武汉邮电科学研究院硕士学位论文

开题报告

专 业： 通信与信息系统

研究方向： 中文信息处理

论文题目：基于多样免疫的SEIRS网络舆情预测

研 究 生： 田小虎 学号： 20150034

导 师： 廖闻剑、彭艳兵

报告日期： 2016-10-24

|  |
| --- |
| 1. 课题的来源、目的及意义 |
| 随着信息传播技术的发展和普及，网络作为信息交流和知识共享的最好平台，被公认为是继报纸、广播、电视之后的“第四媒体”，成为反映社会舆情的主要载体之一。同时，自20世纪90年代以来，各类公共危机和突发事件在我国不断出现，其频发度和危害性呈明显上升趋势，而网络的迅速普及和网民的飞速增长，网络对现实的影响日益增强。因此，南京烽火星空通信发展有限公司，结合公司资源及业务发展，开发了烽火舆情网，能实时获取贴吧、微博、各大论坛等舆情数据，而本课题在此基础上将获取的数据汇总建模分析，实现对热点数据的舆情预测及监控。  本课题的目的：网络舆情的传播同传染病的扩散机理具有相似性，因此通过引入传染病传播模型，利用传染病模型较为丰富和成熟的模型，研究了网络舆情的衍生效应；同时为网络舆情的衍生传播引入WS小世界模型，研究了对网络舆情的衍生效应对舆情传播的具体影响。实现对互联网中重大突发事件的及时监控，为决策者提供可靠的决策导向，给出切实合理的舆情引导建议。  课题的意义：通过引入了类似传染病的机理，将网络舆情群体类似传染病中的病情的各阶段，将网络舆情中传播的各群体状态进行了类似分类，分析出用户在传播模型中担任的成分和传播方向等重要舆情衍生信息，提高了现实网络舆情传播的契合度；通过引入传播群体的WS小世界模型，对网络舆情群体传播过程进行了科学的分析和精密的计算，得到了合理切实的现状下网络舆情传播趋势，从而给出合理有效的引导建议，提高了决策者的响应效率和快速应对能力。 |

|  |
| --- |
| 1. 本课题的国内外研究现状及发展趋势分析 |
| 国内外研究现状：随着Web2.0时代的来临，信息传播由传播媒体“自上而下”的线性传播，转变为网民可以随时随地发布和接受信息的更为方便快捷的传播方式。另一方面，根据Milgram提出的六度分离推断，网络的组织模式是以真实的社会关系为基础，网络中小团体这种“小世界”在目前的网络舆情阵地，网络中这种小团体这种“小世界”的特点决定了小团体中的网民之间相互信任，处于这种信任和跟从的关系，事情一经网络发布，会很快在网络扩散开来，形成热点事件。目前这些小世界团体主要集中在新浪微博、百度贴吧、微信朋友圈、QQ空间、天涯、铁血等舆情阵地。而在网络舆情预测中，基于传染病的动力学模型研究是一个重要的研究课题。由于网络舆情的传播方式与病毒的传播方式有很多相似之处，所以早在20世纪60年代，Daily和Kendall就借鉴传染病模型提出了谣言传播的DK模型，此后有学者构建了如MT等DK的改进模型、Sudbury借鉴传染病模型中的SIR模型探究谣言的传播规律、Leskovec引入的SIS模型、Gruhl等人扩展的SIR模型、Zhang等人基于SIR模型及群组动力学，构建了博客网络中的热点话题传播模型；国内学者也做了很多尝试，如张彦超将传染病动力学与复杂网络结合提出了一个基于SNS网络的信息传播模型、朱恒民等以无标度网络为载体提出了舆情话题传播的SIRS模型、陈波将潜伏期引入网络舆情传播的带直接免疫的SEIR模型、陈福集针对舆情话题免疫人员个体记忆的免疫时效性问题的SEIRS模型、林晓静针对群体中舆情感染者与舆情未感染者的接触率提出的具有饱和接触率的SEIR模型。  上述研究工作大部分从理论的层面对舆情衍生效应产生的原因和影响因素做了简单分析，并未从建模的角度研究舆情的衍生效应；有些研究虽然提出了类似传染病的动力学模型，但是并未考虑到并非所有的康复者都具有舆情再免疫的等情况；而且以上提出这些模型，仅仅对模型进行了简单的定性分析，并没有对深层次的舆情相关因素和衍生效果进行深入的定量分析。  为了弥补以上工作的不足，本文在以前学者研究的基础上，建立了基于多样免疫的SEIRS网络舆情传播模型，将免疫分为先天免疫、后天免疫及康复无免疫这三种免疫情况，更加符合网络舆情传播的现实情形，并从网络舆情传播源的传播方向上引申出整个网络舆情传播群体的WS小世界模型，分析小世界模型中的各舆情阵营的具体舆情人员，通过研究各阵营流动的稳定状态和流动阈值，得到具体的舆情衍生条件，为后续舆情预测提供有力的数据支持。  参考文献：  [1] 胡亦清. 舆情系统中倾向性分析与实现[D]. 北京:北京邮电大学, 2012.  [2] 周胜臣,瞿文婷,石英子,施询之,孙韵辰. 中文微博情感分析研究综述[J], 计算机应用  与软件,2013,30(3):161-164,181.  [3] 霍良安，黄培清. 突发事件中的谣言扩散问题研究[J].情报杂志，2011,30(10): 77-81  [4] 陈波，于泠，刘君亭，等．泛在媒体环境下的网络舆情传播控制模型[J]. 系统工程理  论与实践，2011,31(11):2140-2150  [5] 马力,宫玉龙. 文本情感分析研究综述[J]. 电子科技, 2014,27(11):180-184.  [6] 马知恩. 传染病动力学的数学建模与研究[M]. 北京 :科学出版社，2004.  [7] 刘常昱,胡晓峰,司光亚,等. 基于小世界网络的舆论传播模型研究[J]. 系统仿真学  报,2006,18(12): 3608—3610.  [8] Zanette D H.Criticality of Rumor Propagation on Small—world Networks [EB/OL] .[2009—  03—01] .http arxiv.org/PS eache/cond—mat/pdf/0109/0109049v1.Pdf [8]  [9] Sehulze C.Sznajd opinion dynamics with global and local neighbourhood[J].International  Journal of Modern Physics C,2004,15(6) :867—872.  [10] Sudbury A J. The proportion of the population never hearing a rumour[J] .Journal of  Applied Probability ，1985(22) 443—446.  [11] Gruhl D,Guha R liben—Nowell D,et al.Information diffusion through blogspace[C]  Proceeding of the 13th International Conference on World Wide Web,New York :ACM Press  2004 :491—501. |

|  |
| --- |
| 三、课题中待解决的关键问题 |
| 本文主要针对网络群体在微博贴吧论坛中的评论、点赞、转发等，并对评论进行情感分析归类，带入传染病模型中各个舆情阵营。建立基于WS小世界的演化模型，分析舆情衍生的演化趋势，分析出各舆情阵营的稳定和流动状态，提出舆情的后续走向及调控建议。在该过程中主要解决如下关键问题：   1. 舆论衍生网络的构建   为了构建合理的网络舆情衍生网络，如何构建WS小世界模型，从某一原始的舆  情源头，衍生出整个舆论体系，拓扑出合理有效的整个舆情网络，这是本文一切可靠数据的重要基础。  2. 适合处理相关事件的传染病模型  对于舆情网络中的某一相关舆情事件，梳理得出该事件的主体脉络，分析出该事件可以划分的舆情阵营，建立并逐步修缮契合该事件的传染病动力学模型，以模型来分析该舆情事件的衍生效应及后续走向。  3. 舆论阵营的划分及个体的归属  本文采用的是传染病动力学模型，该模型在同一时间并行存在多个模型，因此需要  学习各阵营的样本，构建不同阵营的特征词库，将整个舆情网络分成不同的舆论阵营；  对舆情网络中的舆情个体，需要对该个体的行为动作、文本内容进行捕捉分析，得出该舆情个体正确的舆论态度，以将该个体纳入正确的舆论阵营。 |
| 四、课题的研究内容 |
| 本文主要工作是基于网络舆情各群体各阶段现状，运用传染病动力学模型分析各群体用户对网络舆情传播走向的正负影响及大小，进而预测后续整个舆论的导向。传染病动力学模型的核心是各个阶段群体之间在相应条件下的相互转化。本文在已有的网络舆情传染病模型基础及生理医学传染病模型的基础上，考虑对相关舆论带有先天免疫和后天免疫及康复后无免疫能力的舆情传播者对整个舆情导向的影响。在这些过程中主要有下面五个研究内容：  1、WS小世界模型构造算法中的随机化过程可能破坏网络的连通性，因此随机化概率系数将直接影响整个舆情网络的聚合程度，选择合适的概率系数将避免过聚合或过松散的问题。  2、对网络舆情中的文本进行精准的中文切词、构建针对该事件的特殊领域词汇的自定义词库是衡量出该舆情个体的情感倾向的文本分析基础，再结合行为动作对该个体进行正确的舆论阵营划分。  3、分析各舆论阵营人员的稳定条件和流动阈值，研究个舆论阵营的人员流动及舆情的衍生是本文研究的重点问题。  4、针对舆情网络中的各舆情阵营，建立类似传染病的动力学模型，利用传染病的传染机理，对网络舆情的衍生做合理的分析预测是本文的核心问题。  5、对稳定条件和网络舆情阈值进行验证，设置不同的对照组，尤其是设置实时数据的外部对照组，对系统参数及试验结果进行验证。 |
| 五、拟采取的研究方法、实验方案、技术路线 |
| 1. 从舆论源头出发，根据舆情个体的关系网络，逐层抓取个体间的社交关系以及该用户在采样时间内的行为动作及文本内容，建立集群性良好的WS网络舆情衍生小世界模型。 2. 对中文切词采用机械切词和统计切词组合切词的方法，为了准确的划分舆论阵营对切词结果进行情感倾向分析。 3. 舆情衍生的稳定和流动，主要取决于系统传播的内部非零平衡点，建立传染病模型的微分方程组模型，求出该方程组的导数，即为该系统传播的内部非零平衡点。 4. 对不同阵营间传播系数分别设不同的传播系数对照组，分析求解不同传播系数下原始组和对照组的结果。   技术路线：  1. WS小世界的研究。对WS小世界模型的构造进行变化，将原模型中从一个环形规  则网络开始，改为从一个源头发散。先发散出N二次个节点，之后的发散依然遵循WS小世界的构造规则，每个节点向它最邻近的K个节点连出K条边，满足N>>K>>ln(N)>>1  2. 基于多样免疫的SEIRS网络舆情传播模型。网络个体随时间加入舆情网络的输入率为A，输入率A与WS小世界模型的随机化概率系数p有关，根据动力学模型可建立如下微分方程组模型：    ( 1 )  ε      以上阵营间转移系数都是[0,1]且:  ( 2 )    **S**  **E**  **I**  **R**  ρ  β  γ  潜伏者直接变为免疫者（相当后天免疫）ε  康复后无免疫能力μ  相当于先天免疫δ  3. 系统平衡点的研究。  除了无意义的初始无传播平衡点，本文重点研究内部非平衡点。因为系统为有边界系统，且系统初值均在有界区域，则  ( 3 )  令(1)中导数为0，结合(2)、(3)即可求出系统平衡点（即阈值）  4. 系统预测结果的验证  对舆情网络中各阵营传播系数，设置不同的值作为对照组，重要的是实时舆情数据的对照。  5. 中文切词及提取情感特征词的方法研究  为了达到更精准的切词效果，本文采用如下组合分词算法：首先，采用词典分词，先构建专用词典（人名、地名、专业术语等），将字段中的专用名称切出，剩下的部分采用双向分词匹配策略，若两者切分结果相同，说明没有歧义，直接输出。如果不一致，则输出最短路径的那个结果，如果路径长度也相同，则输出单字词少的那一组结果。如果单字词也相同，则选择正向切词结果；此外，采用统计分词，因为词是稳定的汉字组合，在上下文中汉字与汉字相邻出现的概率能够较好的反映成词的可信度（可采用N-Gram元分词模型）。以以上组合模型来构建该舆情话题情感特征词。  6. 对文本评论人员的阵营归属划分研究  阵营划分主要根据文本的情感倾向来划分，采用统计词频的同时去掉高频中的常用词（主要是代词、介词等），剩余高频词作为特征词；同时再用TF-IDF算法统计出词频不高但是重要的特征词；情感倾向一般是由形容词或者副词来决定，因此提取出距离特征词最近的形容词或副词来作为辅助词，对于出现在特征词前面的辅助词选取一段文本内（通常指用数字、标点等非中文符号划分的整段文本）距离最远的特征词、出现在特征词后面的辅助词选取距离最近的辅助词、前后同时出现时选取距离最远的那个辅助词。对于具有否定意义的辅助词，将情感倾向设为反极性。 |
| 六、论文的创新点或实用性 |
| 1. 本文首次综合了数学模型和社会模型，结合了社会模型的适应性和数学模型的严谨性，使预测结果更加合理准确，这在业内属于领先。传统的舆情预测一般基于线性的ARMA、指数平滑，非线性的隐马尔科夫、K近邻、支持向量机等模型，从纯数学模型的角度出发；只有极少数从社会模型角度出发，提出了SEIR、SEIRS等基于传染病传播的动力学模型。但是这些传染病模型很少深入定量分析，带入数学模型的分析方法。  2. 本文首次在传染病动力学模型中加入了康复后无免疫能力的情况，即在感染者不再传播舆情后的一段时间视为康复，但是之后又开始传播舆情则视为康复后无免疫能力，再次被感染。传统的传染病网络舆情传播中，都认为感染者在辟谣后都直接成为免疫者，但并不符合现实舆情状况，毕竟并非所有人在辟谣之后都是“从一而终”再不相信或感兴趣于某舆论。  3. 本文首次组合了词典分词中的所有机械分词方法、基于N-Gram的统计分词方法，能开发出一套精准有效的切词方法。配合各舆情阵营样本，能准确的匹配出舆情人员对应的阵营。为后续的模型提供精确的原始数据，提供定性、定量分析。  4．本文在构建整个舆论体系，采用WS小世界模型，从某一原始舆论源头出发，循环扩展该用户的行为动作、好友关系，能很好的符合舆论传播的社会关系，该舆情衍生模型属于业内前端。 |
| 七、研究计划进度和预期成果 |
| 计划进度：  1、2016年10月-2016年11月，确定论文题目，进行相关理论研究；  2、2016年12月-2017年2月，收集资料，熟悉舆情预测各种线性、非线性模型，重点是基于传染病的动力学模型及相关算法；  3、2017年3月，熟悉相关切词方法和情感分析的相关算法；  4、2017年4月-2017年6月，数据采集并整理，从用户的各种动作行为、文本等对舆情中的用户个体进行舆情阵营的划分，并基于WS小世界模型，不断扩展该阵营；  5、2017年7月-2017年8月，对获取的各阵营人员及流动人员数据进行整理，并进行数学分析，得出舆情传播走势曲线，并作相应预测和验证；  6、2017年9月-2017年11月，完成论文写作、修改定稿。  预期成果：  能基本实现对热点网络舆情的监测汇集，并划分该舆情中各人员的舆情阵营归属，并使用传染病的动力学模型和WS小世界模型，演化该舆情的后续推进，在现有时段数据处理的基础上，能对未来某一时刻的舆情走向进行准确的预测，并给出合理高效的引导建议。 |
| 八、导师意见  （请对项目的意义、具体内容、创新点和取得预期成果的可能性等进行评价） |
| 导师签字： |
| 九、研究生部审核意见 |
| 审核老师签字： |