国际计算机应用杂志（0975 – 8887）
第 112 卷–第 5 期,2015 年 2 月

社交媒体的情绪分析和文本挖掘使用开源工具的微博:实证研究学习

Eman MG Younis埃及米
尼亚大学计算机与信息学院

抽象的

最近,社交媒体不仅作为个人通信媒体出现,而且作为用户之间交流有关产品和服务甚至政治和一般事件的意见的媒体。由于其广泛和受欢迎,每天都会产生和分享大量的用户评论或意见。 Twitter 是使用最广泛的社交媒体微博网站之一。从社交媒体数据中挖掘用户意见并不是一项简单的任务;它可以通过不同的方式来完成。在这项工作中,提出了一种开源方法,在该方法中,使用开源工具收集、预处理、分析和可视化 Twitter 微博数据,以执行文本挖掘和情感分析,以分析用户对两家大型零售商店贡献的在线评论在英国,即 2014 年圣诞节期间的 Tesco 和 Asda 商店。使用调查等传统方法收集客户意见可能是一项昂贵且耗时的任务。对客户意见的情感分析使企业更容易了解自己在不断变化的市场中的竞争价值,了解客户对其产品和服务的看法,这也有助于洞察未来的营销策略和决策政策。

一般条款

自然语言处理、观点挖掘。

关键词

文本挖掘、情感分析、开源、Twitter 数据分析、社交数据挖掘、R 包。

1. 简介由于信息技术的巨大而快速

的进步 ,社交媒体已成为一种新兴现象。

人们每天都在使用社交媒体来交流他们对各种主题、产品和服务的看法,这使其成为文本挖掘和情感分析的丰富资源。社交媒体通信包括 Facebook、Twitter 等。

Twitter 是使用最广泛的社交媒体网站之一。
图 1 显示了全球每秒发送的 Twitter 消息数。在文献中,没有挖掘和分析社交媒体业务数据的标准方法。

本文介绍了一种使用一组 R 包 [2、6 和 7] 进行文本挖掘和情绪分析的开源方法,该方法用于挖掘 Twitter 数据和进行情绪分析,适用于其他社交媒体网站。本文以两家英国商店为例,展示了分析用户从微博中生成的在线意见的重要性。这有助于企业从客户的角度监控其绩效,而不是进行成本高昂且耗时的客户调查。

本文的其余部分组织如下:第 2 节简要介绍文本挖掘。第 3 节概述了情绪分析领域和相关工作。第 4 节展示了针对 Twitter 微博进行挖掘和情绪分析的建议方法。第 5 节提供实验细节并展示结果。第 6 节介绍研究的结论和含义,并展示了未来的研究可能性。

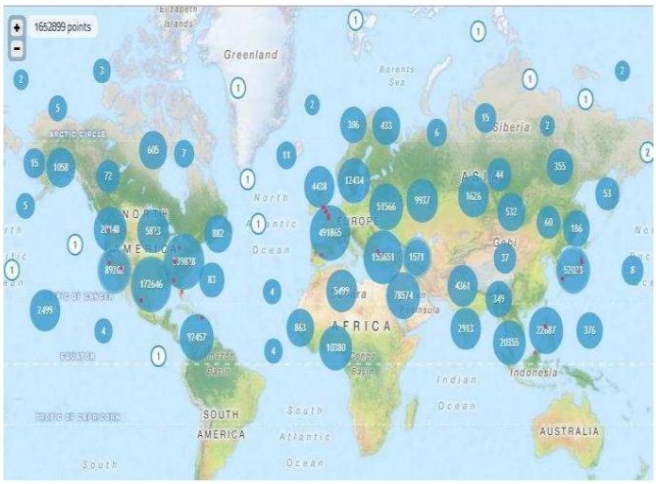

图 1:绘制全球每秒推文数量图。正如[8]中捕获的那样。

2. 文本挖掘文本挖掘是检测和

揭示非结构化文本数据资源中新的、未发现的知识以及相互关系和模式的自动化过程。文本挖掘的目标是大量文本中未被发现的知识。而搜索引擎和信息检索（IR)系统具有特定的搜索目标,例如搜索查询或关键词,并返回相关文档[1]。该研究领域利用数据挖掘算法,例如分类、聚类、关联规则等来探索和发现文本源中的新信息和关系。它是一个集信息检索、数据挖掘、机器学习、统计学和计算语言学于一体的跨学科研究领域[1]。图 2 总结了文本挖掘过程。首先,收集一组非结构化文本文档。然后,对文档进行预处理,去除噪音和常用词、停用词、词干。此过程生成文档的结构化表示,称为术语文档矩阵,其中每一列代表一个文档,每一行代表整个文档中出现的术语。最后一步是应用数据挖掘技术,例如聚类、分类、

关联规则来发现文本中的术语关联和模式,最后使用词云或标签云等工具可视化这些模式。

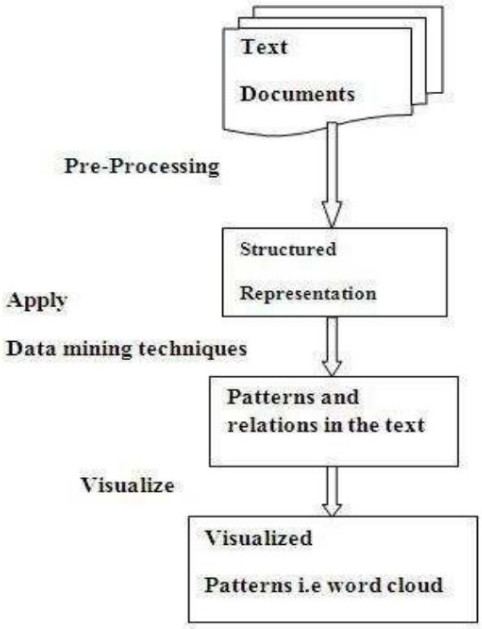

图 2:文本挖掘过程。

3. 情感分析情感分析是由Liu, B [4]首先提出的。它

也称为意见挖掘和主观性分析,是确定人类撰写的意见或评论的态度或极性以对产品或服务进行评级的过程。

情感分析可以应用于任何文本形式的观点,例如博客、评论和微博。微博是那些类似于推文的小文本消息,是一条不能超过 149 个字符的短消息。这些微博比其他形式的观点更容易进行情感分析[11]。情感分析可以在文档级别或

句子级别。在第一种情况下,需要评估整个文档来确定观点极性,其中首先需要提取描述产品/服务的特征。而在第二种情况下,需要将文档分成句子,然后分别评估每个句子以确定观点极性 [4]。

3.1 相关工作Twitter 已在许多

研究中用于不同目的的情感分析[5、10、11、12、13、14 和21]。例如,周等人。 Tumasjan 等人[10, 13]提出了一种从 Twitter 中挖掘有关总统选举候选人的意见并预测选举结果的方法。阿苏尔等人。 [12],提供了一个通过分析社交媒体数据来预测电影预期收入的模型。达斯等人。 [17],实现了一个挖掘公众推特对三星 Galaxy 手机意见的框架。同样,Mostafa, MM [21] 使用 Twitter 数据来评估诺基亚、三星、IBM 等大品牌和埃及航空等航空公司的消费者情绪。情感分析最近已应用于许多其他领域,以分析和预测公众对各种产品、服务、社会和政治事件的行为和感受。情感分析本身已在[20]中得到应用,以探索电子邮件和故事书中的情感。

可以使用两种方法进行情感分析。第一种是基于意见词典的方法[14],其中词典由一组正面和负面意见词组成,用于对意见句子进行评分,正面、负面或中性。这种方法非常流行,需要一个评分函数根据正面或负面单词的存在对每个句子进行评分。图3显示了基于词典的情感分析方法。基于词典的方法使用词典（一组正面和负面单词)结合评分函数来确定情感极性。第二种方法是使用机器学习技术来训练分类器,使用一组预分类意见作为训练集。然后,使用经过训练的分类器将新意见分类为正面、负面或中立[15]。 Pak, A 等人 [5],使用监督技术使用词性标注器和 N-gram 方法构建分类器,并使用该分类器对意见进行分类。研究证实基于词典的方法优于机器学习方法[15]。这项工作利用基于词典的方法进行情感分析。

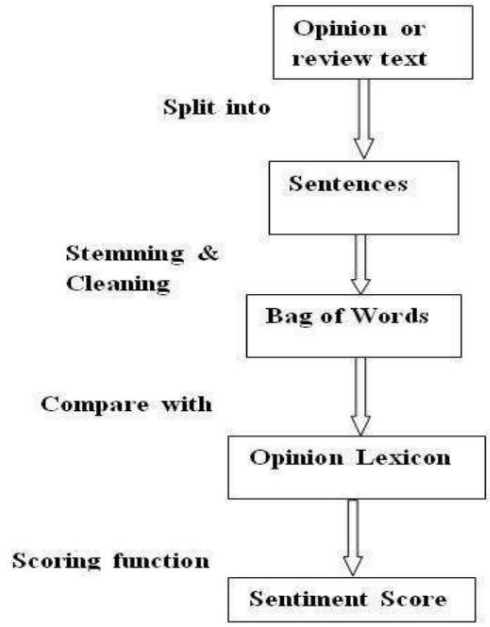

图 3:基于词典的情感分析方法。

4. 提出的方法 用于挖掘 Twitter 微博的方法如图 4 所示。该方法所涉及的步骤描述如下：

1. 数据访问:使用TwitteR包进行关键字搜索来访问twitter消息。

2. 数据清理:使用一些额外的 tm 包获取推文文本,然后清理数据中的停用词（无功能性）,删除空格、标点符号、URL 并执行词干提取（获取词根）。此步骤生成推文的结构化表示,称为术语文档矩阵。

3. 数据分析:上一步生成的结构化表示可用于执行挖掘任务,例如查找关联规则、查找更频繁的术语以及使用基于词典的方法执行情绪分析,该方法使用一组正面和负面的词。评分函数用于为每条推文分配分数。

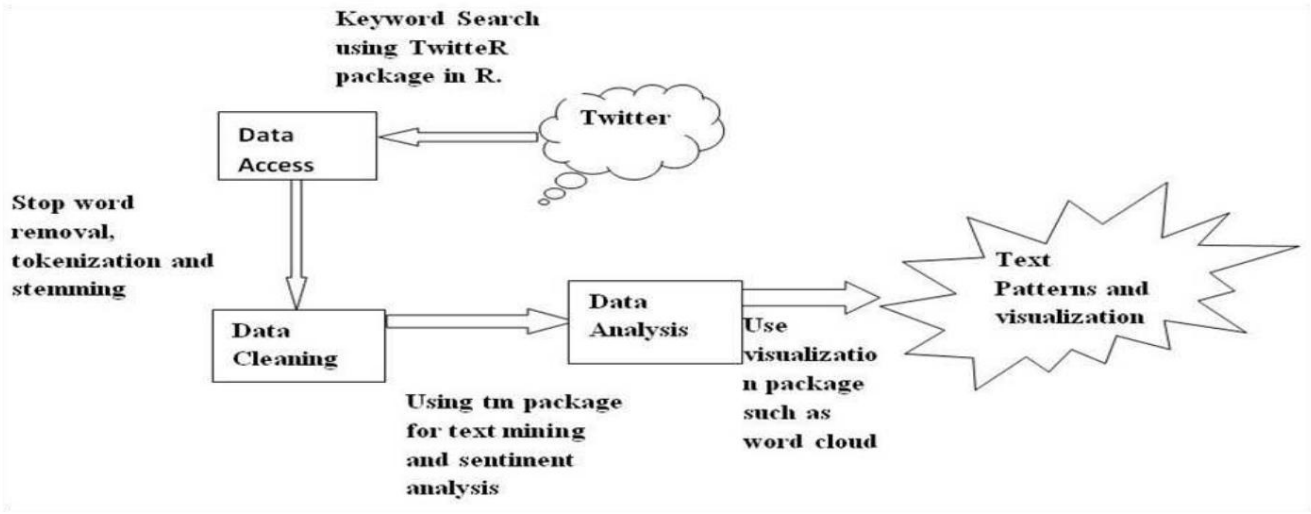

图 4:所提出的 Twitter 微博文本挖掘和情感分析方法。

4. 可视化:wordcloud包[18]和条形图
用于显示客户推文中的词语频率和情感分数。

5.实验与结果实验涉及在 2014 年圣诞节期间收集 Tesco 的 2000 条推文

和 Asda 的 803 条推文。 twitteR [7],R 包已用于访问 Twitter 数据。 该包通过使用关键字搜索查询实现对 Twitter 消息的身份验证和访问 [9]。 获取数据后,使用 tm 包 [19] 提取推文文本并进行清理。 从上一步获得的输出是推文文本的结构化表示,即 tweet-term-matrix。 此结构化表示可用于使用 tm 包执行文本挖掘。 其中一个输出是推文的词云表示。 图 5 和图 6 分别显示了 Tesco 和 Asda 的云。 云中每个术语的大小表示推文中提及该术语的次数,反映了其重要性。 还可以通过特定的出现阈值获得最常见的术语。

此外,还可以计算术语关联。例如,我们可以找到术语“offer”与矩阵其余部分之间的关联。这

显示客户最感兴趣和最不感兴趣的优惠。

与情感分析相关,它是利用图 3 所示的基于词典的方法完成的。在应用文本挖掘后,我们现在有了表示推文的词袋。为了进行情感分析,我们需要意见词典和评分函数来为推文分配分数。我们使用[3]中提供的词典,评分函数如方程 1 所示。

情感得分= Σ 积极词 – Σ 消极词

→ … (公式 1)

如果正面和负面词相等或者文本中不存在任何观点词,则得分可以为0。

(中性的)。

如果正面词汇的数量 > 负面词汇的数量,则分数为正数。(正极性)

如果负面词汇的数量 > 正面词汇的数量,则分数可能为负数。(负极性)

情绪分析的结果如表 1 所示,显示了 Tesco 和 Asda 两家零售商的情绪得分以及具有该得分的推文频率。此外,Tesco 的平均情绪得分为 0.1595,而 Asda 的平均情绪得分为 -

0.00373599,两者的媒体均为 0.0。图 7 和图 8 显示了获得的情绪分数图。

表 2 显示了一些推文示例。很明显,前 4 条推文具有正极性。而最后一条推文具有负极性。

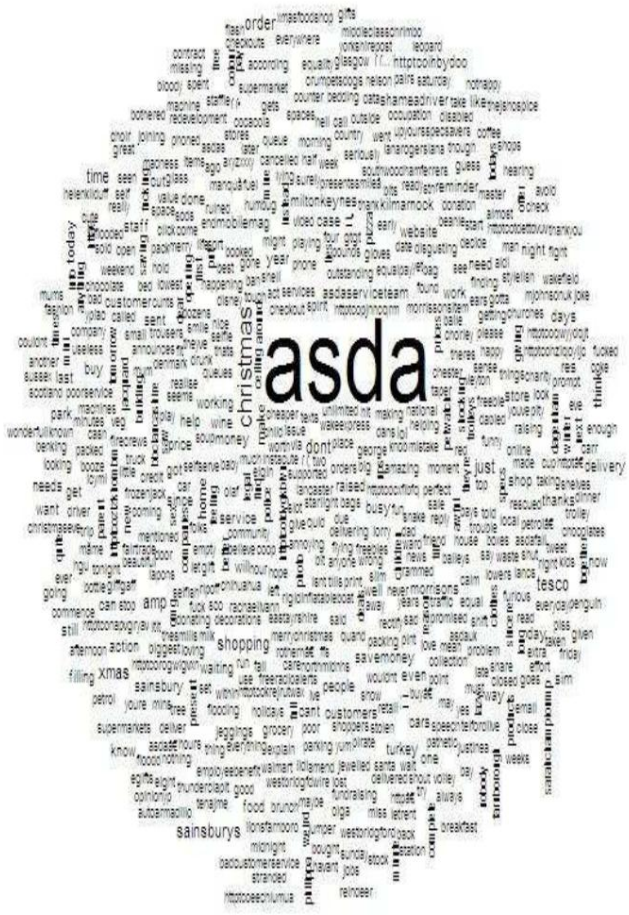

图 5:Asda 推文词云。

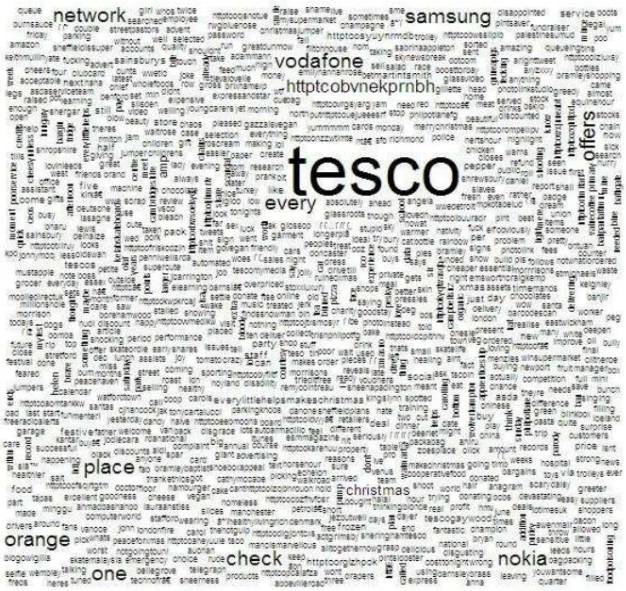

图 6:Tesco 推文词云。

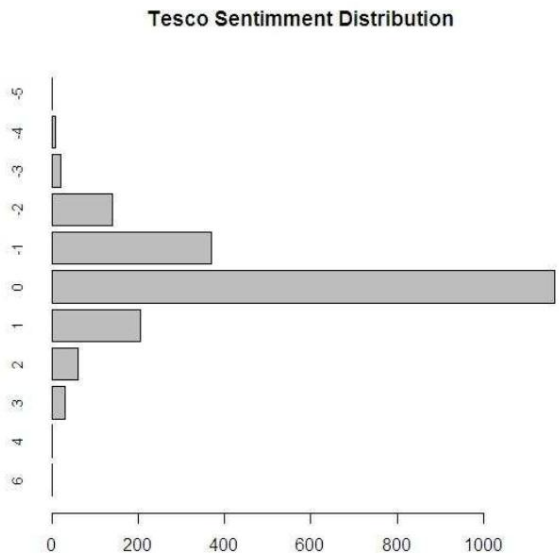

图 7:Tesco 的情绪分数分布。

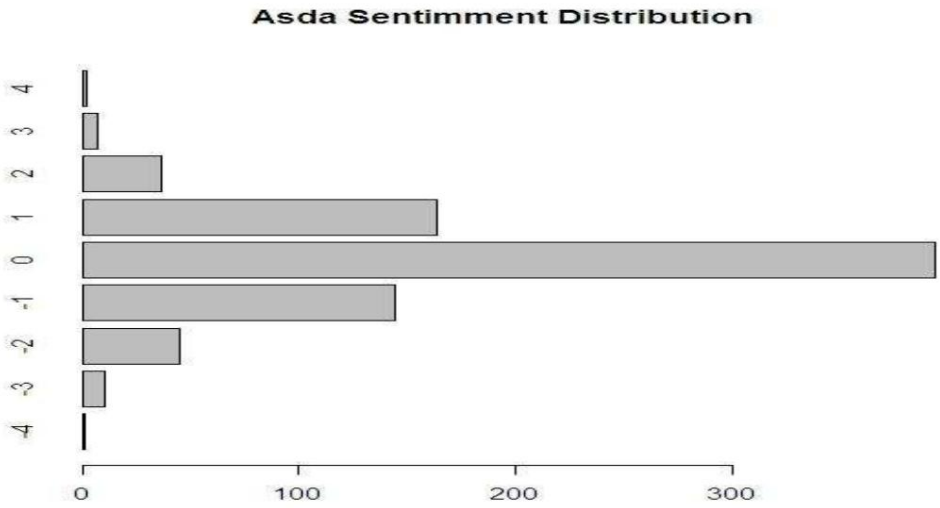

图 8:Asda 推文情绪分布。

表 1:显示 Tesco 和 Asda 数据的情绪极性、分数及其频率。

		消极的					中性 积极					
	情绪 分数	-5 -4	-3 -2	-1			0	1	2	3 4 6		
乐购频率1			1 31	59 205	1166			369 139	21 7 1			
ASDA 频率 0 1				10 45	144 394			164 36		7 2 0		

表 2:显示正向和负向的示例
推文。

“经历了有史以来最糟糕的客户服务体验#tesco Stretford
Tesco 非常感谢您向学校捐款”
“很高兴看到 #Tesco 在地面反击。坚持下去！”
“格兰杰茅斯的阿斯达举办了一场精彩的展示#圣诞节#阿斯达”
“@asda 对你过度收费的那个烦人的时刻#asda”

6.讨论与结论1.本次调查证实,使用R统计软件开源工具包可以收集、预处理、分析和
可视化twitter社交数据。

2. 可以应用文本挖掘任务和情感分析来分析 Twitter 数据,以分析用户对产品或服务的评论。

3. 使用社交媒体数据分析客户对其产品或服务的看法将为商业零售商和服务提供商提供竞争优势。

这将帮助他们提高商业价值并更好地管理客户关系。

4.所描述的方法适用于其他社交媒体数据源,例如Facebook。

可以概括地说,企业可以利用社交媒体跟踪和分析产生的消费者意见,分别调整其营销计划、产品和商业智能。未来工作的一个重要视角可能是建立社交媒体跟踪和监控系统,因为意见随着时间的推移而变化。

此外,在情绪分析和意见挖掘中使用无监督技术对于提高企业竞争价值和客户关系管理也很有价值。除了比较用于意见挖掘的各种情绪分类技术外,社交媒体还可以用作使用意见挖掘进行销售预测的工具。

7. 参考文献[1] Gupta, V., & Lehal,
GS (2009). 文本挖掘技术和应用调查. 网络智能新兴技术杂志, 1(1), 60-76。

[2] Feinerer, I. (2014). tm 包文本简介
俄罗斯采矿业 (nd):n. pag. Web.

[3] 杰弗里·布林,https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/
tree/master/data/opinion-lexicon-English。
访问日期:2014 年 12 月 26 日。

[4] 刘B. (2010)。情感分析和主观性。
自然语言处理手册,2,627-666。

[5] Pak, A. 和 Paroubek, P.（2010 年 5 月）。 Twitter 作为情感分析和意见挖掘的语料库。
在 LREC（第 10 卷,第 1320-1326 页）。

[6] http://www.r-project.org/。访问时间:12 月 26 日
2014年。

[7] http://cran.r-
project.org/web/packages/twitteR/index.html。访问时间:2014 年 12 月 26 日。

[8] http://onemilliontweetmap.com/。访问日期:27 日
2014 年 12 月。

[9] Danneman,N.,&Heimann,R.（2014 年）。社交媒体
与 R. Packt Publishing Ltd 一起采矿

[10] 周鑫,陶鑫,杨建军,杨哲（2013 年 6 月）。
对社交事件推文进行情感分析。计算机支持的协同设计工作 (CSCWD),2013 年
IEEE 第 17 届国际会议（第 557-562 页）。 IEEE。

[11] Milstein, S.、Lorica, B.、Magoulas, R.、Hochmuth, G.、Chowdhury, A. 和 O Reilly, T.
(2008 年)。Twitter 与微型消息革命:沟通、联系和即时性 每次 140 个字符。

奥莱利媒体公司。

[12] Asur, S. 和 Huberman, BA (2010 年 8 月)。利用社交媒体预测未来。在 Web 智能和
智能代理技术 (WI-IAT) 中,2010 年 IEEE/WIC/ACM 国际会议（第 1 卷,第
492-499 页）。IEEE。

[13] Tumasjan, A.、Sprenger, TO、Sandner, PG 和 Welpe, IM (2010 年)。利用 Twitter 预测
选举:140 个字符揭示的政治情绪。

ICWSM,10,178-185。

[14] Taboada, M.、Brooke, J.、Tofiloski, M.、Voll, K. 和 Stede, M. (2011)。基于词典的
情绪分析方法。计算语言学,37(2),267-307。

[15] Zhang, L.、Ghosh, R.、Dekhil, M.、Hsu, M. 和 Liu, B.
（2011 年）。结合基于词典和基于学习的方法进行 Twitter 情绪分析。HP 实验室,
技术报告 HPL-2011,89。

[16] Bhuta, S. 和 Doshi, U.（2014 年 2 月）。 Twitter 数据情绪分析技术综述。智能计
算技术 (ICICT) 的问题和挑战,2014 年国际会议（第 583-591 页）。 IEEE。

[17] Das, TK、Acharjya, DP 和 Patra, MR（2014 年 1 月）。通过分析 Twitter 中的公
共推文来挖掘有关产品的意见。计算机通信与信息学 (ICCCI),2014 年国际会议（第
1-4 页）。 IEEE。

[18] Ian Fellows,http://cran.r-project.org/web/packages/wordcloud/
wordcloud.pdf。

[19] Feinerer, I., & Hornik, K. (2012)。 tm:文本挖掘包。 R 包版本 0.5-7.1。

[20] S. Mohammad (2012)， “从很久以前到幸福快乐的一生:跟踪邮件和书籍中的情
绪”,决策支持系统,53 (2012),第 730 页

741.

[21] 穆斯塔法,MM (2013)。不仅仅是文字:社交网络对消费者品牌情感的文本挖掘。专家系统与
应用,40(10), 4241-4251。