**清 华 大 学**

**综 合 论 文 训 练**

**题目：社交网络中的谣言检测系统**

|  |  |
| --- | --- |
| 系 别： | 软件学院 |
| 专 业： | 计算机软件 |
| 姓 名： | 钟仰新 |
| 指导教师： | 刘世霞 副教授 |

2016年 月 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

近年来，社交网络（如推特、微博）被互联网用户广泛使用。但同时此类平台中存在大量不真实谣言，其传播造成的危害巨大。由于社交网络消息数目庞大，使得人工排查谣言成本很高，因此对谣言的自动检测技术具有重大意义。

本文描述了一个谣言检测系统，此系统首先对社交网络中的消息进行谣言特征匹配，检测出疑似谣言的消息；然后对这些候选消息进行聚类，生成疑似的谣言话题；接着对候选话题进行特征抽取，利用监督学习技术对其进行可疑度排名，最终检测出最有可能为谣言的话题。

在对消息进行聚类时，本文采取了基于相似度矩阵的聚类方法，研究了包括发布时间、消息标签、命名实体、文本单词、用户群体的相似度度量，聚类上采用了k均值、层级聚类以及谱聚类等方法，并对比了不同度量和聚类算法之间的优劣。实验表明，此系统能有效地将属于同一话题的不同消息归为一类，降低人工排查谣言的成本。

在对候选话题进行特征抽取时，本文总结相关工作，选取了45种不同的特征，并引入了名为过滤器和包装器的两种特征选择技术，对比两种技术并将其结合使用，最终选择出有效的特征子集。实验证明，经过选择的特征子集能有效提高系统对谣言的识别准确度。

在对话题进行可疑度排序时，系统选取了决策树和朴素贝叶斯两种分类器算法，并引入多分类器组合投票技术。实验表明，引入组合投票技术能有效提高排名系统的实用性，显著降低人工排查的成本。

**关键词**：社交网络；谣言检测；聚类；特征选择；监督学习

ABSTRACT

In recent years, social network (e.g., Twitter and Weibo) has been widely used on the Internet. But at the same time, there are a large number of false rumors in social network. Their spread may have harmful effect on individuals and society. The large amount of messages on social network make the cost of manual checking extremely high. As a result, automatic rumor detection techniques are of great significance.

This thesis describes a rumor detection system. It first filter out suspected rumor tweets by pattern matching, and then cluster these candidate tweets into topics. After that, it extract features of the candidate topics and rank them by the likelihood of false rumor through supervised learning technique. In this way, the most likely rumor topics can be detected automatically.

The system applies clustering techniques based on similarity matrix. The thesis studies similarity metrics that consider different factors including published time, hashtags, named entities, terms and users. The system employs k-means, hierarchical and spectral clustering algorithms. Experiments demonstrate that the system can lower the cost of rumor checking by effectively clustering similar messages into same topics.

For feature extraction, the thesis chooses 45 different features and introduces two feature selection techniques, Filter and Wrapper, to select effective feature subset for the training of next step. A method of combining these two feature selection techniques have been proposed in this thesis. Experiments show that the new method and the step of feature selection can effectively improve the accuracy of rumor detection.

To rank and find most likely rumor topics, the system employs Decision Tree and Naive Bayes as classifiers and a multiple classifiers combining technique by voting has been introduced. Experiments demonstrate that the combining method can effectively improve practicability of the detection system.

**Keywords:** Social network; rumor detection; clustering; feature selection; supervised learning

目录

[第1章 引言 1](#_Toc452595199)

[1.1 社交网络 1](#_Toc452595200)

[1.2 社交网络中的谣言 1](#_Toc452595201)

[1.3 谣言检测 3](#_Toc452595202)

[1.4 谣言检测面临的挑战 4](#_Toc452595203)

[1.5 相关研究概述 7](#_Toc452595204)

[1.6 论文组织结构 9](#_Toc452595205)

[第2章 系统概述 10](#_Toc452595206)

[2.1 原系统流程概述 10](#_Toc452595207)

[2.2 原系统的不足&新系统的改进 12](#_Toc452595208)

[插图索引 13](#_Toc452595209)

[表格索引 14](#_Toc452595210)

[参考文献 15](#_Toc452595211)

[致 谢 17](#_Toc452595212)

[声 明 18](#_Toc452595213)

[附录 A 外文资料的调研阅读报告 19](#_Toc452595214)

# 引言

## 社交网络

社交网络是近十年来新兴起的一类互联网社交平台，其在国外的代表有推特、脸书（Twitter, Facebook），在中国的代表有微博、人人网等。社交网络，顾名思义是将人们日常的社交活动，推广拓展到了互联网中，或者说是由互联网公司提供一个平台，给互联网用户提供在线交流、分享、交友等服务。

近年来，这类平台的理念、模式日趋成熟，在发展上取得了巨大成功，甚至逐步成为现代人生活的一部分。究其成功的原因，是这类平台将社交这种活动的门槛降低，同时拓展了其形式、主体，大大满足了人类喜爱社交的心理：任何人都能很简单地在此类平台上注册，然后通过简单的操作来发布消息、接收消息、推广消息，用户也能利用平台添加生活中认识的朋友，或者结识素未谋面的新朋友，通过将社交活动移植到互联网，社交网络成功让人们不出门就能完成社交；正是由于社交网络能通过很低的成本拉近人与人的距离，很多文体明星、政治名人（如歌坛巨星、美国总统）都在平台上注册，人们从此可以通过关注他们的账号，与这些原本在生活中很难相遇的名人进行交流、互动，而名人从此也能通过平台很容易地与自己的粉丝分享生活的点点滴滴、对社会事件的看法等等，为自己赢得更高的知名度和更多的人气；同时，社交网络的注册不仅限于个人，任何社会团体（如商业公司、公益组织）也都能注册账号，发布与自己相关的信息，通过平台进行宣传、推广，这就将社交的主体从人与人拓展到了人与团体，甚至团体与团体。

正是由于社交网络的壮大发展，使得这类平台中的信息规模以指数形式爆发增长，每天有越来越多的信息在不同的平台用户间流动、传播，产生各种影响和价值。因此这几年社交网络获得了学术界的广泛关注，越来越多的学者开始研究其中的信息流动模式，或发展相关技术来自动检测有价值的信息和用户，希望能通过研究更好地发挥社交网络的作用，使其产生正面的、更有价值的社会影响。

## 社交网络中的谣言

在现实的社交生活中，存在着一些不真实或真实性有待确定，但却被社交圈中的人们广泛讨论、传播的“小道消息”，学术界称之为“流言”或“谣言”(rumor)。与真正社交活动相似，社交网络中也存在大量的谣言，这些谣言一般分为两种：误传消息和虚假消息。

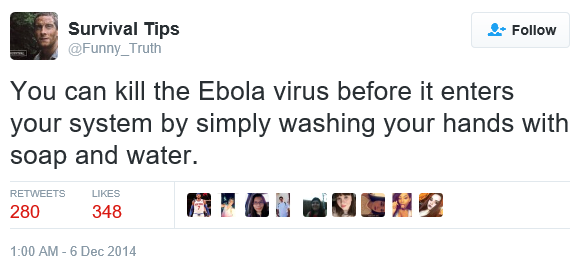
误传消息原本多为源头正规的真实消息，但却在信息的传播过程中被部分用户误解，导致消息在传播中逐渐走样、变形，成为不真实的消息。如图 1.1是推特中的一则误传消息，消息发布于埃博拉病毒盛行的2014年，其大意是我们可以通过用肥皂洗手，简单地杀死手上那些还未进入体内的埃博拉病毒。之所以说这是一则误传消息是因为消息本身并无恶意，是想鼓励更多的人勤洗手来预防埃博拉病毒，事实上这也是世界卫生组织一直提倡的；但是此消息的内容却不具有真实性，原因是用肥皂洗手可以减少手上附着的病毒数目，却不一定能简单地杀死埃博拉病毒（参考信息来源：http://www.killebolavirus.com/does-hand-sanitizer-kill-the-ebola-virus/），而世界卫生组织也从未提到一般的肥皂能有效杀死埃博拉病毒。

图 1.1 推特中的误传消息

（图片来源：https://twitter.com/funny\_truth/status/541155156264382465）

虚假消息，则是另一种更加普遍、潜在危害更大的谣言。这种谣言的特点是其发布时并无可靠的消息来源或有力的证据进行佐证，通常是人为臆测或刻意捏造产生的，其内容不具有真实性且常带有恶意或中伤。但通过包装，此类消息能获得很高的隐蔽性，虚假性很难被人们一眼辨识出来；通过包装，此类消息也能具有很强的轰动性，使得很多用户不顾其是否真实，盲目地争相转发。由于社交网络的用户量庞大、用户网络连通性强，虚假消息很容易被广泛传播：从一开始仅有的零星几个源头，发展成被成千上万人知晓、推送只需要短短几个小时。一旦这类消息被广泛传播，将有可能对某些个人或者团体的利益造成巨大损害，甚至引起社会的恐慌，产生非常负面的社会影响。如图 1.2这则推文谈论的就是一则虚假消息：当时谣言盛传美国摇滚歌手Akon在刚果共和国的一场庆典中将自己包裹在一个巨型的泡泡球内进行表演，其目的是为了避免自己在与群众接触时感染到当时在非洲盛行的埃博拉病毒。这则谣言当时在推特上疯传，部分用户对谣言表示怀疑、不确定（如图 1.2），也有Akon的粉丝站出来为他说话，但也有很多用户因此对Akon的为人表示愤怒、谴责和质疑，这使得其名声受到了很大的冲击。但后续的相关报道证实这只是一则人为臆测的虚假消息：Akon这种表演形式只是为了和观众互动，远在几年前埃博拉病毒未爆发时，他就已经用过同样的形式在世界各地进行表演了（参考信息来源：http://www.independent.co.uk/arts-entertainment/music/news/akon-didnt-perform-in-a-bubble-in-dr-congo-because-of-ebola-he-was-just-having-a-fun-time-for-9772004.html），原谣言缺乏有力论据。

图 1.2 推特中的虚假消息

（图片来源：https://twitter.com/realalexjones/status/517808892437204992）

由于社交网络每天都产生并传播着大量的消息，其中混杂了不少诸如此类的谣言，如果不加以监督管制，尽早发现危害性高的谣言，及时对其进行辟谣并将谣言散播者绳之于法，那么社交网络将成为虚假消息传播的温床，将对个人、团体以及社会造成不可估量的负面影响。

## 谣言检测

因为社交网络存在不少拥有负面影响的谣言，所以必须找到一种有效的监管机制。这就面临着一个很重要的问题：如何找到社交网络中的谣言？

一种朴素的想法是设立“消息审查员”的职位，利用人力去审核推特上的消息，找到其中疑似的谣言消息。但是社交网络的用户规模、消息数量巨大，若要对消息进行逐条的人工审核，那么需要的审查员数量也将极大，其人力成本将极高。因此人工地进行逐条审查并不现实，这不是一个合理有效的解决方案。

正是因为人工检测成本太大，所以研究学者们转而研究开发能自动识别、检测出谣言的技术。这类技术通常会利用计算机强大的处理能力分析海量消息，分析其文本、发布者或传播路径等等的特点，自动找到可疑度高、疑似谣言的消息。但由于机器识别的准确率有一定的限制，所以机器识别出来的可疑消息通常还需要送与审查员进行人工复查，确认为谣言才能采取后续的辟谣和法律相关工作。

以上是当前被广泛认可的，监管社交网络中谣言传播的解决方案之一，学术界称之为谣言检测（rumor detection）。目前还存在一些其他的解决方案，如设立辟谣公众账号，一旦群众用户发现疑似谣言的消息，就举报给公众账号，由公众账号选择举报度高的消息进行审查、辟谣；但本文只专注于谣言检测而略过其它。

本文所指的“谣言检测”，是专指从社交网络的消息中，自动检索出谣言（包括误传消息和虚假消息）相关消息的技术。其中“相关消息”是指对谣言进行传播、讨论、怀疑、质问、辟谣的消息。而有关谣言源头分析、谣言传播轨迹追踪等等的其它相关研究，都不属于本文讨论范围之内。

## 谣言检测面临的挑战

谣言检测技术的研发面临了许多问题和挑战，以下是其中两个主要的挑战：识别准确度不高，检测重复率过高。

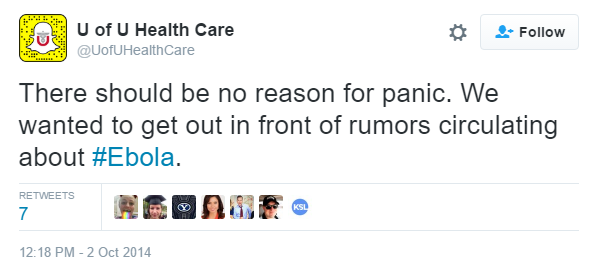
识别准确率不高，是指很多非谣言的消息被机器错误地识别为谣言，或者许多真正的谣言被机器漏过，误认为是普通消息。造成这个问题的因素很多，最主要的一个是分析自然语言难度很高，目前的技术通常只能对文本中的词或者短词组进行语义、情感分析，却很难将文本形成一个整体去分析句子或段落的真正意义。如图 1.3就是一个例子，在这则推文中，存在着如恐慌（panic）、谣言（rumors）和流传（circulating）这些词语，如果仅仅分析它们的语义和情感导向，机器很容易将这则推文误认为是谣言相关的消息。但事实上在恐慌这个词前面还存在没有理由（no reason）这个词组，在谣言和流传这些词前面还存在走出（get out）这个词组。只有将这些词组都串联起来考虑，这个句子的真正意义才能被分析出来：它不是在讲一引起恐慌的谣言，而是在讲我们不应当恐慌，而应该走出来面对那些传播的谣言，这是一则普通消息。

图 1.3 非谣言例子：自然语言分析面临的困难

（图片来源：https://twitter.com/UofUHealthCare/status/517755407100420096）

除此之外，即便自然语言的语义能被机器分析得十分透彻，也不一定能准确地判定一则消息是不是谣言，因为谣言的判定要求对消息的来源、内容的论据进行深入的考证，这就涉及到一定的专业知识和技巧，这不是单一的文本分析就能完成的事情。如图 1.4中的推文称感染了埃博拉病毒的护士Nina Pham的男朋友也被检测出埃博拉病毒的症状，在推文后面作者还附上了佐证的新闻网址，点开后也的确是一则符合推文内容的网页新闻。现在你能告诉我这是一则新闻推广还是一则谣言吗？或许有的读者会认为是前者，因为其附上了消息来源，是一个正规的新闻网站。但事实上我们并不能这么简单地下结论，因为即使是正规的新闻网站也会由于监管不力或为了博取点击率等原因出现一些不实的报道。很遗憾，这则新闻正是这样的报道，后续已有大量媒体对这则谣言进行了辟谣（参考信息来源：http://www.ibtimes.com/ebola-nurse-nina-phams-boyfriend-rumored-admitted-hospital-ebola-symptoms-alcon-releases-1707586）。这个例子说明，光是文本分析并不能完成谣言检测的任务。

图 1.4 谣言例子：谣言判定面临的困难

（图片来源：https://twitter.com/Jody\_Arrington/status/523633560779907074）

因此为了提高谣言检测的准确度，许多研究学者不仅考虑文本语义方面的因素，还综合考虑了如消息发布者信誉度、消息来源可靠度、消息传播路径拓扑结构等等的一些特征，在多元特征上加之以复杂的分析模型（如规则模型、概率模型），最终形成较为鲁棒、识别准确率较高的谣言检测系统。但这其中面临的问题就更多了，例如：如何找出那些有价值的、需要被考虑的特征？不同的分析模型各有优劣，它们之间能否形成互补？本文在接下来的第4章中也在尝试解答以上的两个问题。

检测重复率过高，是谣言识别技术面临的另一个挑战。现在，假设我们的谣言检测系统的识别准确率达到了100%（即便这是不可能的），即便如此，但因为社交网络中的消息数目极大，经常会有成千上万个用户在讨论同一个话题，假设他们都在讨论、推送同一则谣言，这些消息自然都能被我们卓越的系统成功地检测出来。但当我们将检测出来的谣言交给审查员进行人工复查，或者交给相关部门进行辟谣和法律维权时，问题就出现了：这么多的谣言消息，审查人员和相关部门根本看不过来，最重要的是，有大量的消息都是关于同一个谣言的，重复去审查、处理这些消息根本毫无意义。要想利用有限的人力资源审查更多不同的谣言，就必须将那些讨论同一话题的消息归为一类，降低检测的重复率。

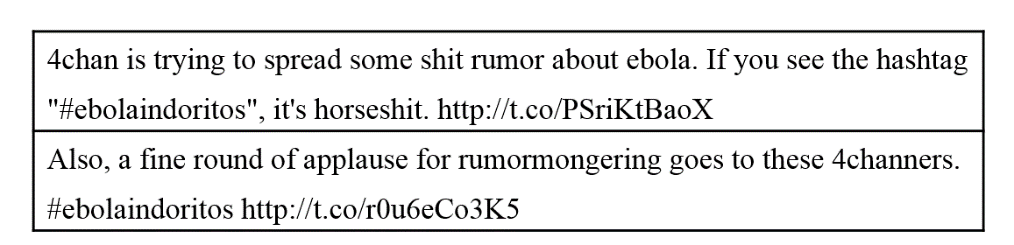
降低重复率，实际上是一个文本聚类的问题。对文本聚类的相关研究已不少，但很多传统的聚类方法对社交网络的文本不适用或不足够准确，因此还需要开发出符合社交网络文本特点的聚类方法。这个问题的主要挑战在于社交网络的消息通常都是短文本，其文字信息不足以完成聚类。如图 1.5的两则推特消息都在讨论同一则谣言：当时在美国的图片网站4chan上盛传一家零食公司Doritos的零食中存在埃博拉病毒。虽然这两则消息在讨论同一谣言，但其文本相似度却很低：由于文本短小，交叠的关键词本就不多，有4chan、rumor、ebolaindoritos，但是在第二则消息中4chan却被表达成4channers，而rumor则与mongering连在一起导致成为了另一个词，所以重复的关键词仅剩下了ebolaindoritos一个；另外文本附上的超链接地址完全不同，尽管它们指向的新闻报道内容相似。如果从传统的聚类方法出发，这两个文本很可能被认为不属于同一类。那么如何才能将社交网络的文本准确聚类呢？研究者们通常会在传统方法的基础上，结合考虑一些社交网络特有的要素，如消息发布时间、消息话题标签、消息提及用户等等。而又因为很多特有要素（标签、提及）不是每个文本都有，所以对应的相似度矩阵经常是稀疏的，所以对聚类方法的选择上也有一定的讲究。本文第3章中将对这个问题进行详细阐述。

图 1.5 两则讨论同一个谣言的消息

尽管谣言检测面临着以上各种挑战，但相关的研究也一直在致力于克服它们。

## 相关研究概述

由于社交网络的兴起是近十年的事情，所以针对它的谣言检测技术目前仍在发展阶段，相关研究不多，大致可以分为两类：监督学习技术、筛选排名系统。

以监督学习技术为主导的谣言检测技术，一般致力于提出针对社交网络消息，适用于谣言识别的有效特征，然后使用带有标记（是否谣言）的社交网络消息数据集，引入各种分类器进行学习、训练，或提出新的分类器模型，用交叉验证的方法评估、找出准确率更高的谣言检测方案。

早期的这类研究选取的消息特征较为简单，大致分为内容特征、用户特征、传播特征[5,10]：内容特征考虑了词语、多元词组，以及社交网络特有的超链接地址（url）、话题标签（hashtag）和提及符号（@[username]）；用户特征考虑了消息发布者粉丝数量、关注数量、影响力、发布消息总量、注册时间、发布消息地区等等；传播特征则考虑了消息是否原创、评论数量、转发数量等。

而后续的研究陆续提出更有效的特征。Shengyun Sun等人的研究[7]加入考虑了多媒体要素，提取社交网络消息中附带的图片，借助搜索引擎评估图片的可疑程度，形成新特征；Carlos Castillo等人的工作[2]虽然时间也很早，但考虑特征高达68类，比前文的基础特征多考虑了标点符号、文本情感评估、消息传播树、话题分布特征等等；Sejeong Kwon等人的研究[4]提出了一种基于时间序列分析的特征，用名为PES的模型拟合消息转发数目与时间的关系，并学习出峰值强度、峰数目、峰周期等10个参数作为新特征；Ke Wu等人的研究[9]则是提出了一种图核（graph kernel）的度量作为新特征，他们将用户分为普通用户和有影响力用户两类，对消息抽取其在这两类用户间转发的传播树，并提出一种基于随机游走的算法计算树的相似度，最终形成新特征。

而在分类器方面以上研究选择各不相同，一部分选用了较为简单的分类器：采用决策规则的研究[2]、采用朴素贝叶斯的研究[5,7]、采用线性回归的研究[5]、采用支持向量机的研究[2,4,9,10]、采用决策树的研究[2,4,7]；也有的研究使用了较为复杂的分类器：采用贝叶斯网络的研究[2,7]、采用随机森林的研究[4]、采用神经网络的研究[7]。以上研究根据消息数据集的不同、特征选取方案的不同，表现最好的分类器也各不一样，目前仍没有定论说哪种分类器最好，也缺乏理论指导在什么类型的数据集中采用哪种分类器更好。

以筛选排名系统为核心思想的谣言检测技术，不是对每条消息直接进行分类，它们首先对消息数据集进行话题提取，然后对同一话题中的消息再抽取统计特征，最后设计出一套可疑度评估规则，或是引入可求后验概率的分类器，对消息话题进行可疑度排名，筛选出最可能是谣言的话题列表。

Tetsuro Takahashi等人的研究[8]提出了一套简单的谣言筛选机制，总共有三个步骤：一，对消息数据集的文本进行命名实体(named entity)抽取，对抽取的关键词做爆发检测（burst detection），找出讨论频率高的话题；二，对话题相关的消息集计算转发率，设置阈值筛选出消息转发率高的话题；三，对话题消息集统计其含有谣言线索词（如“false rumor”）的消息比例，设置阈值筛选出那些含谣言线索词比例最高的话题，作为最疑似谣言的话题列表进行输出。这套系统使用的技术和筛选机制过于简单粗糙，其检测的准确率和召回率也势必不高。但他们的思路值得借鉴：不再是对每条消息进行谣言判定，而是先检测出重要的话题，再检索相关消息，对话题内的所有消息进行统计特征的提取；不再是判定是否谣言，而是对消息话题进行可疑度排名；不再纯粹用分类的思路，而是引入了线索词的概念，尝试用文本匹配筛选出谣言消息。

Zhe Zhao等人的工作[11]借鉴以上思路并进行拓展，设计出了一套更加完整、鲁棒的筛选排名系统。该系统分为五步：一，通过模式匹配检测出谣言信号消息；二，谣言信号消息聚类；三，对信号消息进行关键词组提取；四，利用关键词组比对将普通消息归到信号消息类中；五，使用监督学习技术对每个消息类进行可疑度排名。该工作论证了他们系统在时间上具有高效性，而且能很早地检测出那些刚开始传播的谣言，这两个特点使得该系统能处理大规模的消息数据，还能用于及时阻止谣言的进一步传播，非常满足社交网络的真实需求，实用性高。事实上本文的系统正是在此系统的基础上进行提高、拓展，因此在第2章中还会再详细介绍此系统的流程架构以及本文所做的改进。

更详尽的相关工作介绍可参考附录A。

## 论文组织结构

本文剩余部分将按照以下结构进行组织：第2章将介绍本文实现的谣言检测系统概况和关键改进；第3章将详细介绍各类特征选择技术，并新提出的一种特征选择方法；第4章将简要介绍本系统采用的监督学习技术和组合投票技术，并通过实验比较不同的特征选择技术和分类器选取对系统检测准确率的效果，且论证了新特征选择方法和组合投票技术对系统的提高作用；第5章将进行总结并介绍未来的工作。

# 系统概述

## 原系统流程概述

本文实现的谣言检测系统是基于一项已有工作的谣言筛选排名系统[11]，对其进行拓展和改进，本小节将对原系统的检测流程和设计细节进行介绍。

图 2.1 原系统流程图

（图片来源：参考文献中Zhe Zhao等人的工作[11]）

原系统的流程图见图 2.1，一共有五步，下面逐步进行介绍。

第一步，信号消息识别。该工作所指的“信号消息”（signal tweets）是指那些包含怀疑、惊讶、质疑、辟谣文字的消息，他们认为任何谣言在传播过程中，都一定会受到一部分明智用户的怀疑或质疑，而随着时间的推移也会有用户出来辟谣，因此谣言检测可以从检测这些“信号消息”出发，提取这些消息讨论的话题内容，就是潜在的谣言。而原系统识别信号消息使用的是文本的模式匹配，图 2.2就是他们使用的匹配模式的正则表达式和类型。由于社交网络消息为短文本，文本模式匹配耗时低，即便是对规模很大的消息数据集检测信号消息，其速度也非常快。当然，符合模式的消息中有很多谈论的并不是谣言（例如可能是令人难以置信的真实新闻），所以还需要后面的步骤进一步识别。

图 2.2 原系统使用的匹配模式

第二步，信号消息聚类。对上一步检测出来的信号消息进行聚类，原系统使用的是基于距离阈值的聚类方法，距离函数使用Jaccard相似度（一种广泛用于计算集合相似度的度量），具体方法如下：对需要计算距离的两消息，分别提取文本的单词、双单词词组以及三单词词组形成各自的一个集合，最后对这两个集合计算Jaccard相似度；在聚类过程中对于两个消息，如果其集合的Jaccard相似度超过预定义的阈值，则归为一类。也就是说：1、如果消息a被归入类S，那么S中至少存在另一则消息b，满足a与b提取出来的集合相似度超过该阈值；2、如果S1与S2是不同的类，那么对于S1中的任一消息a和S2中的任一消息b，它们提取出来的集合相似度都不超过该阈值。这里每个信号消息类表示一个话题。

第三步，对信号消息类提取关键词组集。该工作中的关键词组集（statements）的集合元素是那些在信号消息类中出现频率超过预定义阈值的单词、双单词词组和三单词词组。关键词组集的元素将会作为该信号消息类的话题关键词。

第四步，归类非信号消息。这一步的目标是将整个消息数据集中的那些非信号消息尝试归到某一个信号消息类中形成完整的消息类（也允许不归入任何类）。使用的方法与第二步类似，但利用了关键词组集：对于一则非信号消息，先提取它的单词、双单词词组以及三单词词组形成的集合，然后遍历所有信号消息类，计算该集合和信号消息类的关键词组集的Jaccard相似度，如果超过预定义的阈值则将该非信号消息归入此消息类，否则就去看下一个信号消息类是否满足条件；如果所有信号消息类都不满足条件，则该非信号消息归类失败，不列入消息类中。这一步产生的所有消息类（每类必含信号消息，可能含非信号消息），将作为筛选后产生的谣言候选话题，进入最后一步的可疑度排名。

第五步，消息类可疑度排名。对第四步中产生的每一个候选消息类，提取其统计特征（包括信号消息比例、转发消息比例、消息平均长度、消息平均含超链接数目、含话题标签数目、含提及符号数目等），将这些特征输入分类器，利用有标数据集（谣言消息类vs非谣言消息类）进行训练。原系统选择了决策树作为分类器，因为它不仅能对某一消息类进行分类，还能输出一个类似“后验概率”的值，表示该消息类属于谣言话题的可能性。将训练后的分类器对所有消息类计算它们是谣言的“后验概率”，按此值对消息类进行排序，找出可疑度最高的N个消息类，作为系统最终检测出来的谣言话题进行输出。

以上就是原系统完整的流程步骤，本文首先按此进行了完整的实现，然后针对原系统的局限进行了拓展和改进，形成新系统。下一小节将详细阐述。

## 原系统的不足&新系统的改进

在原系统的文章中论证了该系统的高效率，以及其用于谣言防治的实用性。但该系统仍存在以下两个比较显著的问题：

一，检测出的话题重复率过高。由于原系统为了提高速度，在聚类上使用了较为简单的阈值法并设置了较高的阈值，而在距离函数上则选取了计算重复率的Jaccard相似度，因此原系统中每个类内部的所有消息（包括信号与非信号的），其文本内容的重复率都特别高，经常是几乎一模一样的句子，只差了几个单词。这导致的问题是，系统输出的前N个疑似谣言的消息类中，有很多其实谈论的是同一个话题，但是原系统却不能把它们聚成一类，原因只是两类中的文本词组重叠率低。这使得如图 1.5这样的两则消息被归到了两类当中，尽管它们在讨论同一个话题。此问题在原系统的结果集中非常普遍，虽然看上去无伤大雅，但在实际应用中，由于还需要将筛选出来的疑似谣言话题交予人工复核，设想审查员反复看到的几十个候选话题其实都属于同一个话题，那么复核效率必然大大降低。

二，检测准确度不高。

# 插图索引

[图 1.1 推特中的误传消息 2](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584311)

[图 1.2 推特中的虚假消息 3](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584312)

[图 1.3 非谣言例子：自然语言分析面临的困难 4](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584313)

[图 1.4 谣言例子：谣言判定面临的困难 5](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584314)

[图 1.5 两则讨论同一个谣言的消息 6](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584315)

[图 2.1 原系统流程图 10](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584316)

[图 2.2 原系统使用的匹配模式 10](file:///E:\rumor\DataProcess\综合论文训练-钟仰新.docx#_Toc452584317)

# 表格索引

[表 4.1 兔子布偶场景中两种渲染方法的效率比较 14](#_Toc451804965)



参考文献

[1] Cao, N., Shi, C., Lin, S., & Lu, J. (2015). TargetVue: Visual analysis of anomalous user behaviors in online communication systems. IEEE Transactions on Visualization and Computer Graphics, 22(1), 280-289.

[2] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. Proceedings of the 20th International Conference on World Wide, 675-684.

[3] Karamchandani, N., & Franceschetti, M. (2013). Rumor source detection under probabilistic sampling. International Symposium on Information Theory, 2184-2188.

[4] Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. IEEE 13th International Conference on Data Mining, 1103-1108.

[5] Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1589-1599.

[6] Seo, E., Mohapatra, P., & Abdelzaher, T. (2012). Identifying rumors and their sources in social networks. Proceedings of the Society of Photographic Instrumentation Engineers, , 8389

[7] Sun, S., Liu, H., He, J., & Du, X. (2013). Detecting event rumors on sina weibo automatically. Proceeding of the Web Technologies and Applications, 120-131.

[8] Takahashi, T., & Igata, N. (2012). Rumor detection on twitter. Joint 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligent Systems, 452-457.

[9] Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. IEEE 31st International Conference on Data Engineering, 651-662.

[10] Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, Article No. 13.

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

致 谢

感谢张慧老师对我论文的知道。感谢黄经纬同学和我进行的有关光场的讨论，给我的工作拓展了思路。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日 期：\_\_\_\_\_\_\_\_\_\_\_\_

# 附录 A 外文资料的调研阅读报告

1 引言

近年来社交网络如Twitter、新浪微博广泛流行，这类平台的功能随着发展而逐渐扩大，从一开始的朋友间的日常分享、想法交流，到了今天广泛被个人、团体用于消息发布、宣传和推广，社交网络的“社会性”得到了前所未有的扩大。与此同时，社交网络也带来了很多负面的影响，谣言的传播就是其中之一。在此类平台上并不是所有消息都是真实的，部分用户会有意地或者无意地发布一些虚假信息、看似真实的误导信息，抑或是误报信息以及未经证实的可疑信息。这类信息的特点是真实性为假或者待定，但在其发布最初容易被误认为是真实信息，同时其真实性难以被迅速鉴定，我们称这样的信息为“谣言”。有些谣言的传播止于小范围的用户群体，而另一些谣言却由于其内容的轰动性、真实度的隐蔽性，以及平台用户的盲从性得到了广泛的传播。广泛传播的谣言经常会带来不同程度的负面影响：对个人或者团体的形象、利益损害，对大众价值观或认知的错误导向，甚至会引起大面积的恐慌或混乱。而社交网络每天发布的信息量巨大，因此监管部门不可能对消息进行逐条的人工审查，这使得谣言的检测与监控面临巨大的挑战。因此，利用机器来进行社交网络上谣言的自动检测成为了巨大需求。

在这篇调研阅读报告中，作者总结了近年来利用数据挖掘、机器学习等技术进行社交网络中谣言检测的相关研究，对他们进行分类说明。总的来说，这些研究利用了谣言的内容、传播特征等，或是训练分类器对社交网络上的消息进行“是不是谣言”的判断识别，或是提出一套评分机制来筛选出高可疑度的消息或话题名单，由于名单上候选的消息数量有限，这使得后续的人工检测成为了可能。

调研阅读报告的剩余部分将按照相关研究的分类展开，具体如下：章节2将介绍异常检测技术，章节3将介绍谣言源头检测技术，章节4将详细介绍谣言检测中的监督学习技术，章节5将重点说明谣言检测中的筛选排名系统，章节6将进行总结，章节7是本报告的参考文献。

2 异常检测技术

在社交网络中，谣言的散布一定程度上可以视为一种用户的异常行为，而一部分相关研究主要集中于社交网络上的异常检测。Nan Cao等人在他们的文章中提到这类研究一般包括两种：使用带标签（正常vs异常）的训练数据进行有监督的学习，或者提出不需要训练数据的非监督学习模型[1]。在他们的文章当中，他们使用的是一种无监督的学习技术，TLOF模型，这个模型有以下优点：1.这是无监督的学习，因此不需要训练数据；2.此模型考虑了用户的历史行为，而不仅仅是社交网络中某个时刻的情况，因此能更准确地检测出“异常”；3.此模型并非只是给用户进行分类（打上“正常”或“异常”的标签），其还能给出一个“异常”程度的分数评估，这就方便给用户进行排序，筛选出行为最异常的用户名单；4.此模型基于欧几里得进行异常测量，方便于对结果进行可视化（事实上他们设计了一套先进的可视化系统方便监测社交网络上的异常用户）；5.此模型有很高的效率，其时间复杂度仅有O(NlogN)，N为用户数量。

在TLOF模型中，异常检测的核心是一种名为LOF的度量机制，此机制基于于欧几里得空间，假设每个用户都被表示为欧几里得空间中的坐标点，此度量对某个坐标点综合考虑了它距离其K近邻邻居点的远近，以及K近邻邻居点距离它们各自的邻居的远近，计算出所谓的“近邻密度”，从而估算出该坐标点是否属于以及有多大程度属于“离群点”[1]。Nan Cao等人在模型中综合考虑了该用户当前时刻的LOF度量在所有时间的所有用户的数据中的离群程度，以及该用户当前时刻的LOF度量与他自己过去一段时间的历史记录相比的离群程度，进而估算出其当前时刻异常程度的总评分数。他们的研究是通过对每个用户抽取不同的特征向量，形成一个高维的向量空间，在此空间中使用LOF度量。其抽取的特征包括用户的行为特征（发布率、回复率）、内容特征（话题类型、情感评分等）、互动特征（传播消息的广度）、时间特征（发布或回复的时间间隔以及熵等）、网络特征（好友网络出度入度的突变）、用户简历特征（修改用户名的频率等）。

如前所述，这类研究检测的重点是用户的异常行为。而谣言识别经常是需要检测出被广泛传播的谣言或虚假消息，虽然也需要检测出是哪些用户发布谣言以及在谣言的传播中起关键作用，但其研究的直接对象是社交网络的消息或话题，而并非用户。第二，“异常”并不等同于“谣言”，在异常检测中会利用“离群度”来衡量可疑程度，但请试想如果一个谣言被认为是真实消息，从而被广泛传播，那么在短期一段时间中其特征与真实新闻将类似，从而会有较低的离群度而导致无法被识别出来。最后，在此类研究中会将用户的当前时刻行为在历史记录中的离群度作为异常的度量因素之一，但设想一个经常散布谣言的用户，当前行为相较于历史记录的离群度并不能作为衡量“异常”或“谣言”的标准。因此，异常检测技术与谣言识别技术虽然相关，但也有明显差异。

3 谣言源头检测技术

在相关研究中，有一类工作着重于谣言的源头检测，这类工作一般会在社交网络上设置一些“监测节点”（实际上也就是某些用户），让这些节点汇报他们是否在社交网络上听到过某条特定的消息或谣言，然后根据监测节点的汇报情况以及社交网络中的follower-followee关系，用有向图对信息流动建模，进而分析出消息或谣言的源头，或者传播中心。

Ensoo Seo等人在他们的研究中就是采取了这样的方式[6]。他们将监测节点分成两类：正节点（听过谣言的节点）和负节点（没有听过谣言的节点），然后分析每个节点到这两类节点的可达性和总距离，对他们进行排序，找出最可能为谣言源头的节点（用户）。他们基于的基本假设是，如果网络中仅有一个谣言源头，那么它将到所有正节点都是可达的，如果存在多个这样的节点，那么其中到所有正节点总距离越小的节点越可能是源头；理想状态下源头节点到所又负节点都是不可达的，也就是及源头节点到所有负节点的距离都是无穷，而实际分析中的原则是到所又负节点的距离越远越可能是源头节点。而在具体实现上，他们综合考虑了这四个因素对节点进行评分和排序。

Ensoo Seo等人在该工作中同时提出了判断某消息是不是谣言的一种算法。他基于的基本假设是，谣言应该是单一源头或者源头节点数量很小，而真实的消息应该是多源头一起发布、相互佐证。为了选出所有可能的源头节点集，他们在文章中提出了一种贪心算法：每次从网络中找出一个节点，此节点在正节点集中拥有最多的可达节点，把此节点加入源头候选集中，再将它在正节点集中的可达节点都从正节点集中去除，然后进行下一轮筛选，直到正节点集为空。此源头候选集中节点数目越小，则该消息越可能是谣言。此外，为了增加算法准确性，他们提出应当将候选集每个节点到正节点集每个可达节点的总距离作为第二指标，距离越小，则该消息的真实源头数目可能越少，即越可能是谣言。

Nikhil Karamchandani在他们的研究中允许监测节点汇报他们听过某消息的可能性（介于0和1之间的概率，称“感染概率”），而不是他们是否明确听过某消息[3]。另外，他们建立的是无向图模型，更准确地说是在正则树和非正则树两种模型下，通过图中每个时刻节点的感染概率以及节点的连结关系，计算出谣言的“边界”，从而估算出每个节点的“中心度”，从而计算出最有可能的谣言源头。

此类研究的重点在于通过“监测节点”的感染状态和社交网络的图结构，试图还原出谣言的传播轨迹，从而找到最有可能是谣言散布源头的用户，而谣言识别则是着重于分析出哪些消息是谣言，而哪些不是，因此两者在目标和方法上都有相当的区别。虽然Ensoo Seo等人也提出了一种谣言检测的算法，但此算法仅考虑了消息的网络传播特征，而实际上还有很多有价值的特征如内容特征、用户特征等也可以作为谣言识别的考虑因素，此外也还有很多其它有价值的网络特征也值得参考，因此这种算法应当还有很大的提升空间。但是此类研究对于谣言识别技术有相当的参考价值，包括设置监测节点的策略，以及图分析和网络特征的考量都非常有借鉴意义。

4 谣言检测中的监督学习技术

在谣言检测中，有很大比例的研究都采用了监督学习技术。这类研究致力于选择或提出更有效的特征，包括消息的内容特征、传播特征、消息发布者的特征等等，然后使用有标记的社交网络消息数据集，尝试使用、训练出各种分类器，使用交叉验证进行评估，找出分类准确度更优的方案。

Vahed Qazvinian等人的研究不仅致力于检测出谣言，还致力于检测出某消息是在传播或同意谣言的内容，还是质疑、反对它，他们将后一步的目标称为“置信分类”[5]。其选择的特征包括：内容特征，包括消息中的单词（unigram）和二元词组（bigram）在正负样本中的出现情况；网络特征，具体来说是消息发布用户以及消息转发源头用户发布正负样本的比例，实际上也就是消息发布者和被转发者的“信用评价”；Twitter特有特征，包括url和hashtag两类，对附有url的消息抓取url网页内容，然后对其中的单词和二元词组出现情况抽取特征，对附有hashtag的消息，抽取hashtag的单词出现情况作为特征。而在分类器方案上，他们采取的是对每类特征建立贝叶斯分类器，最后将每个分类器的得分线性组合起来作为最终的分类得分，在监督学习时利用训练数据学习出最优的线性函数（即各分类器的权重）。

此工作作为最早的社交网络谣言检测研究之一，其选择的特征比较简单朴素，数量也较少，但是对特征的选取和基本分类却给了后来的相关工作很好的启示和范例作用；其选择的监督学习技术属于朴素贝叶斯和线性回归，也相对简单，有提升空间；其对Twitter数据的收集和标记方式对后来的研究也有借鉴作用，但是比较遗憾的该研究只选取了5个谣言话题，而且在实验设计上存在一定缺陷，每次训练的分类器只是针对某一话题的谣言进行训练和评估，也即这是个狭义的分类器，即使其实验结果达到了90%的F1-score，但该实验并没有表明其具备检测广泛话题的谣言的能力。

Fan Yang等人的研究借鉴了Vahed Qazvinian等人的研究，但研究的对象从Twitter转移到了新浪微博[10]。他们借鉴并拓展了Vahed Qazvinian等人研究的特征选取，修改补充了特征种类，并加入了新浪微博特有的特征：内容特征上考虑了正负情感词语的数目，以及消息是否附有URL和多媒体链接等；传播特征上考虑该消息是原创的还是转发的，以及其评论和转发数目；用户特征上考虑了用户是否实名认证、性别、有无个人描述、头像类型、用户名类型、注册时间、粉丝数量、关注数量、所发微博数量等。另外，他们提出了两种“有效的”新特征，包括目标微博的发布所在地以及发布所用的客户端类型（是移动设备APP，还是网页客户端等），他们的研究发现谣言中有很大比例是使用非移动设备客户端发布的，而且是在国外发布的。而在分类器的选择上，他们使用了SVM技术，核函数采用了径向基函数（RBF）。

此研究的数据收集部分很有参考价值，由于研究的对象是新浪微博，研究者有效地利用了一个新浪微博官方账号“微博辟谣”作为谣言的ground truth收集标准，此官方账号的功能就是对一些传播广泛的微博谣言进行辟谣，因此研究者收集了此账号从2010年到2012两年间的辟谣话题，并使用微博API收集话题相关微博，最后再通过两个标记者人工标记相关微博是支持谣言还是质疑或反对谣言。这样的数据收集和标记方案大幅度的增加了数据集中谣言话题的数目和类型广度，而且在实验设计上该研究并非针对单一话题进行分类，而是对所有话题的数据集进行是否谣言（或说支持谣言）的分类，虽然最后的F1-score仅有78%，但实验表明该分类器可以被用于广泛话题的谣言检测。

Shengyun Sun等人的研究对象也是新浪微博中的谣言，他们借鉴了Fan Yang等人的数据收集方式，也是通过“微博辟谣”账号来收集谣言话题（时间范围是2011下半年，共104个谣言微博），但他们的研究着重于事件谣言（谣言内容是某件事情）的检测[7]。在Fan Yang等人研究的特征选取基础上，他们又进行了扩充：内容特征加入考虑了“@”符号、雷同微博数目、微博中关于事件的动词的数量，以及是否含有强否定词；用户特征加入考虑了“声望”的概念（用户粉丝数目占用户关注加粉丝数目的比例），用户发布的所有微博中含有事件动词的微博比例以及含有强烈否定词的微博比例；他们原创性地加入考量了多媒体（图片）特征，对于带图片的微博，他们利用搜索引擎搜索该图片，没有结果（原创图片）视为一类，如果有结果，则计算图片时间戳与微博时间戳的时间跨度并以某阈值进行二分类，以此作为三类特征；此外他们进一步利用搜索引擎，提出了一种检测图文无关谣言的方法。

在实验中他们尝试使用朴素贝叶斯、贝叶斯网络、神经网络和决策树作为分类器进行训练，实验表明他们提出的时间动词特征、强否定词特征、以及图片时间跨度特征对分类准确度的提升有明显帮助，其中贝叶斯网络和神经网络的F1-score最高（都接近74%）。由于使用了“微博辟谣”进行数据收集，所以他们数据集中的谣言话题也具有多样性，训练出来的分类器也能检测出广泛话题中的谣言，唯一的不足是他们的数据仅仅收集自104个谣言话题的参与用户在2011下半年的所有微博，而且在数据集中正例（谣言微博）的比例只有不到0.4%。

Carlos Castillo等人在他们的工作中对Twitter数据设计了两个分类任务，一是消息是否具有新闻价值，二是消息是否可信[2]。他们的第二个任务实际上也就是谣言检测[2]。此工作的难得之处在于虽然它是最早的谣言检测研究之一，但它考虑的特征种类非常多，共68种基本分为四类：消息特征、用户特征、话题特征、传播特征。其中消息特征考虑了很多符号（问号、叹号、表情符）、消息长度以及情感得分等；话题特征则是同类推特中各类消息特征和用户特征的比例；传播特征考虑的是消息传播树的出度、大小、深度等因素。他们在文章中尝试过使用SVM、决策树、决策规则和贝叶斯网络作为分类器，称这几类分类器有近似的分类准确度，其中最好的是J48决策树（89%）。而位于J48决策树顶层的特征有：话题特征中，含URL的推特比例、含强否定词的推特比例以及含感叹号的推特比例；用户特征基本都位于顶层，发布过的推特数目和好友数目最为重要；传播特征中，传播树中最大层的节点数目位于决策树顶层。

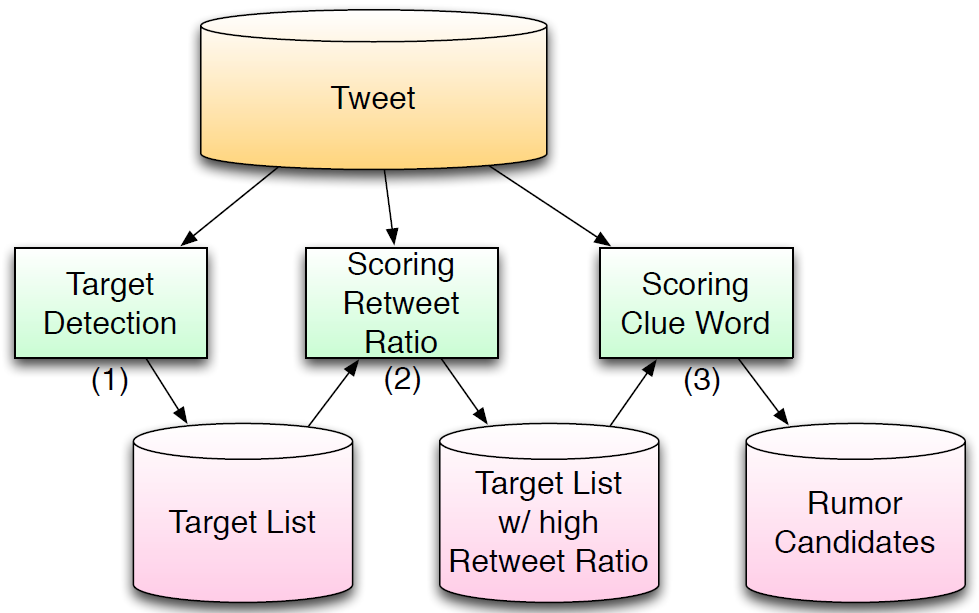
Sejeong Kwon等人的研究提出了一种基于时间序列分析的特征选取，他们基于的基本假设是谣言和非谣言相关的推特数目表现在时间轴上会有诸如峰值强度、峰数目、峰周期、反应延迟等的差异[4]。因此，他们提出了一种名为PES（Periodic External Shocks）的模型去拟合某话题推特数目与时间的函数关系，并采用参数学习技术学出该模型的10个相关参数，作为特征输入分类器，最终对该话题是不是谣言进行分类。在分类时它们还考虑了话题传播树的结构特征（传播特征）和语言特征（内容特征），尝试使用了SVM、决策树和随机森林作为分类器，最终随机森林的F1-score最高（接近90%），并分析出最有效的11个特征，其中包含了PES模型的3个参数特征，但语言特征还是占最高比例。值得赞许的是，他们将所使用的带标记Twitter数据集发布了出来。

Ke Wu等人在他们的研究对象是新浪微博中的谣言，他们提出了一种graph kernel的新度量作为特征[9]。具体做法是，对每一条微博提取出它的传播树（转发评论树），树上节点是转发的用户，将节点分成两类——普通用户或大V用户，分类标准是给用户的“声望”（粉丝数目对关注数目的比值）设置阈值，树上的每条边用一个长度为3的向量表示，分别是转发微博对原微博评论文字的赞许程度、质疑程度以及总的情感得分。他们的研究称过去的研究所选择的传播特征都是诸如出入度、深度、广度这样的广泛值而忽略了一些更具体的传播特征，他们相信消息在普通用户和大V用户间的流动方式对谣言的检测将有很大帮助。但由于传播树巨大，Ke Wu等人提出了一套规则来合并传播树中满足条件的同类节点，以及他们的边，得到一棵约减的传播树，最后他们提出一种基于随机游走的度量计算两棵传播树的相似度，并使用相关急速将计算复杂度降至O(n4i)，其中n是树的节点数目，i是迭代次数。他们在特征选取上除了上述的graph kernel还选择了如内容特征、用户特征等的传统特征，最终将这两者用权重进行配比作为混合内核，采用SVM技术训练分类器。而在“传统特征”中他们其实也提出了一些有参考意义的新特征，如使用LDA分析消息所属的话题分布形成话题特征，以及使用搜索引擎进行简单的谣言分析的搜索引擎特征等。他们的实验最终达到了90%的F1-score，实验显示graph kernel在所有特征中对分类准确率的提升帮助最大。但此graph kernel的问题在于其虽然经过加速，但复杂度依然很高，而且他们的方法需要对两两微博都进行相似度计算，也就是O(N2)的外循环复杂度，则训练集的数目不能过大，这与微博数目巨大、种类繁多的特点是矛盾的。

总的来说，此类工作不断研究社交网络中谣言的内容、发布、传播等各种特性，提出新的更有效的特征，并尝试使用各种分类器找到最优的分类方案。关于其局限性，谣言检测中的监督学习技术非常依赖于数据集的收集和标记质量，而社交网络中大规模带标记的高质量数据集却往往是难以获得的；虽然此类工作公布的分类准确度都普遍很高，甚至能达到90%，但其解决方案往往只能对消息进行逐条分类，即便是10%的错误率在大规模微博下也是非常难以接受的比例；再者，这类工作无法系统地检测出共同讨论一个话题的消息集，并利用消息集的特征判断该话题是否谣言；最后，此类研究的很多特征提取和训练过程都相当耗时，这在数据膨胀的社交网络中无疑面临着规模性的挑战。

5 谣言检测中的筛选排名系统

谣言检测还有一小部分研究提出的解决方案是筛选排名系统。这类研究多是对社交网络的消息数据集进行话题提取，然后对某一类话题的消息抽取特征并进行可疑度与重要度评分和排名，最后筛选出最有可能为谣言的话题候选名单，以此缩小谣言检测范围，最后将有限的候选名单；交予人工判断。

Tetsuro Takahashi等人在他们的研究中提出了一套筛选谣言话题的机制[8]。图1是他们所设计的系统，总共包含三个步骤：第一步，他们采用了一种名词实体提取技术，从所有Tweets的文本中提取出名词实体，然后对每个名词实体，计算出每天含该词的微博数的最大值，并设置最低阈值，筛选出高于阈值的那些实体词形成Target List，这样每个名词实体代表一个话题，这一步实际上做的是一种“爆发检测”，筛选出的是那些被大众热烈讨论的话题；第二步，他们对Target List中的话题做进一步筛选，选取出含有该名词实体的所有Tweets，并计算出其中属于Retweet的比例，并设置最低阈值进行过滤，这一步基于的基本假设是影响重大的谣言或话题都拥有很高的转发率；而第三步则是进一步筛选，计算出含有某名词实体的Tweets集中，含有“线索词”的Tweets比例，而经他们的研究分析，为了不失一般性，最后仅保守地选取了“false rumor”作为线索词，也就是计算出Tweets集中含“false rumor”这个词组的Tweets比例，然后设置最低阈值，筛选出高于阈值的话题作为最后的谣言候选名单，交予人工进行验证。最终他们在数据集中筛选出了10个候选话题，经人工检测其中真正属于谣言的有6个。

图A-1 Tetsuro Takahashi等人的系统

此研究虽然提出了一套筛选系统，但其采用的技术过于简单，设计上也存在很多问题：将实体名词作为话题的代表不尽合理，因为通常一个话题可能包括多个实体名词；仅用关键词的出现次数和retweet率作为筛选条件过于粗糙，并且只能筛选出一些被广泛讨论的问题；而关键的谣言检测步骤也仅是使用了一个线索词“false rumor”进行过滤，显然也是过于简单粗暴的，其提升空间很大；而他们的实验只汇报出前10名的候选话题的准确度，这样的评估也是远远不够的。但是他们的思路值得借鉴，整个系统首先经过“爆发检测”筛选出重要的讨论话题，然后利用谣言的特征线索作进一步筛选或排名，这样的方案不需要训练数据集，执行效率极高，却能从Tweets集中检测出话题，并通过排序选取出最可能是谣言的候选名单。如果将其采用的话题检测、筛选和排名技术进行提升，很可能做出一套实用性、规模性很高的系统。

Zhe Zhao等人的研究设计出的筛选排名系统具有高规模性，并可以用于谣言的早期检测[11]。图2是他们的系统，一共分为五步：第一步，他们先从Tweets集中检测出反映谣言的“信号tweets”，他们的基本假设是在谣言的传播过程中或多或少会伴有质疑、反对或纠正的tweets，这类tweets就是谣言话题的“信号tweets”，先检测出此类tweets可以很大程度帮助谣言检测，而他们采用的检测的方法是模式匹配，图3是他们用于识别信号tweets的模式规则，前三个是质疑的模式，后两个是反对和纠正的模式，如果某个tweets满足其中一个模式则将它识别为信号tweet；第二步，在识别出所有信号tweets之后，对他们进行聚类，该研究非常重视tweets的“重叠率”，他们认为识别讨论同一话题tweets的步骤应当做得尽量谨慎，因为谣言的传播有很大程度都是通过retweet进行的，在此过程中原始tweet和retweet的内容重叠率应当很高，并且由于微博数量众多，聚类的步骤的复杂度应当尽量低，因此他们对信号tweets两两计算内容的单词、二元词组以及三元词组的集合重叠率（Jaccard coefficient）并设置一个较高的最低阈值（0.8），高于此阈值则将两信号tweets聚为一类，此聚类过程有点像图模型中的community detection，由于阈值设置较高，所以每个cluster不会过大，然后筛选出其中规模超过一定阈值cluster；第三步，对每个信号cluster提取出现率较高的单词、二元词组和三元词组作为其statement集合，此集合表示的就是每个cluster的真正的话题内容；第四步，将非信号tweets逐一与每个信号cluster的statement集合进行重复率计算，将超过一定阈值的非信号tweets聚类到该信号cluster中，形成一个话题cluster；第五步，对话题cluster中的信号和非信号tweets提取特征，并采用有监督学习技术训练出一个可疑度排名器，利用此排名器对话题cluster进行可疑度排序，找出最可疑的话题作为谣言候选列表交予人工检测。

图A-2 Zhe Zhao等人的系统

图A-3 信号tweets检测采用的模式规则

值得注意的是此研究中的第五步用到了有监督学习技术，这表明此系统仍需要一定的标记训练集，但为了规模性，该研究选取的特征都比较简单，如信号tweets占所有tweets的比例、话题cluster中词分布的混乱程度（熵）、tweets的平均长度、retweet所占比例以及所含url、hashtag和mention的数目等，并没有第四章的监督技术选取的特征那么复杂和有效。其尝试采用的分类器（排名器）是SVM和决策树，最终发现决策树的准确度更高。最后实验表明，该系统能从每天产生的tweets中筛选出三百个左右的谣言候选话题，大大减少了人工筛选的工作量，其排名前10的候选者中真正的谣言有60%，而前50的候选者中真正的谣言有33%，而所有候选者（共350个）的准确率有26%。相较于采用监督学习的谣言检测技术（最高90%的F1-score），这个准确率显得十分低，但是监督学习技术采用的分类器仅能对每条tweets进行分类，所以错误率即便仅有10%，在巨大的tweets基数下也是不能容忍的；相对的，此系统能对tweets进行话题检测，将一个月的12亿条tweets缩减为350个候选话题，并最终有92个是真正的谣言，其实用性远远高于前者；由于该系统的核心检测步骤是对信号tweets的检测，又因为此研究发现谣言消息的信号tweets都出现得特别早（一般10min到4h就出现），所以此系统可用于谣言的早期检测（此结论经过了实验说明）。最后，由于系统采用的算法复杂度不高，所以其可用于真实的社交网络大规模消息的谣言检测（该系统处理每天产生tweets的10%只需要30min）。

当然，目前该系统使用的聚类算法、信号检测技术、话题抽取技术和可疑度排名技术都较为简单，如果舍弃一定的规模性，采用更精细的技术，比如借鉴第四章监督学习技术中抽取的特征和分类器选取，设计出更好的排名器，那么很有可能提升其候选集的准确度，得到更为让人满意的结果。

总结而言，谣言检测中的筛选系统一般采取的技术都很简易，系统设计也依赖于很多谣言相关的先验知识和观察（如检测信号tweets的模式规则），准确率也不及监督学习技术，但其具有高效率、高实用性的特点。

6 总结

本报告一共介绍了与谣言检测相关的四类技术，其中相关度较高的是监督学习技术以及筛选排名系统。监督学习技术有高准确性，但依赖于高质量的带标记训练集，其算法复杂度通常很高，而且分类只能逐条消息进行，实用性低；而筛选排名系统虽然准确度低而且最后需要一定的人工检测，但它能检测出可能的谣言话题，并对可疑度进行排名，其规模性与实用性都很高。如果能参照筛选排名系统的设计思路，同时借鉴监督学习的特征提取和分类技术，将两者的长处结合起来，那么很可能会取得更好的成果。

最后，社交网络的谣言检测目前还处于不成熟阶段，相关研究数目极少，也还有很多没有尝试引入但可能有效的技术（如半监督、无监督学习，模式识别等），因此有待相关学者进行更多有价值的研究，使该领域的理论更加丰富。

7 参考文献

[1] Cao, N., Shi, C., Lin, S., & Lu, J. (2015). TargetVue: Visual analysis of anomalous user behaviors in online communication systems.*IEEE Transactions on Visualization and Computer Graphics, 22*(1), 280-289.

[2] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide,*675-684.

[3] Karamchandani, N., & Franceschetti, M. (2013). Rumor source detection under probabilistic sampling. *International Symposium on Information Theory,*2184-2188.

[4] Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. *IEEE 13th International Conference on Data Mining,*1103-1108.

[5] Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,*1589-1599.

[6] Seo, E., Mohapatra, P., & Abdelzaher, T. (2012). Identifying rumors and their sources in social networks. *Proceedings of the Society of Photographic Instrumentation Engineers, , 8389*

[7] Sun, S., Liu, H., He, J., & Du, X. (2013). Detecting event rumors on sina weibo automatically. *Proceeding of the Web Technologies and Applications,*120-131.

[8] Takahashi, T., & Igata, N. (2012). Rumor detection on twitter. *Joint 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligent Systems,*452-457.

[9] Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. *IEEE 31st International Conference on Data Engineering,*651-662.

[10] Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics,*Article No. 13.

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

综合论文训练记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学生姓名** |  | **学号** |  | **班级** |  |
| **论文题目** |  | | | | |
| **主要内容以及进度安排** | **指导教师签字：**  **考核组组长签字：**  **年 月 日** | | | | |
| **中期考核意见** | **考核组组长签字：**  **年 月 日** | | | | |
| **指导教师评语** | **指导教师签字：**  **年 月 日** | | | | |
| **评阅教师评语** | **评阅教师签字：**  **年 月 日** | | | | |
| **答辩小组评语** | **答辩小组组长签字：**  **年 月 日** | | | | |

**总成绩：**

**教学负责人签字：**

**年 月 日**