基于社交媒体的事件感知与多模态事件脉络生成

徐程浩[1] 郭 斌[1] 欧阳逸[1] 翟书颖[2] 於志文[1]
(西北工业大学计算机学院 西安 710129)[1] (西北工业大学明德学院 西安 710129)[2]

摘 要 随着信息技术的发展和社交媒体的流行,普通用户已经完成了从信息接受者到信息产生者的转变,每个人都可以实时分享自己身边的信息,也可以转发自己感兴趣的内容,这使得社交媒体的数据量迅速增长。在海量数据中蕴含着丰富的社会事件发生和发展的记录,如何有效地从这些数据中挖掘出有价值的信息成为了当前信息领域的重要问题。针对该问题,介绍了基于社交媒体的事件感知与多模态事件脉络生成。基于社交媒体的事件感知与多模态事件脉络生成旨在通过分析社交媒体中的文本、时间、图像、评论、观点、情感和用户交互等多模态数据,感知事件并刻画事件的关系,从而实现对事件的总结。讨论了基于社交媒体的事件感知与多模态事件脉络生成的描述模型、概念、发展历史、关键技术与挑战以及其广泛的应用领域,综述了社交媒体分析在事件感知和事件总结方面的研究进展,并对其未来发展进行了展望。

关键词 社交媒体,事件感知,多模态数据,事件脉络,跨媒体
中图法分类号 TP391 **文献标识码** A

Event Sensing and Multimodal Event Vein Generation Leveraging Social Media

XU Cheng-hao[1] GUO Bin[1] OUYANG Yi[1] ZHAI Shu-ying[2] YU Zhi-wen[1]
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)[1]
(Northwestern Polytechnical University Ming De College, Xi'an 710129, China)[2]

Abstract With the development of information technology and popularity of social media, normal users have become information producers from receivers and everyone can share what happened around them and repost what they are interested in, which makes the information stored in social media increase rapidly. The large amount of data contains abundant and valuable records of social events. How to get valuable informations from these data has become one of the most important problems in information field. This paper introduced the new research field, including crowd-powered event sensing and multimodal summarization to solve this problem. Crowd-powered event sensing and multimodal summarization aim at sensing and analyzing events by analyzing multimodal data existed in social media to predict and summarize events effectively. This paper described the modal of event, the history of sensing, the key technology, challenges and wide application field, summarized the development of event sensing and summarization based social media analysis and looked into the future.

Keywords Social media, Event sensing, Multimodal data, Storyline, Cross media

1 引言

社交媒体是人们之间用来分享见闻、经验和观点的网络平台,目前国内外常用的社交媒体主要有微博、Twitter 和 Facebook 等。随着 Web2.0 的兴起和社交媒体的发展,Twitter 和微博等社交应用允许人们发布和分享自己的所见所闻,越来越多的信息以多种数据形式出现在各大社交媒体中。大量的实时数据吸引了很多研究者分析这些社交媒体的数据,并从中挖掘出有用的知识。文献[1]把 Twitter 当作社会事件的传感器来实时感知地震的方位并通知当地的民众,这种通知方式甚至比日本气象局的通知还要迅速。文献[2]研究了 Twitter 在突发事件报道上的高时效性,显示在官方发布本拉登死亡消息之前,Twitter 上关于这一消息的转发量已经达到上百万。

社交媒体数据分析的应用前景十分广泛。例如,有些研究致力于通过社交媒体来侦测突发事件,包括恐怖袭击、疾病和地震[4-6]等灾难事件;有些研究工作用于对一些有规律的事件进行预测,比如政治选举和体育比赛[7];还有些研究通过更加直观的方式对社交媒体中的事件进行总结,然后以不同的可视化方式呈现给用户,不同的总结和分析方法衍生了不同的结果呈现方式,比如检索排序[18-20]和时间轴[21-27]。

对社交媒体数据的处理面临着许多问题:1)社交媒体数据作为个人用户发布信息的渠道,其内容往往具有随意性[3],这与经过编辑的新闻报道是不同的,这种随意性给事件分析

本文受国家重点基础研究发展计划(973 计划)(2015CB352400),国家自然科学基金(61332005,61373119)资助。
徐程浩(1994—),男,主要研究领域为普适计算、社交媒体挖掘,E-mail:haochengxu@mail.nwpu.edu.cn;郭 斌(1980—),男,博士,教授,CCF 高级会员,主要研究领域为普适计算和移动群智感知,E-mail:guob@nwpu.edu.cn(通信作者);欧阳逸(1994—),男,博士生,主要研究领域为普适计算;翟书颖(1981—),女,讲师,主要研究领域为物联网和移动社交网络;於志文(1977—),男,博士,教授,CCF 高级会员,主要研究领域为普适计算和社会感知计算。

带来困难;2)由于微博本身的字数限制,单条微博往往难以提供有效的信息[3],这使得对长文本分析的传统方法不能简单地应用到社交媒体分析中;3)多模态的数据也给社交媒体分析带来新的挑战,由于信息的随意性和每条信息携带了一定数据量,充分利用每条微博的信息成为必然,而如何统一处理这些不同模态的信息是研究者们现在关注的重点之一。

本文第 2 节介绍社交媒体的事件模型和概念模型的一些研究成果;第 3 节和第 4 节分别介绍事件感知和事件总结的发展概况和方法;第 5 节介绍面向社交媒体的事件感知和多模态事件脉络生成所面临的挑战和针对这些问题的方法;第 6 节介绍面向社交媒体的数据分析的应用和前景;最后总结全文。

2 社交媒体的事件描述模型

社交媒体的普及使得每天有千万级以上的用户在这些平台上分享和传播信息,面对海量数据,需要一个统一的模型来分析隐藏在数据之中的信息和知识。对社交媒体的事件描述模型中涉及的相关概念进行介绍。

子事件:社会事件通常会包含不同的子事件,每个子事件反映的是一个事件的不同侧面,比如对疾病事件的感知、对疾病的传播的描述和对疾病的防治的描述可以分别作为“疾病”这一事件的子事件。

线索:事件之间往往都不是孤立的,根据关注事件的用户群体是否相似、事件之间的时空关系以及事件包含的关键词可以得到事件之间的关系图,这个体现事件变化和事件相关性的图就是一个线索,即事件发展的脉络。

关联:两个事件有关联是指两个事件之间存在某种关系,通常这种关系是因果关系或者是互补关系,即一个事件经过一段时间发展成为另一个事件,或者两个事件同时是另一个事件的子事件。

情感:社交媒体中的信息除了像新闻报道那样包含了事件的发展信息,还包含了民众对这些事件的评价和看法,可以将这些看法简单地分为积极的看法和消极的看法,或者以其他标准划分看法中带有的情绪,这就是社交媒体信息中包含的情感。

3 事件检测

随着社交媒体的发展,越来越多的用户在社交媒体上分享和讨论热点事件,这些事件可以是物理世界中发生的事件,也可以是社交媒体中的热点话题。社交媒体拥有丰富、多维度的信息,因此,通过分析这些海量数据可以实时发现热点事件[36]。然而,对于某一事件,与之相关的微博量非常庞大,同时,其中充斥着大量的冗余信息,因此社交媒体的事件检测也面临着巨大的挑战。目前事件检测的方法可以分为两类。

第一类方法以文档为中心,通过分析文档间的相似度进行聚类,从而检测事件[9-14]。文献[12]提出将事件划分为不重合的片段,通过对这些片段进行聚类来生成不同的事件。文献[13]采用 LDA 模型来发现时间和空间属性不同的事件。文献[14]使用一种层次和非层次的聚类算法来发现时间轴上的不同事件。由于微博的长度较短,单条微博中包含的信息往往很少,使用以文档为中心的方法进行事件发现时会生成大量的稀疏向量从而影响微博相关性的测量,所以更多的事件检测工作建立在以特征为中心的方法上。

第二类方法以特征为中心,通过分析与事件相关的关键词来发现事件[15-17]。文献[15]使用无限状态自动机对数据流进行建模,采用层次结构来分析数据流,从而发现数据流中的热点事件。文献[16]自动提取微博中突发事件的关键词,并通过对这些关键词进行聚类来发现热点事件。文献[17]不仅通过识别微博文本特征和用户特征进行事件发现,同时还对热点事件的发展进行了预测。

4 多模态事件脉络生成

通常在检测到事件发生之后,用户会进一步关注事件的发展过程,一个清晰的事件脉络能够告知用户整个事件的发展过程[37]。在一个事件中,用户可以通过文本、图片、评论、转发等形式来描述事件的发展,这些多模态的数据提供了丰富的信息,同时也给事件脉络生成带来了挑战。因此,很多研究者致力于多模态事件脉络生成,目前的方法可以分为 3 类。

第一种方法只选取有代表性的微博或关键词作为事件总结[18-20]。文献[18]针对体育比赛,通过分析用户对体育比赛的观点对事件进行总结。文献[20]针对灾难事件,提取微博文本中关于不同的灾难事件的情景信息,然后总结这些信息。文献[19]的方法不同于前两种,不仅使用了微博的文本信息,还使用了其中的图片信息来总结事件。然而,这些工作都只着眼于选取有代表性的的微博,却没有一个清晰的事件发展脉络,用户很难理解事件的发展过程。

第二种方法将事件总结的结果按照时间排序,以时间轴的方法来呈现事件脉络[21-27]。文献[23,25-26]使用峰值检测的方法来找出一个事件的重要时刻,通过选取在这些时刻的微博来生成一个按时间顺序排列的事件脉络。然而,Meladianos等[21]认为使用峰值检测的方法并不能很好地找出重要的时刻,并提出了一种基于图论的算法来发现事件的重要时刻。Xu 等[22]使用了文本以及用户间的交互信息来创建图结构,从而选出有代表性的图片作为事件总结。Yan 等[24]通过文本与图片信息的分析来选取重要的微博或者句子。Chang 等[27]通过分析用户间的交互信息的方法来选取微博。虽然第二种方法按照时间顺序生成了事件脉络,但是不能更全面地展现一个事件的各个侧面。

第三种方法也需要选取有代表性的微博或者句子,但与之前的方法不同的是其重点研究了事件中不同侧面之间的关系[28-31]。文献[28,30-31]把微博挑选问题看作是寻找图中最小支配集的问题。文献[29-31]在建立图的时候只使用了文字信息,而 Wang 等[28]同时使用文字和图片信息来建立图。文献[28,30-31]采用 Steiner 树算法构建事件的脉络和分支。Lee 等[29]提出了一种基于语境信息的搜索方法来跟踪上下文结构,从而建立事件脉络。

5 关键技术与挑战

5.1 多模态信息交互融合的事件要素

一条微博往往包含丰富、多维度的信息[19,32-33],比如:时间、文本、图片、地理位置、评论、转发等。通常单模态的数据只能刻画事件的一个侧面,不能展现事件的全貌,而多模态的数据能够起到互补的作用,能够发现事件的各个侧面以及他们之间的关联,从而呈现事件发展的全貌。因此越来越多的研究者开始关注如何将多模态的数据进行交互融合。

当前很多工作都致力于对 2 种或者 3 种信息进行融合,

而其中如何对文字和图像进行跨媒体分析成为了多模态信息分析的重点和难点。文献[32]提出了一种对微博中的文字、URL 中的文字和图片信息进行联合分析的方法，通过对微博转发量进行统计，从而发现热点事件，但这种方法并没有对图像信息进行深入分析，而且对 URL 中的信息进行分析时无法通过转发数来计算权重。文献[19]通过对传统的 LDA 方法进行改进，深入分析了文本和图像之间的关系，选取出有代表性的文本和图像信息作为一个事件的子事件；这种方法考虑了图像本身携带的信息，但只对文本和图像进行了分析，没有对其他信息进行进一步分析。文献[34]使用卷积神经网络的方法对图像和文本进行统一处理和检索，但文本和图像必须成对输入。相比之下，文献[33]更加全面地考虑了微博中的文字、图像、时间和用户之间关系的信息，通过传统图论的方法来生成对事件的总结。以上方法对微博或推特中的两种或两种以上多模态的信息进行了深入的挖掘和分析，但尚未有任何一种方法可以对事件中的所有信息进行统一的建模分析。

5.2　事件之间的关系刻画

在社交媒体发现热点事件之后，对事件本身的分析也成为了一个值得研究者们关注的问题。同一个时间段内的事件往往是具有相关性的，而对事件之间的相关性的刻画则取决于很多因素，比如事件是否发生在同一地点、关注事件的人群是否有某一特点以及事件的主体是否相同等。对一个事件来说，其本身又可以划分为一个个子事件，不同的子事件刻画了事件的不同侧面。通过对事件的划分和事件之间关系的刻画，可以生成一个易于用户理解的故事脉络。

文献[19]对同一事件的子事件进行了总结和评分，生成了对同一事件的不同侧面的报道；采用 LDA 的方法对同一事件的相关微博进行聚类，但并没有考虑到事件之间的关系。文献[29]通过图论的方法总结了在同一时间段的微博流中发生的事件，并对事件之间的关系进行了刻画，从而生成事件的脉络，但对同一事件的子事件并没有做更多的考虑；而且在分析事件之间的关系时，只分析了微博的相似性，然后利用图论来进一步分析事件关系，没有利用事件中的用户交互信息。文献[33]分析了事件之间的用户和内容的相关性，但工作只局限于选取不同事件中有代表意义的微博，并不能生成一个便于用户理解的事件发展脉络。CrowdStory[35]考虑了时间、文本、图片和用户交互 4 种属性，可以认为图片和用户交互是时间和文本属性的补充，并且还能发现微博之间的关系。CrowdStory 不仅通过时间和文本层的融合发现了事件的不同线索，而且通过图片和用户交互层的融合生成了一个细粒度、多侧面的事件发展脉络。

6　未来的应用和研究方向

社交媒体的事件感知具有广泛的应用领域。由于社交媒体信息的发布者是广泛的个人用户，因此基于社交媒体的事件感知往往要比正式的新闻报道更加及时[2]，因此，通过对社交媒体信息的分析，可以更快地得到热点事件的发展情况。

一段时间内的社交媒体的数据反映了当前时间段内社会的变化，通过对事件的分类，可以看到在这段时间内发生在政治、体育和科学等领域的事件[7]，这使得我们可以分析在某一时间段内事件的变化和它们之间的关系，从而得到有用的信息。同时，对于很多有规律可循的事件，比如公民选举和疾病[5,20]等事件，对社交媒体的分析可以对事件的发展进行预测，分析事件的发展趋势和起因，这对疾病管控和突发事件的处理都具有重要意义。

随着社交媒体用户量的增多和信息的爆炸式增长，对社交媒体数据的有效分析也具有广泛的应用前景。仅仅对事件的感知还不能形成一种便于用户理解的模式，因此要对事件做进一步分析。通过对信息的相关性和重要性评估，可以展示给用户最有价值的微博。将同一事件进一步分解为一个个子事件，不同的子事件体现了事件的不同侧面。除了对事件进行进一步分析以外，对事件之间的关系也需要分析。利用微博用户之间的关系以及事件本身的相关性可以得到一段时间内事件的关系网，在关系网中加入时间信息后可以看到事件的发展脉络[25,28,30]。同时，基于多模态的分析方法并不局限于社交媒体分析的领域，这种方法可以广泛应用于含有多种形式信息分析的情况，比如信息检索和信息过滤。

新兴领域的多模态分析和事件脉络总结有着广泛的发展前景。如何尽可能地将社交媒体中多种形式的数据进行统一的分析以及找到更加有效的事件关系分析方法是未来社交媒体数据分析的研究重点。

结束语　本文综述了基于社交媒体的事件感知的发展概况和应用前景，以及新兴领域的多模态事件脉络总结的发展概况和关键技术。随着存储在社交媒体中的数据呈指数型增长，使用行之有效的方法来挖掘和分析这些数据已经成为迫切的需求。个人用户通过社交媒体成为了对周遭社会信息的实时传感器，实时感知这些数据可以对突发事件进行侦测以及预测这些事件的发展状况。然而需要看到社交媒体信息的发布者是个人用户以及社交媒体平台本身的限制，社交媒体的信息充满了随意性，这给分析数据带来了新的挑战，对社交媒体中多种形式的数据进行综合分析是解决信息随意性问题的关键技术之一。对感知到的事件进行进一步总结也是重要的研究课题之一。仅仅对事件的感知不足以给用户一个便于理解和分析的结果，因此对事件之间的关系进行进一步的分析从而获得更有价值的知识，充分利用事件信息中的用户关系和图论的知识来分析事件关系是解决这一问题的关键技术。

参 考 文 献

[1] SAKAKI T，OKAZAKI M，MATSUO Y. Earthquake shakes Twitter users：real-time event detection by social sensors [C]// Proceedings of the 19th International Conference on World Wide Web. ACM，2010：851-860.

[2] HU M D，LIU S X，WEI F R，et al. Breaking news on twitter [C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM，2012：2751-2754.

[3] PSALLIDAS F，BECKER H，NAAMAN M，et al. Effective Event Identification in Social Media[J]. IEEE Data Eng. Bull.，2013，36(3)：42-50.

[4] ZIN T T，TIN P，HAMA H，et al. Knowledge based social network applications to disaster event analysis [C]// Proceedings of the International Multiconference of Engineers and Computer Scientists. 2013.

[5] LAMPOS V，CRISTIANINI N. Tracking the flu pandemic by monitoring the social web[C]// 2010 2nd International Workshop on Cognitive Information Processing. IEEE，2010：411-416.

[6] SAKAKI T，OKAZAKI M，MATSUO Y. Tweet analysis for

real-time event detection and earthquake reporting system development[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(4):919-931.

[7] ZHAO S Q,ZHONG L,WICKRAMASURIYA J,et al. Human as real-time sensors of social and physical events:A case study of twitter and sports games[J]. arXiv preprint arXiv:1106.4300,2011.

[8] LI R,LEI K H,KHADIWALA R,et al. Tedas:A twitter-based event detection and analysis system[C]// 2012 IEEE 28th International Conference on Data Engineering. IEEE,2012:1273-1276.

[9] AGARWAL P,VAITHIYANATHAN R,SHARMA ARMA S,et al. Catching the Long-Tail:Extracting Local News Events from Twitter[C]// ICWSM. 2012.

[10] PATHAK N,DELONG C,BANERJEE A,et al. Social topic models for community extraction[C]// The 2nd SNA-KDD workshop. 2008.

[11] ZHANG H Z,GILES C L,FOLEY H C,et al. Probabilistic community discovery using hierarchical latent gaussian mixture model[C]// AAAI. 2007:663-668.

[12] LI C L,SUN A X,DATTA A. Twevent:segment-based event detection from tweets[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM,2012:155-164.

[13] PAN C C,MITRA P. Event detection with spatial latent Dirichlet allocation[C]// Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. ACM,2011:349-358.

[14] YANG Y M,PIERCE T,CARBONELL J. A study of retrospective and on-line event detection[C]// Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1998:28-36.

[15] KLEINBERG J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery,2003,7(4):373-397.

[16] FUNG G P C,YU J X,YU P S,et al. Parameter free bursty events detection in text streams[C]// Proceedings of the 31st International Conference on Very Large Data Bases. VLDB Endowment,2005:181-192.

[17] ZHANG X M,CHEN X M,CHEN Y,et al. Event detection and popularity prediction in microblogging[J]. Neurocomputing,2015,149:1469-1480.

[18] CORNEY D,MARTIN C,GÖKER A. Two Sides to Every Story:Subjective Event Summarization of Sports Events using Twitter[C]// SoMuS@ICMR. 2014.

[19] BIAN J W,YANG Y,ZHANG H W,et al. Multimedia summarization for social events in microblog stream[J]. IEEE Transactions on Multimedia,2015,17(2):216-228.

[20] RUDRA K,GHOSH S,GANGULY N,et al. Extracting Situational Information from Microblogs during Disaster Events:a Classification-Summarization Approach[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM,2015:583-592.

[21] MELADIANOS P,NIKOLENTZOS G,ROUSSEAU F,et al. Degeneracy-based real-time sub-event detection in twitter stream[C]// Ninth International AAAI Conference on Web and Social Media. 2015:248-257.

[22] XU J J,LU T C. Seeing the Big Picture from Microblogs:Harnessing Social Signals for Visual Event Summarization[C]// Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM,2015:62-66.

[23] NICHOLS J,MAHMUD J,DREWS C. Summarizing sporting events using twitter[C]// Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces. ACM,2012:189-198.

[24] YAN R,WAN X J,LAPATA M,et al. Visualizing timelines:evolutionary summarization via iterative reinforcement between text and image streams[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM,2012:275-284.

[25] HUANG L F,HUANG L E. Optimized Event Storyline Generation based on Mixture-Event-Aspect Model[C]// EMNLP. 2013:726-735.

[26] CHAKRABARTI D,PUNERA K. Event Summarization Using Tweets[C]// ICWSM. Spain,July 2011:66-73.

[27] CHANG Y,WANG X H,MEI Q Z,et al. Towards Twitter context summarization with user influence models[C]// Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. ACM,2013:527-536.

[28] WANG D D,LI T,OGIHARA M. Generating Pictorial Storylines Via Minimum-Weight Connected Dominating Set Approximation in Multi-View Graphs[C]// AAAI. 2012.

[29] LEE P,LAKSHMANAN L V S,MILIOS E. CAST:A Context-Aware Story-Teller for Streaming Social Content[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM,2014:789-798.

[30] LIN C,LIN C,LI J X,et al. Generating event storylines from microblogs[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM,2012:175-184.

[31] ZHOU W B,SHEN C,LI T,et al. Generating textual storyline to improve situation awareness in disaster management[C]// 2014 IEEE 15th International Conference on Information Reuse and Integration (IRI). IEEE,2014:585-592.

[32] MCPARLANE P J,MCMINN A J,JOSE J M. Picture the scene:Visually Summarising Social Media Events[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM,2014:1459-1468.

[33] SCHINAS M,PAPADOPOULOS S,KOMPATSIARIS Y,et al. MGraph:multimodal event summarization in social media using topic models and graph-based ranking[J]. International Journal of Multimedia Information Retrieval,2016,5(1):51-69.

[34] LYNCH C,ARYAFAR K,ATTENBERG J. Images Don't Lie:Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank[J]. arXiv preprint arXiv:1511.06746,2015.

[35] ZHANG J F,GUO B,HAN Q,et al. CrowdStory:multi-layered event storyline generation with mobile crowdsourced data[C]// Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing:Adjunct. ACM,2016:237-240.

[36] **张佳凡,郭斌,路新江,等. 基于移动群智数据的城市热点事件感知方法[J]. 计算机科学,2015,42(6A):5-9.**

[37] **欧阳逸,郭斌,何萌,等. 微博事件感知与脉络呈现系统[J]. 浙江大学学报(工学版),2016,50(6):1176-1182.**