融合微博热点分析和LSTM模型的网络舆情预测方法

Network public opinion forecasting method fusing Microblog hotspot analysis and LSTM model

刘定一　沈阳阳　詹天明　刘亚军　应　毅

三江学院　南京审计大学　东南大学

二等奖

摘要

当前互联网已成为公众获取信息、表达观点的重要平台，也带来社会舆情事件易发生的风险，通过对网络舆情走势的提前预测，能够准确判断热点事件的发展态势，为政府相关部门应对舆情危机提供参考。针对单一预测模型预测精度不高和社交媒体对舆情走势影响较大的问题，提出融合微博热点分析和LSTM神经网络的舆情预测方法。首先利用网络爬虫和PyTorch机器学习平台构建了用于舆情时序数据分析的网络舆情预测系统；在此系统内，考虑微博的强时效性，采用网络热点分析技术，计算微博热度分值；改进LSTM网络，设计由2个隐含层组成的MH-LSTM预测模型。将MH-LSTM模型用于舆情事件百度指数的定量预测中，通过实验验证了模型的正确性，并证实了该预测模型拥有较好的预测效果。

提出的方法

1. 网络舆情预测系统架构

基于爬虫技术和PyTorch平台构建用于舆情时序数据分析的网络舆情预测系统。系统由3部分组成：数据采集与存储、模型训练、模型部署与应用。

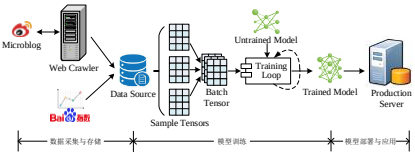

图1. 网络舆情预测系统架构

2. 微博热度计算

在已知热点信息的前提下，定义微博热度分值由转发数、评论数、点赞数的权重累加得到：

$$\text{HotScore}_i = \alpha * \text{转发数} + \beta * \text{评论数} + \gamma * \text{点赞数}$$

舆情热点的形成基本服从二八定律，即对于网络舆情传播造成重大影响的是头部20%的意见领袖；决定整个舆情事件传播规模的，是为数不多的人气媒体和权威网络平台。因此，根据关键词采集热点微博无须穷尽所有，对50个关键词匹配的微博进行热点分析，根据热度分值排序，取前10个累加，得到微博热度总分值，如下：

$$\text{HotScore} = \sum_{i=1}^{10} HotScore_i$$

3. MH-LSTM模型结构

MH-LSTM神经网络模型（Microblog Hotspot-Long Short-Term Memory）由单向LSTM和双向LSTM（Bidirectional LSTM，Bi-LSTM）两个隐含层组成，在保留LSTM网络特性的同时，降低由于训练样本较少而产生过拟合的风险。

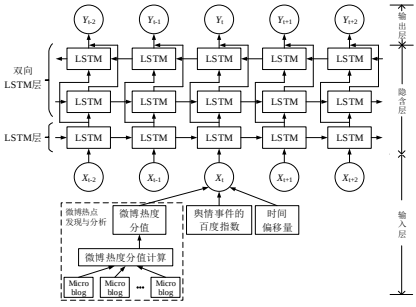

图2. MH-LSTM模型结构

根据舆情预测的特点，模型第一个隐含层的输入包括四部分：百度指数BaiduIndex、微博热度总分值HotScore、时间偏移量△T、上一时该隐含层的输出。

$$h_t^1 = \sigma(W_1[h_{t-1}^1, \text{BaiduIndex}_t, \text{HotScore}_t, \Delta T])$$

其中：h_t^1表示第1个隐含层t时刻的输出，W_1表示第1个隐含层的权重向量，BaiduIndex_t表示t时刻的百度指数，HotScore_t表示t时刻微博热度总分值，△T表示时间偏移量。

模型第2个隐含层的输入包括两部分：同一时刻上一隐含层的输出和同一隐含层上一时间片的输出：

$$h_t^2 = \sigma(W_2[h_{t-1}^2, H_t^1])$$

其中：h_t^2表示第2个隐含层t时刻的输出，W_2表示第2个隐含层的权重矩阵，H_t^1表示t时刻第1个隐含层到第2个隐含层的输入向量。

模型的损失函数是预测误差平方和与模型权值参数的平方和之和：

$$L = \sum_{i=1}^{n}(h(x_i) - y_i)^2 + \alpha\sum_{j=1}^{m} w_j^2$$

其中：n为样本个数，$h(x_i)$表示输入样本x_i时模型的预测输出，y_i为样本x_i的标签，m为模型权重个数，w_j^2表示第j个权重的平方。

4. MH-LSTM模型的训练步骤

(1) 定义MH-LSTM网络结构，根据公式定义损失函数，设置每一层网络节点的舍弃率为0.2，设置优化器为自适应矩估计；

(2) 根据公式计算HotScore，BaiduIndex来源于百度网站；

(3) 每一步处理时间序列中的一个时刻。将当前输入（BaiduIndex_t、HotScore$_t$、△T）和前一时刻输出（h_{t-1}）传入MH-LSTM网络结构，计算得到当前输出（h_t）；

(4) 根据当前输出和实际值计算误差，通过优化器反向传播求解，更新模型参数；

(5) 重复上述步骤直至算法收敛。

实验

1. 数据介绍

以2019年发生的“重庆保时捷女车主打人事件”、“996工作制事件”、“黑洞照片首发事件”三起热点事件为训练样本，以“山东大学学伴事件”为测试样本，用训练样本训练模型优化参数，用测试样本验证模型的有效性和准确性。实验数据主要包括：百度指数、微博热度分值、时间偏移量。

2. 实验结果

将MH-LSTM神经网络模型与粒子群优化BP神经网络模型（PSO-BPNN）、传统LSTM神经网络模型进行对比分析。

利用MATLAB 2015a构造PSO-BPNN模型；依托网络舆情预测系统，构造LSTM模型、MH-LSTM模型进行对比实验。从拟合性和预测精度来看，相比于PSO-BPNN模型，LSTM模型和MH-LSTM模型的预测结果表现得更加优秀，LSTM模型和MH-LSTM模型的预测结果与真实值拟合得较好，预测结果更接近真实值；由于BP神经网络存在容错性差、学习不稳定等缺点，粒子群算法的自身局限性容易陷入局部最优，导致PSO-BPNN模型的预测结果稳定性较差，浮动较大，预测精度上也有所欠缺。

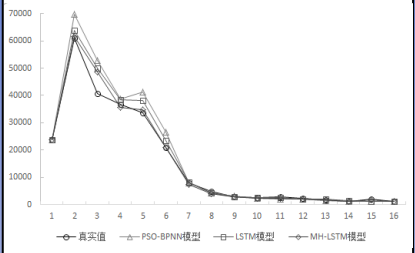

图3. 各模型预测结果对比

计算预测值y_t与实际值t_t的相对误差（Relative Error，RE）和平均相对误差（Mean Relative Error，MRE）。LSTM模型的MRE为0.1027，MH-LSTM模型的MRE为0.0725，都要优于PSO-BPNN模型的MRE（0.1513）。相比于LSTM模型，MH-LSTM模型加入了微博热度分值来修正模型，平均能够降低相对误差3.02个百分点，说明微博热点分析对舆情趋势预测有积极的影响。从相对误差曲线图中可以看出MH-LSTM模型的大部分相对误差都在10%以下，部分时间点的预测值误差接近0，预测结果整体较为稳定。因而得出结论：MH-LSTM模型能够更好的描述时间序列发展过程，具有较强的非线性拟合能力，能够很好地对舆情事件发展进行定量预测。

表1. 各模型MRE数据

	PSO-BPNN模型	LSTM模型	MH-LSTM模型
平均相对误差	0.1513	0.1027	0.0725

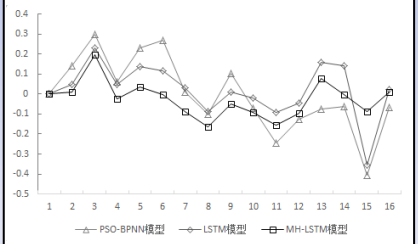

图4. 各模型相对误差对比

结论

网络舆情的发展受到多重因素的影响，呈现出非线性且复杂的变化特点，传统基于统计学和数学方程的方法很难取得较好的预测结果。在互联网快速发展的今天，社交媒体信息成为提高预测精度的积极补充。基于网络爬虫技术和PyTorch机器学习平台构建了针对舆情时序数据分析的网络舆情预测系统。从预测模型和数据扩充两方面进行改进，提出融合微博热点分析和深度学习的新的预测方法——MH-LSTM，该方法结合实时性的微博数据和权威性的百度指数进行网络舆情发展趋势预测，与PSO-BPNN模型、LSTM模型的对比实验验证MH-LSTM模型的正确性和优越性。预测结果有助于政府对舆情信息的控制和引导，有利于社会发展的和谐稳定。

主要参考文献

[1] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[2] 蔡英凤, 朱南楠, 郜康盛, 等. 基于注意力机制的车辆行为预测[J]. 江苏大学学报(自然科学版),2020,41(02):125-130.

[3] Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C]//Advances in Neural Information Processing Systems. 2019: 8024-8035.

[4] 卢栋, 李华康, 孙国梓. 一种基于P2P技术的分布式微博爬虫系统[J]. 江苏大学学报(自然科学版),2016,37(03):296-301.

[5] 应毅, 李晓明, 桑晶. 时间敏感的微博热点爬取与发现模型研究[J]. 淮海工学院学报(自然科学版),2019,28(02):25-28.

[6] 曾子明, 万品玉. 基于双层注意力和Bi-LSTM的公共安全事件微博情感分析[J]. 情报科学,2019,37(06):23-29.

联系方式

联系人：刘定一

手机：15345187522

邮箱：zgzswan@sina.com