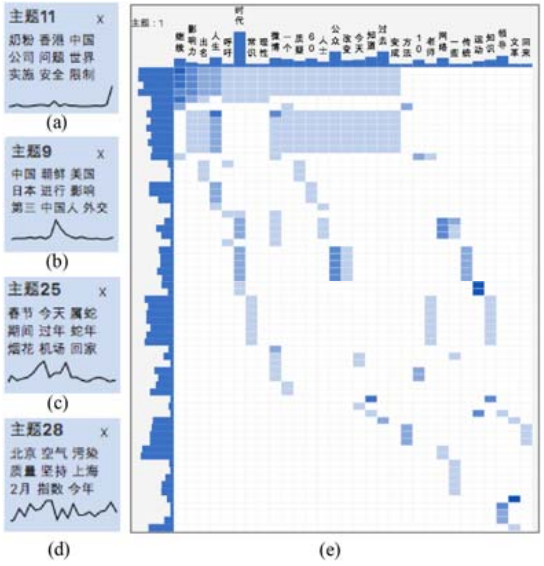
十分接近两主题区域的边界.我们通过观察两个主题内容,发现主题 18 是关于企业将污水排入地下的新闻报道,而主题 26 是中国的地下水污染问题,两者相关性十分大,可将它们合并为同一个主题.



(a)~(d) 主题 11、主题 9、主题 25、主题 28 的主题卡片;(e) 主题 1 的主题矩阵

Fig.4 Result of case study

图 4 系统案例分析结果

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 4(e)所示)与主题 17(如图 1(b)所示).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.我们对图 4 所示左上角的聚集区域进行选择,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.

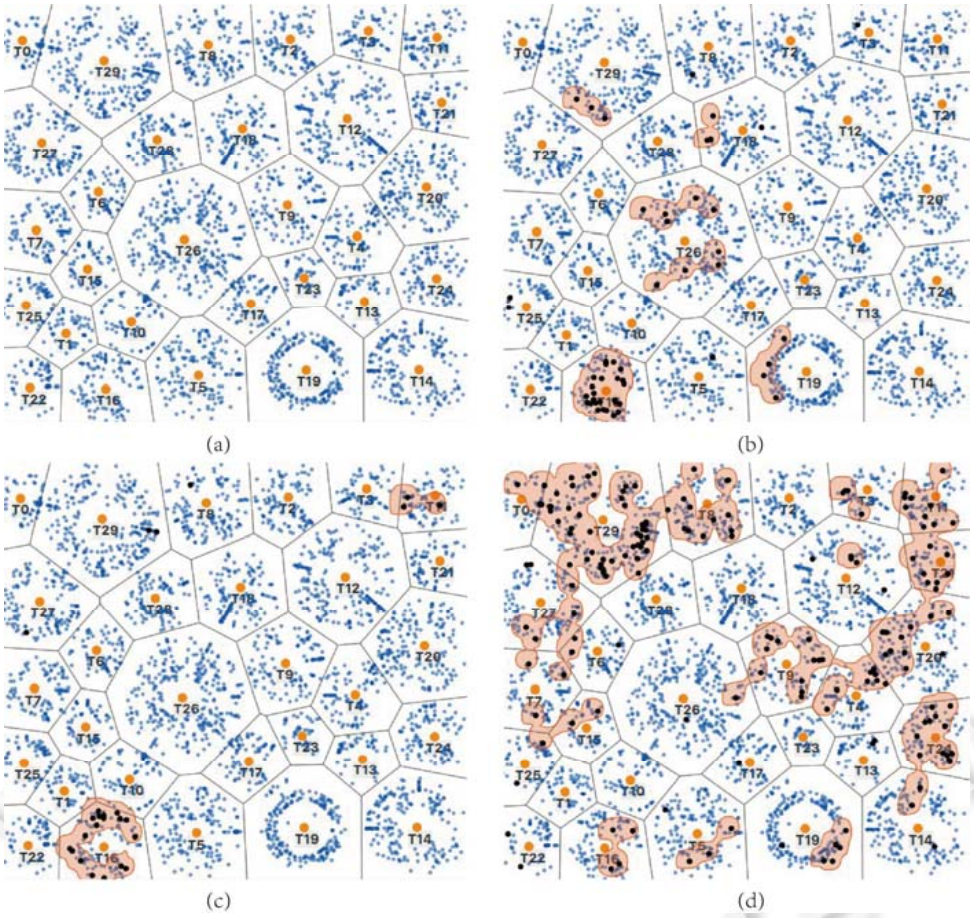
对主题 17 的探索包含了两个步骤.在默认的基于关键词权重排序的情况下,我们较难看出该主题的特征分布(如图 2(a)所示),用户通过对主题的排序,观察出主题 17 包含比较明显的两个分块:左上角与右下角.进一步地,我们通过刷选这两部分区域的内容,相关的微博在微博内容视图中被高亮选中,我们归纳出左上角谈论的是当前中国的贪腐问题(关键词:贪腐等),右下角则是关于微博网友邀请环保局长下河游泳的事件(关键词:局长、环保、网友、游泳等).通过主题矩阵的分析,我们认为该主题应被拆分为两个主题.

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员探索其参与的主题.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 5 中可以看出:雷军的活跃区域主要集中在主题 16 和主题 11(如图 5(c)所示),周鸿祎则主要集中在主题 16、主题 26、主题 18、主题 29 等(如图 5(b)所示).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的(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.从图 5(d)可以看出:李开复关注最多的是主题 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(朝鲜核实验)等,这些话题都是当时的社会热点话题,表明李开复经常在人们关注的热点事件上发表意见.



(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

Fig.5 Active regions of different users in the topic map

图 5 主题地图中用户活跃区域

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,主要是其个人感悟,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的,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-关键词这 3 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[41]和 iVisClustering[46],本系统针对微博数据的特征,考虑了时间因素和用户因素,以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地进行修正操作,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, Jordan MI. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003,3(Jan.):993−1022.

[2] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. In: Proc. of the 1st Workshop on Social Media Analytics. Washington: ACM Press, 2010. 115−122. [doi: 10.1145/1964858.1964874]

[3] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-Time event detection by social sensors. In: Proc. of the 19th Int'l Conf. on World Wide Web. Raleigh: ACM Press, 2010. 851−860. [doi: 10.1145/1772690.1772777]

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

[5] Hu M, Liu S, Wei F, Wu Y, Stasko J, Ma KL. Breaking news on Twitter. In: Proc. of the ACM CHI. Austin: ACM Press, 2012. 2751−2754. [doi: 10.1145/2207676.2208672]

[6] Ren D, Zhang X, Wang Z, Li J, Yuan X. WeiboEvents: A crowd sourcing Weibo visual analytic system. In: Proc. of the IEEE PacificVis (Notes). Sydney: IEEE, 2014. 330−334. [doi: 10.1109/PacificVis.2014.38]

[7] Chen S, Lin L, Yuan X. Social media visual analytics. Computer Graphics Forum, 2017,36(3):563−587.

[8] Starbird K, Palen L, Hughes AL, Vieweg S. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In: Proc. of the 2010 ACM Conf. on Computer Supported Cooperative Work. Savannah: ACM Press, 2010. 241−250. [doi: 10.1145/1718918.1718965]

[9] Marcus A, Bernstein MS, Badar O, Karger DR, Madden S, Miller RC. Twitinfo: Aggregating and visualizing microblogs for event exploration. In: Proc. of the ACM CHI. Vancouver: ACM Press, 2011. 227−236. [doi: 10.1145/1978942.1978975]

[10] Itoh M, Yoshinaga N, Toyoda M, Kitsuregawa M. Analysis and visualization of temporal changes in bloggers' activities and interests. In: Proc. of the IEEE PacificVis. Songdo: IEEE, 2012. 57−64. [doi: 10.1109/PacificVis.2012.6183574]

[11] Bosch H, Thom D, Worner M, Koch S, Puttmann E, Jackle D, Ertl T. Scatterblogs: Geo-Spatial document analysis. In: Proc. of the IEEE VAST. Providence: IEEE, 2011. 309−310. [doi: 10.1109/VAST.2011.6102488]

[12] Chae J, Thom D, Bosch H, Jang Y, Maciejewski R, Ebert D, Ertl T. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: Proc. of the IEEE VAST. Seattle: IEEE, 2012. 143−152. [doi: 10.1109/VAST.2012.6400557]

[13] Thom D, Bosch H, Koch S, Worner M, Ertl T. Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In: Proc. of the IEEE PacificVis. Songdo: IEEE, 2012. 41−48. [doi: 10.1109/PacificVis.2012.6183572]

[14] Chen S, Yuan X, Wang Z, Guo C, Liang J, Wang Z, Zhang X, Zhang J. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. IEEE Trans. on Visualization and Computer Graphics, 2016,22(1):270−279. [doi: 0.1109/TVCG.2015.2467619]

[15] Battista GD, Eades P, Tamassia R, Tollis IG. Graph Drawing: Algorithms for the Visualization of Graphs. Upper Saddle River: Prentice Hall PTR, 1998. https://dl.acm.org/citation.cfm?id=551884

[16] Herman I, Melanon G, Marshall MS. Graph visualization and navigation in information visualization: A survey. IEEE Trans. on Visualization and Computer Graphics, 2000,6(1):24−43. [doi: 0.1109/2945.841119]

[17] Heer J, Boyd D. Vizster: Visualizing online social networks. In: Proc. of the IEEE InfoVis. Minneapolis: IEEE, 2005. 32−39. [doi: 10.1109/INFVIS.2005.1532126]

[18] van Ham F, van Wijk JJ. Interactive visualization of small world graphs. In: Proc. of the IEEE InfoVis. Austin: IEEE, 2014. 199−206. [doi: 10.1109/INFVIS.2004.43]

[19] Shi L, Cao N, Liu S, Qian W, Tan L, Wang G, Sun J, Lin CY. Hi-Map: Adaptive visualization of large-scale online social networks. In: Proc. of the IEEE PacificVis. Beijing: IEEE, 2009. 41−48. [doi: 10.1109/PACIFICVIS.2009.4906836]

[20] Abello J, van Ham F. Matrix zoom: A visual interface to semi-external graphs. In: Proc. of the IEEE InfoVis. Austin: IEEE, 2004. 183−190. [doi: 10.1109/INFVIS.2004.46]

[21] Henry N, Fekete JD. Matrixexplorer: A dual-representation system to explore social networks. IEEE Trans. on Visualization and Computer Graphics, 2006,12(5):677−684. [doi: 0.1109/TVCG.2006.160]

[22] Henry N, Fekete JD, McGuffin MJ. Nodetrix: A hybrid visualization of social networks. IEEE Trans. on Visualization and Computer Graphics, 2007,13(6):1302−1309. [doi: 0.1109/TVCG.2007.70582]

[23] Chen S, Chen S, Wang Z, Liang J, Yuan X, Cao N, Wu Y. D-Map: Visual analysis of ego-centric information diffusion patterns in social media. In: Proc. of the IEEE VAST. Baltimore: IEEE, 2016. 41−50. [doi: 10.1109/VAST.2016.7883510]

[24] Chen S, Chen S, Lin L, Yuan X, Liang J, Zhang X. E-Map: A visual analytics approach for exploring significant event evolutions in social media. In: Proc. of the IEEE VAST. Phoenix: IEEE, 2017.

[25] Blei DM, Lafferty JD. A correlated topic model of science. The Annals of Applied Statistics, 2007,1(1):17−35. [doi: 0.1214/07-AOAS114]

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

[27] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006, 101(476):1566−1581. [doi: 0.1198/016214506000000302]

[28] Blei DM, McAuliffe JD. Supervised topic models. In: Proc. of the 20th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2007. 121−128. http://papers.nips.cc/paper/3328-supervised-topic-models

[29] Wang Y, Bai H, Stanton M, Chen WY, Chang EY. PLDA: Parallel latent Dirichlet allocation for large-scale applications. In: Proc. of the 5th Int'l Conf. on Algorithmic Aspects in Information and Management. San Francisco: Springer-Verlag, 2009. 301−314. [doi: 10.1007/978-3-642-02158-9_26]

[30] Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2009. 248−256. https://dl.acm.org/citation.cfm?id=1699543

[31] Petinot Y, McKeown K, Thadani K. A hierarchical model of Web summaries. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011. 670−675. https://dl.acm.org/citation.cfm?id=2002866

[32] Perotte A, Bartlett N, Elhadad N, Wood F. Hierarchically supervised latent Dirichlet allocation. In: Proc. of the 24th Annual Conf. on Neural Information Processing Systems. Granada: Curran Associates Inc., 2011. 2009−2617.

[33] Blei DM, Lafferty JD. Dynamic topic models. In: Proc. of the 23rd Int'l Conf. on Machine Learning. Pittsburgh: ACM Press, 2006. 113−120. [doi: 10.1145/1143844.1143859]

[34] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. In: Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence. Irvine: AUAI Press, 2004. 487−494. https://dl.acm.org/citation.cfm?id=1036902

[35] Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X. Comparing Twitter and traditional media using topic models. In: Proc. of the 33rd European Conf. on Advances in Information Retrieval. Dublin: Springer-Verlag, 2011. 338−349. [doi: 10.1007/978-3-642-20161-5_34]

[36] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models. In: Proc. of the Int'l AAAI Conf. on Weblogs and Social Media. Washington: AAAI Press, 2010. 130−137. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1528

[37] Chuang J, Manning CD, Heer J. Termite: Visualization techniques for assessing textual topic models. In: Proc. of the Int'l Working Conf. on Advanced Visual Interfaces. Capri Island: ACM Press, 2012. 74−77. [doi: 10.1145/2254556.2254572]

[38] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: Proc. of the 43rd Hawaii Int'l Conf. on System Sciences. Hawaii: IEEE, 2010. 1−10. [doi: 10.1109/HICSS.2010.412]

[39] Eisenstein J, Chau DH, Kittur A, Xing E. Topicviz: Interactive topic exploration in document collections. In: Proc. of the ACM CHI. Austin: ACM Press, 2012. 2177−2182. [doi: 10.1145/2212776.2223772]

[40] Gretarsson B, O'donovan J, Bostandjiev S, Hollerer T, Asuncion A, Newman D, Smyth P. TopicNets: Visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology, 2012,3(2):23−49. [doi: 10.1145/2089094.2089099]

[41] Choo J, Lee C, Reddy CK, Park H. UTOPIAN: User-Driven topic modeling based on interactive non-negative matrix factorization. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):1992−2001. [doi: 0.1109/TVCG.2013.212]

[42] Alexander E, Kohlmann J, Valenza R, Witmore M, Gleicher M. Serendip: Topic model-driven visual exploration of text corpora. In: Proc. of the IEEE VAST. Paris: IEEE, 2014. 173−182. [doi: 10.1109/VAST.2014.7042493]

[43] Dou W, Yu L, Wang X, Ma Z, Ribarsky W. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2002−2011. [doi: 0.1109/TVCG.2013.162]

[44] Cui W, Liu S, Wu Z, Wei H. How hierarchical topics evolve in large text corpora. IEEE Trans. on Visualization and Computer Graphics, 2014,20(12):2281−2290. [doi: 0.1109/TVCG.2014.2346433]

[45] Alexander E, Gleicher M. Task-Driven comparison of topic models. IEEE Trans. on Visualization and Computer Graphics, 2016, 22(1):320−329. [doi: 0.1109/TVCG.2015.2467618]

[46] Lee H, Kihm J, Choo J, Stasko J, Park H. iVisClustering: An interactive visual document clustering via topic modeling. Computer Graphics Forum, 2012,31(3):1155−1164. [doi: 10.1111/j.1467-8659.2012.03108.x]

[47] King JR. Machine-Component group formation in group technology. Omega, 1980,8(2):193−199. [doi: 10.1016/0305-0483(80)90023-7]

[48] King JR, Nakornchai V. Machine-Component group formation in group technology: Review and extension. The Int'l Journal of Production Research, 1982,20(2):117−133. [doi: 10.1080/00207548208947754]

[49] Balzer M, Deussen O. Voronoi treemaps. In: Proc. of the IEEE InfoVis. Minneapolis: IEEE, 2005. 49−55. [doi: 10.1109/INFVIS.2005.1532128]

[50] Balzer M, Deussen O, Lewerentz C. Voronoi treemaps for the visualization of software metrics. In: Proc. of the ACM SoftVis. St. Louis: ACM Press, 2005. 165−172. [doi: 10.1145/1056018.1056041]

[51] Collins C, Penn G, Carpendale S. BubbleSets: Revealing set relations with isocontours over existing visualizations. IEEE Trans. on Visualization and Computer Graphics, 2009,15(6):1009−1016. [doi: 10.1109/TVCG.2009.122]

[52] MonogoDB. https://www.mongodb.com

[53] NLPIR, word cutting system for Chinese. http://ictclas.nlpir.org/

王臻皇(1991－),男,福建福州人,软件工程师,主要研究领域为文本可视化,社交网络可视化.

陈思明(1989－),男,博士生,CCF 学生会员,主要研究领域为信息可视化,社交媒体可视分析,时空可视分析.

袁晓如(1975－),男,博士,研究员,博士生导师,CCF 杰出会员,主要研究领域为可视化,可视分析.