

文章编号: 2095-2163(2019)04-0327-05

中图分类号: TP391.41

文献标志码: A

社交网络公众意见分析研究

靳 锐, 张宏莉

(哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

摘要: 随着社交网络在世界范围内的蓬勃发展, 社交网络公众意见分析成为了一个重要的研究领域, 通过分析社交网络产生的数据研究公众意见。社交网络中公众意见在许多领域, 如政治、经济、商业、金融、贸易、公共政策实行等, 都扮演着重要的角色。然而, 相关的理论与技术并不成熟, 面临着很多挑战, 如不确定性问题的定量化计算、机器学习方法与自然语言处理在社交网络公众意见分析中的应用、公众意见分析理论体系的构建等。本文对社交网络公众意见分析的研究现状进行了综述。

关键词: 公众意见分析; 社交网络; 社会计算; 机器学习; 自然语言处理

Research on public opinion analysis in social networks

JIN Rui, ZHANG Hongli

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the worldwide proliferation of social networks, the public opinion analysis of data generated by social network has become an important field of research. The public opinion from social networks plays an important role in many fields such as politics, economy, commerce, finance, trade, public policy implementation and so on. However, the related theory and technology is still inadequate, it faces a lot of challenges, such as, the quantitative computing of the uncertainty problems, the application of machine learning methods and natural language processing in social network public opinion analysis, the construction of theoretical system about public opinion analysis. This paper reviews the current research status of public opinion analysis in social networks.

【Key words】 public opinion analysis; social networks; social computing; machine learning; natural language processing

0 引言

公众意见分析领域的研究由来已久, 自从政府形态出现以来, 公众意见分析的相关研究与应用即已日渐突显其重要的实用价值。政府在执行政策的时候, 如收税等, 需要了解公众对政府政策的态度, 这也是早期开展公众意见分析的宗旨意图, 不过系统的学术研究却是近代才见到、并得到发展的^[1-3]。

随着社交网络在世界范围内的广泛应用, 人们越来越习惯于通过社交网络进行公众意见表达。近年来, 存在于社交网络空间内的公众意见对政治选举、政策施行、金融交易、自然灾害避险、群体性事件的形成与爆发、经济发展、以及商业、贸易、市场营销等领域产生重大影响, 相关领域的网络公众意见成为重要的、甚至决定事件成败的情报信息, 为此准确把握公众意见的走向已然成为当下亟待获取有效成果的研究热点课题之一。

随着社交网络的发展, 社会计算^[4]思想开始出现, 这是社交网络数据分析的理论基础之一。2007

年底在哈佛大学举办了计算社会学研讨会, 2008年4月, 美国军方在亚利桑那州立大学举办了社会计算、行为建模和预测研讨会。在此基础上, 2009年Lazer等人^[5]在《Science》杂志上提出计算社会学的概念, 指出社交网络上的大量信息, 如博客、论坛、聊天、消费记录、电子邮件等, 都是对现实社会的人及组织行为的映射, 网络数据可用来分析个人和群体的行为模式, 标志着计算科学和社会科学在社交网络数据分析领域的研究出现交叉与融合, 社会计算的相关研究正成为世界范围内新的前沿科学探讨和应用焦点。

社会计算思想提出后, 受到学界研究人员的充分重视, 相关的研究人员依据该思想开始对社交网络信息进行研究与分析, 获得一系列的研究成果, 并将其陆续融入应用在计算机产业当中, 创造出可观社会价值^[6-8]。

在911恐怖事件的影响推动下, 2003年美国首创情报与安全信息学的概念, 其核心思想则立足于研究如何开发智能算法通过数据信息处理技术、安

基金项目: 国家自然科学基金(61202457, 61472108, 61402149); 国家重点基础研究发展计划(973计划)(2013CB329602)。

作者简介: 靳 锐(1976-), 男, 博士研究生, 主要研究方向: 信息安全; 张宏莉(1973-), 女, 博士, 教授, 博士生导师, 主要研究方向: 网络安全、网络测量和网络计算等。

收稿日期: 2017-01-05

哈尔滨工业大学主办 ◆ 科技创新与应用

全策略的集成等使情报采集和安全分析更加系统化、科学化,保障国际安全、国家安全、社会安全、商业安全和个人安全。美国亚利桑那大学关于国家社会安全问题做出了周密考查,进而开启了“情报与安全信息学(ISI)”^[6]的全面深入研究,卡内基梅隆大学也开展了公共卫生事件等领域的学术研讨。

当前,社会计算方法广泛用于社区发现、用户商业推荐分析、金融交易倾向分析、社交媒体挖掘,如社区与意见领袖发现、社交网络用户行为分析、网络观点与态度分析等^[7-9]。

公众意见分析领域的研究仍然处于发展初期阶段,理论体系还没有完全建立起来,在各个领域的实际应用还有很大的潜力有待挖掘。本文首先讨论了网络公众意见分析的相关研究与关键技术,然后综述当前研究现状,并指出该领域后续可能的研究方向。对此可得研究论述如下。

1 公众意见分析研究的背景

公众意见分析领域的研究,起源于政府机构为了执行政府政策而进行的民意调查,并在人类社会的各个时代一直发挥着重要的作用,著名公众意见研究专家 V.O. Key, Jr. 曾经说过一句很经典的话,“即使是最残酷的暴君也需要知道民众的想法,即使仅仅是为了更好地镇压他们。”^[10],这句话透彻地指出,在国家政府事务运行过程中,对公众意见进行调查分析与研究是不容回避的工作,而且可以尽量避免社会冲突,从而降低社会成本。

在社交网络时代到来后,民众开始使用社交网络进行公众意见的表达,大量的网络公众意见汇集在一起并相互交织,从而形成了一个带有鲜明感情色彩、意见多样化并相对统一的主流观点,而且往往代表普通民众的广泛意见,在此基础上则对社会中的其它人群产生重大影响^[7-9]。

1.1 公众意见分析研究的起源

公众意见研究和政府形式是同时出现的,最早期的公众意见表达形式是叛乱与起义反抗,在人类历史进程中,不断有农民起义事件发生,当大范围的农民起义涌现时,统治者可以得到了一个明确的信息:政府正在丧失民众支持。拒绝纳税是另外一个明显的表现,当国王看到自己派出去的纳税官员被杀死的时候,即会知晓民众正在反对其高压统治。之后,统治者们逐渐认识到对公众意见施以调查分析的重要性,开始针对公众意见进行调查与分析研究,并即时调整政府政策,而不是继续使矛盾趋于激

化。于是对公众意见进行处理研究的各种方法即已转入活跃发展期^[10]。

1931年, Droba 发表了一篇名为《用于测量公众意见的5种方法》的学术论文,在文中系统总结了5种对公众意见进行分析与度量的方法。一般认为,这篇文献是近代公众意见分析学术研究领域起点之一^[1]。1962年, Key 发表了另外一篇非常重要的学术论文,在该文章中阐述了公众意见与政治制度之间的关系^[2]。1973年, Mueller^[3]发表了题为《战争、总统、公众意见》的文章,进一步强调了公众意见在国家政治活动中的重要地位。

1.2 社交网络时代公众意见的社会角色

2010~2012年间,在非洲大陆与阿拉伯世界国家中,经历了一次举世瞩目的社会变革运动,史称 Arab Spring 运动^[7]。社交网络在此次社会变革中发挥了关键性作用,其显著的社会功能受到相关研究人员的广泛重视。研究可知,这些国家有着一些共性,均存在一定的社会矛盾,如贫富差距大、失业率高;主流媒体的社会覆盖面与作用不完备,社交网络媒体在一定程度上承担了信息传播的功能;民众利用社交网络渠道选择发布信息或个人观点,形成了强大的公众意见;网络公众意见在很大程度上决定了社会现实中的民意走向。有学者指出,这些国家利用社交网络技术所带来的便利,可以提高社会发展动力并降低社会成本,但同时也必须准备面对社交网络广泛使用带来的社会现实局面^[11-12]。在社交网络时代,网络公众意见似乎获得了更重要的社会角色。

1.3 社交网络时代公众意见的重要性

在当今的社交网络时代,公众意见发挥了重要作用,政府工作运转与政策实行等必然受到网络舆论监督;在其它领域,如金融交易、国际贸易、商业实体发展、市场营销等,网络公众意见也开始占据至关重要的位置。例如,当一个商业实体的产品在社交网络空间中获得了更多的负面评价,那么后果往往是,商业实体可能要面临着巨大的经济损失。在社交网络时代,网络公众意见的社会功能与作用是不能被忽视的,在某些条件下,还可能会起到决定性作用^[7,9,12]。

2 社交网络公众意见分析关键技术

2.1 社会计算

2.1.1 社会计算理论

(1) 社会计算思想。2009年, Lazer 等人在《Science》上发表了著名的学术观点,指出,在社交

媒体,如博客、论坛、聊天记录、日志文件、电子邮件等所发布的内容和相应的网络行为都是社会中个人或组织行为的映射。这个学术思想的提出具有划时代意义,为通过分析和计算社交网络中的用户行为数据来判断社会人群的倾向性、挖掘社区行为、分析公众意见等研究提供了理论依据^[5]。

(2)小世界理论。1998年,Watts和Strogatz在《Nature》上发表了一篇重要的学术论文,文中提出著名的“小世界理论”。该理论与社会计算思想类似,提供了在社交网络上信息传播六步距离的基本理论描述,这是分析社交网络信息传播规律的基础理论依据之一^[13]。

2.1.2 网络结构模型

自从Watts and Strogatz (1998)与Barabási and Albert (1999)研究社交网络结构建模开始,网络建模已经获得了长足进步。一个社交网络的结构示例如图1所示^[13-15]。

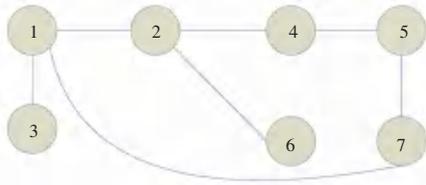


图1 一个社交网络的结构

Fig. 1 The structure of a social network

这里,针对研究中的网络数学模型可具体描述如下。

在一个网络中, V 和 E 表示节点集合与边集合,点的数量为 n ,边的数量为 m ,那么,矩阵 $A \in \{0,1\}_{n \times n}$ 可用来描述网络的邻接矩阵。矩阵中的每一个值(例如, $A_{ij} \in \{0,1\}$)表示在点 v_i 与 v_j 之间是否存在一个链接。点 v_i 与 v_j 之间边表示为 $e(v_i, v_j)$,并且当 $A_{ij} = 1$ 时,2个点 v_i 与 v_j 是邻接的。 N_i 表示所有与节点 v_i 邻接的点。直接与点 v_i 连接的点的数量即称为该点的度,用 d_i 来表示。在2个点(例如, v_i 与 v_j)最短的路径可称为测地线,2个节点之间测地线的跳数则称为测地线距离,表示为 $g(v_i, v_j)$ 。综上全部设计定义,汇总后结果见表1^[16]。

2.1.3 社交网络分析的技术指标

量化的分析方法已全面应用于社交网络分析中,相关的研究人员提出了众多技术指标的数学定义,用来辅助技术分析或表示社交网络特征,对此内容可做解析论述如下。

表1 命名法

Tab. 1 Nomenclature

符号	意义
A	一个网络的邻接矩阵
V	一个网络的点集合
E	一个网络的边集合
n	一个网络点集合的数量($n = V $)
m	一个网络的边集合的数量($m = E $)
v_i	节点 v_i
$e(v_i, v_j)$	点 v_i 与 v_j 之间的一条边
A_{ij}	如果在点 v_i 与 v_j 之间存在一条边,值为1;否则为0
N_i	节点 v_i 的邻接节点集合
d_j	节点 v_j ($d_j = N_j $)的度
<i>geodesic</i>	测地线,2个节点之间最短的路径
<i>geodesic distance</i>	测地线距离,最短路径的长度
$g(v_i, v_j)$	节点 v_i 与 v_j 之间测地线距离

2.1.3.1 聚类系数

社交网络具有强社区结构特性,在一个群体中的人倾向于与群体内的人交流的程度比群体外的人要大,这个特性可用聚类系数表示。

假设点 v_i 有 d_i 个邻接点,并且在这些邻接点之间有 k_i 条边,那么聚类系数计算公式可表示为:

$$C_i = \begin{cases} \frac{k_i}{d_i \times (d_i - 1)/2}, & d_i > 1; \\ 0, & d_i = 0, 1. \end{cases} \quad (1)$$

聚类系数用来计算与一个节点连接的友好节点数量密度。当一个社交网络具有强社区结构时,则往往会有较高的平均聚类系数。

2.1.3.2 中心性

(1)度中心性。度中心性的定义如下:

$$C_D(v_i) = d_i = \sum_j A_{ij}, \quad (2)$$

或者

$$C'_D(v_i) = d_i / (n - 1). \quad (3)$$

其中, n 是一个网络中节点的数量。

(2)紧密度中心性。紧密度中心性评测的是一个节点相对于其它节点的紧密度。其定义可写作如下数学形式:

$$D_{avg}(v_i) = \frac{1}{n - 1} \sum_{j \neq i} g(v_i, v_j). \quad (4)$$

其中, n 是节点的数量, $g(v_i, v_j)$ 表示的是节点 v_i 与 v_j 的测地线距离。

时下,还有其它度量中心性的指标,限于篇幅,这里从略。

2.1.3.3 连接强度

网络中点之间的连接强度有多种度量方法,较为常见的一种方法是使用节点之间的测地线距离的反序来度量。设连接强度为 $r(i, j)$,表示节点 v_i 与 v_j 的连接强度。连接强度的数学公式具体表示如下:

$$r(i, j) = d(i, j). \quad (5)$$

其中,节点 v_i 与 v_j (v_i 与 v_j 为邻接点)的测地线距离为 $d(i, j)$,节点 v_k 与 v_l (v_k 与 v_l 为邻接点)的测地线距离为 $d(k, l)$,如果 $d(i, j) > d(k, l)$,则表示边 $e(v_k, v_l)$ 的连接强度比边 $e(v_i, v_j)$ 的连接强度大。

2.1.4 社会计算方法

(1)确定性问题的计算,可以应用表1中的技术参数与网络数学模型进行计算,当前这种方法获得了大范围的应用,诸如可用于社区发现、网络用户行为分析等。

(2)不确定性问题。社会计算任务中,存在着诸多不确定性问题需要进行各类定量化的计算与分析,如公众事件信息熵的计算,大众心理压力指数计算、网络舆论攻击强度计算、网络社会性话题争议度计算、群体性行为复杂度计算、网络群体社会性攻击行为计算等等。

有相关的研究人员提出一种计算方法,可以用来解决这类不确定性研究中的量化计算问题^[17]。

2.2 社交网络数据挖掘

社交网络数据挖掘技术的任务是对社交网络中的关系型数据进行分析,往往不涉及文本内容,多是由于研究关注用户群体的社区关系,挖掘群体行为等,是当前的研究热点之一。

社交网络公众意见分析有一部分工作需要借助社交网络数据挖掘算法进行分析,如相似意见社区发现,相似或不同意见表达者的用户行为分析、公众意见的社交网络传播模式分析等。

2.3 自然语言处理

自然语言处理技术是网络文本分析的基础,在公众意见分析领域中占有重要地位,有很多工作需要借助自然语言技术对网络公众意见进行过滤与分析,因而成为应用基础性技术之一。

2.4 机器学习

社交网络公众意见分析研究中有很多任务需要进行数据分类处理或计算优化,如社交网络用户分类、公众事件分类、评论类别分类、情感信息分类等等,机器学习是一种有效的方法。机器学习是数据分析中强有力的模型工具,通过使用机器学习算法

可以获得普通方法无法得到的优异结果。

2.5 博弈论

公众意见与生俱来特性决定了其在社会政治活动、经济发展、社会群体行为组织等各项事务中的重要角色^[7,10,12]。因而就必然经常处于多个社会角色的博弈当中,并且最终发挥重要作用。

3 结束语

社交网络公众意见分析研究涉及多个领域的技术,如信息论、最大熵理论、社会计算、社交网络数据挖掘、自然语言处理、博弈论等。尤其在海量社交网络数据分析的背景下,网络公众意见分析研究对相关的研究人员提出了挑战。网络公众意见是社会各项事务能够合理运行的博弈因素之一,对其探索、及付诸应用则属于网络空间安全的研究范畴,在未来工作中可能会加入更多的社会性因素,其涉及的安全问题往往带有社会性。

参考文献

- [1] DROBA D D. Methods used for measuring public opinion [J]. American Journal of Sociology, 1931, 37(3): 410-423.
- [2] SHAPIRO R Y. Public opinion and American democracy [J]. American Journal of Sociology, 1963, 69(1): 982-1017.
- [3] MUELLER J E. War, presidents and public opinion [M]. New York: Wiley, 1973.
- [4] Wikipedia. Social computing [EB/OL]. [2017-01-05]. https://en.wikipedia.org/wiki/Social_computing.
- [5] LAZER D, PENTLAND A, ADAMIC L, et al. Computational social science [J]. Science, 2009, 323(5915): 721-723.
- [6] CHEN H, WANG F Y, ZENG D. Intelligence and security informatics for homeland security: Information, communication, and transportation [J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 329-341.
- [7] LERMAN K, GILDER A, DREDZE M, et al. Reading the markets: Forecasting public opinion of political candidates by news analysis [C] // Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, United Kingdom: ACM, 2008, 1: 473-480.
- [8] AKCORA C G, BAYIR M A, DEMIRBAS M, et al. Identifying breakpoints in public opinion [C] // 1st Workshop on Social Media Analytics (SOMA '10). Washington, DC, USA: ACM, 2010: 62-66.
- [9] LI Juan, ZHOU Xueguang, CHEN Bin. Research on analysis and monitoring of Internet Public Opinion [M] // Du Z. Proceedings of the 2012 International Conference of Modern Computer Science and Applications. Advances in Intelligent Systems and Computing. Berlin/ Heidelberg: Springer, 2013, 191: 449-453.
- [10] BROOKER R, SCHAEFER T. Public opinion in the 21st century: Let the people speak [M]. United States: Houghton Mifflin, 2005.
- [11] Wikipedia. Arab spring [EB/OL]. [2017-01-05]. https://en.wikipedia.org/wiki/Arab_Spring.
- [12] KHONDKER H H. Role of the new media in the Arab Spring [J].

[13] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684):440-442.

[14] BARABÁSI A, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439):509-512.

[15] CHAKRABARTI D, FALOUTSOS C. Graph mining: Laws, generators and algorithms[J]. ACM Computing Surveys, 2006, 38(1):2.

[16] TANG L, LIU H. Community detection and mining in social media [M]//HAN Jiawei, GETOOR L, WANG Wei, et al. Synthesis Lectures on Data Mining and Knowledge Discovery. California, USA: Morgan & Claypool Publishers,2010, 2(1):1-137.

[17] JIN R, ZHANG H L, ZHANG Y, et al. Calculation method of Chinese public event information entropy[J]. Journal of Software, 2016,27(11):2855-2869.

(上接第 326 页)

2 计算特征个数

具体识别人脸的方式简而言之就是将某个 Haar 特征矩形对整个图像遍历一遍,得到所有的特征值。通过积分图的方法计算 Haar 特征,并用这些特征值进行判断是否为人脸特征。而如何计算这些特征的数量如图 2 所示。例如,特征矩形大小为 2×1 , 图片大小为 12×12 。那么对于 2×1 这个 Haar 特征,可以水平移动 11 次,垂直移动 12 次。遍历整个图像后可以得到 $11 \times 12 = 132$ 个特征值。同时为了检测结果的准确性,对于同一个特征,可以沿着水平或竖直方向进行伸缩。就如同刚才的特征 2×1 , 可以进行水平放大变成 4×1 , 也可以进行竖直放大变成 2×2 等等。

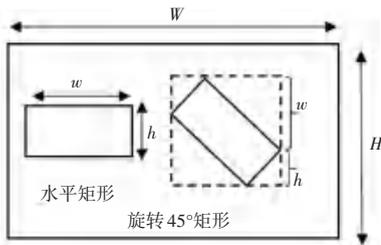


图 2 特征个数计算示意图

Fig. 2 Calculating diagram of characteristic number

3 实验测试

本文是基于复杂环境对人脸进行了动态的识别。经测试在戴帽子、遮住嘴巴、遮住半张脸、距离超过 2 m 等情况下皆能成功检测出人脸,结果如图 3 所示。



图 3 人脸识别结果

Fig. 3 Face recognition results

通过返还特征值的坐标来对目标进行跟踪及飞行器的姿态调整。将识别框的面积与图像面积做比得到一个面积占比,通过与阈值相比进行一个大致的水平复杂环境下的测距。通过返还的水平距离来控制飞行器不至于靠近或远离目标。如图 4 所示。



图 4 跟踪结果及返还坐标

Fig. 4 Tracking results and return coordinates

4 结束语

本文主要实现低空拍摄时基于人脸特征识别的自动跟拍系统,并且可以适应人体小角度或缓慢转体情况下的拍摄角度调整,亦可使用手动控制实现实时调整。可以取代当下手持式摄影机,避免画面抖动的同时可以更好地突破地面条件对摄影人员的移动限制,达到更大的拍摄自由度。

参考文献

[1] 李倩玉, 蒋建国, 齐美彬. 基于改进深层网络的人脸识别算法[J]. 电子学报, 2017, 45(3):619-625.

[2] 姜俊俊, 李明, 张宇. 基于视觉的四旋翼无人机目标跟踪系统设计[J]. 中国科技信息, 2016(14):65-66.

[3] 李永健. 基于机器视觉的四旋翼无人机定点着陆系统设计与实现[D]. 广州:华南理工大学,2015.

[4] 龙云露, 陈洋, 滕雄. 四旋翼飞行器姿态解算与滤波[J]. 计算机测量与控制, 2016, 24(10):194-197,201.

[5] 李航, 王耀力. 四旋翼飞行器中 PID 控制的优化[J]. 电子技术应用, 2017, 43(2):73-76.