

基于把关人行为的微博虚假信息及早检测方法

谢柏林^{1),2)} 蒋盛益¹⁾ 周咏梅¹⁾ 谢逸³⁾ 李霞²⁾

¹⁾(广东外语外贸大学思科信息学院 广州 510006)

²⁾(广东外语外贸大学语言工程与计算实验室 广州 510006)

³⁾(中山大学信息科学与技术学院 广州 510006)

摘 要 目前微博已成为人们获取信息和发布信息的一个重要平台,然而微博也正成为虚假信息滋生和泛滥的温床.现有的方法主要基于分类算法来识别虚假信息,这些方法不能及早发现微博上流行的虚假信息.为了减少虚假信息对公众的影响,使微博在人们的生产和生活中发挥更积极的作用,文中提出一种基于把关人行为的微博虚假信息及早检测方法.该方法利用模型状态持续时间概率为 Gamma 分布的隐半马尔可夫模型来刻画信息转发者和评论者对流行的真实信息的把关行为,基于此来及早识别微博上流行的虚假信息.该方法分为模型训练和虚假信息检测两个阶段,在虚假信息检测阶段,计算每条信息在传播过程中产生的观测序列相对于模型的平均对数似然概率,实时更新每条信息的可信度,从而及早发现虚假信息,降低虚假信息的危害.使用采集的新浪微博数据集和 Twitter 数据集对文中的方法进行了测试,实验结果表明了该方法的有效性.

关键词 微博;虚假信息;把关人;隐半马尔可夫模型;社交媒体;社交网络;数据挖掘

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2016.00730

Misinformation Detection Based on Gatekeepers' Behaviors in Microblog

XIE Bai-Lin^{1),2)} JIANG Sheng-Yi¹⁾ ZHOU Yong-Mei¹⁾ XIE Yi³⁾ LI Xia²⁾

¹⁾(Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006)

²⁾(Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006)

³⁾(School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510006)

Abstract Nowadays, microblog has become a popular social medium for information dissemination. However, it is also one of the main ways of misinformation transmission, due to the lack of an effective scheme for detection and countermeasures. Existing methods mainly detect misinformation based on classification algorithms. They can't identify the popular misinformation early. In order to reduce the harmfulness of misinformation and make microblog more user-friendly, in this paper a new method is presented for detecting misinformation based on gatekeepers' behaviors. In the proposed scheme, hidden semi-Markov model is used to describe the behaviors of the forwarders and reviewers of popularly true information. Gamma distribution is introduced to describe the state duration of the model. The proposed method includes a training phase and a detection phase. In the detection phase, the average log likelihood of every observation sequence is calculated, and the credibility value of information is updated in real time. So this method can identify the popular misinformation early and reduce the harmfulness of misinformation. An experiment based on real datasets of Sina Weibo and Twitter is conducted to evaluate this method. The experiment results validate the effectiveness of this method.

收稿日期:2015-01-18;在线出版日期:2015-11-12. 本课题得到国家自然科学基金(61202271,61572145,61402119)、广东省高等学校优秀青年教师培养计划项目(GWTPSY201403)、国家社会科学基金项目(13CGL130)、教育部人文社会科学研究青年项目(14YJC870021,13YJCZH258)资助. 谢柏林,男,1982年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为微博虚假信息检测、在线社交网络. E-mail: xiebailin96@126.com. 蒋盛益,男,1963年生,博士,教授,主要研究领域为数据挖掘. 周咏梅,女,1971年生,教授,主要研究领域为文本情感分析. 谢逸,男,1973年生,博士,副教授,主要研究方向为网络安全. 李霞,女,1976年生,博士研究生,副教授,主要研究方向为自然语言处理.

Keywords microblog; misinformation; gatekeeper; hidden semi-Markov model; social media; social networks; data mining

1 引言

微博, 又称“微博客”(Microblog), 是一种基于用户关系的信息分享、传播以及获取平台。最近几年微博迅速发展, 已成为人们生活中不可缺少的一部分。美国市场研究公司 eMarketer 发布的最新调查报告显示, 截至 2014 年 2 月, 国外微博网站 Twitter 的月活跃用户数已达 2.41 亿。目前, 国内最大的微博网站新浪微博的日活跃用户数为 6140 万。根据中国互联网络信息中心(CNNIC)发布的《第 34 次中国互联网络发展状况调查统计报告》^①显示, 截至 2014 年 6 月, 我国微博用户数为 2.75 亿。在微博上, 用户主要通过发布博文(即简短文本)来实现信息的即时分享。微博上的博文通常是限制在 140 字以内, 微博用户在发布博文时无需长篇大论, 因此微博具有较低的门槛。借助电脑、手机等设备, 任何一个微博用户都可以在任何时间、任何地点随意地发布信息, 这使得微博上的信息具有很强的即时性, 也使得每个微博用户都可以成为新闻的报道者。目前, 微博已成为新闻报道的利器。例如: 2011 年 5 月, 本·拉登的死讯是由 Twitter 上的一位用户首先报道, 2012 年 9 月, 中国云南彝良地震是由新浪微博上的一位用户首先报道。微博上的信息主要依靠用户的转发进行传播。用户在转发信息时, 也可以附加评论。为了方便表达, 我们把微博上的信息简称为微博信息。微博信息具有极快的传播速度^[1-2]。信息发布的便捷性、信息的及时性和信息传播的快速性, 使得微博已成为网民获取信息、发布信息的重要渠道。例如: 在 2011 年 7 月国内的温州动车事故中, 许多网民通过微博来实时了解事故救援的最新进展, 有些网民则借助微博来发布寻人、请求献血等信息, 微博在这次救援行动中发挥了十分重要的作用。

然而由于以下几个原因, 微博正成为虚假信息滋生和泛滥的温床: (1) 微博上的博文字数简短, 较难全面客观地传递信息, 这使得微博上的信息容易成为虚假信息; (2) 有些用户为了吸引眼球喜欢故意发布一些耸人听闻的虚假信息; (3) 普通网民自身水平、信息量有限, 导致其在使用微博时不经意间就会产生一些虚假信息; (4) 微博信息在传播过程中, 缺乏强有力的把关人(Gatekeeper)^[3]。在传统媒

体(例如电视、报纸等)中, 信息的发布需要经过记者、编辑等专业把关人的重重把关, 因此传统媒体中信息的可信度比较高。微博虚假信息(即微博上的虚假信息)会混淆公众的视线, 轻则使网民被恶搞、被愚弄, 重则引起大众的群体性恐慌, 造成社会的动荡, 给社会带来重大经济损失。例如: 2010 年 12 月, 新浪微博上的一位实名认证用户发布了一条“著名武侠小说家金庸去世”的虚假信息, 在该虚假信息被快速转发的 22 分钟内, 无数网民已开始表达对金庸去世的遗憾和悲痛; 2011 年 3 月, 我国微博上流传的“食用碘盐可预防核辐射”、“日本核泄漏出现的放射性物质可能污染海水, 没法再提炼盐”等虚假信息, 引发了一场百姓排队抢购食盐的风波, 给社会造成了一定的动荡。为了减少微博虚假信息对公众的影响, 同时也为了营造诚信、健康的网络环境, 使微博在网民的信息获取和信息发布中发挥更积极的作用, 有必要及早识别出微博上的虚假信息, 以便及时控制虚假信息的传播, 降低虚假信息对公众的影响。

微博上的信息主要通过用户的转发来进行传播, 当用户在微博上发布一条信息时, 该信息的传播过程大体如图 1 所示(不考虑同一用户多次转发该信息的情况)。其中 S 表示信息发送者, $S^1, S^2, \dots, S^{15}$ 表示信息转发者, 圆圈代表用户的粉丝, 圆圈的大小表示用户粉丝的数量。信息在传播过程中, 其转发者可能是信息发送者的粉丝或粉丝的粉丝, 也可能是其他一些用户, 这些用户通过微博上的搜索引擎或主页发现了信息, 并对信息进行了转发。

微博上每天新增的信息量十分庞大, 例如 Twitter 平均每天新增的博文为 3.4 亿条, 新浪微博平均每天新增的博文为 1 亿多条。在每天新增的信息中, 绝大部分信息都不会在微博上广泛传播, 例如: 用户之间的聊天信息等。因此即便这些信息中存在虚假信息, 也不会混淆公众的视线。微博上绝大部分用户的粉丝数量都比较少^[4-5], 当用户的粉丝在接收到某条虚假信息时, 某些粉丝可能会识别出该虚假信息, 并转告其他用户, 从而使得微博对虚假信息具有一定的自我净化能力。例如在图 2 中, 用户 U_1 发布了一条虚假信息, 然而其粉丝没有转发该信息, 此时微博上的其他用户一般不会知道该虚假信息的

① <http://www.cnnic.net.cn/hlwfzyj/hlwzxbg/hlwtjbg/201407/P020140721507223212132.pdf>

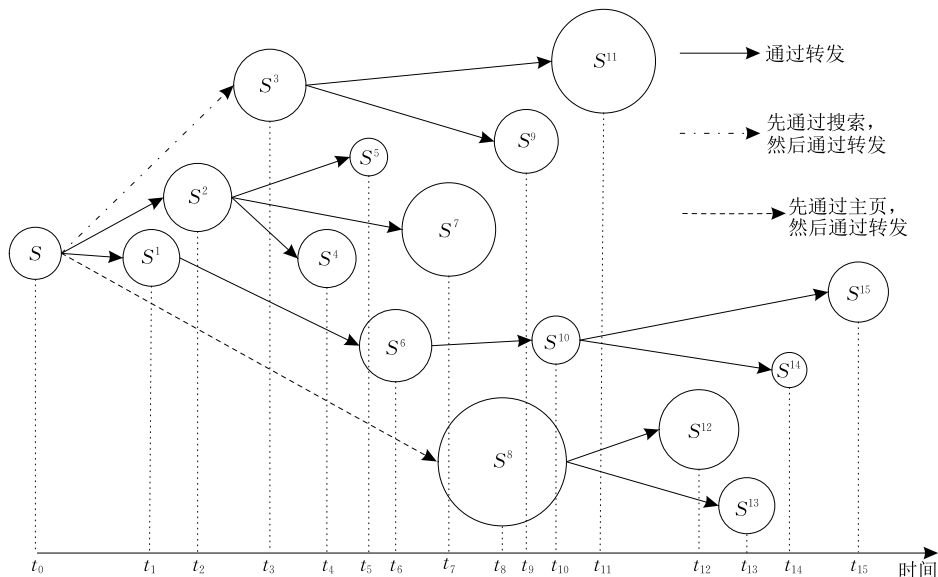


图 1 微博信息传播过程

内容. 因此, 在图 2 中, 用户 U_1 发布的虚假信息不会混淆公众的视线. 由此可知, 我们只需对微博上流行的信息进行虚假性检测, 如果能实时评估微博信息的可信度, 便可实现对微博虚假信息的及早发现.

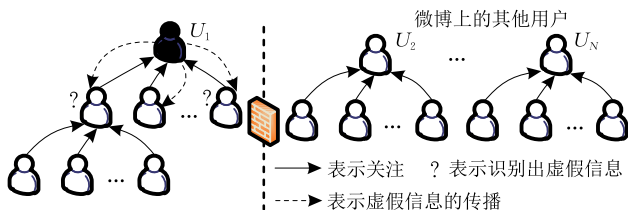


图 2 非流行的虚假信息传播过程

本文提出一种基于把关人行为的微博虚假信息及早检测方法, 该方法利用模型状态持续时间概率为 Gamma 分布的隐半马尔可夫模型 (Hidden Semi-Markov Models, HSMM)^[6] 来刻画信息把关人对流行的真实信息的把关行为. 在虚假信息检测阶段, 该方法实时更新每条信息在传播过程中产生的观测序列相对于模型的平均对数似然概率, 即对每条信息的可信度进行实时评估, 以便及早发现微博上流行的虚假信息. 该方法能够在虚假信息传播的前期就能识别出虚假信息, 从而降低虚假信息带来的危害.

本文第 2 节介绍相关研究; 第 3 节介绍微博虚假信息检测的原理; 第 4 节给出实验结果; 第 5 节讨论把关人态度识别的相关问题; 最后在第 6 节总结并讨论下一步研究工作.

2 相关研究

在微博虚假信息检测方面, Castillo 等人^[7] 提出

利用微博的信息特征 (Message-based Features)、用户特征 (User-based Features)、话题特征 (Topic-based Features) 和传播特征 (Propagation-based Features), 并基于 C4.5 分类算法来识别 Twitter 上的虚假新闻话题. 在微博上, 一个话题一般会包含许多与之相关的博文^[8]. 在 Castillo 等人提出的方法中, 信息特征选为: 博文中包含的不同 URL 的个数; 用户特征选为: 平均每个用户发布的博文的数量、平均每个用户粉丝的数量、平均每个用户关注的用户的个数、平均每个用户注册的时间; 话题特征选为: (话题博文集中) 含有 URL 的博文所占的比例、持肯定观点的博文所占的比例、持否定观点的博文所占的比例、最频繁用户 (即在话题中发布博文最多的用户) 发布的博文所占的比例、含有“?”的博文所占的比例、含有微笑表情的博文所占的比例、提及其他用户的博文所占的比例、含有第一人称代词的博文所占的比例以及博文的平均情感指数等; 传播特征选为: 信息传播树 (Propagation Tree) 的深度. Gupta 等人^[9] 提出通过构建事件图 (Event Graph), 并利用微博的用户特征、博文特征和事件特征 (Event Features) 来检测 Twitter 上的虚假新闻事件. 事件通常是由某些原因、条件引起的, 发生在特定时间、地点, 涉及某些对象 (人或物), 并可能伴随某些必然结果. 在微博上, 一个新闻事件通常会引起很多人的关注, 导致出现大量与之相关的博文^[10]. 在 Gupta 等人^[9] 提出的方法中, 用户特征选为: 用户粉丝的数量、用户是否为认证用户、用户注册的时间、用户关注的用户的个数以及用户个人信息中是

否包含头像/地理位置/URL 等;博文特征选为:博文是否为专家所写、博文中是否含有 URL、博文中的观点是否与整个事件的观点一致、博文中第一/二/三人称代词所含单词的个数以及最频繁用户是否来自新闻事件的发生地等;事件特征选为:事件所含博文的总数目、博文集中不同 URL 的个数、博文集中不同标签的个数、事件不同发生地的个数、事件持续的时间、事件发展到高潮时的当天用户发布的博文所占的比例。该方法在检测虚假新闻事件时,考虑了特征之间的相互关系。Xia 等人^[11]提出利用微博的用户特征、内容特征、话题特征和传播特征,并采用贝叶斯网络分类算法来识别社会紧急情况(Emergency Situation)下 Twitter 上的虚假新闻话题,其中用户特征选为:用户粉丝的数量、用户是否为认证用户、用户使用 Twitter 的次数等;内容特征选为:博文的长度、博文中含有“?”的个数、博文中含有微笑表情个数等;话题特征选为:(话题博文集中)含有 URL 的博文所占的比例、含有褒义词的个数、含有贬义词的个数等;传播特征选为:博文被引用的次数。Yang 等人^[12]提出利用博文的内容特征、客户端特征、账号特征、传播特征和地理位置特征,并基于 SVM 分类算法来判别新浪微博上的新闻事件是否属于谣言。在该方法中,内容特征选为:博文是否含有图片/视频/音频、博文中含有正/负情感词的个数、博文是否含有 URL 等;客户端特征选为:用户发送博文时是采用 Web 客户端程序还是移动客户端程序;账号特征选为:用户是否为认证用户、用户的性别、用户粉丝的数量、用户的注册时间等;传播特征选为:博文是否为原创博文、博文被评论的总次数、博文被转发的总次数;地理位置特征选为:事件发生地。Cai 等人^[13]提出利用信息评论者发表的评论中出现的词,并基于 SVM 分类算法来识别微博上的谣言。Aditi 等人^[14]提出利用微博的用户特征和博文特征,并基于决策树分类算法来识别 Twitter 上的虚假图片,其中用户特征选为:用户粉丝的数量、用户朋友的数量、用户是否为认证用户等;博文特征选为:博文的长度、博文中单词的个数、博文是否包含“?”、博文集中含有 URL 的个数、博文中含有负面情感词的个数等。

上述几种方法都采用分类算法来检测微博虚假信息,即先选用一些特征来训练分类器,然后用待判别的信息与训练好的分类器进行比较,从而识别出虚假信息。这些方法没有考虑信息在传播过程中其转发者和评论者对信息的动态把关过程,另外只有

当微博上存在大量与某条信息相关的博文时,这些方法才能对其进行判别,而此时该信息已在微博上广泛传播。因此,上述几种方法不能实现对微博虚假信息的及早检测。

此外,Al-Khalifa 等人^[15]提出一种用于识别 Twitter 上虚假新闻信息的系统。为了判别某条新闻信息是否可信,该系统首先在一些权威新闻网站(例如 Google News 等)上找到与该信息相关的文章,然后把该信息所对应的博文与新闻网站上的相关文章进行对比,如果两者的相似度比较高,则认为该信息是可信的,反之则认为是虚假的。由于很多新闻事件都是通过微博首先被报道,在新闻信息传播的前期阶段,不一定能在权威新闻网站上找到与之相关的文章,因此 Al-Khalifa 等人^[15]提出的方法对 Twitter 上的大部分虚假新闻信息都不能正确及早识别。Suzuki 等人^[16]提出一种利用维基百科(Wikipedia)来评估 Twitter 上的信息可信性的方法。为了判别某条信息是否可信,该方法首先在维基百科上找到与该信息相关的文章,并对文章的可信性进行评估。如果文章是可信的,则把信息所对应的博文与文章进行对比,如果两者的相似度比较高,则认为该信息是可信的,反之则认为是虚假的。由于 Twitter 上的信息具有很强的及时性,而维基百科上的内容更新的速度比较慢,因此该方法对 Twitter 上的很多虚假信息都不能正确识别,该方法也不能实现对虚假信息的及早检测。Liang 等人^[17]提出利用群体智能和机器智能来识别微博虚假信息,该方法首先在微博上寻找到每条可疑信息所对应领域的专家,然后由专家来判别该信息是否可信。由于微博每天新增的信息量十分庞大,因此该方法需要大量人工的介入,很难适用于实际情况。

在网络信息的虚假性检测其他方法中,Rubin 等人^[18]提出利用自然语言处理技术来评估博客(Blogs)内容的可信性。Olteanu 等人^[19]提出利用 Web 页面的内容特征和社会特征,并基于有监督学习算法来评估传统的 Web 内容的可信性。王腾等人^[20]提出一种针对事实陈述的基于语义相似度的可信判定模型,然后基于该模型来评估传统的 Web 内容的可信度。另外,Lucassen 等人^[21]对影响读者评估维基百科(Wikipedia)内容可信性的因素进行了调查,发现维基百科中的一些文本特征、参考文献和图片是影响读者评估的最重要因素。由于博客、传统的 Web 页面和维基百科中的文本一般都是长文本,而微博中的文本是短文本(通常限制在 140 字以

内),短文本字数很少,往往不具备长文本的很多特性,因此针对博客、传统的 Web 页面和维基百科信息可信性评估的方法很难有效识别出微博上的虚假信息.

本文将采用模型状态持续时间概率为 Gamma 分布的隐半马尔可夫模型来描述流行的真实信息在传播过程中其转发者和评论者对信息的把关行为,基于此来对信息的可信度进行实时评估,以便及早识别出微博上流行的虚假信息.

3 微博虚假信息及早检测方法

当微博用户在转发或评论信息时,用户既是信息的消费者,同时也是信息的把关人(Gatekeeper),即用户会对信息的真实性进行判断.微博信息的把关人可分为 3 类:只评论信息不转发信息的用户、只转发信息不评论信息的用户、转发信息时附加评论的用户.当某条信息在微博上传播时,其每位把关人对该信息真假的识别能力是不同的,例如当与经济学相关的信息在微博上传播时,拥有经济学背景的把关人具有更高的识别能力.我们把把关人对某条信息真假的识别能力称为识别度,其中识别度越高,表明用户越能正确识别出虚假信息.在现有研究和实验结果的基础上,本文使用用户粉丝的数量、用户关注的用户数量、用户是否为认证用户、用户的性别、用户发表/转发的博文数量、用户注册的时间以及用户的标签(Hashtag)在所把关信息中出现的次数来衡量把关人的识别度.

另外,信息在传播过程中,其每位把关人对信息的态度也是不同的,根据用户的评论可以确定把关人对信息的态度,即肯定、否定或中立.对于那些只转发信息不评论信息的用户,则认为其对信息的态度是中立的.我们把把关人对信息的态度简称为把关人的态度.虚假信息在传播过程中,持中立、否定或肯定态度的把关人的识别度的统计分布与真实信息会有比较大的差异,例如持肯定态度的把关人的识别度会普遍比较低.因此,可以利用把关人的识别度和态度来检测微博上流行的虚假信息.

本文使用分类算法来自动判别把关人的识别度等级,然后采用现有的微博意见挖掘(Sentiment Analysis)方法得到把关人的态度,最后由把关人的识别度等级和态度构建观测值.假设把关人的识别度可分为 M 个等级,当某条信息 F 在微博上传播时,令 x_t 表示第 t 个把关人 K_t 对信息 F 的识别度等

级.本文使用分类算法来自动得到 K_t 的识别度等级 x_t ,其步骤如下:

(1)在微博上采集大量真实的流行信息在传播过程中出现的把关人的相关数据,然后提取出每个把关人的特征值,即把关人粉丝的数量、把关人关注的用户数量、把关人是否为认证用户、把关人的性别、把关人发表/转发的博文数量、把关人注册的时间以及把关人的标签在所把关信息中出现的次数;

(2)基于把关人的特征值,使用 k -means 聚类算法把步 1 中的把关人聚成 M 个不同的簇,然后对每个簇中的把关人进行人工检验和筛选,使得每一个簇对应一种识别度等级,且同一簇中把关人的识别度等级都相同,最后把同一簇中的把关人都标记相同的类标号;

(3)使用步 2 中已标记的数据来训练和测试基于不同分类算法的分类器,然后根据测试结果选择出最佳分类器;

(4)从把关人 K_t 的个人资料和信息 F 所对应的博文中提取出特征值,然后使用训练好的最佳分类器对 K_t 进行分类,从而得到 x_t 的取值.

在得到 x_t 的取值后,基于用户的评论,使用现有的一些微博意见挖掘方法便可得到把关人 K_t 的态度,例如 Pak 等人^[22]提出一种用于挖掘 Twitter 上英文微博意见的方法,吴维等人^[23]提出一种用于挖掘中文微博意见的方法.令 y_t 表示第 t 个把关人 K_t 产生的观测值.当 K_t 对 F 持肯定态度时,则 $y_t = x_t$;当 K_t 对 F 持否定态度时,则 $y_t = -x_t$;当 K_t 对 F 持中立态度时,则 $y_t = x_t + M$.因此, $-M \leq y_t \leq 2M$, y_t 取的是整数值.令 $y'_t = y_1 y_2 \cdots y_t$ 表示信息 F 在传播过程中产生的一个长度为 t 的观测序列.

某条流行的真实信息在传播过程中,其把关人的行为一般会发生变化,导致把关人识别度和态度的统计特征会随时间发生变化,例如在信息刚开始传播时其把关人的识别度可能都比较低,且不少把关人持否定态度,而在某个时间段其把关人的识别度可能都比较高,且把关人都持肯定态度.可以把把关人识别度和态度的不同统计特征作为状态,假设不同状态的个数为 I ,即 $q_1, q_2, \dots, q_I$.当流行的真实信息处于不同状态时,表示信息流经了不同类型的把关人,同一类型的把关人具有大体相似的兴趣爱好和识别度等级,他们在把关同一信息时大部分人会具有相似的行为,例如大部分人只转发信息或者转发信息时对信息做肯定的评价等.

状态的转移过程则可以看作是一个马尔可夫过程,即当前状态只与前一个状态有关. 由于把关人产生的观测值与状态不具有一一对应的关系,即给定一个观测值 y_t ,我们不能直接得到此时的状态,因此这是一个隐马尔可夫过程. 令 $\mathbf{A} = \{a_{ij}; 1 \leq i, j \leq I\}$ 为状态转移概率矩阵,它的元素 a_{ij} 表示在把关人把关系流行的真实信息的过程中,状态从 q_i 跳转到 q_j 的概率. 令 $\mathbf{B} = \{b_i(v); -M \leq v \leq 2M\}$ 为观测值概率矩阵,它的元素 $b_i(v)$ 表示在状态 q_i 下把关人产生观测值 $y_t = v$ 的概率. 令 $\boldsymbol{\pi} = \{\pi_i; 1 \leq i \leq I\}$ 为初始状态概率矩阵,它的元素 π_i 表示流行的真实信息在传播过程中其第一个把关人出现时状态为 q_i 的概率.

令 $\mathbf{P} = \{p_i(d); 1 \leq i \leq I, 1 \leq d \leq D\}$ 为状态持续时间概率矩阵,它的元素 $p_i(d)$ 表示在状态 q_i 下,连续出现 d 个把关人的概率,其中 D 为状态持续的最大时间. 流行的真实信息在传播过程中,其把关人的行为是由多种因素所决定的,因此状态持续时间概

率 $p_i(d)$ 可能服从一个比较复杂的分布,不一定是几何分布. 而在传统的隐马尔可夫模型中,模型状态的持续时间概率必须服从几何分布,所以流行的真实信息在传播过程中,其把关人的把关过程实际上是一个隐半马尔可夫模型.

隐半马尔可夫模型是在隐马尔可夫模型(Hidden Markov Model, HMM)的基础上发展起来的,在隐半马尔可夫模型中,模型状态的持续时间概率可以为任意分布,因此隐半马尔可夫模型的应用范围比隐马尔可夫模型更广^[24]. 与隐马尔可夫模型相比,隐半马尔可夫模型更适合描述非稳态和非Markovian分布的随机过程^[6]. 图3为某条流行的真实信息在传播过程中模型状态的转移过程,其中 S 为信息发送者, G^1, G^2, G^3 代表信息的三类把关人, y_t 为第 t 个把关人产生的观测值, q_1, q_2, q_i 为模型状态, d_i 为状态 i 持续的时间, a_{12} 为模型状态从 q_1 跳转到 q_2 的概率.

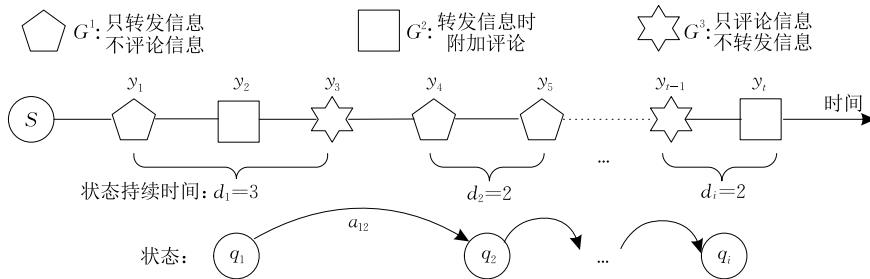


图3 状态的转移过程

3.1 虚假信息检测

本文提出的微博虚假信息及早检测方法的流程如下:

(1) 数据采集、预处理. 在微博上采集大量的流行的真实信息产生的数据集,然后对数据集进行预处理,利用本文提出的把关人识别度等级自动判别方法和现有的微博意见挖掘方法得到把关人的识别度等级和态度,接着由把关人的识别度等级和态度构建观测值,最后得到大量的流行的真实信息产生的观测序列,作为模型训练的数据集;

(2) 模型训练. 利用训练数据集来训练隐半马尔可夫模型,得到模型参数值;

(3) 虚假信息检测. 利用训练得到的隐半马尔可夫模型来实时计算每条信息的可信度,从而及早发现虚假信息.

其中虚假信息检测的具体过程如下:

① 当某个微博用户发布某条信息后,如果出现该信息的把关人,则令 $t=t+1$,并计算出 x_t 的值,其

中 t 的初始值等于 0;

② 求出观测序列 y_t^i , 然后计算观测序列 y_t^i 相对于模型的平均对数似然概率 Q , 我们把 Q 作为信息的可信度, Q 的计算公式如式(1)所示,其中 Δ 表示隐半马尔可夫模型, $\alpha_t(i, d)$ 为前向变量,它的定义和计算公式如式(2)所示^[25]. 在式(2)中 c_t 表示第 t 个把关人出现时模型所处的状态, τ_t 表示在状态 c_t 下将连续出现的把关人的个数. 如果 Q 小于某个阈值,则认为该信息是虚假信息,退出循环,从而实现了对虚假信息的及早检测;反之,则跳转到步①.

$$Q = \log\{\Pr[y_t^i | \Delta]\} / t = \log\left[\sum_{i=1}^I \sum_{d=1}^D \alpha_t(i, d)\right] / t \quad (1)$$

$$\alpha_t(i, d) \equiv \Pr[y_t^i, c_t = q_i, \tau_t = d] =$$

$$\begin{cases} \alpha_{t-1}(i, d+1)b_i(y_t) + \left[\sum_{j \neq i} \alpha_{t-1}(j, 1)a_{ji}\right] \times b_i(y_t)p_i(d), & t > 1 \\ \pi_i b_i(y_1)p_i(d), & t = 1 \end{cases} \quad (2)$$

模型的状态总数 I 和不同观测值的总个数 $3M$ 都比较小,由式(1)和(2)可知,信息可信度 Q 的更新速度主要受状态持续时间概率矩阵 $\mathbf{P}$ 的影响. 状态持续时间概率矩阵 $\mathbf{P}$ 的求解一般比较复杂, D 的取值一般比较大. 当某条流行的信息在微博上传播时,其转发者和评论者会非常多,导致短时间内会产生大量观测值. 为了能快速更新 Q 的值、减少模型参数的个数,我们假定 $p_i(d)$ 服从均值为 θ_i/φ_i , 方差为 θ_i/φ_i^2 的 Gamma 分布, $p_i(d)$ 的计算公式如式(3)所示,其中 $\Gamma(\cdot)$ 为 Gamma 函数,它的定义如式(4)所示. 为了简化模型,本文中 θ_i 的值取为正整数,因此 $\Gamma(\theta_i) = (\theta_i - 1)!$, 如式(5)所示.

$$p_i(d) = \frac{d^{\theta_i-1} e^{-\varphi_i d} \varphi_i^{\theta_i}}{\Gamma(\theta_i)} \quad (3)$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad (4)$$

$$\begin{aligned} \Gamma(\theta_i) &= (\theta_i - 1) \times (\theta_i - 2) \times (\theta_i - 3) \times \cdots \times 2 \times 1 \\ &= (\theta_i - 1)! \end{aligned} \quad (5)$$

通过调整 θ_i 和 φ_i 的取值, $p_i(d)$ 可以成为指数分布、高斯分布等多种分布,图 4 为 (θ_i, φ_i) 取不同值时 $p_i(d)$ 的分布情况,因此假定 $p_i(d)$ 服从 Gamma 分布是合理的.

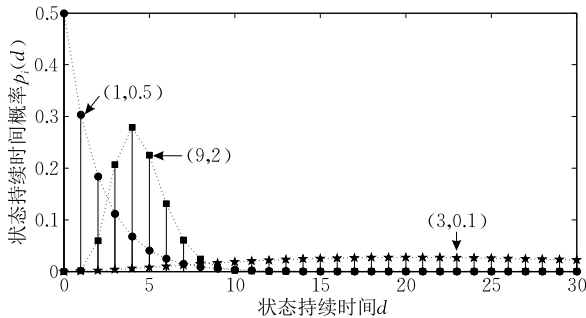


图 4 状态持续时间概率分布

3.2 模型训练

令 λ 表示模型的新参数集, 即 $\lambda = \{a_{ij}, b_i(v), \pi_i, \theta_i, \varphi_i\}$. 假设训练数据集为 U , 其包含 N 个不同的观测序列, 即 $U = \{Y^{(n)}, n=1, 2, \dots, N\}$, 其中 $Y^{(n)} = y_1^{T_n}$ 为第 n 个观测序列, T_n 为对应序列的长度, 并且假定各个观测序列是相互独立的. 在模型训练阶段, 我们采用文献[25]中的前向-后向算法, 并采用多序列来更新参数集 $\{a_{ij}, b_i(v), \pi_i\}$ 的值, 参数更新的公式如式(6)~(8)所示. 其中 ω_n 为训练集中观测序列 $Y^{(n)}$ 相对于模型的平均对数似然概率, 其计算公式如式(1)所示, $\xi_t^n(i, j)$, $\gamma_t^n(i)$ 的定义和计算公式分别

如式(9)、(10)所示, $c_t^{(n)}$ 表示在观测序列 $Y^{(n)}$ 中第 t 个把关人出现时模型所处的状态, $\tau_t^{(n)}$ 表示在观测序列 $Y^{(n)}$ 中模型在状态 $c_t^{(n)}$ 下将连续出现的把关人的个数. $\alpha_t^n(i, d)$ 表示观测序列 $Y^{(n)}$ 中的前向变量, 其定义和计算公式如式(2)所示. $\beta_t^n(i, d)$ 表示观测序列 $Y^{(n)}$ 中的后向变量, 其定义和计算公式如式(11)所示. 在式(7)中, 当 $y_t^{(n)} = v$ 时, $\delta(y_t^{(n)} - v) = 1$, 否则 $\delta(y_t^{(n)} - v) = 0$, 其中 $y_t^{(n)}$ 表示观测序列 $Y^{(n)}$ 中的第 t 个观测值.

$$\bar{a}_{ij} = \frac{\sum_{n=1}^N \frac{1}{\omega_n} \times \left[\sum_{t=2}^{T_n} \xi_t^n(i, j) \right]}{\sum_{n=1}^N \frac{1}{\omega_n} \times \left[\sum_{j=1}^I \sum_{t=2}^{T_n} \xi_t^n(i, j) \right]} \quad (6)$$

$$\bar{b}_i(v) = \frac{\sum_{n=1}^N \frac{1}{\omega_n} \times \left[\sum_{t=1}^{T_n} \gamma_t^n(i) \times \delta(y_t^{(n)} - v) \right]}{\sum_{n=1}^N \frac{1}{\omega_n} \times \left[\sum_{v=-M}^{2M} \sum_{t=1}^{T_n} \gamma_t^n(i) \times \delta(y_t^{(n)} - v) \right]} \quad (7)$$

$$\pi_i = \frac{\sum_{n=1}^N \frac{1}{\omega_n} \times \gamma_1^n(i)}{\sum_{n=1}^N \frac{1}{\omega_n} \times \left[\sum_{i=1}^I \gamma_1^n(i) \right]} \quad (8)$$

$$\begin{aligned} \xi_t^n(i, j) &\equiv \Pr[Y^{(n)}, c_{t-1}^{(n)} = q_i, c_t^{(n)} = q_j] \\ &= \alpha_{t-1}^n(i, 1) a_{ij} b_j(y_t^{(n)}) \times \left[\sum_d p_j(d) \beta_t^n(j, d) \right] \end{aligned} \quad (9)$$

$$\begin{aligned} \gamma_t^n(i) &\equiv \Pr[Y^{(n)}, c_t^{(n)} = q_i] \\ &= \begin{cases} \gamma_{t+1}^n(i) + \sum_{j \neq i} [\xi_{t+1}^n(i, j) - \xi_{t+1}^n(j, i)], & 1 \leq t < T_n \\ \sum_d \alpha_{T_n}^n(i, d), & t = T_n \end{cases} \end{aligned} \quad (10)$$

$$\beta_t^n(i, d) \equiv \Pr[y_{t+1}^{T_n} | c_t^{(n)} = q_i, \tau_t^{(n)} = d]$$

$$\beta_t^n(i, d) = \begin{cases} b_i(y_{t+1}^{(n)}) \beta_{t+1}^n(i, d-1), & d > 1, t \neq T_n \\ \sum_{j \neq i} a_{ij} b_j(y_{t+1}^{(n)}) \left[\sum_d p_j(d) \beta_{t+1}^n(j, d) \right], & d = 1, t \neq T_n \\ \beta_{T_n}^n(i, d) = 1, & d \geq 1, t = T_n \end{cases} \quad (11)$$

通过求解式(12)和(13)来更新 φ_i 和 θ_i 的值^[26], 其中 $\bar{\varphi}_i, \bar{\theta}_i$ 分别表示 φ_i 和 θ_i 更新后的值, $\eta_t^n(i, d)$ 为状态持续联合概率, 它表示在观测序列为 $Y^{(n)}$ 的基础上, 当第 t 个把关人出现时, 模型从其他状态跳转到状态 q_i 并将在状态 q_i 下连续出现 d 个观测值的概率. $\eta_t^n(i, d)$ 的定义和计算公式如式(14)所示.

$$\bar{\varphi}_i = \frac{\theta_i \times \left[\sum_{n=1}^N \sum_d \sum_{t=1}^{T_n} \eta_i^n(i, d) \right]}{\sum_{n=1}^N \sum_d d \times \left[\sum_{t=1}^{T_n} \eta_i^n(i, d) \right]} \quad (12)$$

$$\left\{ \ln \bar{\theta}_i + \ln \frac{d \times \left[\sum_{n=1}^N \sum_d \sum_{t=1}^{T_n} \eta_i^n(i, d) \right]}{\sum_{n=1}^N \sum_d d \times \left[\sum_{t=1}^{T_n} \eta_i^n(i, d) \right]} - \Psi(\bar{\theta}_i) \right\} = 0 \quad (13)$$

$$\begin{aligned} \eta_i^n(i, d) &\equiv \Pr[Y^{(n)}, c_{t-1}^{(n)} \neq q_i, c_t^{(n)} = q_i, \tau_t^{(n)} = d] \\ &= \left[\sum_{j \neq i} \alpha_{t-1}^{(n)}(j, 1) a_{ji} \right] b_i(y_t^{(n)}) p_i(d) \beta_i^n(i, d) \end{aligned} \quad (14)$$

在式(13)中, $\Psi(\cdot)$ 为 Digamma 函数, 其计算公式如式(15)所示, 其中 γ 为欧拉-马歇罗尼常数, γ 的计算公式如式(16)所示。

$$\begin{aligned} \Psi(\bar{\theta}_i) &= \frac{\partial \ln \Gamma(\bar{\theta}_i)}{\partial \bar{\theta}_i} = \frac{\Gamma'(\bar{\theta}_i)}{\Gamma(\bar{\theta}_i)} \\ &= -\gamma - \bar{\theta}_i^{-1} + \bar{\theta}_i \times \left\{ \sum_{h=1}^{\infty} [h \times (h + \bar{\theta}_i)]^{-1} \right\} \end{aligned} \quad (15)$$

$$\gamma = -\Gamma'(1) = \sum_{h=1}^{\infty} \left[\frac{1}{h} - \ln \left(1 + \frac{1}{h} \right) \right] \quad (16)$$

由式(12)和(13)可知, 如果能得到 $\bar{\theta}_i$ 的值, 则根据式(12)就可得到 $\bar{\varphi}_i$ 的值, 然而 $\bar{\theta}_i$ 的求解比较复杂, 文献[26]提出了一种求解 $\bar{\theta}_i$ 的简化方法, 本文采用该方法来求解 $\bar{\theta}_i$ 的值。

4 实验测试及结果分析

我们使用采集到的新浪微博数据集来测试本文提出的虚假信息及早检测方法, 新浪微博是目前国内最大的微博网站, 其注册用户数已超过 5.3 亿, 日活跃用户数已达到 6140 万。我们使用新浪微博的 API 和文献[13]中的数据采集方法, 在热门微博排行榜中采集流行的真实信息。新浪微博运营商会监督热门微博排行榜中的信息, 因此热门微博排行榜中的信息绝大部分是真实信息。我们对热门微博排行榜中采集到的信息进行人工筛选, 得到 100 260 条流行的真实信息。新浪微博上存在很多流行的娱乐信息, 这些信息的绝大部分是娱乐明星在跟粉丝分享自己的生活和工作, 或娱乐明星与粉丝之间的互动, 体现了微博的娱乐性。这些信息由于侧重

于娱乐性, 因此即便存在虚假信息, 也很难引起大众的群体性恐慌。在我们采集到的流行的真实信息中, 娱乐信息所占比例为 10%。

我们在新浪微博“举报处理大厅”的“不实信息”中采集流行的虚假信息, 同时也通过以下方式采集流行的虚假信息, 即首先使用新浪微博上的搜索功能来获取可疑信息, 搜索关键词设置为“求辟谣”、“求认证”, 然后人工的从这些可疑信息中挑选出流行的虚假信息。我们共采集到 4330 条流行的虚假信息。在数据采集过程中, 我们把信息传播过程中出现的所有把关人的相关资料都采集下来, 保存于数据库中。

4.1 数据集统计分析

我们对实验所用的新浪微博数据集进行了一些统计分析, 信息发布者和把关人的认证统计情况如图 5 所示, 其中虚假信息发布者中 12% 为认证用户, 真实信息发布者中 67% 为认证用户, 虚假信息把关人中 11% 为认证用户, 真实信息把关人中 20% 为认证用户。信息发布者和把关人的性别统计情况如图 6 所示, 其中虚假信息发布者中 84% 为男性, 真实信息发布者中 73% 为男性, 虚假信息把关人中 71% 为男性, 真实信息把关人中 58% 为男性。

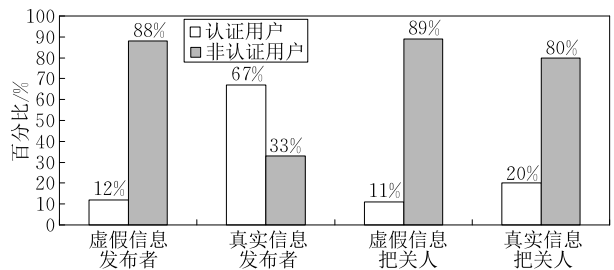


图 5 信息发布者和把关人的认证统计情况

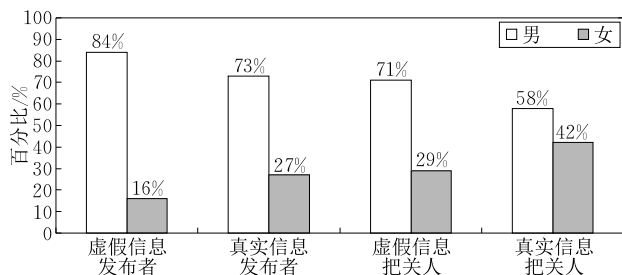


图 6 信息发布者和把关人的性别统计情况

信息发布者的粉丝数量统计情况如图 7 所示, 虚假信息发布者的粉丝数量主要分布在 101~10000 之间, 其中 44% 的虚假信息发布者的粉丝数量在 101~1000 之间, 49% 的虚假信息发布者的粉丝数量在 1001~10000 之间; 而粉丝数量在 100 以

上的真实信息发布者的分布相对比较均匀,其中 25%的真实信息发布者的粉丝数量在 101~1000 之间,29%的真实信息发布者的粉丝数量在 1001~10000 之间,27%的真实信息发布者的粉丝数量在 10000~100000 之间,18%的真实信息发布者的粉丝数量在 10 万以上.信息把关人的粉丝数量统计情况如图 8 所示,其中 57%的虚假信息把关人和 42%的真实信息把关人的粉丝数量在 101~1000 之间,21%的虚假信息把关人和 30%的真实信息把关人的粉丝数量在 1001~10000 之间,10%的虚假信息把关人和 15%的真实信息把关人的粉丝数量在 10000~100000 之间.

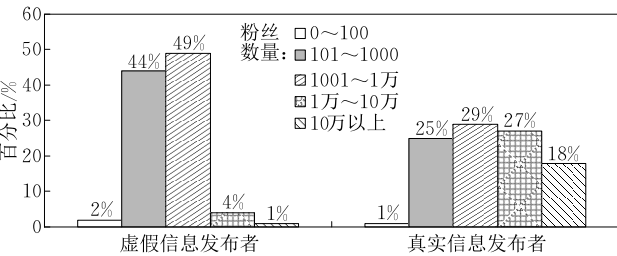


图 7 信息发布者的粉丝数量统计情况

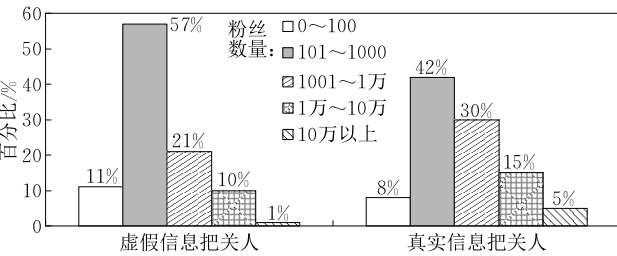


图 8 信息把关人的粉丝数量统计情况

信息发布者关注的用户数量的统计情况如图 9 所示,其中 33%的虚假信息发布者和 45%的真实信息发布者关注的用户数量在 201~500 之间,15%的虚假信息发布者和 6%的真实信息发布者关注的用户数量在 501~1000 之间,11%的虚假信息发布者和 5%的真实信息发布者关注的用户数量在 1000 以上.信息把关人关注的用户数量的统计情况如图 10 所示,其中虚假信息把关人与真实信息把关人关注的用户数量的分布基本相同.

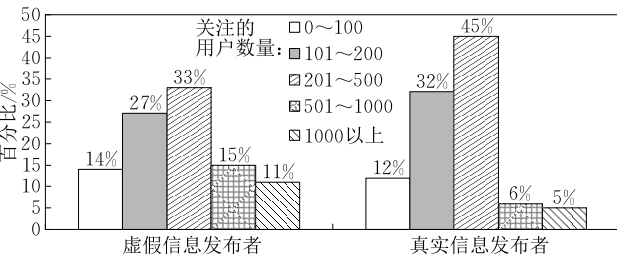


图 9 信息发布者关注的用户数量统计情况

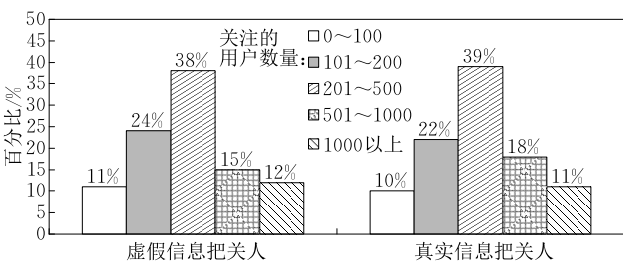


图 10 信息把关人关注的用户数量统计情况

信息发布者注册的时间统计情况如图 11 所示,其中 48%的虚假信息发布者的注册时间为半年以内,81%的虚假信息发布者的注册时间为 1 年以内,而 56%的真实信息发布者的注册时间为 1 年以前.信息把关人注册的时间统计情况如图 12 所示,其中 23%的虚假信息把关人和 15%的真实信息把关人的注册时间为 31~180 天前,32%的虚假信息把关人和 25%的真实信息把关人的注册时间为 181~365 天前,40%的虚假信息把关人和 57%的真实信息把关人的注册时间为 1 年以前.

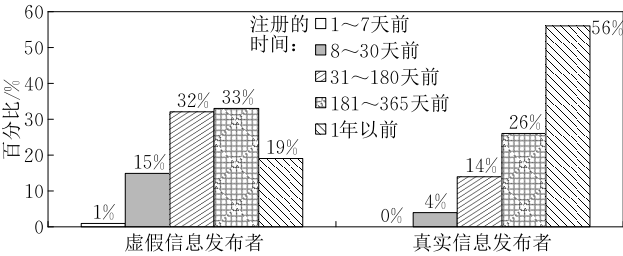


图 11 信息发布者注册的时间统计情况

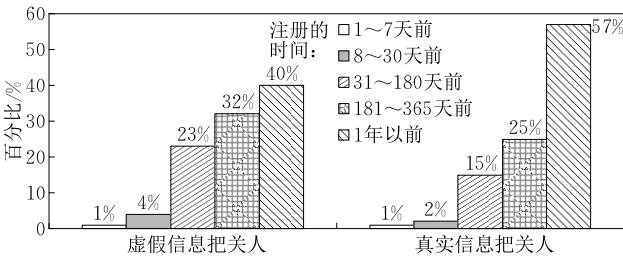


图 12 信息把关人注册的时间统计情况

信息发布者发表/转发的博文数量的统计情况如图 13 所示,其中 27%的虚假信息发布者和 16%的真实信息发布者发表/转发的博文数量在 11~50 之间,36%的虚假信息发布者和 24%的真实信息发布者发表/转发的博文数量在 51~100 之间,25%的虚假信息发布者和 30%的真实信息发布者发表/转发的博文数量在 101~500 之间,10%的虚假信息发布者和 29%的真实信息发布者发表/转发的博文数量在 500 以上.信息把关人发表/转发的博文数量的统计情况如图 14 所示,其中 27%的虚假信息把关

人和 14% 的真实信息把关人发表/转发的博文数量在 11~50 之间,10% 的虚假信息把关人和 24% 的真实信息把关人发表/转发的博文数量在 500 以上.

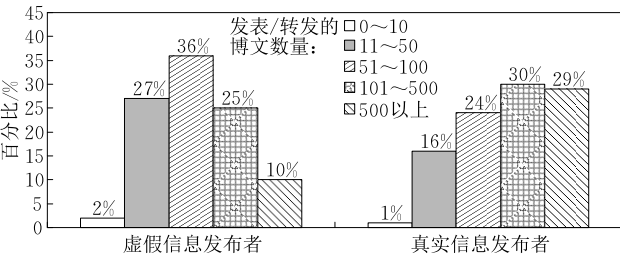


图 13 信息发布者发表/转发的博文数量统计情况

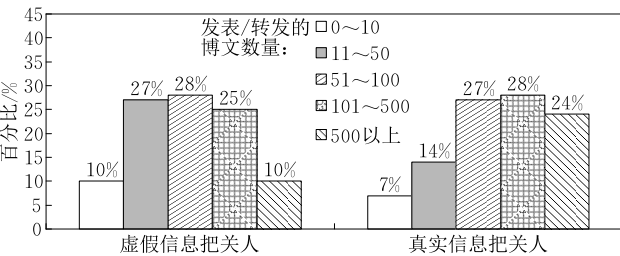


图 14 信息把关人发表/转发的博文数量统计情况

由以上实验结果可知,虚假信息把关人与真实信息把关人的统计特征存在一定的差异,虚假信息发布者与真实信息发布者的统计特征也存在一定的差异.

4.2 把关人识别度等级自动判别方法测试

由于测试数据集的标注需要人工介入,如果把把关人识别度的等级类别过多,则在人工标注把关人识别度等级时其准确率比较难保证,如果把把关人识别度的等级类别过少,则很难反映出不同把关人识别能力的差异.因此我们参照目前流行的网络购物商品评价星级制度,把把关人的识别度等级分为 5 级,这样使得人工在标注把关人识别度等级时具有很高的准确率.我们从采集到的流行的新浪微博真实信息的把关人中随机挑选出大量把关人,并提取

出每个把关人的特征值,即把关人的粉丝数量、把关人关注的用户数量、把关人是否为认证用户、把关人的性别、把关人发表/转发的博文数量、把关人注册的时间以及把关人的标签在所把关信息中出现的次数.基于把关人的特征值,我们使用 k -means 聚类算法把这些把关人聚成 5 个不同的簇,然后对每个簇中的把关人进行人工检验和筛选,使得每一个簇对应一种识别度等级,且同一簇中把关人的识别度等级都相同,接着把同一簇中的把关人都标记相同的类标号,即类 I、类 II、类 III、类 IV 和类 V.最后,我们从已标记的数据集中通过随机抽取样本的方式来生成测试数据集,使得测试数据集包含 5 个类别,并且每个类别包含 1000 个把关人.

基于上述测试数据集,我们采用交叉验证方法,分别测试了 C4.5 分类算法、朴素贝叶斯 (Naive Bayesian, NB) 分类算法、支持向量机 (SVM) 分类算法和 K 最近邻 (KNN) 分类算法在自动判别把关人识别度等级时的性能,测试实验所用主机的配置为 CPU: Intel 酷睿 i7 960 (四核心,主频: 3.46 GHz), 内存: 8 GB. 测试过程中,我们采用不断尝试的方式来确定分类算法的最优参数, C4.5、NB 和 SVM 分类算法的最优参数如表 1 所示. 在最优参数下, C4.5、NB 和 SVM 分类算法在识别每一类把关人时的准确率 (Accuracy) 和召回率 (Recall) 如表 2 所示,其中准确率的计算方法如下:

准确率(Accuracy) = 正确分类的样本数 / 样本总数

表 1 3 种分类算法的最优参数	
分类算法	参数
C4.5	$C=0.25, M=2$
NB	—
SVM	$C=1.0, L=0.001, P=1.0E-12, N=0, V=10, W=1, K=$ “weka. classifiers. functions. supportVector. PolyKernel $E=1.0, C=0$ ”

表 2 3 种分类算法的性能

分类算法	类 I 的 准确率	类 I 的 召回率	类 II 的 准确率	类 II 的 召回率	类 III 的 准确率	类 III 的 召回率	类 IV 的 准确率	类 IV 的 召回率	类 V 的 准确率	类 V 的 召回率
C4.5	0.94	0.92	0.91	0.89	0.93	0.92	0.90	0.89	0.95	0.96
NB	0.91	0.89	0.87	0.86	0.88	0.88	0.86	0.84	0.92	0.91
SVM	0.95	0.93	0.92	0.91	0.93	0.94	0.91	0.92	0.95	0.95

当 K 取不同值时, KNN 分类算法在识别每一类把关人时的准确率和召回率如表 3 所示, 从表 3 可知, 当 $K=9$ 时, KNN 分类算法的效果较好. 由于 KNN 分类算法在判别某个把关人的识别度等级时, 需要匹配训练数据集中所有的样本, 因此其消耗

的时间过长, 很难应用于实际的微博环境. C4.5、NB、SVM 和 KNN 分类算法的建模时间和测试时间如表 4 所示, 其中在 KNN 分类算法中, $K=9$. 从表 2 可知, SVM 分类算法的准确率略高于 C4.5 分类算法, 然而从表 4 可知, SVM 分类算法的

表 3 KNN 分类算法的性能

K 的取值	类Ⅰ的 准确率	类Ⅰ的 召回率	类Ⅱ的 准确率	类Ⅱ的 召回率	类Ⅲ的 准确率	类Ⅲ的 召回率	类Ⅳ的 准确率	类Ⅳ的 召回率	类Ⅴ的 准确率	类Ⅴ的 召回率
K=1	0.82	0.83	0.81	0.78	0.79	0.80	0.85	0.83	0.89	0.88
K=3	0.84	0.86	0.82	0.81	0.83	0.83	0.87	0.85	0.91	0.89
K=5	0.88	0.87	0.85	0.83	0.86	0.85	0.90	0.85	0.94	0.91
K=7	0.89	0.89	0.88	0.85	0.89	0.87	0.92	0.92	0.94	0.92
K=9	0.92	0.93	0.90	0.89	0.89	0.91	0.94	0.91	0.95	0.93
K=11	0.92	0.91	0.89	0.87	0.87	0.90	0.93	0.92	0.93	0.91
K=13	0.91	0.91	0.88	0.86	0.85	0.84	0.93	0.90	0.92	0.92

表 4 4 种分类算法的建模时间和测试时间

分类算法	建模时间/s	测试时间/s
C4.5	0.12	0.026
NB	0.01	1.124
SVM	2.78	1.513
KNN	—	4.832

测试时间多于 C4.5 分类算法. 当流行的信息在微博上传播时, 短时间内会出现大量的把关人, 此时判别把关人识别度等级的速度就显得非常重要. 从表 4 可知, C4.5 分类算法的测试时间最短, 非常适合在线分类, 因此本文采用 C4.5 分类算法来自动判别把关人的识别度等级.

4.3 虚假信息及早检测方法测试

为了测试本文提出的微博虚假信息及早检测方法, 我们使用本文提出的把关人识别度等级自动判别方法对采集到的新浪微博数据集进行预处理, 得到信息传播过程中所有把关人的识别度等级. 然后利用文献[23]中的中文微博意见挖掘方法得到信息在传播过程中产生的观测序列. 最终我们得到 100 260 条真实信息产生的观测序列和 4330 条虚假信息产生的观测序列.

我们从真实信息产生的观测序列中随机抽取出 90 000 条序列来训练隐半马尔可夫模型, 剩下的 10 260 条观测序列和 4330 条虚假信息产生的观测序列则作为模型的测试集, 用于测试模型的性能. 在模型训练过程中, 令模型的初始状态数为 40 个, 然后删除那些很少出现的状态, 最终得到流行的真实信息传播过程中其把关人行为的最终状态数. 图 15 为模型的

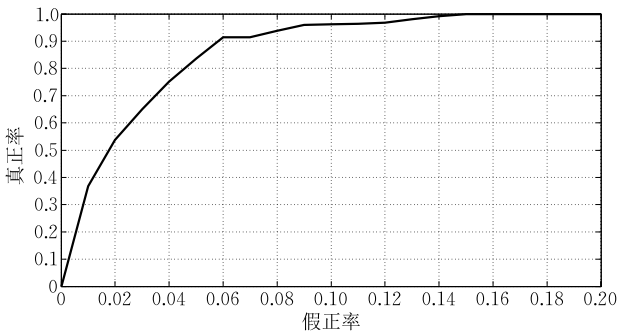


图 15 模型 ROC 曲线

ROC 曲线 (Receiver Operating Characteristic Curve), 当模型的假正率 (False Positive Rate) 为 10% 时, 模型的真正率 (True Positive Rate) 为 96.2%.

在本文中, 隐半马尔可夫模型的不同状态代表信息把关人识别度和态度的不同统计特征. 通过分析训练数据集和测试数据集中各个状态的分布情况, 可以了解真实信息把关人和虚假信息把关人的行为特征. 我们在真实信息和虚假信息产生的数据集中对 5 种频繁出现的状态进行了统计, 结果如图 16 所示. 5 种频繁出现的状态分别对应信息持肯定态度的 5 种等级的把关人, 例如当信息处于状态 1 时, 表明此时信息把关人的识别度等级为 I, 且把关人对信息的态度为肯定; 当信息处于状态 3 时, 表明此时信息把关人的识别度等级为 III, 且把关人对信息的态度为肯定; 当信息处于状态 5 时, 表明此时信息把关人的识别度等级为 V, 且把关人对信息的态度为肯定. 在虚假信息传播过程中, 大部分把关人的识别度都比较低, 因此虚假信息产生的数据集中各个状态的统计分布与真实信息具有明显差异.

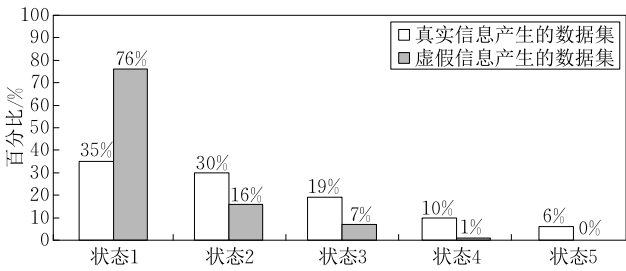


图 16 模型状态统计分布

为了进一步测试本文提出的微博虚假信息及早检测方法, 我们从 Twitter 上采集到了一些数据, 该数据集包含 1970 条流行的真实信息和 485 条流行的虚假信息, 以及这些信息传播过程中出现的所有把关人的相关资料. 我们使用本文提出的把关人识别度等级自动判别方法和文献[22]中提出的方法来获取每条信息产生的观测序列, 然后使用这些观测序列来测试本文提出的微博虚假信息及早检测方法的性能. 图 17 为模型的 ROC 曲线, 当模型的假正率

为 7% 时, 模型的真正率为 99.3%。真实信息和虚假信息产生的数据集中, 5 种频繁出现的状态的统计分布情况如图 18 所示。

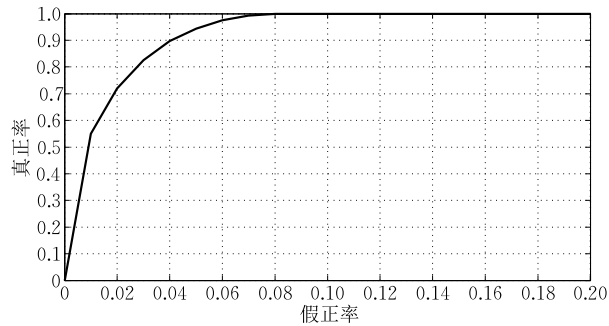


图 17 Twitter 数据集下模型 ROC 曲线

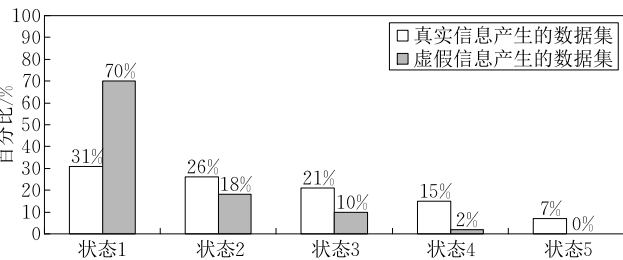


图 18 Twitter 数据集下模型状态统计分布

在 Twitter 上, 用户只能转发信息或评论信息, 不能同时既转发又评论信息, 因此在我们采集到的 Twitter 数据集中只转发信息的把关人所占的比例比较高。另外由于文献[22]中提出的英文微博意见挖掘方法的准确率高于我们采用的中文微博意见挖掘方法, 因此本文提出的虚假信息检测方法在 Twitter 数据集测试中具有更好的性能。

从上面的测试实验可知本文提出的方法具有较好的性能, 另外由于每条信息可信度的每次更新只涉及到几十次乘法和几十次加法, 因此, 本文提出的方法也具有比较快的在线检测速度。

4.4 对比测试

我们使用采集到的新浪微博数据集和 Twitter 数据集对本文提出的方法和文献[7]中提出的方法进行了对比测试, 新浪微博数据集下的测试结果如图 19 所示, Twitter 数据集下的测试结果如图 20 所示, 其中 HSMM 代表本文中的方法, C4.5 代表文献[7]中的方法。从图 19、图 20 可知, 本文提出的方法具有更好的性能。

文献[7]中的方法利用微博的一些特征, 并基于 C4.5 分类算法来识别微博上的虚假新闻话题, 该方法使用的特征包括: 博文中包含的不同 URL 的个数、(话题博文集中) 含有 URL 的博文所占的比例、含有第一人称代词的博文所占的比例、博文的平均

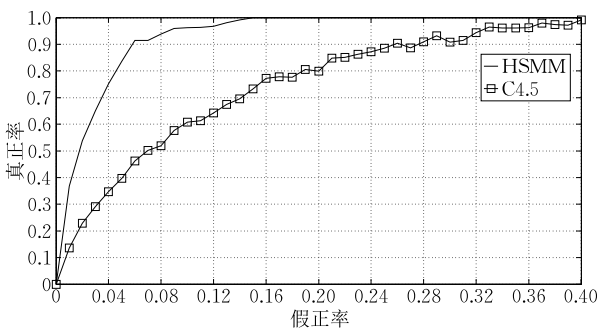


图 19 新浪微博数据集下的对比测试

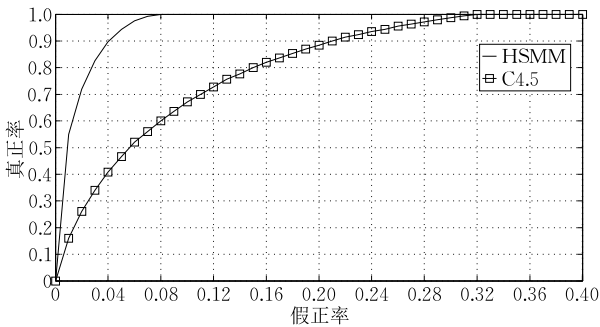


图 20 Twitter 数据集下的对比测试

情感指数、信息传播树的深度等。然而在测试数据集中, 不少信息所属的博文不包含 URL 链接, 另外有些信息产生的话题博文集中含有第一人称代词的博文所占的比例非常低, 因此文献[7]中的方法不能准确区分出这些信息是否为虚假信息。本文提出的方法基于把关人的识别度等级和态度来识别微博上流行的虚假信息, 该方法具有很高的真正率和较低的假正率。

另外, 我们对本文提出的方法和文献[7]中提出的方法在识别虚假信息的及时性方面进行了对比测试。在采集到的测试数据集中, 我们把每条信息传播过程中出现的所有把关人按照时间先后构成一个序列, 然后把这个序列等分为 3 段(大体对应信息传播的前期、中期和后期), 使得每段中把关人的个数都相等。最后, 我们在新浪微博测试数据集中, 把序列前 1/3 段、前 2/3 段、前 3/3 段中把关人产生的相关数据集分别标注为 U_1 、 U_2 和 U_3 ; 同时, 在 Twitter 测试数据集中, 把序列前 1/3 段、前 2/3 段、前 3/3 段中把关人产生的相关数据集分别标注为 Ω_1 、 Ω_2 和 Ω_3 。

我们分别在 U_1 、 U_2 和 U_3 以及 Ω_1 、 Ω_2 和 Ω_3 下测试本文提出的方法和文献[7]中提出的方法的真正率, 在测试过程中, 两种方法的假正率都设置为 10%, 测试结果分别如表 5 和表 6 所示。从表 5 和表 6 可知, 由于本文提出的方法是通过实时计算每

条信息的可信度来检测虚假信息,因此该方法能在信息传播的前期识别出虚假信息,实现对虚假信息的及早检测。

表 5 新浪微博数据集下的及时性测试			
测试方法	U_1 下的真正率/%	U_2 下的真正率/%	U_3 下的真正率/%
HSMM	86.9	92.4	96.2
C4.5	47.7	54.3	60.8

表 6 Twitter 数据集下的及时性测试			
测试方法	Ω_1 下的真正率/%	Ω_2 下的真正率/%	Ω_3 下的真正率/%
HSMM	90.5	97.6	100
C4.5	51.8	60.2	67.5

5 讨 论

本文把把关人对信息的态度分为:肯定、否定和中立,当把关人评论信息时,基于把关人的评论,我们利用现有的微博意见挖掘方法来确定把关人对信息的态度;当把关人只转发信息不评论信息时,我们认为把关人对信息的态度为中立.然而转发本身也代表了一种情感倾向,在下一步的研究中我们将把把关人对信息的态度分为:肯定、否定、中立和转发,当把关人只转发信息不评论信息时,则认为把关人对信息的态度为转发,并将基于此来检测微博上流行的虚假信息。

本文使用现有的微博意见挖掘方法来分析把关人的评论,以便得到把关人对信息的态度,我们在采集到的新浪微博数据集中,测得文献[23]中的中文微博意见挖掘方法的准确率为 58%,在采集到的 Twitter 数据集中测得文献[22]中的英文微博意见挖掘方法的准确率为 74%.在我们采集到的微博数据集中,对 92%的信息而言,其 55%~65%的把关人只转发信息不评论信息,这部分把关人的态度识别(无需使用微博意见挖掘方法)的准确率为 100%,因此本文使用文献[23]中的方法来识别新浪微博数据集中的把关人的态度时,其整体准确率在 79%以上,使用文献[22]中的方法来识别 Twitter 数据集中的把关人的态度时,其整体准确率在 86%以上.现有的微博意见挖掘方法的性能对本文提出的虚假信息检测方法的影响比较小。

6 总结与展望

本文提出一种基于把关人行为的微博虚假信息

及早检测方法,该方法以信息把关人的识别度等级和态度构建观测值,使用 C4.5 分类算法来自动得到把关人的识别度等级,采用模型状态持续时间概率为 Gamma 分布的隐半马尔可夫模型来刻画信息把关人对流行的真实信息的把关行为,以便及早识别出微博上流行的虚假信息.我们使用采集的新浪微博数据集和 Twitter 数据集对本文提出的方法进行了测试,实验结果表明该方法具有很高的真正率和较低的假正率。

今后的主要工作将侧重于使用更多的微博数据集来测试该方法的性能以及在实际微博平台上在线测试该方法的性能。

致 谢 审稿专家为本文提出了宝贵意见,在此致谢!

参 考 文 献

[1] Ye S, Wu S F. Measuring message propagation and social influence on Twitter.com//Proceedings of the 2nd International Conference on Social Informatics. Laxenburg, Austria, 2010: 216-231

[2] Wang D, Li Z, Salamatian K, Xie G. The pattern of information diffusion in microblog//Proceedings of the 7th International Conference on Emerging Networking Experiments and Technologies. Tokyo, Japan, 2011: 1-2

[3] Westerman D, Spence P R, Van Der Heide D. A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. Computers in Human Behavior, 2012, 28(1): 199-206

[4] Kwak H, Lee C, Park H. What is Twitter, a social network or a news media?//Proceedings of the 19th International Conference on World Wide Web. North Carolina, USA, 2010: 591-600

[5] Yuan Wei-Guo, Liu Yun, Cheng Jun-Jun. Research on the user characteristics and growth rates distribution in microblog. Chinese Journal of Computers, 2014, 37(4): 767-778 (in Chinese)
(苑卫国, 刘云, 程军军. 微博网络中用户特征量和增长率分布的研究. 计算机学报, 2014, 37(4): 767-778)

[6] Yu S Z. Hidden semi-Markov models. Artificial Intelligence, 2010, 174(2): 215-243

[7] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 675-684

[8] O'Connor B, Krieger M, Ahn D. TweetMotif: Exploratory search and topic summarization for Twitter//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Georgia, USA, 2010: 384-385

[9] Gupta M, Zhao P, Han J. Evaluating event credibility on Twitter//Proceedings of the SIAM International Conference on Data Mining. Philadelphia, USA, 2012; 153-164

[10] Long R, Wang H, Chen Y, et al. Towards effective event detection, tracking and summarization on microblog data//Proceedings of the 12th International Conference on Web-Age Information Management. Wuhan, China, 2011; 652-663

[11] Xia X, Yang X, Wu C. Information credibility on Twitter in emergency situation//Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics. Kuala Lumpur, Malaysia, 2012; 45-59

[12] Yang Fan, Yu Xiao-Hui. Automatic detection of rumor on sina weibo//Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. Beijing, China, 2012; 1-7

[13] Cai G, Wu H, Lv R. Rumors detection in Chinese via crowd responses//Proceedings of 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Beijing, China, 2014; 912-917

[14] Gupta Aditi, Lamba H, Kumaraguru P. Faking Sandy: Characterizing and identifying fake images on Twitter during hurricane sandy//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013; 729-736

[15] Al-Khalifa H S, Al-Eidan R M. An experimental system for measuring the credibility of news content in Twitter. International Journal of Web Information Systems, 2011, 7(2): 130-151

[16] Suzuki Y, Nadamoto A. Credibility assessment using Wikipedia for messages on social network services//Proceedings of the 9th International Conference on Dependable, Autonomic and Secure Computing. Sydney, Australia, 2011; 887-894

[17] Chen Liang, Liu Zhiyuan, Sun Maosong. Expert finding for microblog misinformation identification//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012; 703-712

[18] Rubin V L, Liddy E D. Assessing credibility of weblogs//Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. Boston, USA, 2006; 187-190

[19] Olteanu A, Peshterliev S, Liu X. Web credibility: Features exploration and credibility prediction. Advances in Information Retrieval, 2013, 7814(1): 557-568

[20] Wang Teng, Zhu Qin, Wang Shan. Fact statements verification based on semantic similarity. Chinese Journal of Computers, 2013, 36(8): 1668-1681(in Chinese)
(王腾, 朱青, 王珊. 基于语义相似度的 Web 信息可信分析. 计算机学报, 2013, 36(8): 1668-1681)

[21] Lucassen T, Schraagen J M. Trust in Wikipedia: How users trust information from an unknown source//Proceedings of the 4th Workshop on Information Credibility. North Carolina, USA, 2010; 19-26

[22] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining//Proceedings of the 7th Conference on International Language Resources and Evaluation. Valletta, Malta, 2010; 1320-1326

[23] Wu Wei, Xiao Shi-Bin. Sentiment analysis of Chinese microblog based on multi-feature and combined classification. Journal of Beijing Information Science and Technology University, 2013, 28(4): 39-45(in Chinese)
(吴维, 肖诗斌. 基于多特征与复合分类法的中文微博情感分析. 北京信息科技大学学报, 2013, 28(8): 39-45)

[24] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257-286

[25] Yu S Z, Kobayashi H. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. IEEE Signal Processing Letters, 2003, 10(1): 11-14

[26] Xie Y, Tang S, Tang C, Huang X. An efficient algorithm for parameterizing HsMM with Gaussian and Gamma distributions. Information Processing Letters, 2012, 112(19): 732-737



XIE Bai-Lin, born in 1982, Ph.D., associate professor. His main research interests include misinformation detection in microblog, online social network.

JIANG Sheng-Yi, born in 1963, Ph.D., professor. His main research interests focus on data mining.

ZHOU Yong-Mei, born in 1971, professor. Her main research interests focus on sentiment analysis.

XIE Yi, born in 1973, Ph.D., associate professor. His main research interests focus on network security.

LI Xia, born in 1976, Ph.D. candidate, associate professor. Her main research interests focus on natural language processing.

Background

Nowadays, microblog has become a popular social medium for information dissemination. Hundreds of millions of users share information on different aspects of life everyday in

microblog. Everywhere and everytime, a user can disseminate any information by a computer or mobile phone. Microblog facilitates real-time propagation of information to a large

group of users. This makes it an ideal environment for the dissemination of news directly from the news source or geographical location of events. However, it could be seen that a lot of misinformation is sent and forwarded everyday in microblog, due to the lack of an effective scheme for detection and countermeasures.

Most of existing methods detect popular misinformation based on classification algorithms, for example C4.5, SVM and so on. They don't consider the information gatekeepers' behaviors, and they can't identify the popular misinformation early. In order to identify the popular misinformation early and reduce the harmfulness of misinformation, in this paper a new method is presented for detecting misinformation based on gatekeepers' behaviors. In the proposed scheme, hidden semi-Markov model is used to describe the behaviors of the forwarders and reviewers of popularly true information.

Gamma distribution is introduced to describe the state duration of the model. The proposed method includes a training phase and a detection phase. In the detection phase, the average log likelihood of every observation sequence is calculated, and the credibility value of information is updated in real time. This method can identify the popular misinformation early and reduce the harmfulness of misinformation.

This work is supported by the National Natural Science Foundation of China under grant Nos. 61202271, 61572145, 61402119, the Training Program for Outstanding Young Teachers in University of Guangdong Province under Grant No. GWTPSY201403, the National Social Science Fund of China under Grant No. 13CGL130, the Humanities and Social Sciences Research Youth Foundation of Ministry of Education of China under Grant Nos. 14YJC870021, 13YJCZH258.