，例如根据 Facebook 上的支持率就能够成功预测 2008 年美国总统大选结果（Williams & Gulati 2008）。更惊人的是，《信号与噪声》（Silver 2012）的作者 Nate Silver 在 2012 年准确预测了美国 50 个州的总统选举结果，虽然他不仅使用社会媒体中的信息，而是充分占据可获得的各类信息来进行预测，但是毫无疑问社会媒体在其中发挥了重要作用。2012 年 Nature 上发表的一篇题为《一个 6100 万人参与的关于社会影响和政治动员的实验》的文章（Bond, et al. 2012），则系统分析了 2010 年美国总统大选期间 Facebook 用户的相关情况，发现通过 Facebook 上的信息递送等社会动员（Social Mobilization），至少影响了现实世界中数以百万计人群的政治自我表达和投票行为。这说明，社会媒体不仅反映了人们的各种立场，可以用于预测，而且社会媒体还会对人们的现实生活产生深远的影响。在未来，如何将预测与干预有效结合，更好地分析、管理和利用社会媒体平台，将是身处于大数据中的每个政府、企业和政策制定者面临的重要课题。

毋庸置疑，由于社会媒体用户属性与现实社会的用户属性存在一定偏置，例如在我国，社会媒体上年轻人居多，收入相对较高，因此他们传达出来的关注与观点，并不能完全反映整个社会的立场和形势。因此，在近年来社会预测与干预研究轰轰烈烈开展的同时，也有人反思其有效性（Gayo-Avello 2012）。但纵观大势，随着移动设备的普及和互联网的发展，越来越多的人成为社会媒体用户，相信只要充分正视在线社会媒体与真实社会之间存在的偏差，我们就能够更好地利用社会媒体做好社会管理工作，更好地为人类生活服务。

- ## 2.2 霸凌现象定量分析

面向社会媒体的自然语言分析不仅可以用来进行社会预测，还可以用来支持解决社会公益问题，其中霸凌（bully）现象就是典型代表。霸凌是社会科学、尤其是青少年研究的经典研究课题。然而传统研究方法中这个课题的数据普遍量小、缺乏、对问题的呈现不够全面。而在社会媒体领域中关注这一话题的人士又普遍把视野局限在了网上欺负他人这个小范围内，没能够把线上线下的霸凌行为进行整合。最近有研究（Xu et al. 2012; Angela et al. 204）开始通过对 Twitter 上与霸凌有关的文本/叙述进行分析，而其关注的范围包括现实和虚拟环境中的欺负行为。

在这个研究中，先是从大量的 Twitter 博文中选取了与霸凌有关的作为原始数据，再主要进行四个方面的分析：文本分类（把含有霸凌关键词但并不相关的文本剔除）、角色判断（判断在欺负行为中是指责者、欺负者、受害者、报告者、还是其他）、感情分析、话题判断。该课题也证明，面向社会媒体的自然语言分析将有助于识别霸凌现象，及时干预，给予儿童更健康的生活环境。

3 未来研究的挑战与展望

关于面向社会媒体的自然语言分析及其应用，已然成为今年的研究热点，呈星火燎原之势，以上简介限于作者所见，难免有顾此失彼、挂一漏万之处，需要感兴趣的读者不断探索更多的研究成果和发现。然而，通过以上研究工作，我们可大致总结出面向社会媒体的自然语言分析及其应用的发展趋势。

（1）自然语言的深度分析。我们可以看到，仅基于词汇层（单词或短语）的简单统计，就已经产生了大量影响深远的研究工作。而近年来，伴随着互联网大数据爆发式增长，自然语言处理和机器学习领域飞速发展，未来将会有更多的自然语言深度分析的技术和工具不断成熟，例如自动根据大规模文档集合进行词汇语义聚类的隐含主题模型（Steyvers & Griffiths 2007），进行情感分析和观点挖掘的相关技术（Liu 2012）、进行跨语言分析的机器翻译技术（Koehn 2010）、对人类知识进行结构化管理和推理的知识图谱（Singhal 2012），等等。这些技术和工具的不断成熟和完善，将使我们面向社会媒体的分析如虎添翼，打开另一双天眼，可以看到以往所无法看到的世界，从而发现以往所不能发现的规律。

（2）跨媒体、跨平台、跨信息源的综合分析。从媒体类型而言，虽然社会媒体的出现对传统主流媒体（如国内各大新闻门户网站）产生重大冲击，但可以看到，主流媒体和社会媒体各有侧重、互为补充、深度交融，均为人们日常生活不可或缺的信息来源，很多情况下，主流媒体的相关新闻事件可以作为社会媒体分析的大背景，是分析人格特质的重要因素，例如探索人们在面临重大事件（如特大自然灾害）时的反应，等等；从社会媒体平台而言，无论是 Twitter 还是 Facebook，都只反映了人们生活的某个切面，例如以 Twitter 为代表的微博平台更具备自媒体特质，而以 Facebook 为代表的社会网络服务更具备好友圈特质，但这些平台背后都是同样的人，他们在不同平台上会有怎样不同的表现，以及这样的表现原因是什么，这既是社会学关心的话题，也是商业服务关心的问题，最近社会计算中的一个热门研究问题就是社会媒体跨平台的相同用户识别（Vosecky, et al. 2009，Liu, et al. 2013）；从信息源而言，社会媒体用户产生的内容非常丰富，包括文本、图像、社会网络以及大量结构化信息（如 Facebook 中的个人属性，虽然往往填写不完整、不准确），其中文本内容固然是重要组成部分，也是本书关注重点，但其他信息源亦扮演重要角色，例如大规模社会网络分析（Leskovec, et al. 2008）、大规模图像标注（Weston, et al. 2010），等等。未来，面向社会媒体的分析及其应用，需要将文本内容与其他信息源充分融合，进行跨媒体、跨平台的融合分析，只有充分进行跨媒体、跨平台和跨信息源的综合分析，才能发现人类社会更复杂、更深层的科学规律。

总之，面向社会媒体的自然语言分析与应用，无论对社会学和信息科学各领域的推进，还是对商业服务的发展，均具有重要意义，日益引起人们的关注。其原因不言而喻，语言是人类区别于其他生物的最大特点，是进化厚赠人类的最珍贵礼物，也是人工智能、神经科学、社会语言学等领域孜孜以求希望真正理解的人类本质，还是人们

进行日常交流、传承文化的重要载体。可以想象，随着社会媒体和互联网产生的海量数据，随着自然语言处理和机器学习等技术的高速发展，面向社会媒体的自然语言分析与应用必将大行其道，大有作为。

4 参考文献

[注：因微信文章长度限制，请点击左下角阅读原文，可查看本文参考文献]

作者简介：刘知远，清华大学助理研究员，中国计算机学会高级会员。2011 年获得清华大学博士学位。主要研究方向为自然语言处理与社会计算，已在相关领域顶级会议和期刊发表论文 20 余篇。liuzy@tsinghua.edu.cn。
https://mp.weixin.qq.com/s?__biz=MzIxNzE2MTM4OA==&mid=413320663&idx=1&sn=a05c37c870fa439676d959ad73d231c8&scene=0#wechat_redirect