软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.005261]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向微博主题的可视分析研究*

王臻皇[1,2], 陈思明[1,2], 袁晓如[1,2]

[1](机器感知与智能教育部重点实验室(北京大学),北京 北京 100871)

[2](北京大学 信息科学技术学院,北京 北京 100871)

通讯作者: 袁晓如, E-mail: xiaoru.yuan@pku.edu.cn

摘 要: 随着微博的发展,微博的影响力日益增大,对微博主题内容进行分析具有重要的价值.主题模型技术能够从文本数据中提取主题,但由于微博文本短、随意性大、信息量小等特点,微博主题的分析具有一定难度.本文提出了一个微博主题可视分析系统,利用多种互相关联的视图与丰富的交互手段,支持用户对主题模型结果进行分析与探索.系统结合了微博数据的特点,引入微博用户与时间因素,支持分析者从多角度对微博主题进行全面分析.系统支持用户在主题可视分析的基础上,通过交互操作对主题进行编辑,从而改进主题模型,提高模型的准确性和可靠性.案例分析结果表明本文提出的系统可以有效地帮助用户分析微博主题和修正主题.

关键词: 文本可视化;主题模型;微博主题分析;微博可视化;微博可视分析

中图法分类号: TP311

中文引用格式: 王臻皇,陈思明,袁晓如.面向微博主题的可视分析研究.软件学报. http://www.jos.org.cn/1000-9825/5261.htm

英文引用格式: Wang Z, Chen S, Yuan X. Visual Analysis for Microblog Topic Modeling. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/5261.htm

Visual Analysis for Microblog Topic Modeling

WANG Zhen-Huang[1,2], CHEN Si-Ming[1,2], YUAN Xiao-Ru[1,2]

[1](Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China)

[2](School of Electoronic Engineering and Computer Science, Peking University, Beijing 100871, China)

Abstract: With the development of social media (e.g. Microblog), the impact of social media increases. It's critical to analyze the topic of the microblog. Topic Modeling can extract topics from text data. However, it's challenging to do so for the microblog data, due to the short content, heavy noises and limited information amount in each microblog message. We propose a Microblog Topic Modeling Visual Analytics System. The proposed system enables the visual exploration and analysis process of the Topic Modeling results of Microblogs with multiple linked views and interactions. We consider user behaviors and time effects in the topic modeling process. Users can analyze topics of microblog from multiple perspectives. Our system supports interactive topic editing to improve the topic modeling results in accuracy and reliability. The case study confirms that our system can effectively help users analyze the Sina Weibo contents interactively.

Key words: text visual analytics; topic modeling; microblog topic analysis; microblog visualization

网络时代带来了海量的文本数据.网页、报告、文档都包含了大量的文本数据,涉及政治、文化、经济、生活等方方面面.计算机技术可以帮助人们自动化地处理文档,过滤噪声信息,提取有价值信息.文本主题分析技术可以高效地在海量的文本数据中发现主题信息,将文档数据集中的主要内容简洁地提供给用户[1].

将文本主题分析技术应用于微博数据具有很重要的意义.微博如今已成为了人类网络生活中一个重要组

* 基金项目: 国家自然科学基金面上项目(61672055);国家"九七三"重点基础研究发展规划项目基金(2015CB352503).
Fund items: NSFC General Project(61672055); the National Program on Key Basic Research Project (973 Program)(2015CB352503)
收稿时间: 2016-10-11; 修改时间: 2016-11-25; 采用时间: 2017-01-16; jos 在线出版时间: 2017-07-12
CNKI 网络优先出版: 2017-07-12 15:33:07, http://kns.cnki.net/kcms/detail/11.2560.TP.20170712.1533.001.html

成部分,人们在微博平台上发布自己的所见所闻,所思所想.微博平台上存储着海量的文本数据,并且每天都在以惊人的速度增长.微博文本内容包含了大量的信息,从中挖掘主题信息可以用来进行突发事件监测、事件发展态势预测、精准营销等.目前,文本主题分析技术应用于新闻、文章等长文本数据上已经具有了较好的效果.但是,微博文本长度短,通常被限制在 140 字以下(除少量长微博以外),每条微博的信息量较小,人们在微博上的文字使用较为随意,这都大大增加了微博文本处理的难度.因此,对海量微博文本进行主题分析是一个具有挑战性的课题.主题模型是一种当前流行的主题分析技术,其代表算法为 LDA (Latent Dirichlet Allocation),最初由 Blei 等人于 2003 年提出[1].LDA 模型建立了三层结构:文档-主题-词语,基本思想是每个文档是主题的混合,文档到主题服从一个概率分布,而同时主题也是词典的一个概率分布.然而,主题模型的三层结构在带来信息表达效果提升的同时也增加了人们对模型的理解和深入分析模型结果的难度.因此,我们在研究中将可视化与可视分析技术与主题模型结合,提供有效的交互手段,可以让人们充分参与到分析主题模型的结果中来,利用人的认知能力从数据中挖掘有效信息.此外,微博数据除了文本信息外,还有时间、微博发布者等其他属性,可以辅助人们进行分析.通过交互技术,人们可以在分析的基础上对模型进行一定程度的修正,提高模型的准确性.

本文提出了一个面向微博数据的交互式主题模型可视分析系统,通过交互的可视分析技术,帮助分析者达到如下目的:

(1) 快速感知文档集的主题概况,并基于主题模型的结果分析文档-主题、主题-词语的关系.
(2) 结合微博数据的时间属性和用户属性,分析主题的时变趋势及微博用户的主题偏好.
(3) 在对微博主题分析进行分析和探索的基础上,支持用户通过交互对主题模型进行进一步改进,从而提高模型准确率.

本文的结构如下:第一节介绍相关研究工作,第二节讨论面向微博数据的主题模型可视分析系统的设计与交互,第三节就数据获取与系统实现进行深入描述,第四节通过案例分析来展示系统的可用性和有效性,最后进行总结并对未来的研究方向进行展望.

1 相关工作

本文的研究工作涉及微博可视分析、主题模型技术及主题模型可视分析等多个方面.因此,本节将从基于微博的可视分析研究、主题模型及其在微博中的应用、基于主题模型的可视分析三个角度来总结已有的相关研究工作.

1.1 基于微博的可视分析研究

自从 Twitter 于 2006 年正式上线面向公众开放以来,微博(Microblog)就受到了广泛关注,并以其简洁性、实时性、社交性等特点在短时间内吸引大量用户.随着微博的发展,如今它对现实社会也产生着越来越大的影响.虽然每条微博的内容只有短短不到 140 字的长度,但是微博平台上每天数亿条的新增微博中潜藏着海量的信息与巨大的价值.诸多国内外学者开展了大量的相关研究工作.目前,已经有不少研究工作验证了海量的微博信息中蕴藏着巨大的价值.Aron Culotta 追踪 Twitter 平台上与流感相关的消息建立起了一个流感预测模型[2],Takeshi Sakaki 等人通过 Twitter 进行地震监测[3],Lotan 等人通过分析突尼斯、埃及革命期间的 Twitter 信息传播,显示了微博信息如何影响人的社会和政治生活[4].Hu 等人也证明了微博在突发事件的传播上比传统的媒体更具有优势[5].Ren 等人提出了 WeiboEvents,针对任意一条微博,可以快速可视化并允许用户探索它的所有转发,对新闻、传媒、舆情领域提供了重要的帮助[6].

时间信息在社交网络分析中一直扮演着重要角色.社交网络中事件的发生、发展到结束总是在时间维度上展开的,时间线可以帮助人们分析事件的发展过程.Eddi Starbird 等人[7]利用 Twitter 上的信息,通过时间线来了解红河洪灾的发展.Marcus 等人[8]的工作是基于 Twitter 数据流进行可视分析,通过时间轴进行事件趋势的可视分析,并通过峰值检测算法发现事件高峰,从而对事件发生过程中的重要时间点进行标注.除了二维的时间轴和热度曲线,三维空间可以展示更丰富的信息,Itoh 等人[9]利用了 3D 可视化,展示了用户的活动和兴趣随时间的变

化.伴随着移动互联网的发展,社交网络中的地理信息也愈加丰富.提取微博数据中的地理信息并加以分析和呈现,可以建立起事件地理分布模型.社交网络的实时性使得事件地理分布模型可以比传统媒体更加快速地感知到事件的发生和发展,为突发事件的应对争取到宝贵的时间.TwitInfo[8]将微博在地图上表示出来,辅以时间轴及其他信息,支持多角度地对微博事件进行全面探索.ScatterBlogs 及其后续的工作[10,11,12]提供了一个基于时空的交互可视分析系统,它在 Twitter 中带有地理信息的微博的数据基础上进行语义上的主题分析和空间上的聚类分析,以支持异常事件的发现与探索分析.Chen 等人[13]对微博上的人群移动模式进行了研究,利用可视分析进行移动特征提取.

社交关系是社交网络的基础,信息的传播和互动都是建立在这些社交关系的基础上.错综复杂的社交关系组成了复杂的网络,然而网络结构本身十分抽象,不易于理解和分析,因此借助可视化方法直观呈现网络结构是一种有效的解决方法.点边图(Node-link Diagram)是最为常见的网络可视化方法,点表示网络中的个体,边表示个体间的关系,用其他特征(如颜色、大小、形状)等来编码其他维度的信息.为了使图排布美观、易于分析,出现了很多基于点线图的网络结构布局算法[14,15]. 然而,随着网络规模的增大,网络结构也愈加复杂,简单的布局方法会产生混乱的网络图形,不利于人直观的理解,从而降低了可视化的价值.因此,突出网络的某些特征能够有效地提高可视化的效果.Vizster[16]着重体现社交网络中的连接性和群聚性.Van Ham 和 Van Wijk[17]则注重交互性,提出了一种 Focus+context 的网络可视化方法,使用户在探索局部结构时还能保持全局的视图.HiMap[18]就是针对层次性的社交网络关系进行的.除了点线图,邻接矩阵是另一种表示网络结构的可视化方法,如 Matrix Zoom[19]和 MatrixExplorer[20].邻接矩阵虽然避免了点的重叠和边的交错问题,但在直观性上不如点边图,同时可视化大规模网络时需要耗费大量空间.NodeTrix[21] 融合了点边图和邻接矩阵的特性,在密度较高的团块部分使用邻接矩阵表示,而在团块之间使用点边图进行连接,利用了两种方法的优点,大量减少了点的覆盖和边的交叉,又有效地利用了空间.进一步地,Chen 等人将微博的转发关系及参与者投影到抽象紧致的二维空间,利用基于地图的隐喻展示了用户之间的转发关系[22].基于社交关系,微博的内容在人与人之间相互扩散,主题不断动态演变.本文着重关注于微博的文本内容,考虑参与用户以及主题的动态变化,进行深入地探索.

1.2 主题模型及其在微博中的应用

LDA 是一种无监督式、对文本文档集主题信息进行识别的一种机器学习技术[1].在结构上,LDA 是一种多层的贝叶斯模型,包含了文档、主题、词语三个层次.在 LDA 模型中,每一个主题是基于词的概率分布,而同时每一个文档也是基于主题的概率分布.模型的基本思想是每个文档是多个主题的混合,文档服从主题的一个概率分布,而同时每个主题也是词典中词的一个概率分布.给定一个包含 M 篇文档的文档集,词典长度为 V,事先设定的主题数为 K,LDA 将会生成 θ 与 β 这两个矩阵.其中,θ 表示了文档在主题上的分布,由 M 行 K 列构成,$\theta_{i,j}$ 表示了文档 i 在主题 j 的概率分布. β 为 K*V 的矩阵,$\beta_{i,j}$ 表示了词 i 在主题 j 中概率分布,$\sum_{i=0}^{V}\beta_{i,j}=1$ 矩阵需要满足 $\sum_{i=0}^{V}\beta_{i,j}=1$.

LDA 模型在提出后被广泛应用于文本分析领域的应用和研究,并出现了许多 LDA 模型的变种.Blei 等人又提出了 Correlated Topic Model (CTM) [23],将 LDA 中的狄利克雷分布改为对数正态分布 (Logistic-Normal Distribution),解决传统的 LDA 无法处理主题间相关性的缺点.一部分 LDA 的变种使得主题模型具有层次结构,如 HLDA[24]和 HDP[25].另一方面,标准的 LDA 是一种无监督学习机制,然而实际应用中,一些数据会包含有标签信息,如人工标注的标签、分类信息、用户标签等,有监督的学习机制将能够更加充分地利用这些信息,得到更好的结果.有监督的主题模型变种有 sLDA[26]、PLDA[27]、Labeled LDA[28]等.而另一部分主题模型的变种既支持层次结构,也属于有监督学习机制,如 hLLDA[29], HSLDA[30].另外,如 DTM(Dynamic Topic Models) [31]在主题模型中加入了时间因素,将主题与时间两个维度结合起来.

作为一种新兴的社交网络形式,微博的影响力日益增大,提取微博主题的需求也与日俱增.然而,微博与普通的文档集合有着许多不同之处:内容短、主题分散、语言随意等,这些特点使得一些适合于传统文档集(如论文、新闻等)的主题模型算法应用于微博文本主题提取时效果不佳.因此,一些可以被应用于微博数据的 LDA 模型变种主要是通过增加微博数据中其他非文本结构的数据以达到增强信息量、提高效果的目的.将作者信息加入到主题模型的工作主要有 Atuhor-Topic LDA[32]和 Twitter LDA[33].Author-Topic LDA (ATM)形成了作者-主题层面的分布,每个作者都对应着一个主题上的分布,而同时在传统的 LDA 算法中存在的文档-主题的分布在 ATM 中并不存在.通过 ATM 模型可以对用户的兴趣点进行描述,但该模型缺少了用作者-主题分布取代了文档-主题分布,因此从 ATM 模型并不能获取到微博与主题之间的关系.Twitter LDA 则是兼顾了作者-主题和文档-主题两种分布,因此 Twitter LDA 模型在回答文档与主题之间的关系时比 ATM 更加有效.另外,Ramage 等人[34]就将 Labeled-LDA[39]的方法应用到微博上.这个工作是将微博的 hashtag(hashtag 是一种用户在发布微博时自主标示主题内容的标签,在新浪微博中通常以“#hashtag”标识)作为标签信息,作用到 Labeled-LDA 模型中.然而,由于新浪微博中带有 hashtag 标签的微博数只占很小的比例,这种方法的适用性并不强[35].

已有的工作对微博主题的探索受限于主题模型本身,用户难以依据现有知识的判断控制其模型的行为.本文提出通过可视分析和交互技术来分析和修正主题模型,提高主题模型的准确率,可以充分结合计算机高速的计算能力和人的分析能力,有更强的灵活性.

1.3 基于主题模型的可视分析

基于主题模型中文档-主题的分布,可以通过如 KL 散度、Hellinger 距离等方式度量文档之间的相似性,再利用力导向布局[36,37]、t-SNE[38]等算法将其排布在二维平面上,从而能够直观地观察文档和主题的分布.

主题模型在原本自然的词语-文档结构中插入“主题”一层,增加了人们理解主题模型的结果的难度.一些工作利用主题模型生成的文档-主题和主题-词语的分布进行了可视化,力求解释这两个分布的关系.如 Termite[39]使用表格的形式展示关键词与主题之间的分布关系,用圆圈的大小编码关键词在主题中的概率分布大小,支持对主题内容、关键词分布进行分析,同时支持进行主题间的比较.Serendip[40]用不同的视图分别提供文档在主题上的概率分布,以及词语在各主题上的排名,使得用户可以在较高层次上了解和探索整个文档集,同时也提供了文本阅读视图来让用户可以在更细节的层次上进行浏览和分析.

文档集的主题和聚类通常会带有层次特征,例如体育相关的话题会包含足球、篮球等子话题,娱乐相关的话题也可能包含电影、音乐、电视剧等子话题.Hierarchicaltopics[41]在 Parallel Topic Modeling 的基础上,使用 Topic Rose Tree 算法构建具有树形层次结构的主题信息,通过树状图和树形时间轴展示主题的层次信息以及主题在时间上的分布.另外,Cui 等人[42]则研究了层次结构的主题随时间的演变.

不同主题模型的比较也具有重要的意义.同样的文档集数据,当设置了不同的主题个数、使用了不同的参数,甚至不同次运行之间由于随机性,会生成不同的主题模型.Alexander 和 Gleicher[43]通过矩阵和二部图的可视化形式,帮助人们理解不同模型之间的匹配性,发现两个模型中相匹配和不匹配(缺失或合并/分裂)的主题,从而对不同的模型进行比较.

通过主题模型生成的结果有时并不能令人满意,此时人们就希望能够让人参与到主题模型的计算过程中,通过交互的手段来干预模型的计算以提高模型的准确率,也即及交互式聚类.一般来说,交互式聚类可视分析首先需要让用户能够感知和分析当前聚类的结果,通过探索发现聚类结果中的不足之处,并且通过交互支持用户对局部进行编辑操作,并且通过重新运行聚类算法来更新聚类结果.例如,iVisClustering[44]在主题模型的基础上提供了多种视图,如平行坐标、点边图、柱状图等,使用户可以从多个不同的角度探索 LDA 的结果.其中,散点图揭示了文档的分布,以及不同主题在平面上的分布,平行坐标对应于每个文档在主题上的概率分布.用户可以通过修改系统修改 LDA 的参数、过滤噪音文档、主题分割、主题合并等操作对主题模型计算结果的简单干预.UTOPIAN[38] 使用 NMF(非负矩阵分解)来获得文档-主题、主题-单词的分布,以支持对主题模型的探索和编辑..

本文的主题模型可视分析系统与 iVisClustering 和 UTOPIAN 有共同的目标,即通过可视化视图分析主题,并通过交互对主题进行编辑.但本文提出的系统使用了不同的可视化设计来帮助人们分析主题模型的结果,包括提供了主题矩阵、新颖的主题地图以及时间线结合的方式.同时我们根据微博数据的特点,引入了时间和微博用户因素,帮助人们更全面地分析数据.

2 面向微博的主题模型可视分析系统设计

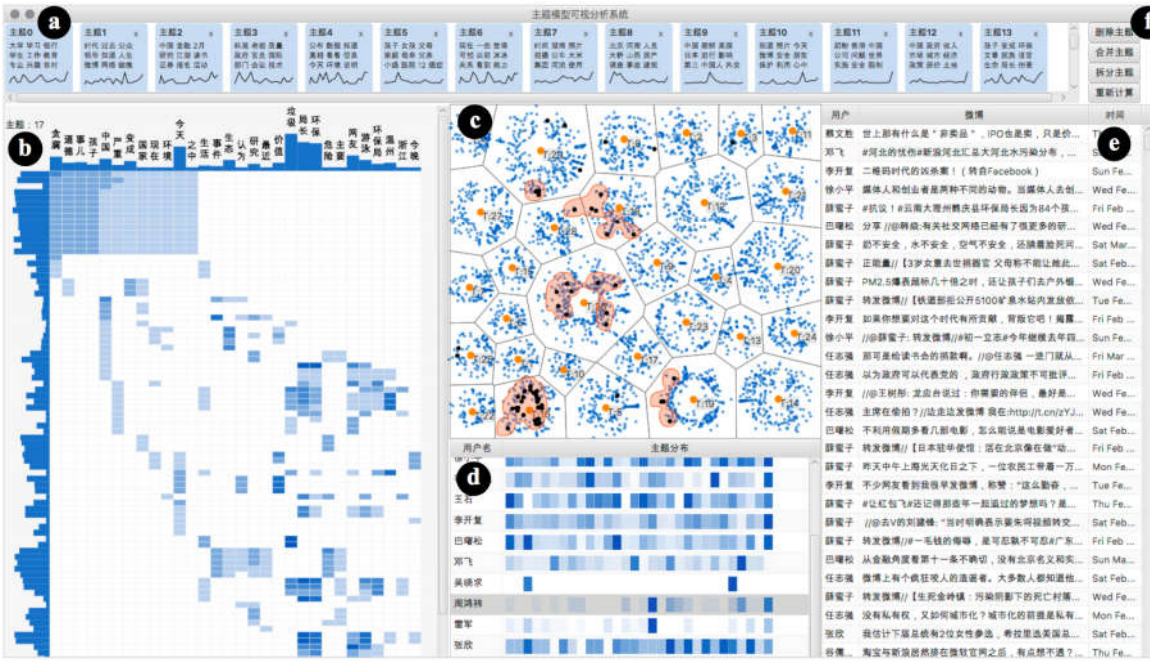

Fig.1 Topic modeling visual analytics system for microblog data. It includes six views, (a) Topic List; (b) Topic Matrix, (c) Topic Map, (d) User List, (e) Detail List, (f) Tool Box.

图 1 面向微博数据的主题模型可视分析系统,系统包含 6 个视图:(a)主题列表,(b)主题矩阵,(c)主题地图,(d)用户列表,(e)微博列表,(f)工具箱.

本节主要介绍面向微博的主题模型可视分析系统,该系统提供了相互关联的可视化视图,提供丰富的交互手段,让用户对由主题模型提取的微博主题结果进行分析与探索.

系统结合了微博数据的特点,引入微博用户因素与时间因素,使分析者可以从更多角度对微博主题进行全面的分析.此外,在深入了解主题内容和主题特征的基础上,用户可以对主题模型进行编辑改进,从而提高模型准确性.系统的界面如图 1 所示,系统主要包含 6 个视图:主题列表(a),主题矩阵(b),主题地图(c),用户列表(d),微博列表(e),工具箱(f).本节将介绍系统的视图设计和交互设计.

2.1 主题矩阵

主题矩阵对一个主题中关键词与微博内容的关系进行可视化.在主题模型中,主题是基于词的分布,我们可以通过观察主题中的关键词的分布来大致地猜测主题内容,判断主题模型结果的可靠性.通常情况下,我们只能依靠经验知识进行判断,但经验知识并不能让人对主题内容进行完全准确理解和判断.如在某个主题中“苹果”、“手机”、“平板”、“天空”等关键词拥有较高的权重,我们或许能够通过前三个关键词进行联想推断出该主题是

在讨论苹果公司的产品,但对“天空”一词我们便难以推测它与该主题的关系.这是因为我们在理解主题内容的过程中缺少了词语的来源——文档内容的参与.主题矩阵的设计目标是通过可视化的手段将词语-主题-文档联系起来,使用户能够细致地查看每个主题内部的文档具体内容,从文档内容的角度分析关键词在主题中具有高权重的原因,并进一步分析关键词之间的关系.

如图 1b 所示,整个主题矩阵表示一个主题,在主题矩阵中,每一行代表一条微博,每一列表示一个关键词,行列交叉处的方格用颜色表示对应的微博中对应关键词出现的次数,深色表示这条微博出现多次对应的关键词.矩阵上方的直方图展示了对应关键词的概率分布值,矩阵左侧的直方图表示的是对应微博的概率值.通过主题矩阵,我们可以对分析主题中微博与关键词的关系.

矩阵的行和列可以简单地依照对应微博和对应关键词的概率值大小进行排序的,但这种排序方式生成的主题矩阵会显得较为杂乱.我们对主题矩阵的行列进行了重新排列,使得关联度较高的微博和关键词被聚集在一起,形成聚类效果,让用户能够更轻松地分析主题中的内容以及微博与关键词的分布关系.矩阵重排列算法采用 Rank Order Clustering (ROC)算法[45,46],通过循环迭代地对行和列进行重新排序,可以使矩阵中关联性较强的元素聚集起来.

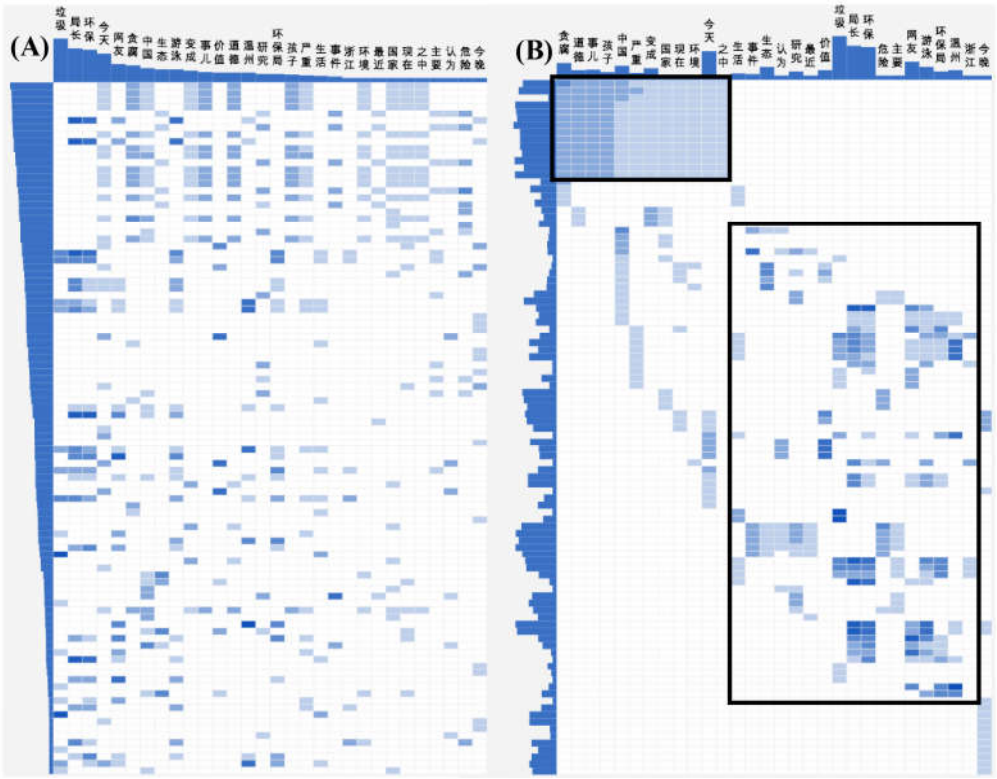

Fig. 2 Comparison of the topic matrix before and after the sorting operation, (A) is the topic matrix before sorting, while (B) is the one after sorting.

图 2 主题矩阵重排列前后对比,(A)为对矩阵行列重排序前的主题矩阵,(B)为对矩阵行列进行重排序后的主题矩阵.

图 2 是矩阵重排列前后的对比效果.经过重排列,我们发现在该主题中人们实际上是在谈论两个话题(在图中被黑色框标出).从关键词的分布可以推测出左侧话题主要是在讨论中国贪腐和道德问题,而右侧话题主要在

谈论网友邀请环保局长下河游泳的事件,并引申讨论环保问题.

从上述示例可以看出,经过重排后的主题矩阵可以帮助用户分析主题的内聚性.有时由于主题模型主题数设置不当等原因,不同内容的主题会被合并在一个主题中,在修正的过程中需要被拆分.而有时是因为多个内聚性较高的话题经过松散的连接组成了一个内聚性并不高的主题.当人们对主题内聚性要求较低时,这多个话题可以被认为是属于同一个主题.但若人们对主题内聚性要求较高时,则需要将这个主题拆分成多个主题.

2.2 主题地图

本节我们着重介绍主题地图的设计原理与生成步骤.

2.2.1 设计原理

上节中提到的主题矩阵主要针对单个主题进行分析,可以让用户了解主题的内容和质量,但是它很难有效帮助用户快速了解整个数据集的概况.因此,我们提供了一个视图来让用户感知和探索整个数据集,了解文档和主题的分布.

利用散点图在二维平面上展示文档的分布是一种常用的可视化手段.散点图将每个文档表示成二维平面上的一个点,相似的文档被摆放在靠近的位置上,因此点的分布一定程度上代表了数据集中文档语义的分布.然而,这样的方式只利用了文档的信息,而不能充分表现主题模型中主题的特性.为了充分表征文档的相似性以及在不同主题上的分布,我们设计了基于地图隐喻的主题地图可视化方法.基于地图的可视化隐喻在时空可视化中较为常用,它允许人们探索地理空间中实体的分布、数量情况.在此,我们利用地图的隐喻来对微博的主题与相关微博进行探索.在主题地图中,我们需要探索主题分布以及其相关微博的分布,了解相似主题的分布以及直观感知各个主题的微博数量并进行比较.具体而言,主题地图需要满足如下需求:

(1) 每个主题拥有一块独自的区域,所有属于某一主题的微博需要全部位于该主题区域内.

(2) 主题区域的大小与主题内微博的数量正相关,他们的相对位置应能体现出主题之间的相似程度.

(3) 主题内微博的位置需要能够体现它与本主题的相关程度.

针对以上原则,我们利用实线将平面划分为若干多边形区域,每个区域代表一个主题.文档用点表示,点所在的区域即代表了对应文档所属的主题,并且所有属于该主题的微博都被放置在该区域内.同时,多边形区域的面积与主题内微博的数量相关,多边形的相对位置体现了对应主题之间的相似程度.

主题地图的效果如图 1 (c)所示.为了便于用户对主题进行识别,我们在每个主题多边形的中心点位置绘制了橙色圆点,并标示出主题序号.图中,每一个多边形区域代表一个主题,区域内的每个点代表属于该主题的一条微博.区域中心橙色的圆点代表了主题中心点的位置,并标示出了主题序号以便于用户识别和分析.

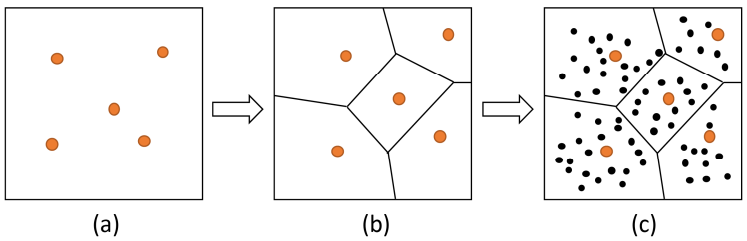

Fig. 3 The illustration of topic map layout process: (1) Calculate the centric point of each topic, (2) Segment the topic region, (3) Layout the weibo nodes in each topic region.

图 3 主题地图排布过程示意图:(1)计算主题中心点位置,(2)划分主题区域,(3)在主题区域内排布微博节点.

2.2.2 主题地图生成

为了让主题地图满足上述提出的三个设计目标,我们采用一种层次性的主题地图生成方法.生成主题地图可以分为以下三个步骤:

(1) 计算主题中心点位置

在主题地图中,主题区域的位置需要体现主题之间的相关程度,越相似的主题应该被放在越相近的位置.因此在生成主题地图时,我们需要首先定位每个主题的理想位置.力导向布局算法可以在给定各主题间的距离时计算出每个主题中心点的目标位置.

基于主题模型的结果,主题间的距离可以使用 KL 散度(Kullback-Leibler divergence)进行度量:

$$D_{KL}(i,j)=\sum_{v}\beta_{i,v}ln\frac{\beta_{i,v}}{\beta_{j,v}}$$

其中, $D_{KL}(i,j)$ 为第 i 个和第 j 个主题间的距离, $\beta_{i,v}$ 表示词典中的第 v 个词在主题 i 中的概率值.由于 KL 散度并不是对称的,即 $D_{KL}(i,j)\neq D_{KL}(j,i)$,为了方便,主题间的距离可以定义为:

$$D(i,j)=\frac{1}{2}(D_{KL}(i,j)+D_{KL}(j,i))$$

(2) 划分主题区域

在已知各个主题的大致位置的情况下,我们需要为每个主题划分出各自的主题区域,该区域的面积大小与对应主题中微博的数量成正比.这里我们采用 Voronoi Treemaps[47,48],它将树图(TreeMap)与 Voronoi 图结合起来,在给定中心点理想位置及各区域面积比例的条件下,划分出子区域,使得子区域的中心点尽可能地在给定中心点的附近,且面积比例尽可能与给定的面积比例相近.此外,所有的子区域都具有相近的长宽比,使得整体分布更加美观.

(3) 主题区域内排布微博点

在这一步的排布中,主要考虑微博与所有主题之间的关系.对概率分布越大的主题,微博与其中心的距离越近.因此在主题地图中,通过微博的位置大致可以表现出微博在本主题的概率分布大小,以及该微博还在哪些主题上具有较高的概率分布.

2.2.3 用户活跃主题区域

用户是发布微博信息的主体,用户个体的兴趣、思想、职业等因素将会影响到用户所关注的微博主题.在微博主题分析中,分析用户关注的主题可以帮助我们发现用户的兴趣和关注点,辅助人们推断主题模型结果的准确性.例如,某用户的主要活动区域出现在两个主题中,则可能的情况有:(1)该用户有两个不同的关注点;(2)这两个主题之间有一定的相关性,通过进一步分析主题的内容,若发现这两个主题的相关性足够强时,应该将它们合并为一个主题.

在主题地图中,我们将用户所发布微博的集中区域的轮廓标示出来,来凸显用户的活跃主题区域,帮助分析者更容易地发现和观察用户所关心的主题,如图 1 (c)所示.在绘制活跃区域轮廓时,我们首先需要在主题地图中提取用户发布微博较为集中的区域,过滤掉零散的点.通常情况下,用户除了关心他所关注的主题外,偶尔也会发布一些与其他主题相关的微博,这些零散的微博并不能代表用户的兴趣点,有时反而会成为噪音点影响人们的分析.之后,我们利用 Bubble Sets[49]方法来绘制提取出的集中区域的轮廓.

在提取集中区域时,我们采用 OPTICS (Ordering Points to identify the clustering structure)聚类算法,它是一种基于密度的聚类算法.OPTICS 可以帮助我们在主题地图上找到用户微博点密度较高的区域,并且它能够发现任意形状的簇,不会使区域轮廓趋于圆形.同时,一些零散的微博点由于区域密度较低,会被视为噪音点被忽略,达到了过滤噪音点的目的.

2.3 其他视图

其他视图主要用于辅助用户的交互探索,提供细节信息.

2.3.1 主题列表

主题列表中每一个“卡片”代表了一个主题(图 1a),卡片上包含三部分信息:主题编号、关键词表、时间线.关键词表中列出了若干该主题中具有高分布概率的关键词,按照概率值由大到小排列,可以让使用者在短时间内对每个主题的内容有一个大致了解.时间线展示了该主题中微博发布时间的分布情况.通过时间线,使用者可以研究不同主题的时间特性,进而探索分析主题内容的变化.当用户点击主题列表上的某一个主题卡片时,主题矩阵将会切换显示对应主题的内容.

2.3.2 用户列表

用户列表列出了数据集中的所有微博用户(图 1d),每一行代表一位微博用户,第一列为用户昵称,第二列则包含了与主题数相同数量的矩形方格,用方格的颜色来展示用户对不同主题的参与程度,参与程度可以通过该用户所发布的微博中属于该主题的微博数比例进行度量.方格颜色越深代表该用户较多地关注和参与了对应的主题.在用户列表中选择某一用户,该用户的主题活跃区域轮廓将会在主题地图中被标示出来.

2.3.3 微博列表

微博列表列出了数据集所包含的所有微博,让使用者可以分析每一条微博的具体内容(图 1e).

2.3.4 工具箱

工具箱提供了“删除主题”、“合并主题”、“拆分主题”和“重新计算”等功能(图 1f),前三个功能是用于对主题进行修改操作,最后一个是应用先前对主题所作的所有修改操作,重新运行 LDA 算法获得新的主题结果.

2.4 对主题模型的修正

在 LDA 主题模型中,是一个 K*V 的矩阵,$\beta_{i,j}$ 表示词 i 在主题 j 上的分布值.LDA 模型可以在给定 β 矩阵的情况下重新进行采样,计算文档在主题上的分布.因此,通过修改 β 矩阵可以实现对矩阵模型结果的修正.用户可以通过四种方式实现对主题模型结果的修正:修改词的权重、删除主题、合并主题、拆分主题.

对于每一种修改操作,β 矩阵的值都将发生改变,下面将具体阐述在不同的操作对 β 的修改方法.假设 β' 为修改后的矩阵.

2.4.1 修改词的权重

假设用户对第 p 个词的权重进行了修改,修改后的权重为 β'_{ir},则其余词的权重修改为:

$$\triangle f = \beta'_{pj} - \beta_{pj}$$

$$\beta'_{ij} = \beta_{ij} + \triangle f * \frac{\beta_{ij}}{1-\beta_{pj}}, \quad i \neq p$$

这样的处理方法即将被修改关键词的权重差按比例分配到其它关键词上.

2.4.2 删除主题

假设用户删除了第 p 个主题,由于 β 矩阵仅需满足 $\sum_{i=0}^{V} \beta_{i,j} = 1$,即同一主题的关键词分布概率加和为 1,但对同一关键词在不同主题下的分布概率加和无要求,因此可以直接删除 β 矩阵的第 p 列,这样新的 β 矩阵

β' 将是一个 V*(K-1)的矩阵.

2.4.3 拆分主题

假设用户在第 p 个主题中选择了部分词 W 和部分文档 D,拆分出第 q 个主题.假设在界面中可供用户选择的词集为 S.定义

$$f_s=\sum_{i\in S}\beta_{ip};\qquad f_{s-w}=\sum_{i\in(S-W)}\beta_{ip};\qquad f_w=\sum_{i\in W}\beta_{ip};$$

则在新的主题 p 和 q 中,

$$\beta'_{iq}=\begin{cases}0,\ if\ i\in(S-W)\\ \beta_{ip}\dfrac{f_s}{f_w},\ if\ i\in W\\ \beta_{ip},\ if\ i\notin S\end{cases}\qquad \beta'_{ip}=\begin{cases}0,\ if\ i\in W\\ \beta_{ip}\dfrac{f_s}{f_{s-w}},\ if\ i\in(S-W)\\ \beta_{ip},\ if\ i\notin S\end{cases}$$

2.4.4 .合并主题

假设用户合并了第 p 和第 q 个主题,新主题为第 r 个主题.合并主题只要将关键词在两个主题下的分布概率进行合并即可,即:

$$\beta'_{ir}=\beta_{ip}+\beta_{iq},\ i=0,1,2,\ldots,K$$

2.5 交互操作

交互是可视化与可视分析的一个重要组成部分.简单的静态可视化只能传达出有限的信息,只有通过交互操作让用户主动地探索和分析数据,才能让用户进一步地了解数据,充分发挥人的分析能力,做出准确有效的分析和决策.本系统为用户提供了丰富的交互操作.

2.5.1 鼠标悬停

当用户将鼠标悬停在主题矩阵上方表示关键词分布的直方图上时,将会提示对应的关键词及其分布权重,鼠标悬停在主题矩阵左侧的代表微博分布概率的直方图上时,对应的微博内容也会被提示.另外,鼠标悬停在矩阵中的每一个矩形方格时,系统会提示对应的微博内容以及对应的关键词.在主题地图上,鼠标悬停于微博点上时,也会显示微博的具体内容.同时,鼠标悬停在主题地图的微博点,主题矩阵中代表对应微博的矩形方格将会被高亮,反之,鼠标悬停于主题矩阵中的某一方格时,主题地图上代表对应微博的点也会被高亮出来.当鼠标悬停在主题列表的某一主题上时,主题地图上的对应主题区域会被高亮提示.如此,通过交互操作让主题列表、主题矩阵和主题地图联系起来,增强了互动性,让用户可以在多种视图中切换探索,进行深入的分析.

2.5.2 选择操作

主题列表粗略地展示了每个主题的内容信息和时间信息,通过鼠标点击选择某一主题,主题矩阵将会切换显示对应主题的内容.当用户选择微博用户列表中的某一用户时,对应用户的活跃主题局域的轮廓将在主题地图上被标示出来.当用户选择微博列表中的某一微博时,主题矩阵中对应的方格及主题地图上代表对应微博的点也会被高亮显示.

2.5.3 修正主题

系统支持用户对主题模型的结果进行修正,共支持四种修正操作:修改关键词权重,删除主题,合并主题和拆分主题.主题矩阵上方的直方图显示了对应主题下高分布概率关键词的权重,用户可以在对应关键词的直方图上进行修改.在主题列表上,每一个主题“卡片”的右上角都有一个“X”标志,点击该标志可以删除对应的主题.在主题矩阵中,用户可以圈选一块区域,这块区域中的关键词将从原主题中被分割出来形成新的主题(即拆分主题).用户可以在主题列表上选中两个主题进行合并.此外,在工具栏中也提供了三个按钮,支持对主题的三种操

作.点击工具箱中的“重新计算”将执行之前的所有对主题模型结果的修改操作,并重新运行 LDA 主题模型,得到新的结果.

3 数据获取与系统实现

本案例的研究对象为新浪微博.新浪微博提供了丰富的 API 接口,让研究者和开发者可以轻松地获取到微博数据,进行进一步研究和开发.利用 statuses/user_timeline 接口,可以获取到指定用户的微博.微博数据存储在 MongoDB[50]中.MongoDB 是一种非关系数据库,每条记录被称为“文档”,以键值对的方式存储,是一种类似于 JSON 的 BSON 格式,它可以支持非常松散的数据结构,字段无需进行事先定义.由于微博 API 的数据可以以 JSON 的格式返回,因此我们可以非常方便地将微博 API 返回的数据完整地存储进数据库中,且因 MongoDB 无需事先定义字段,数据库可以不受 API 返回数据结构变化的影响.

微博的文本内容出现在 text 字段和 retweeted_status 中的 text 字段.其中,text 字段表示该微博的文本信息,retweeted_status 存储了被转发的原微博信息(仅该微博为转发微博时有效).由于很多用户在转发微博时不再输入文字来表达自己的想法,或仅输入少量意义较小的文字,因此转发微博的信息量有很大一部分是蕴藏在被转发的原微博中的.为了丰富文本信息量,在预处理阶段若该微博为转发微博,则首先将 text 字段的文本和 retweeted_status 中的 text 字段的文本合并,以增大文本信息量.

在文本预处理中,我们首先使用 NLPIR/ICTCLAS 工具[51]对中文文本进行分词,之后进行词语过滤.词语过滤部分包括两个步骤:停用词过滤和表情过滤.停用词过滤可以去除如“不仅”、“而且”、“另外”等使用频率高但无实际含义的词;表情过滤是指去除微博文本中含有的表情,表情符号在微博文本中是以“[表情]”记录的,因此使用简单的正则匹配即可进行过滤.经过上述预处理后,微博文本会被处理成由若干个有效词语.若一条微博包含的有效词语过少,则它所含的信息量也很少,就会成为噪声,干扰主题模型的效果.因此,在预处理阶段,需要过滤有效词语少的微博.按照经验测试,本文中案例设置的过滤阈值设为 3.用户也可根据具体需要调节过滤阈值.

可视分析系统使用 Java 实现.系统利用 JavaFX 进行可视化界面和交互的实现.JavaFX 能够直接调用 Java API,同时可以利用类似 CSS 的形式定义界面样式,提高了界面设计和实现的效率.

4 案例分析

本系统对经过 LDA 算法提取主题后微博内容进行了可视化展示,支持用户从微博文本、用户、时间等不同角度对微博主题进行分析,支持使用者对主题模型的结果进行修正.通过交互式的探索和修改,得到一个更为合理的主题结果.本节利用案例来介绍用户使用系统进行微博主题模型可视分析的方法和过程.

4.1 数据描述与主题提取

在案例中,数据集包含 15 个用户从 2013 年 2 月 1 日至 3 月 3 日间发布的所有微博,共 7218 条微博.这 15 个用户为任志强、潘石屹、徐小平、王石、李开复、陈志武、巴曙松、邓飞、张欣、周鸿祎、蔡文胜、雷军、谷儒李晨、薛蛮子、吴晓求.他们大多都为新浪微博平台上较为重要的意见领袖,关注热点话题并发表评论.由于他们很少发布无太多意义的微博,因此他们的微博具有相对较大的信息量,不会使数据集充满噪音点.使用我们的系统,用户可以快速地交互探索出大量数据中蕴含的语义,并进一步对主题理解、修正与探索.

考虑到我们分析了 15 人共 7218 条微博,每个人都是在自己特定领域较有影响力,因此话题数量理应不少于 15.话题的设定粒度太粗或太细都会对探索产生相应的影响,我们做了初步的尝试,设定了话题数量为 15、30、45、60 等参数,综合考虑实际效果中主题划分的合理性以及用户的可感知范围(主题数量过大造成认知负担以及主题过于稀疏),初始状态下我们设定了 30 个话题.正是由于初始划分由于参数的不同会产生不同的主题划分,我们允许用户进一步地对主题进行评估、合并、拆分与删除.

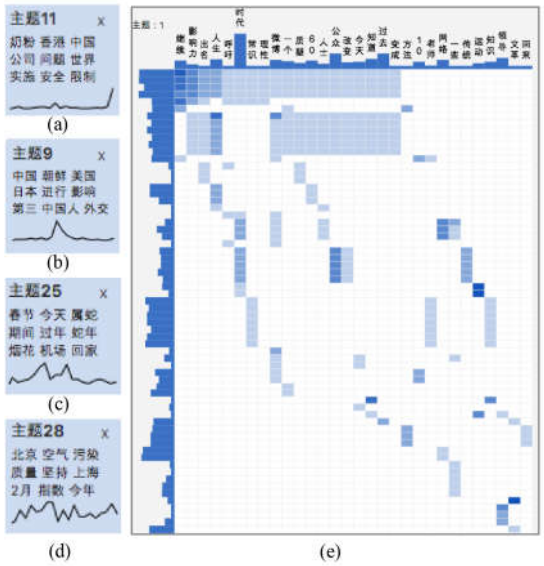

Fig. 4 The result of case study, (a)-(d) are Topic 11, Topic 9, Topic 25 and Topic 28, respectively. (e) is the topic matrix of Topic 1.

图 4 系统案例分析结果,(a) - (d)分别为主题 11、主题 9、主题 25、主题 28 的主题卡片,(e)为主题 1 的主题矩阵.

4.2 主题可视分析

利用系统进行主题可视分析时,可以综合利用多个视图,从主题的内容和时间分布、用户主题分布等角度对主题进行从整体到局部的探索,其中通过主题列表、主题地图对主题的内容及相似性进行分析,进一步探索主题的时变特征.基于以上总体的探索,用户针对感兴趣的主题,可以进一步进行主题内部关键词与微博分布并进行质量分析;针对感兴趣的人,也可以进一步探索他们的主题分布与感兴趣的话题所在.其中对用户、主题、微博以及关键词的探索是相互链接,在系统中联动变化的.

4.2.1 主题内容、分布与时变特征分析

用户首先会对主题的总体特征进行总览式探索,包括主题列表中的内容与时间探索,以及主题地图中的分布与相似性分析.主题列表列举了所有主题及每个主题较为重要的关键词,以及该主题中微博的时间分布.用户通过观察每个主题“卡片”可以大致地对每个主题的主要内容有初步的了解,如主题 21 中具有较高权重的关键词是“国家”、“官员”、“财产”、“信息”、“公开”、“代表”、“地方”等,可以猜测该主题谈论的是官员的财产信息公开问题.又如主题 29 中,“公司”、“互联网”、“产品”、“企业”、“微信”、“投资”、“创业”等具有较高的权重,我们也不难推测该主题讨论的是互联网企业的创业与互联网产品问题.当然,有些主题内容通过关键词并不容易做出推测判断,例如主题 15 的关键词为“第一”、“喜欢”、“分享”、“时间”、“建议”、“今天”、“将军”和“研究”等,这就需要我们进一步深入到主题中去进行探究.

主题卡片上的时间线可以帮助人们对主题的时间分布有大致了解.图 4 (a-d)展示了 4 个主题卡片.主题 11 主要谈论奶粉限购令问题.从时间线上可以看出,该话题在 2013 年 2 月底时突然爆发(香港于 2013 年 3 月 1 日起实施奶粉限购);主题 9 谈论中朝关系问题,而在 2013 年 2 月 12 日,朝鲜进行了第三次核试验,引发网友的讨论;主题 25 在谈论春节的事情,其主要高峰位于 2 月中旬左右,与春节时间吻合;主题 28 谈论了北京的空气污染问题,在 2 月到 3 月初之间不断地呈现高峰,其高峰时间与北京空气污染较为严重的时间同样较吻合.

我们通过对主题卡片上的主题选择,可以高亮出主题地图上的相应区域与微博,感知主题的数量以及相关性.例如我们可以观察发现主题 12、14、19、26 和 29 的占比较大,即微博热度比较大.如以上的分析所示,主题 29 主要关于互联网企业相关的话题讨论,该话题较为火热的原因之一在于我们选择的名人有许多互联网的从业者,以及与互联网创业相关的投资人.更进一步,我们可以探索主题之间的相关性.在主题地图中,主题区域的距离与其对应主题之间的语义距离相关,因此我们需要特别注意互相接壤的主题区域,它们通常在语义距离上比较相近,有可能是同一主题.例如,主题 18 与主题 26 相互接壤,并且有部分微博十分接近两主题区域的边界.我们通过观察两个主题内容,发现主题 18 是关于企业将污水排入地下的新闻报道,而主题 26 是中国的地下水污染,两者相关性十分大,可以将它们合并为同一个主题.

4.2.2 主题细节探索与质量分析

基于对主题的内容、分布以及时变特征的总览性分析,我们可以对具体的主题内容进行深入的探索以及质量分析.主题矩阵可以帮助人们更深入地分析主题.用户可以在主题列表、地图以及时变视图中选择相应的主题,主题矩阵的视图会被更新.例如,我们选择了主题 1(如图 4e)与主题 17(如图 1b).对主题 1 进行分析,发现大部分高权重关键词都被几条微博占据,而剩余的几个高权重关键词也都分布在不同的微博中,它们之间没有相关性.我们对左上角的聚集区域进行进行选择,在用户列表中可以观察发布微博中包含这些关键词的用户,我们发现这些微博其实来自同一作者,他多次转发了自己的某一条关于个人感悟的微博,因此这些微博在文本上具有十分大的重复性,从而形成了一个聚类但并未在微博上形成主流话题,因此该主题应被删去.

对主题 17 的探索包含了两个步骤.在默认的基于关键词权重排序的情况下,我们较难看出该主题的特征分布(如图 2a),用户通过对主题的排序,观察出主题 17 包含比较明显的两个分块,左上角与右下角.进一步地,我们通过刷选这两部分区域的内容,相关的微博在微博内容视图中被高亮选中,我们归纳出左上角谈论的是当前中国的贪腐问题(关键词:贪腐等),右下角则是关于微博网友邀请环保局长下河游泳的事件(关键词:局长、环保、网友、游泳等).我们通过主题矩阵的分析,我们认为该主题应被拆分为 2 个主题.

4.2.3 用户主题分布分析

通过对主题内容分布的了解,我们进一步可以针对特定的人探索其参与的主题.首先针对用户参与主题的列表,可以选择具体的用户,主题地图会相应地高亮出该用户参与讨论主题的轮廓.我们可以利用主题地图进一步研究作者的主题活跃区域,图 5 分别展示了周鸿祎、雷军、李开复所关注的话题.

从图中可以看出雷军的活跃区域主要集中在主题 16 和主题 11(图 5c),周鸿祎则主要集中在主题 16,主题 26,主题 18,主题 29 等(图 5b).两人共同的话题 16,主要讨论的是 360 手机和小米手机,这正好是两人所领导的企业的手机产品.此外,周鸿祎所关注的主题 18 和主题 26 都是关于中国地下水污染问题(上文已提到这两个主题应被合并),而主题 29 是关于互联网公司产品与创业的问题,这与周鸿祎的个人背景相吻合.

相比于上述两人,李开复在微博上更加活跃,他关注的话题也更广泛.图 5d 可以看出,李开复关注最多的为主题 29(关于互联网公司、互联网产品及互联网创业),该主题与李开复的个人职业和背景相符.另外,李开复还关注了一些诸如主题 11(香港奶粉限购)、主题 21(官员财产公开)、主题 20(中国社会改革)、主题 9(朝鲜核试验)等,这些话题都是当时的社会热点话题,表明了李开复经常在人们关注的热点事件上发表意见.

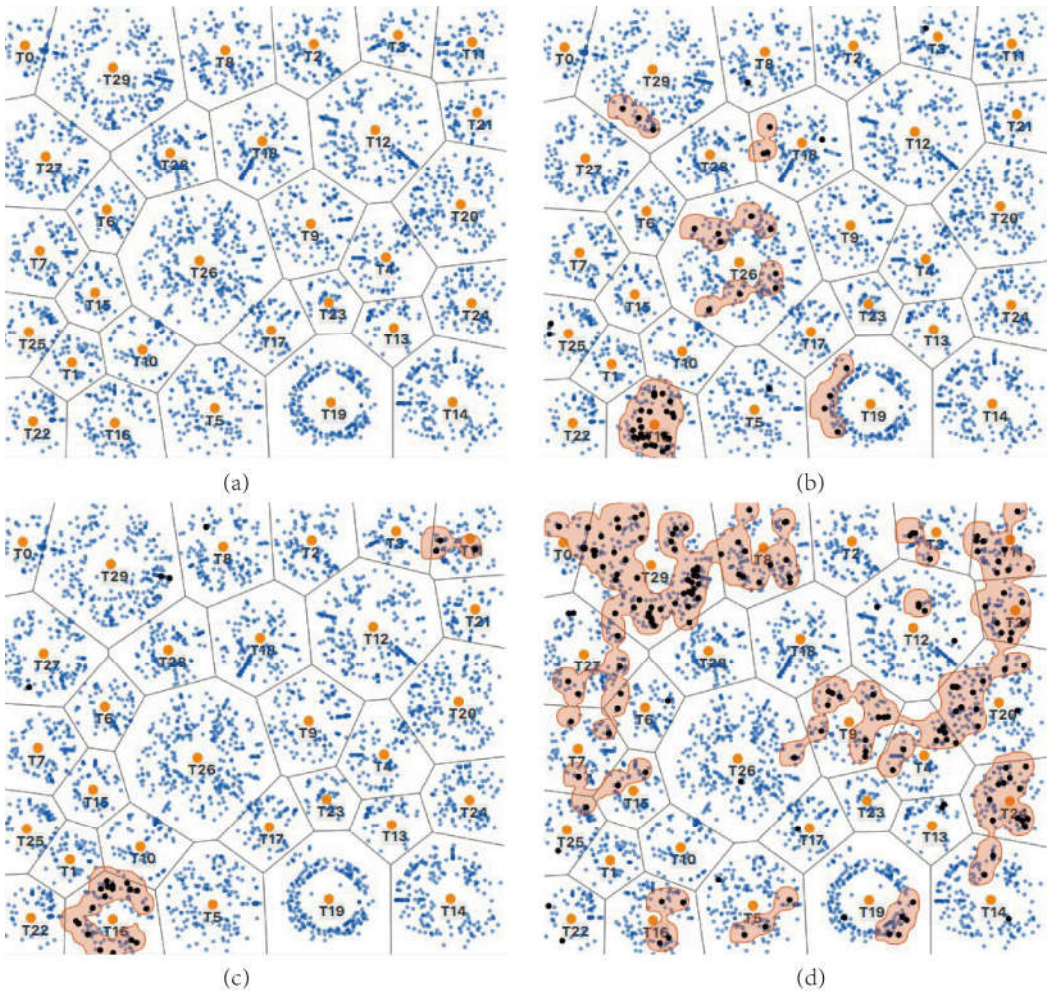

图 5 主题地图中用户活跃区域.(a) 原始主题地图,(b)周鸿祎活跃区域,(c)雷军活跃区域,(d)李开复活跃区域.

4.3 主题修正

在上述的主题可视分析中,我们会发现一些不合理的主题,并进行分析和修正.我们进行了如下操作:

(1) 删除主题 1(关键词:时代,过去,公众,领导,知道,人生,微博,网络,继续,60……).由于该主题内主要微博都由单个作者所发,主要使其个人感悟,不是微博上的主流话题,因此删除.

(2) 拆分主题 17(关键词:垃圾,局长,环保,今天,网友,贪腐,中国,生态,游泳,变成……).该主题主要包括了中国贪腐问题和网友邀请环保局长下河游泳事件,它们之间没有关联性,因此应被拆分为两个主题.

(3) 合并主题 18(关键词:记者、企业、媒体、潍坊、举报……)和主题 26(关键词:污染,地下,水污染,环境,中国……).这两个主题都是关于水污染,应该被合并.

经过相应的交互操作并重新运算主题模型后,修正后的结果中主题 1 不再出现,主题 18 和主题 26 合并为新的主题(关键词:污染、地下、企业、水污染、记者……),主题 17 被拆分为两个新的主题:主题 A(关键词:垃圾、局长、环保、网友、游泳……)和主题 B(关键词:贪腐、中国、变成、事儿、道德……).

通过上述案例分析可以看出,本系统支持用户在主题模型提取的主题的基础上对主题进行有效的探索与

分析,支持用户交互地修正主题模型的结果,得到质量更高的主题,也验证了本系统有助于提高微博主题分析的准确率,具有可用性及有效性.

5 总结与展望

本文提出了基于微博数据的文本主题可视分析系统.该系统通过多种可视化方法,利用主题矩阵、主题地图等视图,展示了由主题模型提取的微博主题信息和微博信息、时间信息、用户信息等,帮助系统用户分析主题信息、时变特征和微博用户特征,进一步探索微博-主题-关键词三个层次之间的相互关系.用户可以通过观察和链接不同的视图,获得对数据集和主题分布的理解,发现自动提取的主题中不合理的地方,对其进行修改关键词权重、删除主题、拆分主题、合并主题等操作,实现对主题模型结果的修正.通过使用案例,我们初步验证了系统的可用性和有效性.

系统注重结合计算机的运算能力和人的分析能力,让用户在自动计算的主题模型结果之上进行分析和编辑修正.相比于大多数利用主题模型结果进行可视化的研究工作,本系统能够通过交互实现对模型的修正.相比于 UTOPIAN[38]和 iVisClustering[44],本系统针对微博数据的特征,考虑了时间因素和用户因素,来帮助用户从不同角度分析微博主题,对主题获得深入的认识,同时也更容易发现自动计算的主题中不合理的地方.

本系统也具有一些缺点.由于采用原始的 LDA 算法,需要用户事先设置主题数目.然而在对数据集不了解的情况下,对目标主题数目做出判断并不容易.在未来的研究工作中,我们计划提供更便捷的交互操作,使用户可以更加方便进行修正操作,同时记录下用户的历史操作记录,支持用户撤销操作等.

References:

[1] Blei DM, Ng AY, and Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993–1022.

[2] Culotta A. Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics. Washington D.C: 2010, ACM.115–122.

[3] Sakaki T, Okazaki M, and Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. Raleigh: ACM, 2010. 851–860.

[4] Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, and Boyd D. The Arab Spring: the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. International Journal of Communication, 2011 5: 31-63.

[5] Hu M, Liu S, Wei F, Wu Y, Stasko J, and Ma KL. Breaking news on twitter. In: Proceedings of ACM CHI. Austin: ACM, 2012. 2751–2754.

[6] Ren D, Zhang X, Wang Z., Li J. and Yuan X. WeiboEvents: A Crowd Sourcing Weibo Visual Analytic System. In: Proceedings of IEEE PacificVis (Notes). Sydney: IEEE, 2014. 330-334.

[7] Starbird K, Palen L, Hughes AL, and Vieweg S. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work. Savannah: ACM, 2010. 241–250.

[8] Marcus A, Bernstein MS, Badar O, Karger DR, Madden S, and Miller RC. Twitinfo: aggregating and visualizing microblogs for event exploration. In: Proceedings of ACM CHI. Vancouver: ACM, 2011. 227–236.

[9] Itoh M, Yoshinaga N, Toyoda M, and Kitsuregawa M. Analysis and visualization of temporal changes in bloggers' activities and interests. In: Proceedings of IEEE PacificVis. Songdo: IEEE, 2012. 57–64.

[10] Bosch H, Thom D, Worner M, Koch S, Puttmann E, Jackle D, and Ertl T. Scatterblogs: Geo-spatial document analysis. In: Proceedings of IEEE VAST. Providence: IEEE, 2011. 309–310.

[11] Chae J, Thom D, Bosch H, Jang Y, Maciejewski R, Ebert D, and Ertl T. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: Proceedings of IEEE VAST. Seattle: IEEE, 2012. 143–152.

[12] Thom D, Bosch H, Koch S, Worner M, and Ertl T. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In: Proceedings of IEEE PacificVis. Songdo: IEEE, 2012. 41–48.

[13] Chen S, Yuan X, Wang Z, Guo C, Liang J, Wang Z, Zhang X, and Zhang J. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1):270–279.

[14] Battista GD, Eades P, Tamassia R, and Tollis IG. Graph drawing: algorithms for the visualization of graphs. Prentice Hall PTR, 1998.

[15] Herman I, Melanon G, and Marshall MS. Graph visualization and navigation in information visualization: A survey. IEEE Transactions on Visualization and Computer Graphics, 2000, 6(1):24–43.

[16] Heer J and Boyd D. Vizster: Visualizing online social networks. In: Proceedings of IEEE InfoVis. Minneapolis: IEEE, 2005. 32–39.

[17] F van Ham and JJ van Wijk. Interactive visualization of small world graphs. In: Proceedings of IEEE InfoVis. Austin: IEEE, 2014. 199–206.

[18] Shi L, Cao N, Liu S, Qian W, Tan L, Wang G, Sun J, and Lin CY. Hi-Map: Adaptive visualization of large-scale online social networks. In: Proceedings of IEEE PacificVis. Beijing: IEEE, 2009. 41–48.

[19] Abello J and F. van Ham. Matrix zoom: A visual interface to semi- external graphs. In: Proceedings of IEEE InfoVis. Austin: IEEE, 2004, 183–190.

[20] Henry N and Fekete JD. Matrixexplorer: a dual-representation system to explore social networks. IEEE Transactions on Visualization and Computer Graphics, 2006, 12(5):677–684.

[21] Henry N, Fekete JD, and McGuffin MJ. Nodetrix: a hybrid visualization of social networks. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6):1302–1309.

[22] Chen S, Chen S, Wang Z, Liang J, Yuan X, Cao N and Wu Y. D-Map: Visual Analysis of Ego-centric Information Diffusion Patterns in Social Media. In: Proceedings of IEEE VAST. Baltimore: IEEE, 2016. 41-50.

[23] Blei DM and Lafferty JD. A correlated topic model of science. The Annals of Applied Statistics. 2007, 1(1): 17–35.

[24] Blei DM, Griffiths T, Jordan M, and Tenenbaum J. Hierarchical topic models and the nested Chinese restaurant process. Advances in Neural Information Processing Systems, 2004, 16(17): 106–114.

[25] Teh YW, Jordan MI, Beal MJ, and Blei DM. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2006, 101(476):1566-1581.

[26] Blei DM and McAuliffe JD. Supervised topic models. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc, 2007, 121-128.

[27] Wang Y, Bai H, Stanton M, Chen WY, and Chang EY. PLDA: Parallel latent dirichlet allocation for large-scale applications. In: Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management. San Francisco: Springer, 2009. 301–314.

[28] Ramage D, Hall D, Nallapati R, and Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009. 248–256.

[29] Petinot Y, McKeown K, and Thadani K. A hierarchical model of web summaries. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011. 670– 675.

[30] Perotte A, Bartlett N, Elhadad N, and Wood F. Hierarchically supervised latent dirichlet allocation. In: Proceedings of 24th Annual Conference on Neural Information Processing Systems, Granada: Curran Associates Inc., 2011. 2009-2617.

[31] Blei DM and Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh: ACM, 2006, 113–120.

[32] Rosen-Zvi M, Griffiths T, Steyvers M, and Smyth P. The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Irvine: AUAI Press, 2004. 487–494.

[33] Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, and Li X. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. Dublin: Springer, 2011. 338–349.

[34] Ramage D, Dumais S, and Liebling D. Characterizing microblogs with topic models. In: proceedings of International AAAI Conference on Weblogs and Social Media. Washington, DC: AAAI Press, 2010. 130–137.

[35] Boyd D, Golder S, and Lotan G. Tweet, Tweet, Retweet: Conversational aspects of retweeting on twitter. In: proceedings of 43rd Hawaii International Conference on System Sciences. Hawaii: IEEE, 2010. 1–10.

[36] Eisenstein J, Chau DH, Kittur A, and Xing E. Topicviz: interactive topic exploration in document collections. In proceedings of ACM CHI. Austin: ACM, 2012. 2177–2182.

[37] Gretarsson B, O'donovan J, Bostandjiev S, Hollerer T, A. Asuncion, D. Newman, and P. Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. ACM Transactions on Intelligent Systems and Technology, 2012, 3(2): 2012, 23-49.

[38] Choo J, Lee C, Reddy CK, and Park H. UTOPIAN: User-driven topic modeling based on interactive non-negative matrix factorization. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12):1992–2001.

[39] Chuang J, Manning CD, and Heer J. Termite: Visualization techniques for assessing textual topic models. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. Capri Island: ACM, 2012. 74–77.

[40] Alexander E, Kohlmann J, Valenza R, Witmore M, and Gleicher M. Serendip: Topic model-driven visual exploration of text corpora. In: Proceedings of IEEE VAST. Paris: IEEE, 2014, 173–182.

[41] Dou W, Yu L, Wang X, Ma Z, and Ribarsky W. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12):2002–2011.

[42] Cui W, Liu S, Wu Z, and Wei H. How hierarchical topics evolve in large text corpora. IEEE Transactions on Visualization and Computer Graphics, 2014, 20(12):2281–2290.

[43] Alexander E and Gleicher M. Task-driven comparison of topic models. IEEE Transactions on Visualization and Computer Graphics, 2016, 22(1):320– 329.

[44] Lee H, Kihm J, Choo J, Stasko J, and Park H. iVisClustering: An interactive visual document clustering via topic modeling. Computer Graphics Forum, 2012, 31(3): 1155–1164.

[45] King JR. Machine-component group formation in group technology. Omega, 1980, 8(2):193–199.

[46] King JR and Nakornchai V. Machine-component group formation in group technology: review and extension. The International Journal of Production Research, 1982, 20(2):117–133.

[47] Balzer M and Deussen O. Voronoi treemaps. In: Proceedings of the IEEE InfoVis. Minneapolis: IEEE, 2005, 49-55.

[48] Balzer M, Deussen O, and Lewerentz C. Voronoi treemaps for the visualization of software metrics. In: Proceedings of ACM SoftVis. St. Louis: ACM, 2005. 165–172.

[49] Collins C, Penn G, and Carpendale S. BubbleSets: Revealing set relations with isocontours over existing visualizations. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6):1009–1016.

[50] MonogoDB, https://www.mongodb.com.

[51] NLPIR, Word Cutting System for Chinese. http://ictclas.nlpir.org/