第44卷第9期
2016年9月

华南理工大学学报(自然科学版)
Journal of South China University of Technology
(Natural Science Edition)

Vol. 44 No. 9
September 2016

文章编号: 1000-565X(2016)09-0047-06

基于神经网络的微博舆情预测方法

何炎祥 刘健博 孙松涛

(武汉大学 计算机学院//软件工程国家重点实验室, 湖北 武汉 430072)

摘 要: 根据微博社交平台特征,提出了一种基于神经网络的微博舆情预测方法. 该方法使用单位时间内的微博发帖量作为事件趋势的量化指标,考虑影响事件发展的因素,根据样本内的数据趋势建模,使用神经网络来预测范围外的事件的未来趋势. 仿真实验结果表明,该方法可以快速地对事件发展的趋势进行量化分析和建模,能够准确地预测事件的爆发点和发帖量.

关键词: 社会计算;神经网络;大数据;趋势预测;微博

中图分类号: TP399 doi: 10.3969/j.issn.1000-565X.2016.09.007

微博(微型博客)是一个基于用户关系信息的分享、传播以及获取平台. 微博因其碎片化、便捷性的特点逐渐成为热点事件的发源地,同时也是推动事件发展的重要力量. 随着移动互联网的普及,一些热点事件在微博平台中能在极短的时间内大范围传播. 人民网发布的《2014年互联网舆情报告》指出,2014年网络舆情总体热度下降,例如位居全年热点舆情榜首的马航MH370航班失联事件,已成为当今世界最大的悬疑案,微博帖文约2500万条;而2011年7·23甬温线动车事故发生后的几天内,微博发帖量约5亿条. 从2013年6月30日到2014年6月30日,网民中微博使用率下降了11.1个百分点,微博的用户流失,热度在下降,但在马航失联、东莞扫黄、山东招远血案等突发事件和热门议题中,微博在信息传播、意见表达上仍然展现出强大的功能. 社会群体通过微博表达各种利益诉求,微博内容不仅包含了对事件的具体描述,而且还包括了用户对事件的观点和立场,这些内容在短时间内的大规模、快速扩散又会影响事件的发展,因此尽早地预测微博上话题的走向和趋势,政府就可以掌握社情民意的脉搏,及时发现基层治理中存在的问题和矛盾,释放社会压力,并澄清其中的谣言,对负面声音进行引导,企业也可以从中获取用户的消费兴趣以及对产品的反馈意见,及时地指定营销与公关策略.

大数据是指所涉及的资料量规模巨大到无法通过目前主流软件或工具在合理时间内获取、管理、处理,并整理成为对日常生产有积极帮助的资讯[1]. 微博平台就是一个典型的大数据平台,在大数据环境下对微博事件的趋势进行快速准确的预测是文中需要解决的问题. 国内外学者在舆情预测领域已经取得了一些研究成果,文献[2]对信息扩散预测的研究现状做了详细的综述,并提出了从宏观角度对信息的扩散范围、广度和深度等宏观特性进行预测的方法,该方法可以应用于突发事件探测和舆情预测. 文献[3]提出的线性影响力模型可用于预测电子商务网站中用户购买行为的信息扩散. 文献[4]根据Lyapunov指数证明了网络舆情的混沌特征,并使用改进的径向基神经网络对两会的舆情发展趋势

收稿日期: 2015-12-25

* **基金项目**: 国家自然科学基金资助项目(61303115,61472290,61472291);武汉市科技攻关项目资助项目(201210421135)

Foundation items: Supported by the National Natural Science Foundation of China(61303115,61472290,61472291)

作者简介: 何炎祥(1952-),男,博士,教授,主要从事可信软件、自然语言处理、分布并行处理和软件工程研究. E-mail:yxhe@whu.edu.cn

进行预测. 文献[5]提出了一种基于灰色支持向量机的网络舆情预测模型,相对于传统预测模型,该模型能有效地提高了网络舆情的预测精度.

微博作为目前国内重要的舆论场,是一个复杂的巨系统,信息在微博中的传播受到了众多因素的影响,用户的特征提取以及网络拓扑结构的分析也十分困难,个体的随机性也为同类问题的分析带来了困难. 文中针对微博信息的特征提出了一种针对微博平台的事件趋势预测方法,该方法以单位时间内微博数量作为衡量话题趋势的指标,采用神经网络预测事件趋势,并在仿真实验环境下对该方法进行了验证.

1 趋势预测方法

在微博平台中,微博的发帖量是各项影响因素的综合反映,是反映事件趋势最直观的指标,因此,文中使用发帖量来衡量话题趋势,预测框架如图1所示. 在该框架中,选取影响因素和预测模型进行组合构成不同的预测方法. 神经网络已在诸多领域取得了较好的预测结果,能有效解决不少传统统计学不能解决或不宜解决的问题[6]. 因此,文中根据微博平台特征提出了影响微博趋势的因素,以影响因素作为神经网络的输入,并选取4种基于神经网络的预测模型对微博事件趋势进行预测.

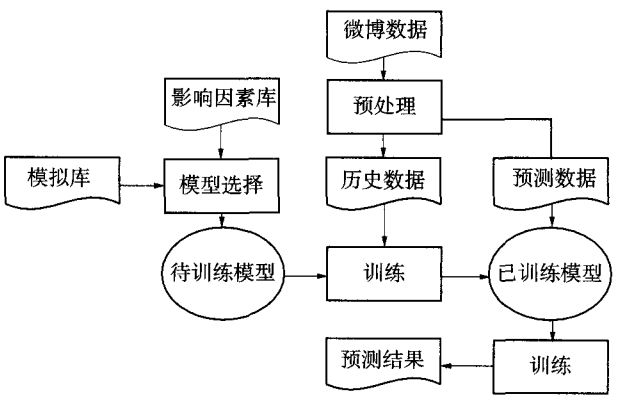


图1 微博事件趋势预测框架

Fig. 1 Framework of microblogging topic trend forecast

1.1 影响因素库

微博用户对事件的关注程度直接体现在微博的发帖量,与话题相关的微博发帖量在单位时间内越多,说明该话题越热门. 影响微博发帖量的因素很多,文中考虑实际环境下预测框架的运行效率与数据预处理的规模,主要讨论5个影响因素即微博的评论转发量、影响力、增长率、参与度、内容热度.

参考文献[7]的描述语言,在时间序列下,微博数据覆盖的时间段由若干个单位时间(ΔT)组成,某一事件的微博内容 $\mathrm{MB}(t)$ 定义为 $\mathrm{MB}(t)=(I_{\mathrm{mb}},I_{\mathrm{u}},N_{\mathrm{rt}},N_{\mathrm{cm}},N_{\mathrm{i}},C)$,其中,$I_{\mathrm{mb}}$ 为微博内容的标识集合,I_{u} 为发布微博内容的用户标识集合,N_{rt} 为微博内容被转发的次数集合,N_{cm} 为微博内容被评论的次数集合,N_{i} 为微博内容的影响力集合,C 为微博内容集合.

事件的某一条微博内容可以定义为 $\mathrm{mb}(t)=(i_{\mathrm{mb}},i_{\mathrm{u}},n_{\mathrm{rt}},n_{\mathrm{cm}},n_{\mathrm{i}},c)$,其中,$i_{\mathrm{mb}}\in I_{\mathrm{mb}}$,$i_{\mathrm{u}}\in I_{\mathrm{u}}$,$n_{\mathrm{rt}}\in N_{\mathrm{rt}}$,$n_{\mathrm{cm}}\in N_{\mathrm{cm}}$,$n_{\mathrm{i}}\in N_{\mathrm{i}}$,$c\in C$. 在第 t 个时间片内微博发帖量 $N_{\mathrm{mb}}(t)=|\mathrm{MB}(t).I_{\mathrm{mb}}|$.

微博的评论转发量 $N_{\mathrm{rtcm}}(t)$ 是指在第 t 个时间片内的微博评论转发量,转发微博评论的这部分人虽然没有直接发布微博内容,但参与了微博话题,因而是最有可能发表新博文的,表示为

$$N_{\mathrm{rtcm}}(t)=\sum^{n_{\mathrm{rt}}\in N_{\mathrm{rt}},n_{\mathrm{cm}}\in N_{\mathrm{cm}}}[\mathrm{MB}(t).n_{\mathrm{rt}}+\mathrm{MB}(t).n_{\mathrm{cm}}].$$

微博影响力由发微博用户的 PeopleRank 值[8]反映,PeopleRank 是用户影响力的表现,用户级别越高,其发出的微博的影响力也越大,因此在第 t 个时间片内的微博影响力 $N_{\mathrm{i}}(t)$ 为所有发布微博用户的 PeopleRank 值之和,即 $N_{\mathrm{i}}(t)=\sum^{i_{\mathrm{u}}\in I_{\mathrm{u}}}\mathrm{PeR}(\mathrm{MB}(t).i_{\mathrm{u}})$.

微博增长率反映当前时间段微博发帖量和评论转发量之和相对于前一时间段的增长率,在第 t 个时间片内微博增长率为

$$N_{\mathrm{g}}(t)=\frac{[N_{\mathrm{mb}}(t)+N_{\mathrm{rtcm}}(t)]-[N_{\mathrm{mb}}(t-1)+N_{\mathrm{rtcm}}(t-1)]}{N_{\mathrm{mb}}(t-1)+N_{\mathrm{rtcm}}(t-1)}.$$

微博参与度是当前时间段的微博量和微博评论转发量的比值,是用户倾向的表现,该值越小,表示用户更倾向于发表新的微博. 在第 t 个时间片内微博参与度为 $N_{\mathrm{r}}(t)=N_{\mathrm{mb}}(t)/N_{\mathrm{rtcm}}(t)$.

微博内容热度是微博文本内容的吸引力的反映,当微博内容热度高时,微博内容对用户的吸引力大,用户会更倾向于发表新的微博,表示为 $H(t)=\sum_{c\in C}\mathrm{Hot}(\mathrm{MB}(t).c)$,其中 $\mathrm{Hot}(\mathrm{MB}(t).c)$ 为单个微博博文的热度,文中用微博文本的情感极性来反映,但现阶段对微博短文本的情感极性分析的准确率不是很理想,因此在文中的仿真实验中未使用此因素,而是使用拟合函数 F 对第 t 个时间片内的微博发帖量 $N_{\mathrm{mb}}(t)$ 进行预测,即 $N_{\mathrm{mb}}(t)=F(\boldsymbol{X}(t-1),\boldsymbol{X}(t-2),\cdots)$,其中,$\boldsymbol{X}(t)=(N_{\mathrm{mb}}(t),N_{\mathrm{rtcm}}(t),N_{\mathrm{i}}(t),N_{\mathrm{g}}(t),N_{\mathrm{r}}(t),H(t))$.

影响因素并不局限于这 5 个因素，在文中提出的框架体系下，可以对影响因素库做进一步的扩充以提高预测准确率.

1.2　模型库

1.2.1　SVM 回归模型

支持向量机(SVM)是一种可训练的机器学习方法[9]，其思想是通过一个非线性映射，把样本空间映射到一个高维乃至无穷维的特征空间中，使样本空间中非线性可分的问题转化为特征空间中的线性可分问题. SVM 神经网络(SVMNN)避免了从归纳到演绎的过程，可实现从训练样本到预报样本的转导推理[10]，适合解决微博舆情预测这类非线性回归问题. SVMNN 结构见图 2，其中 $K(\cdot)$ 为核函数.

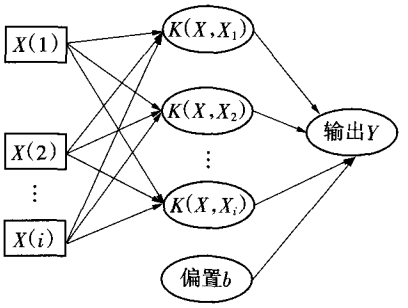


图 2　SVM 神经网络结构

Fig. 2　Configuration of SVM neural network

1.2.2　Elman 神经网络模型

Elman 神经网络(ENN)[11]是一种典型的局部递归网络，该模型在前馈式网络的隐含层增加了一个承接层用于记忆隐含层过去的状态，并在下一时刻连同网络输入作为隐含层单元的输入，这使得局部递归网络具有动态记忆功能. ENN 模型的承接层具有记忆功能，因而该模型具有适应时变特性的能力，解决微博预测问题的准确率更高. Elman 神经网络结构如图 3 所示.

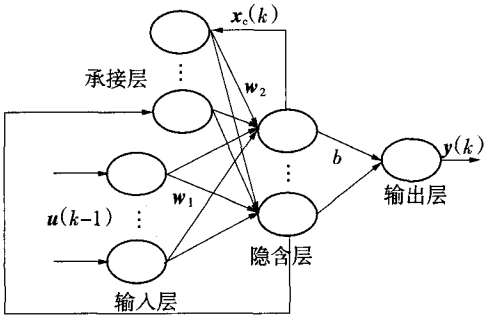


图 3　Elman 神经网络结构

Fig. 3　Configuration of Elman neural network

Elman 神经网络的状态空间表达式为

$$\begin{cases} \boldsymbol{x}_c(k) = \boldsymbol{x}(k-1) \\ \boldsymbol{x}(k) = f(\boldsymbol{w}_2\boldsymbol{x}_c(k) + \boldsymbol{w}_1 u(k-1)) \\ \boldsymbol{y}(k) = g_1(\boldsymbol{w}_3\boldsymbol{x}(k)) \end{cases} \tag{1}$$

式中，$\boldsymbol{x}_c$ 为反馈状态向量，$\boldsymbol{x}$ 为中间层向量，$\boldsymbol{y}$ 为输出向量，$\boldsymbol{w}_1$ 为输入层到中间层的连接权值，$\boldsymbol{w}_2$ 为承接层到中间层的连接权值，$\boldsymbol{w}_3$ 为中间层到输出层的连接权值，$g_1(\cdot)$ 为输出神经元传递函数，$f(\cdot)$ 为中间层神经元传递函数.

1.2.3　广义回归神经网络模型

广义回归神经网络(GRNN)[12]以数理统计为基础进行非线性(核)回归分析，是径向基神经网络的一种特殊形式，具有很强的非线性建模能力和较高的容错性及鲁棒性. GRNN 结构简单，学习速度快，对小样本数据能获得较好的预测结果，因此当微博舆情预测过程中学习样本较少时，可以选用该模型. GRNN 结构如图 4 所示.

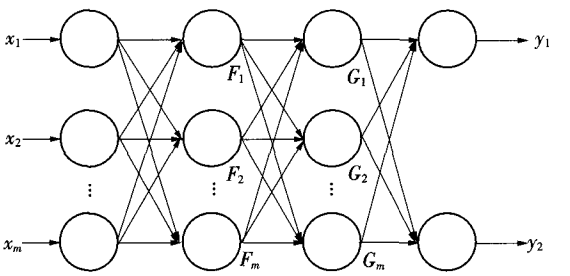


图 4　GRNN 神经网络结构

Fig. 4　Configuration of GRNN neural network

GRNN 由输入层、模式层、求和层和输出层构成. 其中：$F_1, F_2, \cdots, F_m$ 为模式层神经元的传递函数；$G_1, G_2, \cdots, G_m$ 为求和层的传递函数.

1.2.4　小波预测模型

小波分析[13]是针对傅里叶变换在时域内没有分辨能力的缺点发展而来，小波是一种长度有限、平均值为 0 的波形. 小波神经网络(WNN)[14]是小波分析与神经网络相结合的产物. WNN 中的小波基函数可以让神经网络在时域上具有分辨能力，同时结合神经网络的自学习功能，可以有效地解决微博舆情预测问题. WNN 的结构如图 5 所示. 其中，w_{ij} 为输入层和隐含层之间的连接权值，w_{jk} 为隐含层与输出层之间的连接权值，$g_2(x)$ 为小波基函数，文中采用的小波基函数为 $g_2(x) = -xe^{-x^2/2}$.

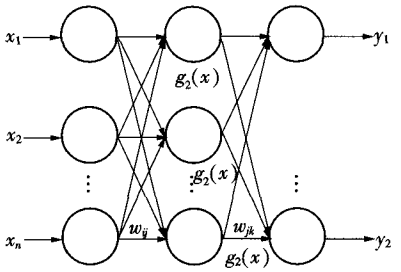


图5 WNN神经网络结构

Fig. 5 Configuration of WNN neural network

2 实验

2.1 实验数据

文中实验采用了WISE2012 Challenge[15]的数据集,其包含了66.3 GB的去除文本内容的微博数据和12.8 GB的用户关系数据.微博数据包含的37个话题(官方文档中共有42个,但有5个话题未出现)的持续周期为2009-08-14T9:7:32—2012-02-17T17:30:16,共有369797719条微博信息,属于37个话题的微博有4474563条,占总量的1.21%.用户关系数据包含了58655849个用户和265108370条用户关系,其中有2819324个用户被其他用户关注(占4.8%),58478875个用户关注了其他人(占99.7%),互相关注的用户有7601842个.

将所有数据看做是对一个大事件的反映,时间片单位为天,发帖量随时间的变化曲线如图6所示,将前600 d的数据作为训练集,后300 d的数据作为测试集,同时选取了3个有代表性的话题进行分析,各事件的发帖量随时间的变化趋势如图7所示,事件的具体情况如表1所示.

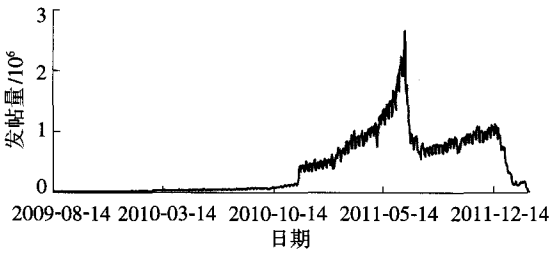
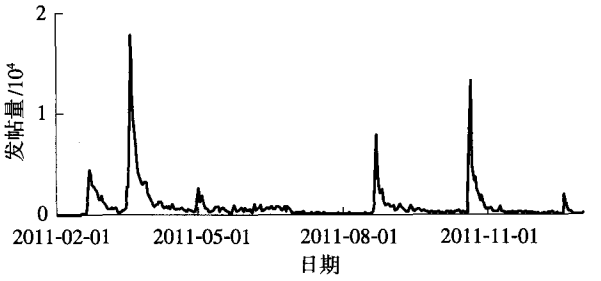
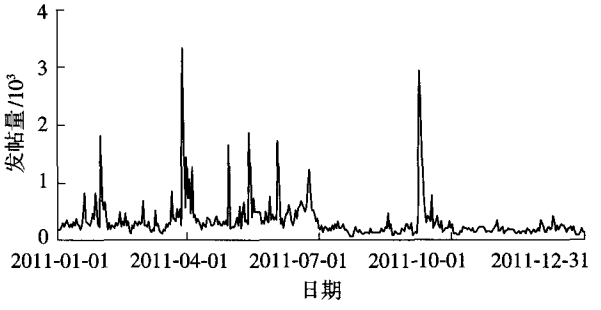


图6 发帖量随时间的变化曲线

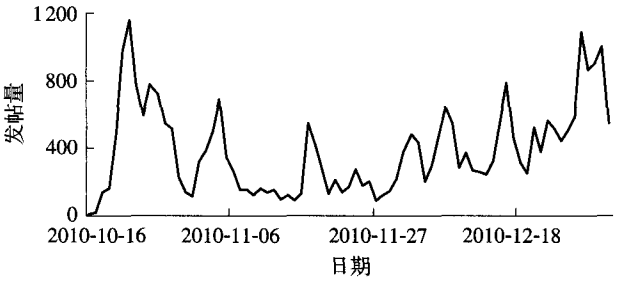
Fig. 6 Changing curve of post count over time



(a)事件1



(b)事件2



(c)事件3

图7 3个事件的发帖量随时间的变化曲线

Fig. 7 Changing curves of post count over time of three events

3个事件在微博平台中都出现了多次爆发的情况,其中事件1和事件2的时间跨度均超过了300 d.这3个事件的演化方式与实际环境中事件的演化方式类似,文中采用4种神经网络模型对这4组数据进行预测.

2.2 评价指标

文中采用平均绝对百分比误差(MAPE)和走势方向准确率来评价文中方法的预测效果.

表1 事件的具体情况

Table 1 Detailed information on the events

编号	事件名称	持续时间	训练集	测试集
1	利比亚内战	2011-02-01—2011-12-31	2011-02-01—2011-08-31	2011-09-01—2011-12-31
2	2011年房价	2011-01-01—2011-12-31	2011-01-01—2011-07-31	2011-08-01—2011-12-31
3	河北大学飙车	2010-10-16—2010-12-31	2010-10-16—2010-10-31	2010-11-01—2010-12-31

平均绝对百分比误差 E_{MAP} 是衡量平均误差的一种比较方便的方法，可以评价数据的变化程度，其值越小说明预测效果越好，其计算公式为

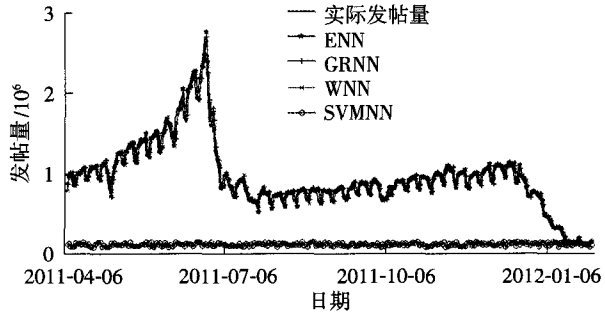
$$E_{MAP}=\frac{1}{\kappa}\sum_{t=1}^{\kappa}\left|\frac{a_{r,t}-a_{f,t}}{a_{r,t}}\right|\times 100\% \qquad (2)$$

式中，$a_{r,t}$ 和 $a_{f,t}$ 分别为第 t 个时间片内发帖量的实际值、预测值，κ 为时间片总数.

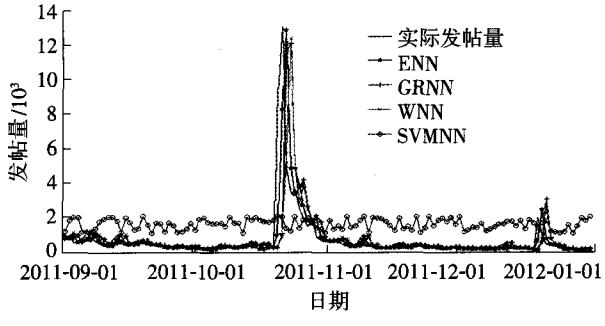
走势方向准确率 r 是指预测结果中预测走势方向和实际走势相符的次数与预测次数的比值，走势方向准确率越高，说明预测效果越好.

2.3　实验结果

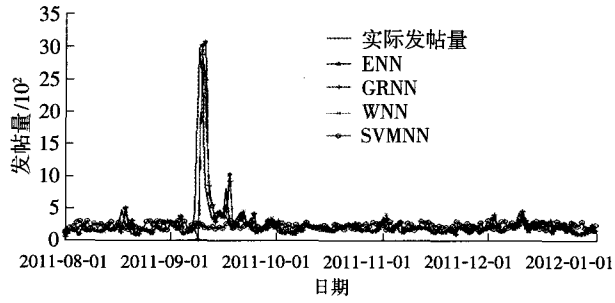
4 种方法对 4 组数据的预测结果如图 8 所示. 在实际环境中，人们更关注爆发点（时间段内发帖量峰值时刻）预测是否准确，文中设定时间段为 3 d，在这 3 d 中第 2 天的发帖量大于第 1 天和第 3 天，则第 2天为爆发点，爆发点的走势方向是指爆发点前一天和后一天的走势情况，其预测准确率 r_1 如表 2 所示. 实验结果表明，ENN 模型在事件趋势的峰值预测上效果最好，GRNN 和 WNN 可以有效预测事件的爆发点和发帖量，理论上 SVMNN 模型可以对微博事件趋势进行有效地预测，但实际结果并不理想，其原因可能是由于未找到合适的初始参数，在实验过程中，即使通过调整归一化区间也未能获得有效的预测结果，图 8 的实验结果是在归一化区间[1,2]上获得的. 可见，文中结合微博平台特征和神经网络的预测方法，可以有效预测微博趋势.



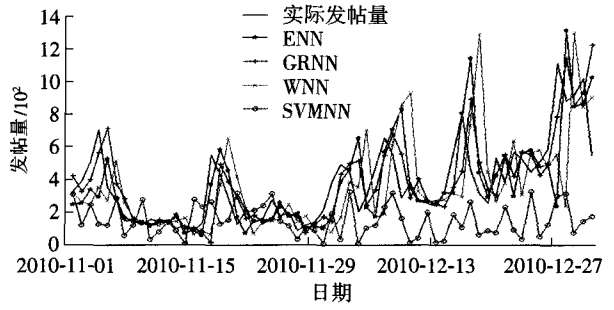
(a)大事件



(b)事件 1



(c)事件 2



(d)事件 3

图 8　4 种方法的预测结果比较

Fig. 8　Comparison of forecasting results among four algorithms

表 2　预测结果评价

Table 2　Evaluation for forecasting results　%

模型	事件	*EMAP*	*r*	r_1
ENN	大事件	7.31	59.87	85.56
	事件 1	15.74	56.20	88.89
	事件 2	12.29	62.75	83.33
	事件 3	16.24	40.00	60.00
GRNN	大事件	7.45	59.20	81.11
	事件 1	23.74	53.72	77.78
	事件 2	13.91	62.5	66.67
	事件 3	20.04	43.33	56.67
WNN	大事件	7.14	62.88	77.78
	事件 1	22.81	53.72	72.22
	事件 2	13.50	59.87	75.00
	事件 3	23.55	58.33	63.33
SVMNN	大事件	84.51	49.83	53.33
	事件 1	213.57	42.15	38.89
	事件 2	31.77	42.11	50.00
	事件 3	47.91	43.33	63.33

3　结论

文中针对微博社交平台特征提出了一种事件趋势预测框架，在该框架下对事件发展的趋势进行量化分析和建模，并在大数据环境下验证了该方法的有效性，为社交网络平台的趋势预测问题提供了一种新的思路. 今后将扩充影响因素库，考虑微博内容

对事件发展趋势的影响,以提高预测准确率,并探索新预测方法的扩充模型库.

参考文献:

[1] SNIJDERS C, MATZAT U, REIPS U D. Big data: big gaps of knowledge in the field of internet science [J]. International Journal of Internet Science, 2012, 7(1): 1-5.

[2] 李栋,徐志明,李生,等. 在线社会网络中信息扩散[J]. 计算机学报,2014,37(1):189-206.
LI Dong, XU Zhi-ming, LI Sheng, et al. A survey on information diffusion in online social networks [J]. Chinese Journal of Computers, 2014, 37(1): 189-206.

[3] YANG J, LESKOVEC J. Modeling information diffusion in implicit networks [C] // Proceedings of the 10th IEEE International Conference on Data Mining. Sydney: IEEE, 2010: 599-608.

[4] 魏德志,陈福集,郑小雪. 基于混沌理论和改进径向基函数神经网络的网络舆情预测方法 [J]. 物理学报,2015,64(11):110503/1-8.
WEI De-zhi, CHEN Fu-ji, ZHENG Xiao-xue. Internet public opinion chaotic prediction based on chaos theory and the improved radial basis function neural networks [J]. Physics, 2015, 64(11): 110503/1-8.

[5] 曾振东. 基于灰色支持向量机的网络舆情预测模型[J]. 计算机应用与软件,2014,31(2):300-302.
ZENG Zhen-dong. Internet pubic opinion prediction model based on grey support vector machine [J]. Computer Applications and Software, 2014, 31(2): 300-302.

[6] 刘豹,胡代平. 神经网络在预测中的一些应用研究[J]. 系统工程学报,1999,14(4):338-344.
LIU Bao, HU Dai-ping. Studies on applying artificial neural networks to some forecasting problems [J]. Journal of Systems Engineering, 1999, 14(4): 338-344.

[7] 何炎祥,刘健博,刘楠,等. 基于改进人口模型的微博话题趋势预测 [J]. 通信学报,2015,36(4):2015094/1-8.
HE Yan-xiang, LIU Jian-bo, LIU Nan, et al. Based on improved malthusian model microblogging topic trend forecast [J]. Journal on Communications, 2015, 36(4): 2015094/1-8.

[8] MTIBAA A, MAY M, DIOT C, et al. PeopleRank: social opportunistic forwarding [C] // Proceedings of International Conference on Computer Communications. San Diego: IEEE, 2010: 111-115.

[9] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[10] GAMMERMAN A, VOVK V, VAPNIK V. Learning by transduction [C] // Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann Publishers Inc, 1998: 148-155.

[11] CRUSE H. Neural networks as cybernetic systems [M]. Stuttgart: Thieme Medical Publishers, 1997.

[12] SPECHT D F. A general regression neural network [J]. IEEE Transactions on Neural Networks, 1991, 2(6): 568-576.

[13] 珀西瓦尔. 时间序列分析的小波方法 [M]. 北京:机械工业出版社,2006.

[14] CHEN Y, YANG B, DONG J. Time-series prediction using a local linear wavelet neural network [J]. Neurocomputing, 2006, 69(4): 449-465.

[15] UNANKARD S, CHEN L, LI P, et al. On the prediction of re-tweeting activities in social networks: a report on WISE 2012 challenge [M]. Berlin/Heidelberg: Springer, 2012: 744-754.

Neural Network-Based Public Opinion Prediction Method for Microblog

HE Yan-xiang LIU Jian-bo SUN Song-tao

(School of Computer // State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, Hubei, China)

Abstract: In view of the characteristics of microblog platform, a public opinion prediction method is proposed on the basis of neural networks. In this method, the post amount in unit time is taken as the quantitative index of event trend. Then, by considering the factors influencing events, the modeling is performed according to sample data, and the neural networks are employed to predict the future trend of the events outside the scope. Simulation results show that the proposed method is fast in terms of the quantization and modeling of the event trend, and it is effective in predicting the outbreak point and the post amount.

Key words: social computing; neural networks; big data; trend forecasting; microblog