文章编号：1003-0077(2018)02-0001-11

网络传播信息内容的可信度研究进展

吴连伟，饶 元，樊笑冰，杨 浩

（西安交通大学 软件学院 社会智能与复杂数据处理实验室，陕西 西安 710049）

摘 要：网络中存在着大量的谣言、偏激和虚假信息，这对网络信息的质量、可信度以及舆情的产生与发展趋势具有严重的负面影响。为实现信息可信度的准确判断与高效度量，该文在大量已有最新研究成果与文献的基础上，将不可信信息分为极端突发事件信息、网络偏激信息、网络谣言、虚假信息、误报信息和垃圾信息等类型，并分别针对这些类型信息从分类定义、内容特征描述、可信度建模以及可信度评测等四个方面进行研究综述，从而为网络传播中信息内容的可信度分析与度量研究奠定坚实基础。最后，进一步对信息可信度研究的发展方向进行展望。

关键词：社交网络；信息可信度；可信度计算；信息特征抽取

中图分类号：TP391 **文献标识码**：A

A Study on the Credibility of Information Spreaded on Social Networks

WU Lianwei, RAO Yuan, FAN Xiaobing, YANG Hao

(Lab of Social Intelligence&Complex Data Processing, School of Software,
Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China)

Abstract: There are a large number of rumors, extreme and fake news in network, which will reduce quality of information, destroy the credible atmosphere of internet, and produce the serious negative effects for the occurrence and development of public opinion. To measure the credibility of information, the paper divides incredibility contents into such types of extreme emergency information, network extreme information, network rumors, misinformation, disinformation, spam information and so on. And the information contents are studied from the following aspects: concept, content features' description, credibility modeling and credibility evaluation, which provides a solid foundation for credibility analysis and measurement of information content in social networks. Finally, we further analyze the directions of development in current research of credibility of information.

Key words: social network; information credibility; credibility calculating; information feature extraction

0 概述

基于用户生成内容的 Web 2.0 社交网络平台极大地促进了信息内容的生成、传播与快速增长，在享受信息的快速获取与传播共享便利的同时，网络中散布着大量的谣言、偏激和虚假信息。在线博客中存在着许多偏激和虚假的内容，微博也被大量的垃圾和谣言信息严重污染，甚至在线新闻媒体也被大量不可靠且没有被证实的新闻所充斥[1]，这种现象直接影响到了主流媒体。Howell[2]将海量数字化虚假内容信息列为影响现代社会发展的重大威胁之一。

Gupta[3]的研究结果表明：在 Twitter 中有将近 52%的内容是确定可信的、35%的内容是大致可信的、13%的内容是确定不可信的。不可信信息将极大地渲染消极和负面的社会情绪，不仅影响社会和谐，而且也会影响国家安全与政治生态。例如，澎湃新闻曾在 2016 年 1 月 4 日发布“江西九江市浔阳区发生 6.9 级地震”的假新闻所引起的社会恐慌，2016 年英国脱欧和美国总统大选事件中所引爆的媒体信任危机，许多类似的新闻使人们开始深刻地

收稿日期：2017-05-23 **定稿日期**：2017-06-07

基金项目：国家社科基金(13&ZD177)；国家自然科学基金(61602370)；陕西省协同创新计划(2015XT-21)；陕西省科技重点项目(2013K06-20)；教育部“云数融合”基金项目(2017B00030)；中央高校基本科研业务费(zdyf2017006)

意识到“阴谋论、假新闻、极端的感情抒发”的信息在网络传播中给社会所传递的负面影响。如何在复杂网络环境下快速识别出信息的真伪，以确保网络中传播信息的真实性与可信性，并对传播信息内容的可信度进行度量，已成为目前学术界、工业界和政府机构共同关注且亟需解决的重要问题。

为了解决上述问题，本文在文献调研分析与总结的基础上，从信息可信性与不可信的特征出发，针对不同类型的信息内容特征进行识别、抽取与比较，在此基础上，系统地梳理和分析当前主要的信息可信度建模与评测方法，为信息内容的可信度分析与研究奠定基础。

1 网络信息可信性分类与可信度定义

1.1 网络信息可信性分类

从可信的角度看，信息可以分为可信与不可信两大类，除了能够证明信息本身的真实性、科学性、客观性以及完整性以外的信息，其余信息均可称为不可信信息。而在网络中传播的这些不可信信息本身也存在着一些明显的差异，根据这些差异将不可信信息进一步归纳为：极端突发事件下的模糊信息、网络偏激信息、网络普通虚假信息、网络谣言、误报信息与垃圾信息等六种类型。

其中，极端突发事件是指具备严重危害性的不可预知的突发性事件，特别是指由于自然灾害、事故灾难、公共卫生事故以及社会安全事件等方面突发且不会重现的事件[4]，由于极端突发事件除了具有爆炸性、不可重复性和严重危害性等特征外，还具有极强的模糊性，从而为虚假信息的快速传播提供了空间；网络偏激信息是指夸大或贬低事实、断章取义或者是以偏概全的信息，这类信息中往往融入了个人的极端情感；网络普通虚假信息包括恶意造假或蓄意欺骗的信息；网络谣言指在网络中传播的一个存在争议或者事实有待检验的信息陈述[5]；误报信息则是由于工作失误而错报的信息，产生的原因包括录入失误、疏忽或者专业能力差等[6]；垃圾信息指与用户无关且无价值、不被关注的信息，也包括失去时效的过时信息等。根据上述定义，表1从特点、发布者、目的性、危害性和可信度等特征的差异对信息进行了比较。

表1 六类不可信信息的特征对比表

信息类型	特点	发布者	目的性	危害性	可信度
极端突发事件下的模糊信息	爆炸性/模糊性/不可重复性/严重危害性	明确	有	极强	未被证实
网络偏激信息	夸张性/失真性/煽动性	明确	有	很强	不可信
网络谣言	待证实性	不明确	不确定	强	未被证实
网络普通虚假信息	蓄意欺骗性	不明确	有	较强	不可信
误报信息	误导性/无意性	明确	无	一般	不可信
垃圾信息	无用性	不明确	有	弱	未被证实

由于信息传播过程中的用户社交网络与兴趣网络交织融合，信息内容的组织形式具有多样性且具有跨媒体特征，使得不可信信息识别的复杂性程度大幅提高，这直接影响到了网络信息传播过程中预测与引导策略的有效性。因此，本文提出了信息可信度概念来对所有信息的可信程度进行统一度量。

1.2 信息可信度的定义

信息可信度是评价信息内容质量的一种关键性指标，它与信息在网络中传播的核心要素相关，即与信息内容、话题、信息传播者和传播媒介及信息接受者等特征相关，因此，可用如下五元组来形式化地定义为：

$$\mathrm{IC} = < C, T, P, M, R > \tag{1}$$

其中，IC表示信息的可信度，C、T、P、M和R分别表示了信息内容、话题、信息传播者、传播媒介及信息接受者的特征集合，该模型所描述的信息在网络中的传播过程如图1所示。尽管该模型在传播要素与内容可信度度量之间建立了一种联系，但并没有解决如何选择不同的特征维度并进行有效的评估测量这一关键问题。

West[7]认为可信度是信息接受者对信源或传播媒介品质的一种主观感受，这种品质不管内容如

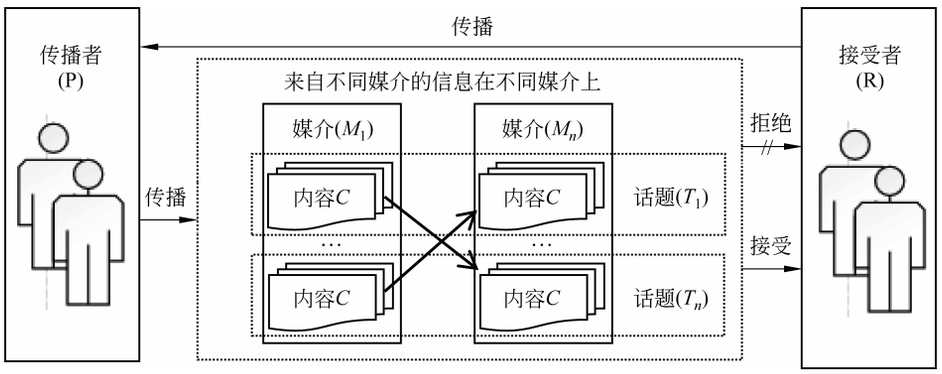

图 1　信息可信度 IC 模型中网络信息的传播过程示意图

何，受众都能毫无保留地对其信赖。而 Fogg[8] 进一步强调受众对信息传播者的信任主要来自于个人特质和信息来源可信程度特征的主观测量。周东浩[9] 将微博看作一个融合了社交图谱和兴趣图谱的关系网络，其中节点之间的结构相似度以及用户对信息的传播兴趣对信息传播概率的影响最大。在此基础上，Metzger[10] 认为信息可信度不仅包括了对信息源的专业性、吸引力以及可信赖性的主观信任度，同时也包括了信息内容质量、精确度的客观评判。而方滨兴等[11] 进一步将信息内容、人员以及行为动机的识别作为信息内容安全判断与控制的核心要素，且通过行为动机的分析来客观地反映人员的主观行为。为了更好地分析信息内容的质量，Miyamori[12] 开发了一个 WISDOM 系统，并从信息的内容、传播者、表面特征和社会价值等四个方面来度量信息的可信度。Castillo[13] 提出了一个基于多级社交网络的信息内容可信度评价指标体系，其中一级特征指标包括信息内容、接收者、话题和传播等四项，二级指标 74 项，为信息内容可信度测量奠定了重要的分析基础。

综上，考虑到信息在传播过程中主观与客观因素对信息可信度测量的影响，为了更有效地建立信息的可信度评价模型，需要进一步深入地对可信信息以及上述六种不可信信息的特征进行分析和量度，因此，本文从 IC 模型的五个维度出发，对信息在传播过程中的可信度特征进行研究与分析。

2　信息可信度特征与指标体系的建立

2.1　信息可信度特征指标

由于网络结构与人们的行为倾向对信息传播会产生非常大的不确定性影响，且传播内容的可信度与网络的结构特征、个体行为以及信息传播的初始状态之间存在着密切关系。同时，在线文本的有用性与价值性以及社交文本（如 Tweets）内容中的 URL、关注数、转发数和内容长度均可以作为信息可信度评估的最佳指标[14]。Metzger[15] 认为信息可视化模式比信息内容以及来源对可信度评估结果的影响更大。而 Lipshultz[16] 却认为在构建公众信任时的参与度、完整性以及目的性才是关键，他利用 TweetLevel 工具对 Twitter 中的信任进行了度量，结果表明网络中的个体愿意信任那些和自己建立联系的用户所发表的、且具有一定转发数量与引用数的信息内容。Castillo[17] 则认为 Tweet 中信息的可信度与信息源、主题、作者的声誉、写作风格、信息传播以及与时间相关特征紧密相关；徐静[18] 针对 Web 信息可信度的时效性、权威性、影响力和关注度四个特征进行验证，并提出了一个多维度加权结合的可信度计算方法。Hardalov[1] 则进一步提出了一个基于语言学（主要指 n-gram）、可信性（大小写、发音、拼写与情感）以及语义（Embedding and DBPedia Data）三者融合的富特征（20 条特征）条件下，语言无关的自动化的英文信息可信度识别方法，实验结果表明在特定的测试集下，内容可信度的识别率竟高达 99.36%。

目前，信息内容的可信度研究主要集中在对信息特征的分析与定义以及基于特征的可信度检测上，本文将 IC 模型中的五个维度作为信息内容可信度特征分析的一级指标，在此基础上，将该指标下所涉及的子特征细化为二级指标，并将具体可度量的细化特征作为三级指标，从而构建了一个信息内容可信度特征分析的指标体系，如表 2 所示。

表 2 信息内容可信度特征指标体系表

一级指标	二级指标	三级指标
信息传播者	传播者影响力	粉丝量/关注量/主页浏览量/消息浏览量/点赞量/用户类型/发送消息量/发送频率/粉丝大 V 量/好友大 V 量
	传播者基本特征	登录名/昵称/手机号/真实姓名/所在地/注册地/常用地/性别/性取向/感情状况/生日/血型/简介/被验证/等级/类型/注册时间/个人网站地址/教育信息/职业信息
	传播者兴趣偏好	转发(评论/浏览)频率/转发(评论/浏览)主题类型/用户简介/兴趣标签
内容	内容平台特征	点赞量/评论量/浏览量/桌面端内容量(和比例)/移动端内容量(和比例)/内容地理位置
	内容语法特征	语法是否正确/文字大小写/标点符号/人称代词的使用/内容长度/包含 URL 数量/感叹词(问号)数量/标签数量/积极(消极)情感表情/脏话量/是否包含多媒体
	内容语义特征	内容关键词特征/情感词特征/内容所属主题
话题	话题基本特征	话题名称/话题类别/话题阅读量/话题关注量/话题内信息数量/话题热度/话题摘要
	话题用户特征	话题主持人/大 V 用户量/用户性别比/用户兴趣标签/用户常用地/用户教育信息/用户职业信息
媒介	内容来源媒介	网址是否被备案/被熟知/真伪/被认证;界面布局是否规整/符合大众审美/可读性
	特征	广告数量/广告真实度/网址内其他信息可信度/曝光度
	媒介平台特征	平台的日(月)活跃用户数/平台用户总数/平台影响力/平台移动端(桌面端)用户数量和比例
	媒介内容特征	内容语法特征/语义特征/来源用户可信度
	媒介信息传播	传播子树结构/传播平均深度/最大深度/传播路径/内容分散结构/内容传递结构/强连通分量
	网络特征	网络密度/平均聚集系数/平均路径长度/度分布/幂律分布/匹配模式
接受者	用户感知特征	内容包含的视频(图片)/字体大小(加粗高亮)等内容识别/标点符号数量/段落首句内容/内容语义主题类别/情感词数量/标签
	兴趣偏好特征	接受者转发(评论和浏览)频率/转发(评论和浏览)主题类型/用户简介/兴趣标签/接受者关注用户类型/接受者粉丝用户类型
	评论/转发特征	转发(评论)内容语法语义特征/评论内容与信息情感正反比较/评论是否支持内容

2.2 不可信信息特征描述与定义

在上述特征分析的基础上,为了对不同类型的不可信信息的可信程度进行分析,下面对这些类型的信息可信特征以及度量方法进行分析。

2.2.1 极端突发事件下模糊信息特征描述

突发事件是指突然发生的超常规的自然灾害、事故灾难、公共卫生事件和社会安全事件的总称。它具有突发性、无重复性、无章可循但又能产生重大次生影响的特征。特别是突发事件产生后,相关信息的模糊性,使得心理处于恐慌状态的人群对信息的渴求强烈,从而成为了各类不可信信息滋生与传播的土壤[19]。例如,在日本福岛核泄露事件发生后,Thomson[20]研究发现与这场危机现场距离越近的传播者(即传播者可信度中的地理位置)越能增加共享信息的可信度。Mendoza[21]定义了基于词共现的主题(话题)抽取方法,通过抽取包括信息、内容(内容在平台上的特征、语法特征、语义特征)、信息媒介(媒介平台特征)和接受者(评论转发特征)等一系列特征研究极端突发事件信息可信度,尤其是在新闻可信度识别领域获得了较高的准确度。薛传业[22]从信息内容、传播来源(内容来源媒介特征)、传播渠道(媒介平台特征)、传播者影响力、传播者可信度、网络依赖性等方面对突发事件中信息可信度进行研究,发现网络的使用和信息的完整性对突发事件信息的可信度影响不显著,但其他因素对突发事件中信息的可信度则存在明显影响。

2.2.2 网络偏激信息特征描述

网络偏激信息本质上是人们对现实社会认知和情绪的反映,它包括了夸大事实信息、断章取义和以偏概全等言论,并影响公众对社会生活审视的立场和价值判断。网络偏激信息往往会带来严重的煽动效应,并对个人及社会造成严重的不良影响。Lewandowsky[23]指出:人们对事件的信任源自于其

大脑中所形成的未被大众质疑的信仰和观念。特别是当信息中包含与用户兴趣偏好一致的信仰与观念时，即使信息内容中存在着夸大或带有某些偏激的言论，人们可能也会不加验证地选择接受。另外，由于偏激信息整体言论是可信的，只有部分言论是不可信的，导致了仅从文本内容语法特征以及浅显的语义特征两个方面将无法完成对偏激信息的判断与识别，所以目前的研究采用了深度学习技术从信息内容本身的深层语义理解方向进行分析与研究，进而识别与判定偏激信息。

2.2.3　网络谣言特征描述

谣言是一种在人群之间私下流传，对公众感兴趣的事物、事件或问题未经证实的阐述或诠释[24]。因此，谣言往往是一个有争议的与事实有待检验的陈述[5]。Turner[25]指出，通过是否有可靠的消息来源、是否是人们所预期与希望知道的信息以及听起来似乎是真的这三个方面的特征可以有效验证谣言的真实性。Bessi[26]在研究 Facebook 中的谣言信息时，发现用户更倾向于和自己观点相同的好友（话题的用户特征）进行交流与传播。Hamidian[27]利用了包括时间、标签、URL 和转发等特征和内容的一元、二元模型的语法特征以及 100 维的 Twitter 潜在语义向量（TLV）特征进行谣言检测。Yang[28]收集了大量被新浪官方正式辟谣的新浪微博谣言数据集，进行了基于时间与地理位置以及客户端程序发送微博情况在内的 19 种信息内容特征的谣言检测与识别。Liu[29]将信息来源媒介的可信度、媒介身份和媒介平台等特征相结合，并采用实时性算法来实现谣言信息的实时检测。周东浩[9]利用传播者与接受者的兴趣偏好相似度特征来研究信息的传播，并指出传播者与接受者的兴趣偏好相似度越高，接受者越倾向于接受传播者所传播的信息，且信息是否契合用户的兴趣偏好也是决定用户是否接受并传播的重要因素。

2.2.4　网络普通虚假信息特征描述

虚假信息是指“故意制造的不真实信息”，它具有传播速度快、传播范围广和传播结构呈散布型网状结构的特点。Fallis[30]将虚假信息的主要特点概括为：它是一款精心策划的产品，从技术上看是一个复杂的欺骗过程，但虚假信息的来源并不一定是虚假的，也就是说通过虚假信息的来源特征并不能准确评判虚假信息。因此，仅利用信息的来源则无法区分虚假信息与其他信息，同时，虚假信息的传播对象往往针对特定的人群或组织。Karlova[6]从真实性、准确性、完整性、时效性和欺骗性五个维度测量与区分误报信息、虚假信息和政治宣传信息，并指出这三种信息在本质上只有虚假信息带有蓄意欺骗性质。Kumar[31]从认知心理学角度，采用信息传播所涉及的信息的一致性、相关的消息、信息接受者的总体可接受性和消息来源的可靠性等四种相关的欺骗线索来分析和评估社交媒体上误报信息、虚假信息和政治宣传信息的差异，并提出了阻止虚假信息传播的不同解决方案。

2.2.5　误报信息特征描述

误报信息是指错误的或误导性的信息，它常常具有被官方或影响力高的人员来发布、扩散传播迅速、存活时间短和较易被证实的特点。Ratkiewicz[5]研究了 2010 年美国总统竞选活动中在 Twitter 上的选举造势的内容数据，发现具有很强传播感染性的误报信息用错误观念影响了民众的支持倾向，并对投票选举结果产生了严重影响。Karlova[32]认为误报信息很难检测，但是采用基于群体智慧的众包方式则可以有效地对其进行识别和控制。Neys[33]和 Lewandowsky[23]认为误报信息的存在是极其危险的，需要对网络中的误报信息进行检测、识别，并尽可能使其在早期得到及时的预防与控制。表 3 将信息、误报信息与虚假信息从真实性、完整性、时效性和欺骗性这四个维度进行比较分析，其中误报信息和虚假信息均不真实，但只有虚假信息具有欺骗性。

表 3　信息、误报信息与虚假信息比较表

	信息	误报信息	虚假信息
真实性	Y	N	N
完整性	Y/N	Y/N	Y/N
时效性	Y	Y/N	Y/N
欺骗性	N	N	Y

注：Y＝Yes，N＝No；Y/N＝可能是 Yes 也可能是 No，取决于信息的本身和时间

2.2.6　垃圾信息特征描述

网络中垃圾信息通常是指由网络水军创造的随意且无用的信息，以及各种无效广告等与用户无关的信息。由于网络垃圾信息无用且干扰了正常的信息获取，因此，用户往往不会主动传播这类信息，同时也希望识别并过滤掉这些信息对人们产生的负面影响。Ratkiewicz[5]利用 meme 的节点数量、边的数量、平均度、平均强度、最大连接组件的平均边权

重、最大最小出入度以及六类情绪统计维度等共18种特征对网络中的垃圾信息进行了分类。Wang[34]利用基于用户粉丝和关注的有向图特征以及Tweets内容本身特征如重复Tweets、评论与@用户(接受者评论/转发特征)、URL和话题等四种特征对Twitter进行了垃圾信息检测。Tan[35]抽取了网站评论信息中的垃圾内容与URL之间的连接关系,并通过社交图谱定义了垃圾信息散布者的节点特征、分享信息的URL和用户链接图谱的节点度、边特征等在内的九个相关特征,从而为垃圾信息的识别与过滤奠定了实现基础。

综上,通过对上述六类信息的可信度特征描述与分析,本文进一步将这六种不可信信息的特征指标进行对比,形成的整个特征指标体系如表4所示。

表中的对勾号(√)代表"存在",比如第一个(√),表示在"极端突发事件信息"中存在着"传播者影响力"特征。

表4 六类不可信信息的特征比较表

特征抽取指标 \ 不可信信息类型		极端突发事件信息	网络偏激信息	网络谣言	网络普通虚假信息	误报信息	垃圾信息
一级指标	二级指标						
信息传播者	传播者影响力	√	√	√	√	√	√
	传播者可信度	√		√	√	√	√
	传播者兴趣偏好		√	√			
内容	内容平台特征	√	√	√	√	√	√
	内容语法特征	√	√	√	√		√
	内容语义特征	√	√	√	√		√
传播	话题基本特征	√	√		√		
	话题用户特征	√	√		√		
媒介	内容来源媒介特征	√		√	√	√	√
	媒介平台特征	√		√		√	
	媒介内容特征					√	
	媒介信息传播网络特征				√		√
接受者	用户感知特征	√	√	√	√	√	
	兴趣偏好特征		√	√	√		
	评论/转发特征	√	√	√	√	√	√

3 内容可信度分析建模

根据IC信息可信度模型以及信息可信特征指标体系,如何建立基于特征的信息内容可信度分析与评估模型则成为了关键。图2显示了网络信息可信度分析的基本过程,即主要包括信息获取、话题识别与跟踪、特征抽取、可信度模型的建立与分析以及计算结果的评估。在此基础上,本文从传统的信息可信度基本模型、基于浅语义特征的可信度模型、基于媒体融合的深层语义理解研究以及其他相关模型[38]等方面来分别进行介绍。

3.1 信息可信度基本模型

Fogg[36]提出一个评判互联网信息可信度过程的"关注—释义"模型,该模型认为人们对信息往往是先关注后释义,即:当评判在线信息的可信度时,人们首先会观察到一些需要关注的信息要素,然后再对这些元素进行解释和释义。其中,有五个关键因素直接影响到了"关注"的程度:用户的参与程度(即审查网页内容的动机或能力)、网站的话题(新闻或娱乐)、用户的任务动机(寻找信息)、用户的经验

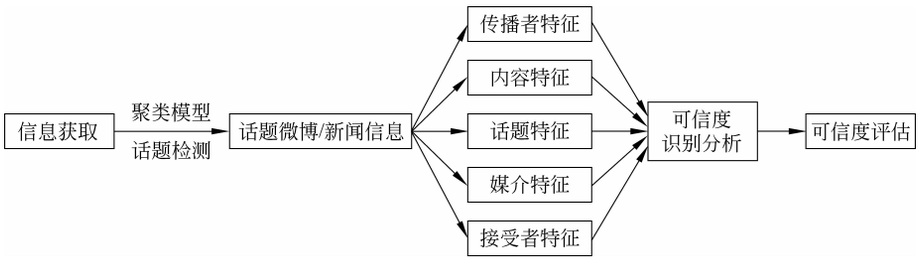

图 2　信息内容可信度分析过程示意图

(新手或专家)以及个体差异(认知的需求、学习方式)。而在“释义”阶段,主要的影响因素包括用户的期望假设(文化、过去经历等)、用户能力与知识以及外部环境等因素。Sunder[37]提出了由信息传播的媒介、代理、交互性和可操纵性等四个要素组成的MAIN模型。其中,信息媒介的差异会采用不同的方式将信息传播给不同的受众;代理则表示媒介的具体代表;交互性反映了人们的信息交流方式,不同的互动方式所采用的启发式评测规则也会存在差异;可操纵性反映了获取信息的操作方式,例如,网站的层次结构、大纲和超链接的设计会直接影响到人们获取信息的难易程度。通过分析上述四个要素来获取对信息质量评估的启发式判断规则。

高雅[38]在新闻信息可信度评测要素研究的基础之上,结合微博的传播学特征和社会网络结构特征,在多级信息分层传播条件下,建立了一个微博新闻事件信息可信度模型,即一级传播提供了对事件信息可信度分析的基准,而多级多次传播则为信息可信度分析和度量提供了一个基于网络节点特征以及传播动力学特征的新视角,并利用因子分析法和层次分析法,采用主客观相结合的方式来确定指标体系的权重,实现了微博事件信息的可信度评判。郭国庆[39]在研究消费者在线评论可信度的影响时,在霍夫兰德信息传播模型的基础上,从信源、信息内容、接受者以及社会影响这四个角度对在线评论内容的可信度进行研究,特别是将在线评论作为一个重要社会影响因素,提出了一个在线评论的可信度影响模型。Lucassell[40]利用信息语义(semantics)、表面特征(surface)和信息源(source)三者组成的3S模型来判断信息的可信度,并展示了信任判断的形成过程,通过实验验证了该模型具有较好的信息可信度识别率。

Wu[41]利用新浪微博官方公布的谣言库建立了网络信息可信度评估平台(NICE),并用来评估社交媒体上未被检验可信性的信息可信度。该平台首先从用户特征、内容特征、时间特征和评论特征四个方面对事件信息进行可信度表示(the credibility representation),事件可信度表示如式(2)所示。

$$r^{e_i}=\frac{1}{n_{e_i}}\sum_{m_j^{e_i}\in M^{e_i}}(T_t^{e_i}+C_j^{e_i}+B_j^{e_i})\,u_j^{e_i},\qquad(2)$$

其中,r^{e_i}表示第 i 个事件e_i的可信度,n_{e_i}指事件e_i中所包含的信息数量,M^{e_i}指事件e_i中信息的集合,$m_j^{e_i}$指事件e_i所包含的一条信息 j,$u_j^{e_i}$是用户发送或转发事件e_i的相关信息向量,$B_j^{e_i}$是用户是否转发事件e_i的行为矩阵,$C_j^{e_i}$是评论中对事件e_i的态度观点矩阵,$T_t^{e_i}$是事件发生后,在一定的时间间隔内的操作矩阵。

随后,利用常规的逻辑回归分类算法将信息划分为谣言信息和非谣言信息,如图3所示。基于该思路,NICE模型在评估信息可信度和检测谣言方面具有了较好的性能。

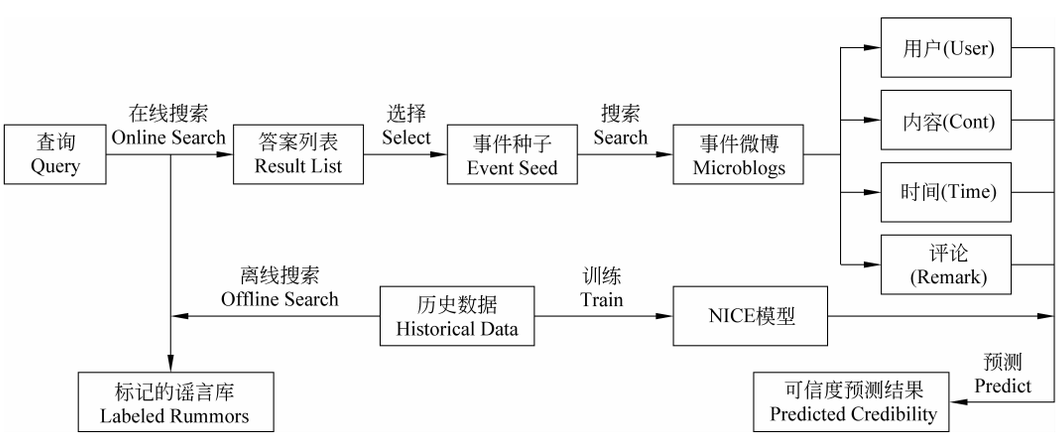

图 3　NICE 平台整体结构示意图

3.2 基于浅语义特征的可信度模型

Gupta[42]在对信息可信度特征进行抽取的基础上,采用有监督机器学习的RankSVM方法对微博信息进行可信度评分排序;同时,利用基于PageRank和事件图相结合的算法来计算Twitter内容的可信度,并获得较高的准确率[43]。Chang[44]则利用谣言特征所建立的五种结构和时间特征规则来对Twitter中的政治谣言以及极端用户进行识别和检测。为了进一步检测具有多信息来源的网络信息内容可信度,Pasternack[45]提出了一个LCA概率模型,该模型利用来自真实世界的两个无监督数据集和两个半监督的数据集,通过对内容的语义扩展来获取更有价值的可信度特征因素,并提高了可信度判别的准确率。而Unankard[46]利用基于文本相似度和位置相关性聚类模型对Tweets的内容进行聚类,从而获取更多信用语义特征,用来对Twitter中所发生的事件的信用特征进行评估。而Kyoto大学研发的基于聚类模型的WISDOM[47]信息可信度综合评估系统,则通过对搜索引擎搜索结果中的一个或多个特征属性进行内容聚类,如按照信息发送者、主要观点和对立观点等方面进行聚类,为用户提供了多个角度的信息可信度评价。

3.3 基于媒体融合的深层语义理解研究可信度

深度学习技术快速发展,使得人们从信息内容表层浅显语义研究过渡到了深层语义理解问题的研究上。例如,由于网络偏激信息中存在一部分夸大的言论或者是贬低的事实,而另一部分信息往往是可信的,从简单的浅层语义特征无法对该类信息做出准确的判断与识别,这就需要进一步采用深层语义理解以提高对偏激信息可信度的识别准确率。而Takahashi抽取了谣言内容的日期、地点、人物和组织等信息特征,并对这些特征信息进行过滤,实现谣言检测,利用浅层特征对谣言的识别率仅维持在34%左右。而Hamidian[27]加入了深层语义特征后,利用谣言内容的时间、标签、URL、转发等Twitter信息以及特定网络特征和内容的一元模型、二元模型等语法特征信息,首次利用了100维的Twitter潜在向量(TLV)的语义特征进行谣言检测,使得谣言的识别精确率提高到了97.2%。

另一方面,网络中的信息越来越呈现出多媒体融合的新特征,大量的图片、视频和音频等多媒体信息与本文信息相互结合且相互影响,通过不同媒体信息的可信特征的抽取与识别,以及语义特征的补充,例如,网络中常说到的"有图有真相",就是将图片中的可信特征与文本的可信特征进行了结合,提高了内容可信度的识别准确率,但同时也增加了计算处理的复杂程度。其中,如何对信息中存在的多媒体内容的真实性与可信性进行度量,以及多媒体信息内容与文本内容之间的可信度特征的融合策略等方面仍然存在着关键性的挑战。

4 信息内容可信度测量评价方法

通过对信息可信度特征的抽取与信息可信度建模,可以对信息内容可信度进行计算和测量,但是如何评价测量结果的质量好坏与有效性,则是一个关键。一般地,可信度测量评价分为可信度的客观评测和用户感知评测两方面,其中,客观评测是指通过常规量化的客观指标评测信息内容的可信度,这些指标包括真正率(TP-Rate)、真负率(TN-Rate)、假正率(FP-Rate)、假负率(FN-Rate)、准确率(Accuracy Rate)、精确率(Precision Rate)、召回率(Recall Rate)和F1度量(F1-measure)等。而用户感知测评主要是从用户自身感受的角度所形成的Checklist标准,其中包括时效性、权威性、客观性、准确性以及信息覆盖范围等。这两个方面从不同的侧面和角度对信息内容的可信度进行了评测。此外,本文还对在线信息内容可信度的实时性测评以及基于实证的测评方法进行了介绍。

4.1 内容可信度客观评测

内容可信度的客观评测主要通过常规量化的客观指标进行评测。Castillo[13]对文本特征子集、用户特征子集、传播特征子集和最常见特征子集等四个特征子集从真正率、假正率、精确率、召回率和F1度量等客观指标进行内容可信度评估。Hardalov[1]利用信息内容的语法、内容和语义等特征,在三种不同的数据集上验证虚假信息检测的准确率,并在此基础上对信息可信度进行评测,结果表明语法特征比内容特征的评价准确率更高,而基于语义特征的评价准确率最高。Zhang[48]使用精确率作为唯一客观评估指标,利用GPPooled Brown、GPPooled Bow和Majority三种方法对Tweets的内容进行了谣言检测,发现GPPooled Bow法的精确度明显高于其他两种算法。Liu[29]利用准确率从Tweets数量和时间上对文中提到的四种方法进行了评估,发现特定的算法

组合将会在数量和时间上获得最佳的准确度。

4.2　内容可信度的用户感知评测

客观指标是从内容的基本物理特征出发，来研究信息内容的可信度，缺少用户主体自身对信息可信度的感受。因此，美国图书馆协会（ALA）主导的可信度评测系统则从信息的权威性、时效性、客观性、公开性、准确性以及信息覆盖范围等标准，并采用 Checklist 方法来对网站信息的质量进行自动评估。Gupta[2]在基于半监督排序模型的基础上，开发了一个实时可信度评分的浏览器插件 TweetCred，它可以利用用户打分和用户问卷调查两种反馈方式获取用户对信息的主观评价，并实现对信息内容可信度质量的度量。Rieh[49]认为信息内容的可信度不是一个离散的评估事件，而是一个持续不断的迭代过程，因此，通过主观判断来实现信息的真实性、可靠性、准确性和完整性的分类，以及信息可信度与信息质量的评价，这也是一个动态的过程。综上，本文根据相关文献[50]总结出与用户感知评测相关的指标，如表 5 所示。

表 5　信息可信度用户感知指标

评测指标	评测内容
相关性	内容与用户需要相关性强 信息内容的深度符合需要 信息内容的广度符合需要
客观性	观点具有客观性 立场中立，不掺杂个人偏见 结论有理论、事实作为支撑的依据
准确性	文字没有语法和逻辑错误 表格、图示等辅助描述适当、准确 文字意思表达清晰 观点、理论正确
信息覆盖范围	信息内容详实 信息内容规模具体
可靠性	信息来源可靠 信息渠道可靠 用户感官可靠
影响力	被引用次数多 在本领域中具有很强的影响力
时效性	发布的时间近 信息表述的内容新 理论、数据、引文等时效性强
可证实性	数据、论据、引文来源明确 结论科学，且可以被证实 研究推导方法可以被证明
通用性	文本格式通用性强 媒体形式通用性强

4.3　信息可信度实时自动检测

在线网络信息往往具有极强的时效性特征，特别是为了消除由于大量不准确或者虚假信息的广泛传播，对社交网络以及社会和谐所造成的危害与负面的影响，如何对信息内容可信度进行实时的分析与检测也成为了目前的研究关键与挑战之一。传统的谣言检测方法一般是对每条 Tweet 内容进行可信度分析，但大多数情况下我们仅记得某个事件的关键字，很难完整地描述一条 Tweet 所叙述的事件。Gupta[2]利用开发的 TweetCred 插件，对 540 万条 Tweets 信息的可信度进行计算，实验结果显示，82％的用户检测到系统中存在的不良信息，所需要的响应时间为 6 秒，99％的用户检测到不良信息的响应时间控制在 10 秒以内，从而保证了实时的可信度评分。Zhao[51]利用 BOSTON 数据集进行谣言检测，利用改进算法来进行谣言检测，其中检测出 46 个谣言所使用的时间为 4.3 小时，而利用主题趋势算法检测出 71 个谣言的平均时间为 3.6 小时，利用标签追踪算法检测 35 个谣言所需要的时间为 2.8 小时。由于信息传播过程的复杂性与信息特征的差异性，面对海量的在线信息，在实时检测的基础上，提高信息可信度识别的准确率，仍然是未来研究的热点之一。

4.4　信息可信度的实证研究

实证研究能够为网络传播中的信息在可信度检测方面提供有效的佐证，并通过问卷调查来获得信息可靠性与可信性的评判依据。目前，信息可信度的实证研究主要是针对特定的热点事件，而网络中不同类型的信息可信度的实证研究并没有形成通用方法或架构，如汤志伟[19]选取了汶川大地震作为网络公共危机案例，采用实证方法研究信息的可信度问题。结果显示，网民对政府与传统媒体所发布的信息的可信度评价显著高于普通网民发布的信息，而对网络新闻的可信度要高于论坛信息和即时通信工具所传播的信息。此外，网民在公共危机时对网络信息的可信度评价与其所具有的网络经验、信任倾向显著相关，但与性别、年龄等因素不存在明显相关性。

5　研究展望

本文针对目前网络传播信息内容的可信度研

究进行了分析与综述。首先，通过对信息特征的梳理，将信息分为可信信息与不可信信息，且不可信信息根据可信的程度又进一步分为：极端突发事件信息、网络偏激信息、网络谣言、虚假信息、误报信息和垃圾信息等六种类型，并结合信息在网络中的传播特点与要素从内容、话题、媒介、传播者和接受者等维度对不同类型的信息进行了特征描述与定义。其次，从信息内容与信息传播等浅层语义特征、基于多媒体的信息融合以及深层语义理解等角度，对信息的可信度建模工作进行了梳理与归纳总结。在此基础上，本文对信息内容可信度的评价方法进行了分析，并通过从客观评测、用户感知评测、信息实时性和实证评价等多个方面对信息可信度的评测标准与方法进行了分析与介绍。

此外，本文针对网络传播信息可信度分析过程中存在的关键技术与挑战也进行了介绍和分析，特别是在目前社交网络正在呈现出海量实时交互条件下的跨语言、跨媒介以及跨媒体的新特征，也为网络传播的信息的可信度识别带来了前所未有的新挑战。例如，当考虑到来自新闻、微博、微信、论坛等不同类型的跨媒介信息交叉扩散传播的过程中，由于信息产生的来源、环境、传播者、接受者、媒介等要素都发生了不同程度的变化，从而导致了原有的单网络媒介信息传播过程中的信息可信度研究方法无法应用于跨媒介情况，因此，需要创建一些全新的跨域条件下的网络传播信息可信度的分析与建模方法与策略。同样，当考虑到多语言信息之间的关联、自动翻译与聚类跟踪，以及多媒体条件下的语义映射，都为信息可信度的分析提出了更高的要求与挑战。解决这些问题，不仅需要通过知识图谱与知识推理，同时也需要对信息的传播动力学机制进行深入研究，在此基础上，结合目前的深度学习以及强化学习的策略，逐步寻找到一个更好的信息可信度的识别与分析方法，而这些工作与挑战也不断激励着人们向更高的研究目标前进。

参考文献

[1] Hardalov M, Koychev I, Nakov P. In search of credible news [C]//Proceedings of the AIMSA 2016, Springer, LNAI9883, 2016: 172-180.

[2] Howell L. Digital wildfires in a hyperconnected world [R/OL]. http://reports. wetorum. org/global-risks-2013/risk-case-11digital-wildfires-in-a-hyper connected-world1, 2013.

[3] Gupta A, Kumaraguru P, Castillo C, et al. TweetCred: Real time credibility assessment of content on Twitter[C]//Proceedings of the SocInfo 2014, 2014: 228-243.

[4] 中国灾害防御协会. 中华人民共和国突发事件应对法[2007][G]. 中国突发公共事件防范与快速处置 2008 优秀成果选编. 2008.

[5] Ratkiewicz J, Conover M, Meiss M, et al. Detecting and tracking the spread of astroturf memes in microblog streams[J]. Computer Science, 2010: 249-252.

[6] Karlova N A, Lee J H. Notes from the underground city of disinformation: A conceptual investigation [C]//Proceedings of the ASIST 2011,2011,48(1): 1-9.

[7] West M D. Validating a Scale for the measurement of credibility: A covariance structure modeling approach [J]. Joumalism Quarterly,1994,71(1): 159-168.

[8] Tseng S, Fogg B J. Credibility and computing technology[J]. Communications of the ACM, 1999, 42(5): 39-44.

[9] 周东浩，韩文报，王勇军. 基于节点和信息特征的社会网络信息传播模型[J]. 计算机研究与发展，2015，52(1): 156-166.

[10] Metzger M J. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research[J]. Journal of the American Society for Information Science and Technology, 2007, 58(13): 2078-2091.

[11] 方滨兴，贾焰，韩毅. 社交网络分析核心科学问题、研究现状及未来展望[J]. 中国科学院院刊，2015，30(2): 187-199.

[12] Miyamori H, Akamine S, Kato Y, et al. Evaluation data and prototype system WISDOM for information credibility analysis[J]. Internet Research, 2008, 18(2): 155-164.

[13] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter[C]//Proceedings of the 20th international conference on World wide web. ACM, 2011: 675-684.

[14] J O'Donovan, B Kang, G. Hllerer, et al. Credibility in context: An analysis of feature distribution in Twitter[J]Prjuacn, Searity, Risk & Trust, 2013, 545(3): 293-301.

[15] Metzger M J, Andrew J F. Credibility and trust of information in online environments: The use of cognitive heuristics[J]. Journal of Pragmatics, 2013, 59(112): 210-220.

[16] J H Lipschultz, Social Media Trust, Credibility and

Reputation Management [EB/OL], https://www.huffingtonpost.com/jeremy-harris-lipschultz/soliul-media-trust-credib_b_3858017.html,2013.

[17] Castillo C, Mendoza M, Poblete B. Predicting information credibility in time-sensitive social media (+ supplementary material). Internet Research [J]. 2013,23(5): 560-588.

[18] 徐静,杨小平,柳增.基于内容信任的 Web 信息可信度验证方法研究[J].北京理工大学学报,2014,34(7): 710-715.

[19] 汤志伟,彭志华,张会平. 网络公共危机信息可信度的实证研究——以汶川地震为例[J]. 情报杂志,2010,29(2): 45-49.

[20] Thomson R, Ito N,Suda H, et al. Trusting tweets: The Fukushima disaster and information source credibility on Twitter[C]//Proceedings of the 9th International ISCRAM Conference,2012: 1-10.

[21] Mendoza M, Poblete B, Castillo C. Twitter under crisis: Can we trust what we RT? [C]//Proceedings of the 1st Workshop on Social Media Analytics. ACM Press, 2010: 71-79.

[22] 薛传业,夏志杰,张志花, 等. 突发事件中社交媒体信息可信度研究[J]. 现代情报,2015,35(4): 12-16.

[23] Lewandowsky S, Ecker U K, Seifert C M, et al. Misinformation and its correction continued influence and successful debiasing[J]. Psychol Sci Public Interest,2012,13(3): 106-131.

[24] Peterson W A, Gist N P. Rumor and public opinion [J]. American Journal of Sociology, 1951, 57(2): 159-167.

[25] Turner R H, Kapferer J N, Fink B. Rumors: Uses, Interpretations and Images[J]. Contemporary Sociology,, 1990,20(5): 794.

[26] Bessi A, Coletto M, Davidescu G A, et al. Science Vs. conspiracy: Collective narratives in the age of Misinformation [J]. Plos One, 2015, 10(2): 1-17.

[27] Hamidian S, Diab M T. Rumor identification and belief investigation on Twitter[C]//Proceedings of the 7th WASSA,2016: 3-8.

[28] Yang F, Liu Y, Yu X, et al. Automatic detection of rumor on Sina Weibo[C]//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. ACM, 2012: 13.

[29] Liu X,Nourbakhsh A, Li Q, et al. Real-time rumor debunking on Twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 1867-1870.

[30] Fallis D. A conceptual analysis of disinformation [C]//Proceedings of the 4th Amual iConference, 2009.

[31] Kumar K P K,Geethakumari G. Detecting misinformation in online social networks using cognitive psychology[J]. Human-centric Computing and Information Sciences, 2014, 4(1): 1.

[32] Karlova N A, Fisher K E. Plz RT: A social diffusion model of misinformation and disinformation for understanding human information behaviour[J]. Inform Research, 2013, 18(1): 1-17.

[33] DeNeys W, Cromheeke S, Osman M(2011) Biased but in doubt: Conflict and decision condence[J]. Plos One 6(1): e15954.

[34] Wang A H. Don't follow me: Spam detection in Twitter[C]//Proceedings of the 2010 International Conference on Security and Cryptography. IEEE, 2010: 1-10.

[35] Tan E,Guo L, Chen S, et al. Unik: Unsupervised social network spam detection[C]//Proceedings of the 22 nd ACM internulind conference on information & knowle dge management 2013: 479-488.

[36] Fogg B J. Prominence-interpretation theory: Explaining how people assess credibility online[C]//Proceedings of the ACM Chi Lauderdle Florida Usa ACM,2003 722-723.

[37] Sundar S S. Technology and credibility: Cognitive heuristics cued by modality, agency, interactivity and navigability[J]. Digital Media, Youth, and Credibility. MacArthur Foundation Series on Digital Media and Learning, 2007: 73-100.

[38] 高雅.微博新闻事件信息可信度评价[D].吉林: 吉林大学硕士学位论文,2013.

[39] 郭国庆,陈讯,何飞.消费者在线评论可信度的影响因素研巧[J],当代经济管理,2010(10): 17-23.

[40] Lucassen T, Schraagen J M. Factual accuracy and trust in information: The role of expertise[J]. Journal of the Association for Information Science and Technology, 2011, 62(7): 1232-1242.

[41] Wu S, Liu Q, Liu Y, et al. Information credibility evaluation on social media[C]//Proceedings of the 13th AAAI Conference on Artificial Intelligence, 2016.

[42] Gupta A,Kumaraguru P. Credibility ranking of Tweets during high impact events[C]//Proceedings of the 1st workshop on Privacy and security in Onlire Social Media. 2012: 2-8.

[43] Gupta M,Zhao P,Han J. Evaluating Event Credibility on Twitter[C]//Proceedings of the 2012 SIAG/DM, 2012: 153-164.

[44] Chang C, Zhang Y, Szabo C, et al. Extreme user and political rumor detection on Twitter[C]//Proceedings of the Advanced Data Mining and Applications. Springer International Publishing, 2016: 751-763.

(下转第 21 页)

刘扬(1971—)，博士，副教授，主要研究领域为语言知识工程、中文信息处理。
E-mail：liuyang@pku.edu.cn

林子(1997—)，主要研究领域为应用语言学、语言知识工程、中文信息处理。
Email：zi.lin@pku.edu.cn

康司辰(1993—)，硕士研究生，主要研究领域为语言知识工程、中文信息处理。
E-mail：1008_frank@sina.com

(上接第 11 页)

[45] Pasternack J, Dan R. Latent credibility analysis [C]//Proceedings of the International Conference on World Wide Web. 2013: 1009-1020.

[46] Unankard S., Li X, Sharaf M A. Emerging event detection in social networks with location sensitivity [J]. World Wide Web-internet & Web Information Systems, 2015, 18(5): 1393-1417.

[47] Akamine S, Kawahara D, Kato Y, et al. WISDOM: A web information credibility analysis system[C]// Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. Association for Computational Linguistics, 2009: 1-4.

[48] Zhang Y, Szabo C, Sheng Q Z, et al. Classifying perspectives on twitter: immediate observation, affection, and speculation[C]//Proceedings of the 16th International Conference on Web Information Systems Engineering, Part I, 493-507.

[49] Rieh S Y. Credibility and cognitive authority of information[N]. Bates M Maack M N. Encyclopedia of library and information sciences: 3rd ed. New York: Taylor and Francis Group, LLC, 2010: 1137-1344.

[50] 冯晓硕. 大数据时代信息可信度分析及可信度评估计算[C]. 全国计算机信息管理学术研讨会, 2013.

[51] Zhao Z, Resnick P, Mei Q. Enquiring minds: Early detection of rumors in social media from enquiry posts [C]//Proceedings of the 24th International Conference on World Wide Web. ACM, 2015: 1395-1405.

吴连伟(1992—)，博士研究生，主要研究领域为自然语言处理、信息可信度识别与分析。
E-mail：wlianwei@qq.com

饶元(1973—)，博士生导师，主要研究领域为社会智能与复杂数据处理。
E-mail：yuanrao@163.com

樊笑冰(1993—)，硕士研究生，主要研究领域为自然语言处理、可信信息传播动力学机制研究。
E-mail：fanxiaobing212@outlook.com