校样前的最后一个作者版本

Berendt, B.（出版中）。用于新闻和博客分析的文本挖掘。出现在 C. Sammut 和 GI Webb（编辑） 《机器学习和数据挖掘百科全书》中。柏林等：施普林格。

用于新闻和博客分析的文本挖掘

贝蒂娜·贝伦特

比利时鲁汶天主教大学

2015 年 2 月 15 日

注意：所有红色术语都是机器学习百科全书第一版中的索引术语；假设索引术语保持不变，它们可以用作新百科全书中的链接。如果条目列表增加，当然可以添加更多链接。（我假设这将在本章的处理过程中进行修改）。

定义

新闻是“对时事进行选择性传播”，其选择取决于“新闻价值”或“公众的兴趣”。新闻也是故事，读者通常希望从中得到五个问题的答案：谁、什么、何时、何地和为什么，其中通常会加上“如何”。新闻写作 与评论写作相反 通常力求客观和/或中立（表达对事件的不同看法）。

在这种以内容为中心的意义上，新闻可以由专业记者和新闻媒体（例如报纸或广播电台或电视台）撰写/创作和出版，但也可以由任何其他人以任何其他形式撰写，通常称为公民新闻： “在主流媒体机构之外运作的另类和积极的新闻采集和报道形式，通常作为对专业新闻领域缺陷的回应，使用类似的新闻实践，但由不同的目标和理想驱动，并依赖于传统或传统媒体之外的替代合法性来源主流新闻业。” （Radsch，2013 年，第 159 页）。然而，新闻或主流（媒体）新闻通常也被认为是以来源为中心的：由主流媒体机构的专业记者撰写的报道，而不是通常在网上发布的公民记者（或任何其他人）的报道。网络，以具有一定周期性的博客形式出现。

博客是网络上经常更新的出版物，按组成博客帖子的时间倒序排列。博客内容可以反映任何兴趣，包括新闻、个人、企业等。早期的博客文章（20 世纪 90 年代末）往往发布在内容管理平台上，没有长度限制；随着 Twitter 和类似的微博平台的成功，许多博客（以及博客挖掘）已转向短帖（例如 Twitter.com 和 Weibo.cn 上的 140 个字符，尽管后者的中文字符允许更复杂的消息）。Twitter 尤其在新闻（或简短的摘要和声明，通过超链接到更多文本和其他内容）的快速传播方面发挥了重要作用

媒体），公民记者、主流媒体本身、政治家和其他人是出版商（Kwak 等，2010）。当前博客挖掘的研究和本文的其余部分反映了（a）新闻或新闻相关内容和（b）微博格式的主导地位。此外，博客挖掘与社交媒体挖掘重叠（Zafarani et al., 2014）。特别是，微博作者的社交图允许挖掘分析师跟踪博主的来源和读者/“关注者”以及内容。

新闻和博客由文本和（在某些情况下）图片内容组成，并且基于网络时，可能
包含任何其他格式（例如视频、音频）和超链接的附加内容。它们已编入索引
按时间划分，并划分为更小的单元：新闻媒体划分为文章，博客划分为博客文章。在大多数
新闻和博客中，文本内容占主导地位。因此，文本分析是最常应用的形式
知识发现。这包括数据/文本挖掘、信息检索和相关领域的任务和方法。随着这些领域的日益融合，本文将它们统
称为文本挖掘。本条目将说明这些字段的重叠/使用，并强调从该领域派生的细节，包括数据、任务、用户和使用

案例。

动机和背景

新闻和博客是当今了解时事的最常见来源，博客也是发表对时事看法的渠道。此外，它们还可能涉及更长期关注的话题。两者都反映并形成了社会、团体和个人对世界的看法，这些看法随着引发报道的事件而迅速甚至即时地发生。然而，这两种媒体在创作、内容和形式方面存在差异。新闻通常由受过新闻训练的人撰写，他们遵守有关报道风格和语言的新闻标准。报道的主题和方式受到社会普遍共识和新闻提供商政策的限制。相比之下，每个能上网的人都可以开设博客，而且对内容和风格没有任何限制（除了适用的审查类型）。因此，博客为最终用户提供了更广泛的主题和观点。

这些应用特征给新闻和博客的文本挖掘分析带来了各种语言和计算挑战：

-索引、分类、部分冗余和数据流：新闻按时间和来源（新闻机构或提供商）建立索引。在多源语料库中，大约在同一时间（以相同或其他语言）发表的许多文章描述相同的事件。随着时间的推移，文章中可能会发展出一个故事。博客中的热门主题也观察到这种多重报告和时间结构。

-语言和含义：新闻以清晰、正确、“客观”的方式撰写，并在某种程度上
图式化的语言。通常，新闻文章的开头总结了整篇文章（提要是博客中的部分类似）。来自新闻机构等外部来源的信息通常会被重复使用，而不是被引用。总之，与许多其他文本相比，新闻对读者的背景和上下文知识的假设较少。

-非标准语言和主观性：博客中的语言范围从高质量的、
“新闻式”语言，质量低劣、代码受限且有很多拼写和语法错误的语言，以及富有创意的、有时是文学性的语言。博客可能使用高质量的

高质量的语言，但在新闻类型之外或跨新闻类型（例如将时事报道与评论和背景信息结合起来）。行话在博客中非常常见，新的语言发展的采用速度远远快于词汇等外部资源所反映的速度。许多博客作者不追求客观性，而是追求主观性和情感性。

-主题多样性和新的分类形式：新闻通常按主题分类领域（“政治”、“商业”等）。相比之下，博客作者可以选择不同的、任意的主题。博客被标记时，通常不是参考稳定的分类系统，而是使用任意数量的标签：用户选择的自由形式、通常是非正式的标签。

-上下文及其对内容和意义的影响：博客（帖子）的内容通常不仅仅包含在文本中。相反，博客软件支持“网络”和“社交网络”行为，博主们也实践了这种行为：多向交流而非广播，以及对内容和人的语义诱导引用。具体来说，指向其他资源的超链接不仅提供上下文，还提供内容；指向和来自引用的链接也是如此。

人员/来源。后者是从“blogrolls”演变而来的。早期博客软件中的“引用链接”分别为“关注者”和“转发”链接。 Twitter 等平台中的“关注者”。

学习系统的结构

任务

从文本挖掘的角度来看，任务可以按照不同的标准分组：

-基本任务和结果类型：描述、分类和预测（监督或无监督，包括主题识别、跟踪和/或新颖性检测；垃圾邮件检测）；搜索（临时或过滤）；推荐（博客、博客文章或（哈希)标签）；总结

-需要提取高阶特征：尤其是主题或事件；观点或情绪

-时间维度：非时间维度；时间（流挖掘）；多个流（例如，不同语言，请参阅跨语言文本挖掘）

-用户适应性：无（没有明确提及用户问题和/或一般受众）；可定制；个性化

现实世界的应用程序越来越多地根据其目标用户和用例来选择或更常见地组合这些任务，特别是：

-新闻聚合器允许普通用户和专业用户（例如记者）查看“新闻内容”，并比较不同来源对同一新闻的文本。新闻聚合器的来源通常是新闻（尤其是主流新闻），这反映了这样的假设：

虽然“白名单”聚合器大多是客观/中立的，但它们专注于主题和事件。现在所有主流搜索引擎都提供新闻聚合器。

-社交媒体监控工具使外行和专业用户不仅可以跟踪关键字或命名实体（例如人物、品牌）的话题提及，还可以汇总对其的情绪。对情绪的关注反映了这样一种看法，即即使与新闻相关，社交媒体内容也往往是主观的，因此研究博客圈是进行市场研究或舆论研究的一种廉价方式。这里的白名单是

通常是平台（例如 Twitter、Tumblr、LiveJournal、Facebook）而不是来源本身，这反映了博客圈/社交网络的巨大规模和动态结构。商业和免费社交媒体监控工具的前景广阔且变化频繁；您可以在网络上轻松找到最新的概述和比较。

- 新兴的应用类型包括文本挖掘，特别是针对新闻文本

在具有高度系统化事件结构和报道的领域中生成自然语言，例如体育和财经报道（例如 Allen 等人，2010 年；narrativescience.com）以及帮助记者寻找消息来源的社交媒体监控工具（Diakopoulos 等人，2012 年））。

有些工具具有仪表板式界面和复杂的数据图形，这对于一些专业用户来说可能是最有趣的。然而，特别是休闲用户越来越多地转向小屏幕移动设备，导致大多数应用程序显示原始内容和挖掘输出，其中包含（特别是短的）文本和少量（特别是数字）分析。

解决方法

标准化：任务、数据集、API

新闻、博客和社交媒体挖掘方法的开发总体上受益于标准数据集、标准任务和竞赛。突出的例子是Reuters-21578数据集，它不仅是新闻专线文章的集合，而且还是一般文本挖掘的最经典数据集（https://archive.ics.uci.edu/ml/datasets/Reuters-21578+文本+分类+收藏），更大且多语言的 RCV1、RCV2 和 TRC2 数据集(http://trec.nist.gov/data/reuters/reuters.html)，国际博客和社交媒体会议（ICWSM， http://www.icwsm.org)提供的博客数据集，和 SNAP 数据集(https://snap.stanford.edu/data)。主题检测和跟踪（TDT）研究计划和研讨会（http://www.itl.nist.gov/iad/mig/tests/tdt； Allan，2002 年）对新闻挖掘作为研究主题的形成至关重要。正在进行的重要任务和竞赛也提供了重要的数据集，包括文本检索会议 (TREC， http://trec.nist.gov)以及文本分析会议（TAC，http://www.nist.gov/tac），前身为文档理解会议 DUC (http://duc.nist.gov)。这些会议中轨迹/任务的历史也说明了各个领域如何成熟或变得不再重要；例如，自 2010 年以来“博客轨迹”已被“微博轨迹”取代，“主题检测”已让位于“事件检测”。

标准数据集是解决新闻、博客和社交媒体挖掘中一个核心问题的方法。由于大多数平台都是商业平台，因此它们限制访问其当前版本或存档版本。其他平台提供免费 API，但会返回代表性和/或采样标准未知的样本；这可能会严重影响挖掘结果（Morstatter 等人，2013 年）。此外，使用条款对创建可重复使用的数据集提出了挑战（有关解决方法，请参阅 McCreadie 等人，2012 年）。

另一个警告涉及所有社交媒体挖掘结果：一般来说，API 仅允许访问“公开”帖子，而不能访问用户设置为“私人”或仅限于受限受众的帖子。

此外,获得个人在线通信的访问权限并不意味着可以使用或处理它。因此,出于隐私和数据保护的考虑,限制了社交媒体用于研究的使用;它们需要对结果进行仔细解释:这些可能代表用户的公开言论,但并不代表他们所有的在线交流。

文本挖掘的建模阶段

解决方案方法基于通用数据挖掘方法,并适应新闻和博客及其挖掘任务的概念细节(请参阅上面的任务列表)。方法包括(文档)分类和聚类、潜变量技术,例如 (P)LSA 或 LDA(参见特征构建;具体有关主题模型的概述,请参见 Blei,2012)、混合模型、时间序列和流挖掘方法。

命名实体识别(例如 Atkinson & Van der Goot,2009;Ritter 等人,2011;Li 等人,2012)是主题检测或文本丰富(例如 tajner 等人,2010)等任务的重要组成部分或伴随任务。主题跟踪和事件线索用于跟踪新闻故事随时间展开的过程(例如

Shahaf 和 Guestrin,2010),特别是为了随时间进行总结,特别关注突发性主题或事件(Kleinberg,2002 年引入的术语;有关进一步参考和实证比较,请参阅 Suba i 和 Berendt,2013),即那些在某些时间点,报告的频率或其他权重以“峰值”为标志。

信息提取有助于提取新闻报道中的事件。事件涉及命名实体(例如人物和地点)、时间和事件特征。信息提取可以利用背景本体(例如 Kuzey 等人,2014 年)。这涵盖了新闻报道的“五个 W”中的前四个;目前,“为什么”和“如何”仍有待人类读者从原文中提取(因此通常可以通过平台访问原文,请参阅下面关于半自动意义建构的评论)。聚类可用于从多语言来源中提取事件(Leban 等人,2014 年)。报道(或世界?)演变的规律也被用于预测新闻事件(Radinsky & Horvitz,2013 年)。微博的简洁性及其信息流的速度和容量对事件检测提出了特殊挑战(McCreadie 等人,2013 年)。

情感分析和意见挖掘对于分析博客和其他社交媒体尤其重要(参见 Feldman,2013;Pang & Lee,2007;Potts,2013 中的概述),并且它们正在朝着采用句法结构和背景知识的更复杂的方法发展/考虑语义(例如 Gangemi 等人,2014)。情感分析和意见挖掘旨在检测和分类“主观”内容,从而很好地描述(某些)社交媒体内容。它也适用于“主观”新闻类型,例如评论。然而,这并不意味着新闻是真正的 或者永远是真正的 客观的。在新闻故事的讲述方式中表达出来的通常微妙且通常是潜意识的结构、背景和信念被称为媒体偏见或框架,文本挖掘已经开始解决这些问题(例如Recasens等人,2013年;Pollak等人) .,2011;奥迪克等人,2013)。

与新闻和博客特别相关的进一步分类任务通常使用具有领域特征和/或可从其数据中轻松提取的特征来解决。它们包括(a)地理位置(例如 Hale 等人,2012 年);(b)推荐(例如,在新闻中随时间跟踪多个主题,针对兴趣可能随时间变化的用户进行个性化推荐,由 Pon 等人开发

等人,2007 年； Ren et al., 2013 提出了一种微博方法； (c) 垃圾邮件检测和拦截（Kolari 等人,2006 年;一般概述请参见 Castillo & Davison,2011 年）。

文本摘要（概述见 Fiori,2014;具体到微博,见 Mackie 等人,2014)是帮助用户概览（a)单个文档的关键信息或（b)大量不同文档的关键技术,这些文档通常来自不同的来源,而这些来源又可能相互抄袭。如今,大多数摘要都是提取式的,要么提取关键句子,要么提取非句子结构（如图表）。在实际应用中,更简单的形式仍然占主导地位,包括基于频率提取单个术语及其在标签云中的显示,以及使用新闻文章的第一句,根据新闻写作惯例,这些句子旨在总结文本。抽象摘要涉及自然语言的生成,这仍然是一个难题。如今,它主要用于高度模式化的文本类型,这样就可以使用模板并填充与手头故事相关的实体/常量（见上文“新兴应用类型”）。

文本或文本摘要不仅可以表示为词袋、主题或事件集,还可以表示为单词和/或命名实体彼此之间存在多种关系的图表（有关示例和更多参考资料,请参阅 Berendt 等人,2014 年）。(浅层)语义解析通常用于提取三元组（例如主语-谓语-宾语语句）(例如 tajner 等人,2010 年;Sudhahar 等人,2015 年）。

基于文本的建模可以通过（例如社交)网络结构来增强（例如 Mei、Cai、Zhang 和 Zhai,2008）(参见链接挖掘和链接发现）。分析网络中的参与者如何相互影响对于新闻和社交媒体领域非常重要（Guille 等,2013）。此类分析不仅适用于单个文本生成者,而且更经常适用于整个领域。一个普遍的问题是,从总体上看,博客和新闻如何相互引用和关联（例如,Gamon 等人,2008 年;Berendt & Trümper,2009 年;Leskovec、Backstrom 和 Kleinberg,2009 年）。

数据理解、数据清理和数据准备的细节

数据清理与其他在线文档的清理类似;具体来说,它需要提供或学习用于删除标记元素的包装器。专注于文本挖掘的分析方法通常会忽略照片和视频等超媒体元素,或仅使用其元数据。

虽然新闻文本采用标准语言并且可以用通用文本分析软件处理,但（微)博客的语言需要特定的词汇（例如,包含常用表情符号）、缩写扩展和语法规则以及类似的技术（参见“诺亚方舟” http://www.ark.cs.cmu.edu/TweetNLP/一套工具和参考）;语言学家发现,微博并没有“错误”和不合语法,而是正在向类似于口语并指示地理区域等细微差别的新系统发展。

（爱森斯坦,2015）。与其他社交媒体一样,它们经常包含讽刺和其他间接使用语言来表达赞赏或不满（例如 Veale &hao,2010）,这仍然是机器理解这些文本的主要障碍。

博客和新闻的半结构化性质可以为理解提供有价值的线索。例如,
格式元素“时间戳”和“评论数量”可以分别被视为主题相关性增加和固执己见可能性的指标（Mishne,2007）。 A

文本聚类和标签分析的结合可用于识别主题以及与主题相关且可能长期保持关注的博客（Hayes 等,2007）。例如， Twitter 标签已被用作情绪指标（Wang 等,2011）。

与其他在线文本一样,新闻和博客经常使用超链接,链接材料的内容甚至可能对人类读者理解帖子也是必不可少的。对于通常只是指向 URL 的指针或 URL 加简短评论的微博来说尤其如此。因此,许多挖掘方法通过引用 URL 的内容等来丰富文本（例如 Abel 等人,2011 年）。语义丰富还可以利用外部（半）结构化数据;例如,Wikification 可以通过借鉴 Wikipedia 或 DBPedia 为微博添加上下文信息（例如 Cheng & Roth,2013 年）。所有这些方法都有助于丰富和消除歧义。

交互式工具对于半自动意义建构的重要性

与大多数文本挖掘一样,无论对于新闻消费者还是新闻生产者来说,对新闻、博客和其他社交媒体的机器分析都是人类意义建构过程的第一步。因此,必须为他们提供支持进一步步骤的接口。因此,新闻消费者的工具（例如新闻聚合器）通常提供原始文章的链接。新闻制作人的工具将统计数据（例如“人群”的总体意见或一个潜在来源的属性）显示为记者的信息,并且在语料库中检测到的主题或事件通常是故事的起点,但不是故事的起点和他们自己。阅读、理解和撰写新闻和博客可能永远不可能完全自动化。造成这种情况的一个原因是,不同的人对给定文本的阅读方式不同,这在社会科学媒体研究中众所周知,但在计算研究中仍然经常被忽视 也许是因为它要求我们质疑文本挖掘的关键方法论概念,例如“地面真相”。人们提出了用于故事检测和跟踪的交互式工具来解决这一困境（Berendt et al., 2014）,并且拖放故事编辑器用于创建自己的新故事（storify.com）。

此外,文本挖掘作为一种处理大量数据的方法,经常与人类智能竞争或结合。例如,许多（通常是无偿的）志愿者的贡献和投票等界面元素构成了“社交新闻聚合器”reddit.com,而Twitter的“转发”是一种主要的、由人类主导的方式,推文被输入到平台用户形成的多个子网络中,并在这些子网络中产生影响。然而,在这些人机协作中,平台采用的算法并不是中立的伙伴,而是塑造了用户如何看待他人的观点,进而影响他们进一步的发布行为。例如,Twitter的“热门话题”算法奖励突发话题

（参见威尔逊,2013）。这意味着,如果兴趣随着时间的推移保持稳定,即使是许多推文中包含的主题也可能从热门主题中消失,从而从公众视野中消失。此类算法决策对用户选择和感知以及公共决策和政策的影响是一个新的研究主题,不仅与文本挖掘相关。

参考

Abel, F.、Gao, Q.、Houben, G.-J. 和 Tao, K. (2011)。Twitter 帖子语义丰富，用于社交网络上的用户资料构建。在 Proc. ESWC (2) 中（第 375-389 页）。

Allan, J. (Ed.)。(2002)。主题检测和跟踪：基于事件的信息组织。Norwell，马萨诸塞州：克鲁维尔学术出版社。

Allen, ND、Templon, JR、McNally, PS、Birnbaum, L. 和 Hammond, K. (2010)。 StatsMonkey：数据驱动的体育叙事作家。在过程中。 2010 AAAI 秋季研讨会系列。 AAAI出版社。 http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2305

Atkinson, M. 和 Van der Goot, E. (2009)。多语言新闻中的近实时信息挖掘。第 18 届万维网国际会议 (WWW 09) 论文集。ACM，纽约，纽约，美国，1153-1154。

Berendt, B.、Last, M.、Suba i , I. 和 Verbeke, M. (2014)。多文档新闻摘要的新格式和界面及其评估。载于 Fiori (2014)（第 231-255 页）。

Berendt, B. 和 Trümper, D. (2009)。基于语义的异构文本语料库分析和导航：Porpoise 新闻和博客引擎。在I.-H。丁和H.-J。 Wu（编辑），电子商务和电子服务中的 Web 挖掘应用程序，柏林：Springer。

布莱，DM (2012)。概率主题模型。 ACM 通讯，55(4), 77-84。

Castillo, C. 和 Davison, BD (2011)。对抗性网络搜索。信息检索基础与趋势。4、5（2011 年 5 月），377-486。DOI=10.1561/1500000021 http://dx.doi.org/10.1561/1500000021

Cheng, X. 和 Roth, D. (2013)。维基百科的关系推理。在进程中EMNLP 2013（第 1787-1796）。

Diakopoulos, N.、De Choudhury, M. 和 Naaman, M. (2012)。在新闻业背景下寻找和评估社交媒体信息来源。在 Proc. CHI 2012（第 2451-2460 页）。ACM。

Eisenstein, J. (2015)。识别在线社交媒体中的区域方言。即将在《方言学手册》中发表。

Feldman, R. (2013)。情绪分析技术和应用。ACM 通讯，56(4)，82-89。

Fiori, A.（主编）（2014 年）。创新的文档摘要技术：彻底改变知识理解。 IGI 全球。

Gamon, M.、Basu, S.、Belenko, D.、Fisher, D.、Hurst, M. 和 König, AC (2008)。 BLWS：使用博客为新闻文章提供上下文。 E. Adar、M. Hurst、T. Finin、N. Glance、N. Nicolov、B. Tseng 和 F. Salvetti（编），第二届网络博客和社交媒体国际会议记录（ICWSM 08）。华盛顿州西雅图市。加利福尼亚州门洛帕克。 http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-015.pdf

Gangemi, A.、Presutti, V. 和 Reforgiato Recupero, D. (2014)。基于框架的意见持有者和主题检测:模型和工具。 IEEE 计算智能杂志,9(1), 20-30。

Guille, A., Hacid, H., Favre, C., & Zighed, DA (2013). 在线社交网络中的信息传播:一项调查。SIGMOD 记录,42(2)。

Hale, S.、Gaffney, D. 和 Graham, M. (2012)。你到底在哪里? Twitter 中的地理位置和语言识别。 ICWSM 12 会议记录(第 518-521 页)。

Hayes, C.、Avesani, P. 和 Bojars, U. (2007)。对博客推荐系统的博主、主题和标签进行分析。 B. Berendt、A. Hotho、D. Mladeni 和 G. Semeraro(编者),从网络到社交网络:发现和部署用户和内容配置文件。 LNAI 4737。柏林:施普林格。

克莱因伯格,JM (2002)。流中的突发和分层结构。在过程中。 SIGKDD 2002(第 91 页-101)。

Kolari, P.、Java, A.、Finin, T.、Oates, T. 和 Joshi, A. (2006)。检测垃圾博客:机器学习方法。第 21 届全国人工智能大会论文集。波士顿:AAAI。

Kuzey, E.、Vreeken, J. 和 Weikum, G. (2014)。重新审视知识库:从新闻中提炼命名事件。在 Proc. CIKM 2014(第 1689-1698 页)。

Kwak, H.、Lee, C.、Park, H. 和 Moon, S. (2010)。什么是 Twitter,社交网络还是新闻媒体?在 Proc. WWW(第 591-600 页)。ACM。

Leban, G.、Fortuna, B.、Brank, J. 和 Grobelnik, M. (2014)。事件登记:从新闻中了解世界事件。在过程中。 WWW 2014(配套卷)(第 107-110 页)。

Leskovec, J.、Backstrom, L. 和 Kleinberg, J. (2009)。模因追踪和新闻周期的动态。 JF Elder IV、F. Fogelman-Soulié、PA Flach 和 MJ Zaki(编辑),第 15 届 ACM SIGKDD 知识发现和数据挖掘国际会议记录,法国巴黎。纽约,

纽约。

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. 和 Lee, B.-S. (2012)。TwiNER:有针对性的推特流中的命名实体识别。第 35 届国际 ACM SIGIR 信息检索研究与开发会议 (SIGIR 12) 论文集。ACM,美国纽约州纽约,721-730。

DOI=10.1145/2348283.2348380 http://doi.acm.org/10.1145/2348283.2348380

Mackie, S.、McCreadie, R.、Macdonald, C. 和 Ounis, I. (2014)。比较微博摘要算法。在 Proc. CLEF 2014(第 153-159 页)。

McCreadie, R.、Macdonald, C.、Ounis, I.、Osborne, M. 和 Petrovic, S. (2013)。Twitter 的可扩展分布式事件检测。在 2013 年 BigData 大会论文集(第 543-549 页)。

McCreadie, R.、Soboroff, I.、Lin, J.、Macdonald, C.、Ounis, I. 和 McCullough, D. (2012)。关于构建可重复使用的 Twitter 语料库。第 35 届 ACM SIGIR 国际信息检索研究与开发会议 (SIGIR 12) 论文集。ACM,美国纽约州纽约,1113-1114。

DOI=10.1145/2348283.2348495 http://doi.acm.org/10.1145/2348283.2348495

Mei, Q.、Cai, D.、Zhang, D. 和 Zhai, C. (2008)。主题建模与网络正则化。J. Huai 和 R. Chen (Eds.),第 17 届万维网国际会议论文集 (WWW 08)
中国北京。纽约州,纽约州。 10.1007/978-0-387-30164-8_827

米什尼,G.（2007）。使用博客属性来改进检索。在 N. Glance、N. Nicolov、E. Adar、M.
Hurst, M. Liberman 和 F. Salvetto (Eds.),国际博客和社交媒体会议 (ICWSM) 论文集。科罗拉多州博尔德。http://www.icwsm.org/papers/paper25.html

Morstatter, F.、Pfeffer, J.、Liu, H. 和 Carley, KM (2013)。样品足够好吗?将 Twitter 的 Streaming API 与 Twitter 的 Firehose 中的数据进行比较。在过程中。 ICWSM 2013。 http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071

Odijk, D.、Burscher, B.、Vliegenthart, R. 和 de Rijke, M. (2013)。自动主题内容分析:在新闻中查找框架。在社会信息学 2013 年（第 333-345 页）。柏林等:Springer。 LNCS 8238。

庞 B. 和李 L. (2007)。意见挖掘和情感分析。信息检索的基础和趋势 2(1-2): 1-135 (2007)。

Pollak, S.、Coesemans, R.、Daelemans, W. 和 Lavra , N. (2011)。通过结合话语分析和文本挖掘来检测报纸文章中的对比模式。语用学,21 (4), 647-683。

Pon, RK、Cardenas, AF、Buttler, D. 和 Critchlow, T. (2007)。跟踪多个主题以查找有趣的文章。收录于 P. Berkhin、R. Caruana 和 X. Wu (Eds.),第 13 届 ACM SIGKDD 国际知识发现和数据挖掘会议论文集。加利福尼亚州圣何塞。纽约州纽约市。

Potts (2013)。情感分析简介。（幻灯片集）。http ://www.stanford.edu/class/cs224u/slides/2013/cs224u-slides-02-26.pdf [检索日期 2015-02-15]

Radinsky, K. 和 Horvitz, E. (2013)。挖掘网络来预测未来事件。在过程中。 WSDM 2013（第 174 页）255-264)。

Radsch, CC (2013)。数字异议与政治变革:埃及的网络行动主义和公民新闻。美国大学国际服务学院博士论文。可在 SSRN 上获取:http://ssrn.com/abstract=2379913

Recasens, M.、Danescu-Niculescu-Mizil, C. 和 Jurafsky, D. (2013)。用于分析和检测有偏见的语言的语言模型。在 ACL 诉讼中。

Ren, Z.、Liang, S.、Meij, E. 和 de Rijke, M. (2013)。个性化时间感知推文摘要。第 36 届 ACM SIGIR 国际信息检索研究与开发会议 (SIGIR 13) 论文集。ACM,美国纽约州纽约,513-522。DOI=10.1145/2484028.2484052 http://doi.acm.org/10.1145/2484028.2484052

Ritter, A.、Clark, S.、Mausam 和 Etzioni, O. (2011)。推文中的命名实体识别:一项实验研究。自然语言处理实证方法会议论文集 (EMNLP 11)。计算语言学协会,美国宾夕法尼亚州斯特劳兹堡,1524-1534。

Shahaf, D. & Guestrin, C. (2010)。连接新闻文章之间的点。在 Proc. SIGKDD 2010 中（第 623-632 页）。

Sudhahar, S.、de Fazio, G.、Franzosi, R. 和 Cristianini, N. (2015)。大型语料库中叙述内容的网络分析。自然语言工程,21(1),81-112。

tajner, T.、Rusu, D.、Dali, L.、Fortuna, B.、Mladenic, D. 和 Grobelnik, M. (2010)。用于自然语言文本丰富的面向服务的框架。Informatica（卢布尔雅那）,34 (3), 307-313。

苏巴西奇,I.,贝伦特,B. (2013)。故事图:使用动态图跟踪文档集演变,智能数据分析,17 (1), 125-147。

维尔,T. 和郝,Y. (2010)。检测创意比较中的讽刺意图。 2010 年 ECAI 会议记录:第 19 届欧洲人工智能会议,Helder Coelho、Rudi Studer 和 Michael Wooldridge（主编）。 IOS Press,荷兰阿姆斯特丹,荷兰,765-770。

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). Twitter 中的主题情绪分析:基于图的标签情绪分类方法。在第 20 届 ACM 信息和知识管理国际会议 (CIKM 11) 论文集上,Bettina Berendt、Arjen de Vries、Wenfei Fan、Craig Macdonald、Iadh Ounis 和 Ian Ruthven (Eds.)。ACM,美国纽约州纽约,1031-

1040.DOI=10.1145/2063576.2063726 http://doi.acm.org/10.1145/2063576.2063726

威尔逊,R.（2013）。 Twitter 上的趋势:热门话题背后的算法一探究竟。点燃社交媒体博客。 http://www.ignitesocialmedia.com/twitter-marketing/trending-on-twitter-a-look-at-algorithms-behind-trending-topics/ [检索于2015年2月15日]

Zafarani, R.、Abbasi, MA 和 Liu, H. (2014)。社交媒体挖掘:简介。剑桥:剑桥大学出版社。