社交媒体健康信息的语义分析：以推特上癌症相关推文为例

韩纲 朱丹 蔡承睿 王文

摘要

本研究报告了对主流微博客社交媒体——推特（Twitter）16天内与癌症相关话题的语义分析。研究共收集了269万余条与癌症有关的推文（tweets），并创建了包含223 条关键词的分类法（taxonomy）。依照推文的频率、周期、同步出现和情绪因素，分析了超过 113万条由该分类筛选的推文并进行可视化呈现。研究结果发现：(1) 可以从推特社交信息中检测到的、有助呈现癌症相关议题的最显见的关键词；(2) 癌症相关推文的“每周两天” 的频率高峰；这种节奏在很大程度上受到突发新闻或新闻事件的影响；(3) 由与乳腺癌、肺癌和前列腺癌相关的推文中的关键词汇的同步呈现构成的语义网络（semantic network），以及 (4) 表达对癌症的积极或消极情绪的情感网络（sentiment network）。同时，本文对研究的潜在理论意义和实际应用进行了讨论。

关键词

社交媒体、推特、健康信息、语义分析、大数据

作者简介

韩纲，美国爱荷华州立大学Greenlee新闻与传播学院副教授，电子邮件：ghan@iastate.edu。

朱丹，美国爱荷华州立大学商学院信息系统系教授。

蔡承睿，美国爱荷华州立大学工程学院博士。

王文，美国爱荷华州立大学人类科学学院人类发展与家庭研究系博士研究生。

Talking about Cancer on Twitter: Health Semantics and Social Media

HAN Gang (Kevin), ZHU Dan, CAI Chengrui, WANG Wen

Abstract

This study reports a semantic analysis of cancer-related conversation in Twitter during a 16-day period. More than 2.69 million tweets related to cancer were collected. Taxonomy consisting

of 223 cancer-related key terms were created and developed. More than 1.13 million tweets filtered with the taxonomy were analyzed and visualized, in terms of the frequency, periodicity, co-occurrence and sentiments. Findings report (1) the most visible keywords, which partially illustrate the topics and message relevant to cancer, detectable from social streaming in Twitter; (2) a two-day-of-week rhythm with frequency of cancer-related tweets, which was highly influenced by breaking news or news events; (3) the key terms co-occurrence in tweets concerning breast cancer, lung cancer and prostate cancer, and (4) a sentiment network that comprises both positive and negative feelings or concerns about cancer. The potential theoretical contributions of this project and its practical implications are also discussed.

Keywords

social media, Twitter, health informatics, semantic analysis, big data

Authors

HAN Gang (Kevin) is an associate professor at Greenlee School of Journalism and Communication, Iowa State University. Email: ghan@iastate.edu.

ZHU Dan is a professor at College of Business in Iowa State University.

CAI Chengrui is a Ph.D. at College of Engineering in Iowa State University.

WANG Wen is a doctoral student at College of Human Sciences in Iowa State University.

在美国，癌症是仅次于心脏病的第二致命的疾病。2008 年全美死亡人数中，有23%与癌症相关 （美国国家卫生统计中心，2012 ）。虽然长期趋势显示，在过去 35 年里癌症总体死亡率持续下降 （美国国家癌症研究院，2013 ），但特定类别癌症的死亡率，如男士的皮肤黑素瘤，以及肝癌、胰腺癌、以及子宫癌等癌症的死亡率，从 2000 年开始不断增长。

与此同时，年龄在18 岁以上的成人癌症病患发生率的总百分比由从2000 年的4.9% 增加到2008 年的5.8%，并在2010 年达到6.3%（国家卫生统计中心，2012 ）。全国性的调查报告还显示有 8%年满 18 岁的成年人曾经从医生或医护保健专业人员那里得知他们患有某类癌症（Schiller，Lucas & Peregoy，2012 ）。

有鉴于癌症对人类健康的威胁和对生活质量的影响，了解人们如何在互联网时代寻找癌症相关信息，已成为健康传播和公共卫生学者感兴趣的主要话题之一（例如，Kim & Kwon，2010 ）。尽管传统上癌症患者的首要信息来源是他们的医疗团队，但越来越多的病人和他们的熟人如今已经转由互联网中获取信息。例如有调查显示，在寻找健康信息的人群中，有47.9%已将互联网作为他们的主要信息来源（Arora，Hesse，Rimer，Viswanath，Clayman & Croyle，2007 ）。尽管学界已有许

多侧重于如何在网上搜索或寻求（seeking）癌症相关信息的研究（例如，Ofran，Paltiel，Pelleg，Rowe & Yom-Tov，2012），但从另一个角度检视何种癌症信息通过互联网获得分享、以及如何被分享的有关研究仍然有限。在社交媒体（social media）或社交网络（Social networking sites，SNS）越来越被广泛使用的今天，这个问题显得尤其引人关注。

本研究以国际主流微博客形态的社交媒体——推特——为例，对癌症相关推文进行探索性的语义分析尝试，并报告初步发现。以期更好地了解社交媒体上癌症相关信息的内容、涵义、语境及其关联。本研究还粗略探讨语义分析对网络化公众通过社交媒体集体建构癌症相关知识的意义，以及对健康传播实践的可能影响。

这项初步研究的重要性或体现在两方面。首先，此前还没有针对基于社交媒体的与健康相关的“社交对话”（social conversations）进行语义分析的系统研究。通过展示经由社交网络平台进行健康知识社会建构的一个侧面，本研究对健康传播文献应有所贡献。其次，本研究试图提供一个辨析推特上卫生医疗保健相关社交信息的语义分析框架，从而有助于我们进一步理解社会媒体在传播和分享健康信息方面的角色。通过探索健康信息学和大数据的交叉领域，本研究所报告的初步发现以及未来对数据的进一步分析将对网络化公众在健康传播中的劝服效果提供更为深入地进行学术探讨的可能性。

一、健康信息的在线寻求

近十多年以来，互联网已成为搜寻以及传播健康信息和医疗保健服务的重要渠道（例如，Koch-Weser，Bradshaw，Gualtieri，& Gallagher，2010）。美国皮尤（Pew）研究中心2011 年 的调查发现，80%的美国互联网用户，即约为 59%的美国成年人，会通过互联网查找有关特定疾病或治疗方法的健康相关信息。Ofran 等研究者（Ofran， Paltiel， Pelleg，Rowe， & Yom-Tovet ，2012）比较了232681位匿名用户和他们的联系人在三个月内使用雅虎搜索引擎查找癌症相关信息的方式。他们发现，查找恶性肿瘤信息的用户搜索时间短，侧重于疾病及治疗信息。而查找惰性肿瘤信息的用户使用搜索引擎的时间更长，搜索主题多样，对疾病信息分享、互助等社会支持团体更感兴趣。在网上查找恶性肿瘤信息时，相比惰性肿瘤信息，病患的熟人比病患本人使用更多时间。此外，Kim和Kwon（2010）的研究利用美国国家癌症研究院的“全美健康信息趋势调查”（HINTS）数据，区分了癌症“电子病

人”（e-patients）和其他癌症病人的不同的信息搜寻行为，并注意到对前者来说，互联网是其首要健康信息来源。

近年来随着各类社交媒体的出现与发展，通过社交网络分享医学和健康信息的用户日见增多（Scanfeld，Scanfeld & Larson，2010）。一个较早的调查发现约有10%的美国成年互联网用户曾在社交网站上访问过或搜寻过健康相关信息（皮尤，2011）。与健康传播中通常关注信息寻求不同，本研究侧重于社交媒体用户直接构建的与癌症相关信息的语义模式的生成。

二、通过社交媒体的信息扩散

“社交媒体”一词描述人们用来相互分享信息、意见、经验和观点的在线工具或平台（Russell，Flora，Strohmaier，Poschko，Perez & Rubens，2011）。至2015 年9月，世界上第一个微博客平台——推特——已拥有近3亿活跃用户；网站月独立访问量 1亿9千万；日均发送推文5800万条（ statisticbrain，2016 ）。在推特上，用户可以通过原创、回复、转发等方式用140个以内的字符交流私人或公众话题。在线社交网络、尤其是微博客的成功，为大规模信息扩散开辟了新的领域。随着新传播渠道的激增和非结构化传播方式的到来，普通用户获得广泛参与为单一或大规模受众直接创制内容的机会。向来为大众传播所主导的一对多的传播模型由此经历了重大的变化（Fogg，2003)。

研究信息如何通过一定的传播渠道获得扩散在包括传播学、社会学、市场营销和流行病学等学科领域都已有相关文献积累（Bakshy，Hofman，Mason & Watts，2011；Rogers，1962)。例如，口碑传播，长期以来一直被认为是信息能够向大规模人群扩散的一种重要机制（Katz & Lazarsfeld，1995）。对影响到越来越多个体受众的传播的个人化过程的了解往往取决于多个学科领域的研究发展，包括人工智能、机器学习、心理学、社会学和传播学。比如，计算机科学近年已有多项研究检视推特这个相对松散的网络如何有助于各类信息向大规模用户的传递和渗透 （例如，Lerman & Ghosh，2010）。

同时，无论是在线上还是线下的口口相传对健康传播中知识和行为的创新扩散也是至关重要的。在这方面，推特等社交媒体提供了刺激“用户生成劝服”（user-generated persuasion）的可能性。而这种经由传播技术强化了的口口相传，透过美国社会各经济阶层人群中社交媒体的快速普及（Chou，Hunt，Beckjord，Moser & Hesse，2009），正在不断改变和重塑着今天的卫生和医疗保健领域（例如，

Hawn，2009；Jain，2009）。

三、推特和公共卫生信息学

随着计算机科学领域中有关推特信息流模式的研究文献的增加（例如，Cha，Haddadi，Benevenuto & Gummadi，2010；Baskshy，Hofman，Mason & Watts，2011），对推特上的健康信息的研究在近年来也得到越来越多的关注（也参见 Paul & Dredze，2011）。例如，Scanfeld等（Scanfeld，Scanfeld & Larson，2010）的研究分析了1000条推文，用以确定推特上分享的健康相关信息的总体类别并探讨人们对抗生素的误解或误用。这个例子同时也显示，健康传播领域中对推特的早期研究尚未引入大数据或数据挖掘的概念和方法。

随着获取和分析社交媒体大数据变得可能，近年更多有关推特的研究开始试图寻找推文信息共享和交换所呈现的总体模式。其中不少研究致力于利用推特进行实时疾病监测或预测健康现象，包括流感疫情（Lampos & Cristianini，2010；Quincey & Kostkova，2010），或疾病传播（Sadilek，Kautz & Silenzio，2012）。例如，Ritterman，Osborne和Klein（2009）使用推特预警甲型流感（H1N1）；而 Chew和Eysenbach（2010）则评估了公众在推特上反映的对于甲型流感的认知。Lampos 和 Cristianini （2010）以及 Culotta（2010）的研究还进一步把人们在推特上对于流感及相关症状的谈论与疾病历史数据建立关联。

就目前涉及社交媒体的健康传播研究来看，尚未触及推特上与癌症有关的话题。社交媒体的赋权，使网络化公众在分享健康知识、经验和情绪感受上得以承担更重要的角色。公众的参与也因此会对随后促进健康信息的知晓、激励健康行为、强化健康行为决策，并保持健康行为的忠诚度方面发挥更大的作用。因此，从语义分析角度研究推特上的信息动态对健康传播具有一定的理论和实践上的意义。

四、社交媒体的语义分析

具有跨学科性质的语义分析既是一种概念，也是一种方法。在语言学中，语义分析是与语法结构相关的一个分析过程，从语言或写作的单位，如短语、从句、句子和段落，到作品整体，然后到文本独立于语言的意义。在机器学习中，对一个语料库（corpus） 的语义分析的任务是建立能够近似表达大批量文档蕴含概念的结构。传统上，语义分析往往是探索性的而并不必要涉及对文档的先验知识。语义分析因此是评估——社交网络中大量信息流的结构和意义、自然语言界面、在线人类

行为、网络传播等——的有用的分析工具。

在计算机科学领域已有若干推特语义分析的研究。例如，对在推特信息中涉及到的与能源有关内容的语义分析（Russell，Flora，Strohmaier，Poschko，Perez & Rubens，2011）。另有一项研究（Saif，He & Alani，2012）结合语义分析和情感分析来测量推特文本中某些概念与消极/积极情绪的相关性。然而目前在健康传播中尚未有关于社交网络中健康相关信息的语义性质和结构的研究。当越来越多的互联网用户使用社交媒体作为信息搜寻和沟通渠道时，研究推特用户如何通过语义网络对与健康有关的知识或信息进行社会化创造和分享就成为了一个有趣的学术课题。

以推特为例，本研究检视社交网络虚拟社区中有关癌症的对话，并结合可视化工具图示推特信息的语义结构。具体而言，本文报告的研究初步发现将回答下面两个研究问题：

研究问题之一：推特上谈论癌症的推文信息出现的频率如何，以及在一定时间段内其周期性是如何变化的?

研究问题之二：这些推文信息是如何通过语义模式连接的，包括关键词同步出现和情感网络结构?

根据分析结果，本研究还探讨了了解语义模式对于通过网络化公众在推特等社交媒体上促进健康传播的意义。

五、方法和发现

为回答以上研究问题，本项目采用语义分析方法检视推特上关于癌症的公共对话（推文），从而区分和突显社交信息流中传递的最重要的词汇、概念和它们之间的关联。

时间框架。鉴于项目研究初期的尝试性，数据采集集中于一段相对较短的时间，从2013年2月26日（周二）至3月13日（周三），共计16 天。

数据/语料采集。本研究使用基于Python的爬虫程序通过推特API实时抓取推特上包含有与癌症有关的关键字词（包括“cancer”和 “#cancer”）的流信息。之后采用MATLAB对数据集进行解析。初步数据集总共包含269万2千286条推文。这些推文再通过MATLAB和 Excel 进行数据筛选和分析，并使用NodeXL对推文中的语义和情感网络进行可视化呈现。

分类法（词库）创建和数据筛选。基于文献回顾，我们创建了关键词（术语）

分类法（词库）来描述与癌症相关的单词、短语、情感和行为。包含那些术语的推文从语料库中被挑选出来后进行进一步分析和可视化。

为了减少在对庞大数据集进行语义分析过程中的干扰，创建描述与癌症有关的目标信息，包括行为、议题、治疗等的分类是研究中的一项关键性任务。分类法整合两种方式来确定数据筛选条件。首先，根据相关的推特研究（例如，Russell，Flora，Strohmaier，Poschko，Perez & Rubens，2011）以及与癌症及健康相关的信息来源（例如，美国癌症协会，2013），我们建立了一个包括与“癌症语言学”八个面向有关的暂定分类，包括：(1) 基础知识；(2) 病因；(3) 检测；(4) 症状和诊断；(5) 治疗；(6) 就医经验；(7) 情感和 (8) 研究。这些类别被细化为若干个包括具体术语的子类别（见附录）。其次，我们选择了在初步数据集之中最频繁出现的词汇，即所有推文中被提及至少 1000次以上的单词。将词频和上述暂定分类相结合作为数据过滤条件。

在数据过滤阶段，根据上述标准暂时列入分类法的搜索词又通过对照在源文档中实际出现的词语进行标准化。事先列入分类法但因未在源文档出现故而不完全一致的关键词则被剔除。由此，最终共有223个关键词或术语用来过滤从推特API中提取的数据集。如表1所示，对过滤后总共113万零627条含有这些关键词的推文进行了语义分析。其中，43万8千590条是原创推文，8万9千180条为回复，60万2千857条为转发推文。

语义分析。我们对抓取的包含癌症相关词语的推文进行的语义分析包括关键词提及频率、周期、同步出现和情感因素四个方面。据此大致描述并识别社交对话中的语义模式。

频率。频率计算的是在此研究时间段内，癌症相关关键词在推文中被提及的数量。附录中在单词或短语旁列出的数字即为其提及频率。同时，图1显示了在这16天期间，八个癌症相关语言类别中最频繁被提及的的关键词，包括死（die）、导致或原因（cause）、诊断（diagnose）、受苦（suffer）、搏斗（fight）、爱（love），不（No），和研究（research）。这些关键词部分体现了用户在推文中谈论癌症时涉及、关注及回应的主题。

周期性。虽然这项研究只涵盖了16天的数据，但通过对推文发布频率的变化的观察将有助于我们对推文对话模式的理解。图2显示在这段时间内的所有包含了与癌症相关关键词的推文，至少呈现了两种有趣的周期性特点。一是“每周两天”的频率节奏（a two-day-of-a-week rhythm）。癌症相关推文数量在每周的某两个工

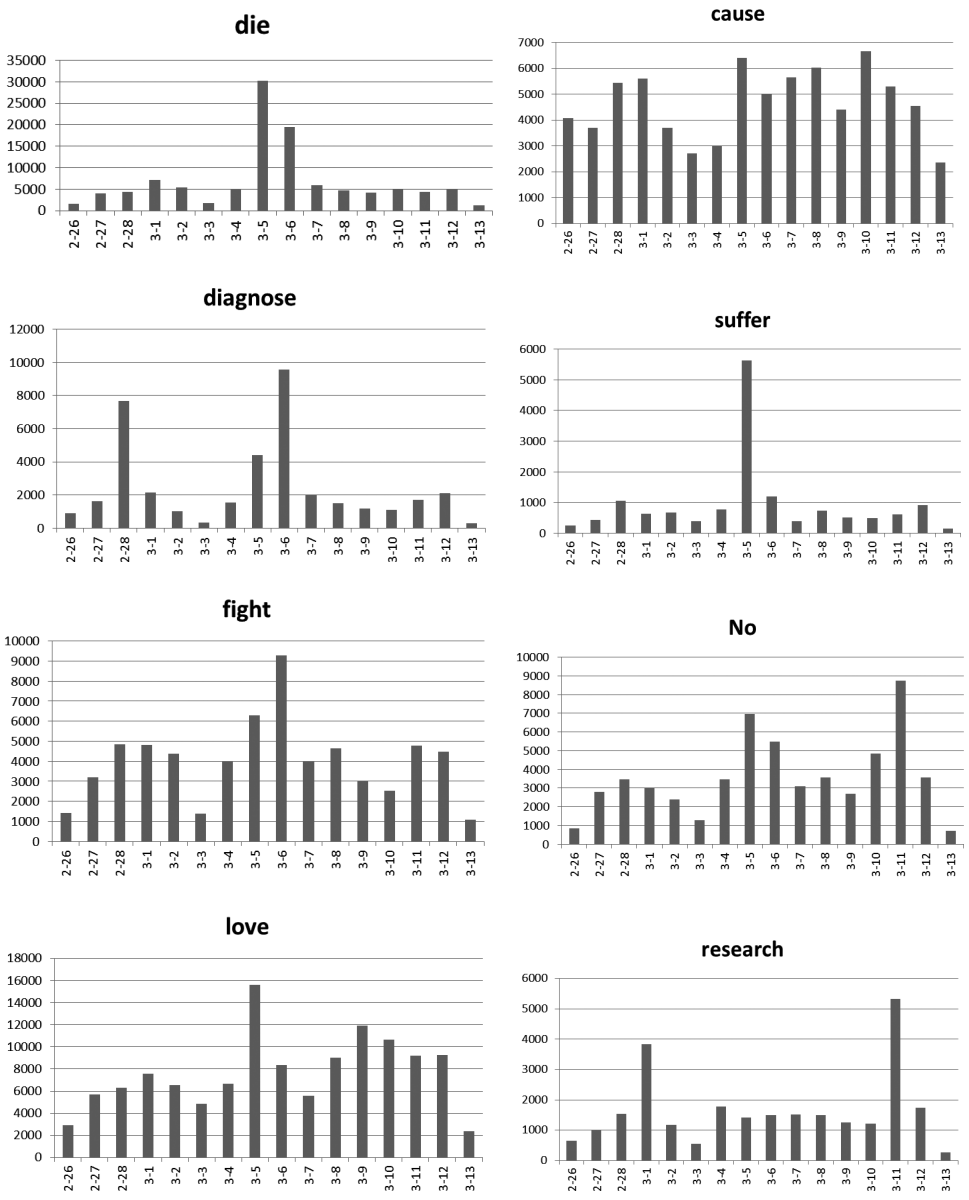

图1：癌症相关语义分类词库各类别最常见关键词每日出现频率

作日会出现高峰，而周六（3月2日和9日）和周日（3月3日和10日）则相对为低点（尽管亦有例外）。这一结果部分呼应先前学者有关推特中能源议题语义分析的研究（例如，Russell，Flora，Strohmaier，Poschko，Perez & Rubens，2011）。二是新

闻事件对推文数量的影响。图2显示，较多的癌症相关推文分别出现于 3月5—6日以及12日。前者出现的原因可能与当时委内瑞拉总统 Hugo Chavez被认为因癌症去世有关。而后者恰逢日本福岛第一核电站核灾难两周年。这两个新闻事件推动了与癌症相关的推文数量的增加。同样地，图1中揭示的与特定关键词有关的推文在此16天中呈现的高峰期也基本和上述两个时段吻合。这种特征意味着突发新闻或有新闻价值的事件可能会成为“引爆”或推动社交网络上健康相关话题的引擎。

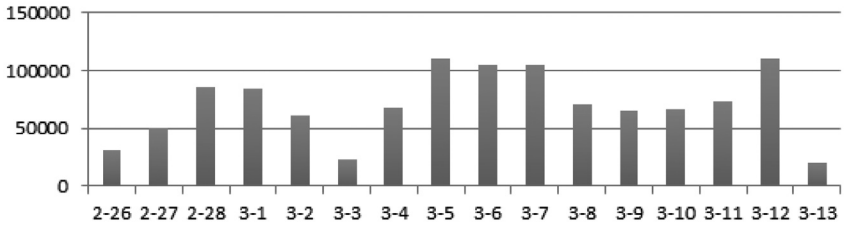

图2: 癌症相关推文在研究日期内每日推文频率
（2013年2月16日至3月13日）

同步出现。图 3 显示了由癌症相关推文中呈现的所有223个关键词所连接的整体语义网络。图4具体到推文中最常提及的特定类型的癌症，包括乳腺癌、前列腺癌和肺癌的关键词之间的联系。我们注意到，乳腺癌是推特上最常被提到或讨论的癌症，涉及185个相关联的关键词。其中与“乳腺癌”（breast cancer）同步出现的最多的九个关键词是吸烟（smoke）、怀孕（pregnancy）、滥用 （abuse）、帮助（help）、诊断（diagnose）、发现（found）、搏斗（fight）、支持 （support）和生活（life）。其次，有 168 关键词与肺癌相关联。抽烟（smoke）、雪茄（cigar）、香烟（cigarettes）、死（die），死亡 （death），原因或导致（cause），治疗（treat）、 希望（hope）和研究（study）是与“肺癌”（lung cancer）同步出现最多的九个关键词。此外，有149个关键词与前列腺癌相关联。其中，加工肉类（processed meat）、减少（reduce）、预防（prevent）、治疗（treat）、疗法（treatment）、帮助（help）、胜过（beat）、生活（life）和乐趣（fun）是与“前列腺癌”（prostate cancer）同步出现最多的九个关键词。同步出现词模式表明人们在推特上谈论的特定癌症与最频繁出现的或与其最相关的对话信息之间有着语义上的紧密联系。

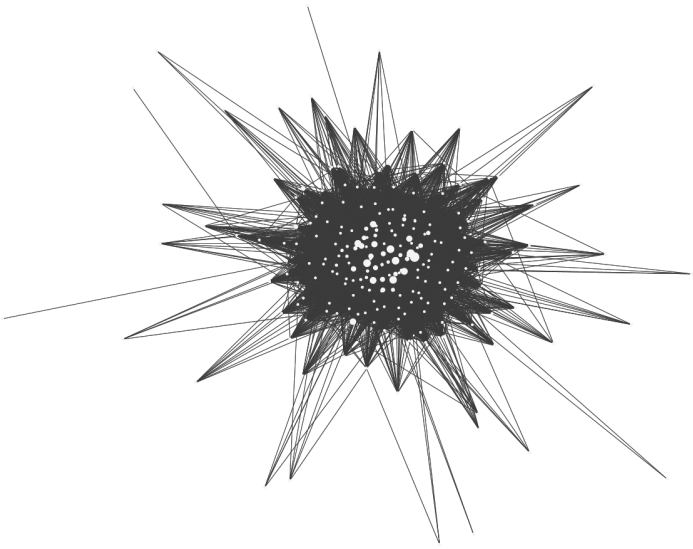

图3：癌症相关关键词的同步出现语义网络

情感。情感展现了一套复杂的以使用词语、句法引用和样式作为指标的语言评价系统（Russell，Flora，Strohmaier，Poschko，Perez & Rubens，2011)。它可以提供与研究主题有关的在线对话和社交连接的重要信息（例如，关于品牌、产品或服务；Pennebaker，Mehl & Niederhoffer，2003）。本研究提供了对癌症相关推文的情感价值的初步观察。图5展示包括积极和消极情绪在内的情感网络。在消极情绪方面，共有六个词语在所有推文中被提及超过2万次：不（No）、不好（bad）、恨（hate）、困难（hard）、疯狂（mad）和悲伤（sad）。这些词语之间的联系比它们与其他词语之间的联系更加紧密；形成情感网络中消极词语的中央组成部分。出现超过1千次但少于2万次的词语，包括严重（serious）、愚蠢（stupid）、恐惧（fear）、抑郁（depressed）、愤怒（angry）、可怕（horrible）、消极（negative）和抱怨（complaining），则围绕在中央组成部分的周围。而其它出现少于1千次的词语则趋于占据网络更边缘的位置。

同时，好（good）、希望（hope）、乐趣（fun）和棒极了（great）则为情感网络中最主要的积极词语，均被提及2万到4万次不等；占据网络中心。其它12个积极词语，如快乐（happy）、美好（nice）、健康（healthy）、勇敢（brave），分别出现1千到1万次，组成网络的下一个圆环。此外，还有其他五个词被提到少于1000次，散落在网络中的边缘地区。

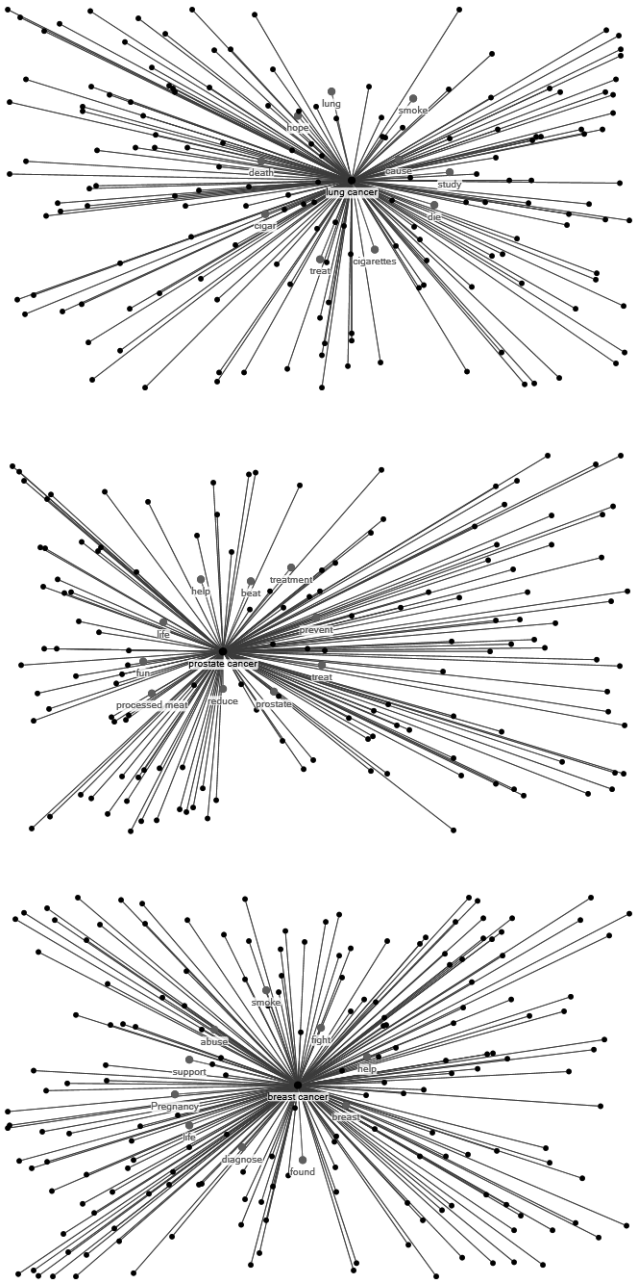

图4：三种主要癌症相关关键词同步出现可视图
（依序为肺癌、前列腺癌、乳腺癌）

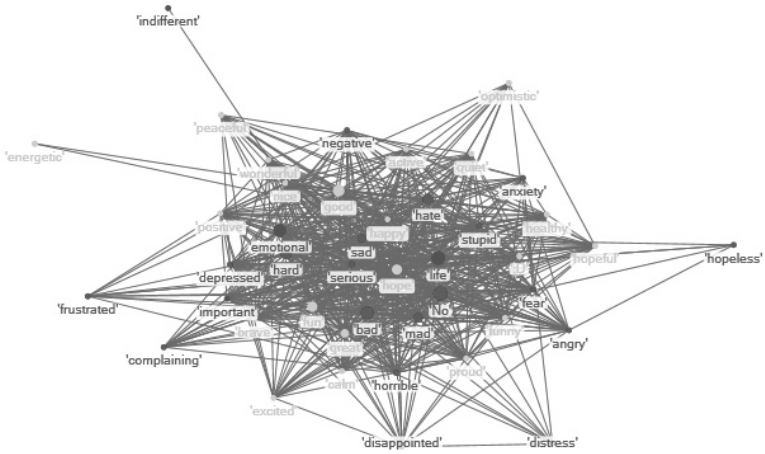

图5：癌症相关语义情感网络

六、讨论

本文报告了对推特上与癌症相关信息的语义分析尝试。本研究收集了超过269万条癌症相关推文；建构了包含223条与癌症相关的关键词分类词库。对经由该分类法筛选的、超过113万条的推文从频率、周期、同步出现和情感等四个方面进行了分析和可视化。

分析结果首先揭示了推特上癌症相关信息中出现的最主要关键词；这些关键词能够部分揭示推特上分享的癌症相关信息的语义特点。其次，分析结果发现了癌症相关推文的“每周两天”的频率高峰；这种节奏在很大程度上受到突发新闻或新闻事件的影响。再次，分析结果显示了与美国三种最常见的癌症，即乳腺癌、肺癌和前列腺癌相关的推文中同步出现的关键词。此外，分析结果还报告了推文中表达的对癌症的积极或消极情绪的情感网络。通过对推特上癌症相关用语呈现的的语义结构以及情感网路的分析，可以更好地理解人们如何使用社交媒体来提供和分享有关癌症的病因、状态、治疗、关切、情绪和后果等信息。

了解社交媒体上健康相关信息的语义性质和模式对健康传播和健康信息学及分析学具有一定理论价值。在传统大众传播和人际传播汇流于网络化平台的今天，“大数据”的兴起为我们研究这种模式提供了可能性，同时也带来了挑战。社交媒体的语义分析探寻文本与大数据以及人文学科与传播科学在互动与数字世界中的交集。利用数据挖掘的研究也能够提供有效检视社交媒体上健康传播信息的分析框

架。也因此，包括本项目在内的此类研究将为未来进一步探讨如何通过社交媒体或社交网络进行健康知识的社会建构奠定基础。

这种努力对医疗保健和健康传播同样具有重要的实践意义。当前对于策略传播者和医疗保健专业人士来说，充分认识到社交媒体在健康信息传播中的作用变得极为必要和迫切。分析和掌握使用社交媒体在扩散公共卫生信息中的利弊，可以使他们通过参与社交网络互动来促进更有效的健康传播。比如从社交媒体中提取和筛选信息的方法有助于从业人员更主动地和有针对性地处理网络化社会的健康风险和应急响应情况（例如，Vance，Howe，& Dellavalle，2009）。与此同时，研究社交媒体中的对话还将有助于专业人士深入了解特定个人有关疾病、保健和治疗措施的想法及其在线交流中的词汇或言语表达特点，从而对需要共享的健康信息进行自定义或个性化，并据此为患者或大众量身定制今后相应的健康促进活动（Chew & Eysenbach，2010）。

七、局限性和未来研究

本研究主要探讨语义分析的四个方面。作为了解推特上健康相关信息的具有探索性质的初始尝试，本文的局限性需要在未来的研究中加以解决。

第一，本研究只涵盖了相对较短的时间周期。尽管关于癌症相关语言如何产生和发展由此可窥一斑，但分析结果有其暂时性，故而无法推而广之到整个推特或其他社交媒体平台作为某种确定性的结论。例如上文报告的推文频率周期性，只总结了特定两周的特点，尚无法概化到其他时段或所有推文的发布特征。未来研究需要涵盖一段更长的合理时间以更加全面地检视社交网站上关于癌症的信息和对话的语义特性。

第二，本研究创建了一个语义分类法或分类词库（taxonomy）。当大量拥有相似关注点或兴趣的用户通过社交媒体集体组织和结构化某种知识或信息时，分类法可以标示出在推特的信息流中可能显现的潜在的语义结构。随着公众在推特上的交流方式可能的变化或更新，今后的研究需要构建更为宽泛的分类法或分类词库。由“社会标签系统”（social tagging systems）（例如，Helic，Strohmaier，Trattner，Muhr，& Lerman，2011）衍生的这种“推特分类词库”（Tweetonomy）（Wagner & Strohmaier，2010）可以在对推特参数，包括共享的网址（ URL）、主题标签（hashtags）、搜索标签（slashtags） 或 “@ 答复”分析的基础上进一步加以更新。由此可以获得更具体的语义结构来建立和确定健康信息的关联及内在含义。

最后，本研究报告的语义网络和情感网络也需要更多的细化分析。这些网络主要是根据同步出现的与癌症相关的推文用语建立的。下一步研究必须通过解释关键词（在网络中的对应节点）之间的关系及它们的意义进行模式解析。同时还需要通过比较来建立和识别关于不同类型的癌症的线上对话的异同点。此外，为了更好地理解推特用户之间的直接交流情景，也应对含有@用户名的推文进行仔细检视。

（责任编辑：李艳艳）

引用文献 [Reference]

American Cancer Society. (2013). Cancer A-Z. Retrieved from http://www.cancer.org/.

Bakshy, E., Hofman, J. M. Mason, W. A. & Watts, D. J. (2011, February). *Everyone's an influencer:quantifying influence on Twitter.* WSDM'11, Hong Kong, China.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, K. P. (2010). *Measuring user influence in Twitter: the million follower fallacy.* Association for the Advancement of Artificial Intelligence. Proceedings of the fourth international AAAI conference on Weblogs and social media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1538/1826

Chew, C. & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 Outbreak. *PLoS ONE,* 5(11), e14118. doi:10.1371/journal.pone.0014118.

Chou, W., Hunt, Y.M., Beckjord, E. B., Moser, R. P. & Hesse, B. W. (2009). Social media use in the United States: implications for health communication. *Journal of Medical Internet Research.* Retrieved from http://www.jmir.org/2009/4/e48/.

Fogg, B. (2003). *Persuasive technology.* Amsterdam: Morgan Kaufmann Publishers.

Hawn, C.,(2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health Affairs,* 28, 361-368.

Helic, D., Strohmaier, M., Trattner, C., Muhr, M. & Lerman, K. (2011, March 28 - April 01). *Pragmatic valuation of folksonomies.* ACM. Proceedings of 20th International World Wide Web Conference (Hyderabad, India).

Jain, S. H. (2009). Practicing medicine in the age of Facebook. *New England Journal of Medicine,* 361(7), 649-651.

Katz, E. & Lazarsfeld, P. (1955). *Personal influence: the part played by people in the flow of mass communications.* New York: The Free Press.

Kim, K. & Kwon, N. (2010). Profile of e-patients: analysis of their cancer information-seeking from a national survey. *Journal of Health Communication,* 15(7), 712-33.

Koch-Weser, S., Bradshaw, Y.S., Gualtieri, L. & Gallagher, S.S. (2010). The internet as a health information source: findings from the 2007 Health Information National Trends Survey and

implications for health communication. *Journal of Health Communication,* 15 (Suppl. 3), 279-293.

Lampos, V. & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In CIP (Ed.), IAPR 2nd Workshop on Cognitive Information Processing.

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages.* In KDD Workshop on Social Media Analytics.

Lerman, K. & Ghosh, R. (2010). *Information contagion: an empirical study of the spread of news on Digg and Twitter social networks.* Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM-10). Retrieved from https://arxiv.org/abs/1003.2664

National Center for Health Statistics. (2012). *Health, United States, 2011: With special feature on socioeconomic status and health.* Hyattsville, MD. Retrieved from https://www.cdc.gov/nchs/data/hus/hus11.pdf

National Cancer Institute (2013). *NCI press release: report to the nation shows U.S. cancer death rates continue to drop; special feature highlights trends in HPV-associated cancers and HPV vaccination coverage levels.* Retrieved from http://www.cancer.gov/newscenter/newsfromnci/2013/ReportNation.

Ofran, Y., Paltiel, O., Pelleg, D., Rowe, J.M. & Yom-Tov, E. (2012) Patterns of information-seeking for cancer on the Internet: an analysis of real world data. *PLoS ONE* 7(9): e45921. doi:10.1371/journal.pone.0045921.

Quincey, E. & Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: the potential of twitter. *Electronic Healthcare.* Springer Berlin Heidelberg (pp. 21-24).

Paul, M.J. & Dredze, M. (2011). *You are what you tweet: analyzing Twitter for public health.* Retrieved from http://www.cs.jhu.edu/~mpaul/files/2011.icwsm.twitter_health.pdf

Pennebaker, J. W., Mehl, M. R. & Niederhoffer, K. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology,* 54, 547-577.

Pew Research Center (2011). *The social life of health information,* 2011. Retrieved from http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx.

Ritterman, J., Osborne, M. & Klein, E. (2009). *Using prediction markets and twitter to predict a swine flu pandemic.* Retrieved from http://homepages.inf.ed.ac.uk/miles/papers/swine09.pdf

Rogers, E. M. (1962). *Diffusion of innovations.* New York, NY: Free Press.

Russell, M. G., Flora, J., Strohmaier, M., Poschko, J., Perez, R. & Rubens, N. (2011, June 3-6). *Semantic analysis of energy-related conversations in social media: A Twitter case study.* Paper presented to the International Conference of Persuasive Technology, Columbus, Ohio.

Saif, H., He, Y. & Alani, H. (2012). *Semantic sentiment analysis of Twitter.* Conference paper, retrieved from iswc2012.semanticweb.org/sites/default/files/76490497.pdf.

Sadilek,A., Kautz, H. & Silenzio, V. (2012). *Predicting disease transmission from geo-tagged micro-blog data.* Association for the Advancement of Artificial Intelligence. Proceedings of the 26th AAAI Conference on Artificial Intelligence. Retrieved from www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/download/4844/5130

Scanfeld, D., Scanfeld, V. & Larson, E. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control,* 38,182-188.

Schiller, J.S., Lucas, J.W. & Peregoy, J.A. (2012). Summary health statistics for U.S. adults: national health interview survey, 2011. National Center for health statistics. *Vital Health Statistics,* 10(256).

Statisticbrain. (2016). Twitter statistics. Retrieved from http://www.statisticbrain.com/twitter-statistics/.

Vance, K., Howe, W. & Dellavalle, R. P. (2009). Social Internet sites as a source of public health information. *Dermatologic Clinics,* 27(2),133-136.

Wagner, C. & Strohmaier, M. (2010). The wisdom in tweetonomies: acquiring latent conceptual structures from social awareness streams. In *Proceedings of the 3rd International Semantic Search Workshop.* ACM,1-10.

附录: 癌症相关语义分类词库及关键词被提及频率

类别	次类别	关键词	频率	关键词	频率
基本信息	什么是癌症	die	109132	life threatening	84
		cell	38241	invade	54
		death	10969	abnormal cell	24
		attack	5671	body cell	3
		killed	1708	grow out of control	1
		affected	1621		
	癌症肇始	colon	8750	HIV	3408
		colon cancer	5379	DNA	592
		growth	4548	anemia	388
		destroy	4227	DNA damage	42
	癌症扩散	pregnancy	25370	metastasis	258
		spread	21085	not contagious	22
		tumor	6596	bloodstream	16
		travel	839	lymph vessels	2
		contagious	334		
	良性肿瘤	tumor	6596	not invade	2
		benign	128	not life threatening	1

	癌症种类	breast	73196	stomach cancer	2852
		breast cancer	66989	bone	2825
		brain	15798	throat	2268
		prostate	14389	liver cancer	1066
		lung	13783	leukemia	656
		prostate cancer	12718	kidney cancer	529
		lung cancer	11271	pancreas cancer	50
		testicular	6850	non-Hodgkin lymphoma	10
		ovarian	5184	pharynx	14
		stomach	3630	rectum cancer	9
癌症起因	烟草	abuse	23551	cigarettes	3714
		smoke	12417	tobacco	1110
		cigar	7106	quit smoking	256
		quit	4313	pipe	160
		drink	4621	beer	527
	酒精	alcohol	1415	drunk	479
		wine	909	liquor	31
	紫外线和日照	sun	6406	sunscreen	385
		UV	3122	sunshine	137
		tanning	3095	sunlight	45
		radiation	1672	X-rays	29
		tanning beds	457	Ultraviolet	7
	放射	cause	74580	nuclear	1629
		Fukushima	8520	exposure	611
		chemical	7304	carcinogen	289
		sun	6406	sunshine	137
		UV	3122	sunlight	45
		tanning	3095	X-rays	29
		burn	2649	Ultraviolet	7
		radiation	1672	expose to	1
	基因	family	22557	DNA	592
		genes	731	inherit	115
		DNA	592		
癌症检测	早期检测	test	24157	detect	5362
		check	5742	early detection	859
	诊断	diagnose	39072	biopsy	274
		judge	2215	diagnostic test	14
		blood test	310		

症状	诊断	suffer	14738	fatigue	324
		pain	6127	fever	251
		symptom	3378	bleeding	135
		weight loss	726	skin changes	1
治疗	疗法	cure	54612	cancer-fighting	1092
		treat	26326	operation	587
		treatment	19852	stem cell	333
		advanced	5137	radiation therapy	112
		therapy	4884	immunotherapy	96
		freeze	4423	bone marrow transplant	49
		surgery	3894	anticancer	36
		chemotherapy	2916	transplantation	26
		medical	2645	hormone therapy	14
	药物	cancer drug	3660	drugs	2101
		medicine	2439		
	饮食	processed meat	2799	whole grain	63
		diet	2505	fresh vegetable	7
		red meat	138		
	疫苗	inocula	12497	vaccination	399
		vaccines	863		
	其他	fight	64117	prevent	14823
		battle	28131	against	12733
		reduce	18224	prevention	2054
		beat	15094		
	副作用	hair	11588	bleeding	135
		pain	6127	dizzy	40
		diabetes	3052	pale skin	20
		hairless	1428	lose hair	10
		infection	1083	tiredness	7
		weight loss	726	shortness of breath	4
		anemia	388	bruising	3
		fatigue	324	skin changes	1
就医经验		love	122623	benefit	6197
		care	65878	afford	4554
		help	52316	nurse	1212
		support	30035	nursing	222
		aid	23007	hospitalize	48
		charity	10236	hospital care	9

	副作用	respect	7728	social support	7
		doctor	6892	outpatient care	3
情感	消极	No	56836	angry	2356
		bad	44771	horrible	2348
		hate	34512	negative	2311
		hard	34152	complaining	1490
		mad	25096	disappointed	863
		sad	20170	anxiety	582
		serious	8679	frustrated	240
		stupid	7691	distress	128
		fear	4191	hopeless	56
		depressed	3584	desperate	5
	积极	good	38994	calm	3640
		hope	30837	positive	3594
		fun	30693	brave	2954
		great	21166	proud	2911
		happy	9384	wonderful	1507
		funny	5307	hopeful	906
		nice	4642	excited	838
		quiet	4394	peaceful	177
		healthy	4053	optimistic	104
		active	4010	energetic	65
		:D	3744		
	其他	life	56488	important	6634
		emotional	49142	indifferent	650
研究		research	26250	fundraiser	2667
		fund	17633	finding	2427
		found	14776	scientists	1729
		study	14249	technology	1479
		lab	6551	researcher	1382