一种基于 FA-SVM 的热门微博特征选择及预测方法研究

周 剑 峰
(广东外语外贸大学图书馆 广东 广州 510420)

摘 要 针对自媒体舆情研究中热门微博预测的问题,以新浪微博为研究对象,分析热门微博的影响因素。从微博属性、微博文本、微博博主、微博受众四个维度构建一套可量化的微博热度评价指标体系。采用因子分析法(FA)对各维度的指标进行处理,提取公共因子;以公共因子作为特征向量,采用 SVM 方法构建热门微博预测模型。实验采集了新浪微博中的热门微博数据验证其可行性和合理性。结果表明,该方法能有效地降低特征维度,消除噪声,提高热门微博预测的准确率。

关键词 新浪微博 热门微博预测 热门微博特征 因子分析 支持向量机

中图分类号 TP393 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.12.021

FEATURES SELECTION AND FORECASTING METHOD FOR HOT MICRO-BLOG BASED ON FA-SVM

Zhou Jianfeng
(*Library, Guangdong University of Foreign Studies, Guangzhou 510420, Guangdong, China*)

Abstract Aiming at the prediction of hot micro-blog in the research of we-media public opinion, this paper took Sina micro-blog as the research object, analyzed the influencing factors of hot micro-blog. We constructed a quantifiable evaluation index system for micro-blog popularity from four dimensions: micro-blog attributes, micro-blog text, micro-blogger and micro-blog audience. Factor analysis (FA) was used to process the indicators of each dimension and extract common factors. Taking the common factors as the feature vector, the hot micro-blog prediction model was constructed by SVM. We collected the hot micro-blog data in Sina micro-blog to verify its feasibility and rationality. The experiment results show that the method can effectively reduce the feature dimension, eliminate the noise and improve the accuracy of hot micro-blog prediction.

Keywords Sina micro-blog Hot micro-blog prediction Characteristics of hot micro-blog Factor analysis Support vector machine

0 引 言

据新浪微博数据中心的《2017 微博用户发展报告》显示,截至 2017 年 9 月,新浪微博月活跃用户共 3.76 亿,相对 2016 年增长了 27%,其中移动端比例达到了 92%。2017 年中的“#杭州保姆纵火案件#、#校园欺凌#、#厉害了我的国#”等事件均凸显了微博作为网络新兴媒体在社会舆情传播中的重要地位及其对国家和社会稳定的深远影响力。相对普通微博,热门微博更容易成为网络舆情的发酵源,本文分析了热门微博的影响因素,构建量化微博热度评价指标体系,采用因子分析法进行特征选择,最终结合 SVM 算法获取热门微博预测方法。研究对于网络舆情监控研究、企业营销、政府舆情监控具有重要意义。

1 研究现状

目前国内外的微博舆情研究方法主要分为两个方向:

收稿日期:2018-06-23。2014 年广东省教育厅青年创新人才项目;广东省教育厅科技创新项目(2013KJCX0067)。**周剑峰**,馆员,主研领域:舆情分析,自然语言处理,大数据挖掘。

1）一类是对微博文本及其评论内容的研究，主要表现为基于微博文本内容的话题发现以及基于评论情感倾向及强度的热点挖掘两种。Puvipadaw 等[1]针对 Twitter 中的文本特征，提出一种的突发新闻检测、排列及跟踪算法。杨亮等[2]提出情感分布语言模型 ELM（emotion distribution language model）来发现微博中的热点事件。吴青等[3]基于微博短文本特点，根据高频微博词实现微博聚类，并分析热点话题的情感强度，跟踪及预测微博舆情。叶成绪等[4]结合最长公共子串和维基百科知识，基于中文微博主题词进行热点话题发现研究。

2）另一类主要基于微博传播路径中的用户、转发等因素进行分析研究。在文献[5]中提出了一种基于地理空间信息的热点事件检测方法，但是该方法基于用户的位置信息，在用户不允许分享位置时容易失去效用。文献[6－7]针对 Twitter 提出基于粉丝、转发帖数、回复数、被转发数等因素计算个人用户的影响力，发现话题的关键用户，为热点话题发现提供参考。上述研究没有针对单条微博热度评价的研究，并且特征覆盖并不全面，受到一定局限。

在单条热门微博预测研究方面，郑志蕴等[8]从微博内容特征、传播特征、博主特征出发，利用信息增益算法对微博热度进行度量结合神经网络算法预测微博的传播特征从而预测微博是否能成为热门微博。陈梦秋等[9]结合微博博主特征、微博传播特征，采用 SVM 模型进行热门微博预测研究。其成果忽略了微博受众特征，且没有对特征进行进一步的选择研究。

针对上述问题，本文提出一套多层级多维度可量化的微博热度评指标体系，全面考虑热门微博影响因素，采用因子分析法进行特征选择研究，降低特征维度，消除噪声，获取公共因子；以公共因子作为向量特征，采用支持向量机算法训练热门微博预测模型，对单条微博是否能成为热门进行预测，为微博舆情研究提供参考。

2 微博热度评价指标体系

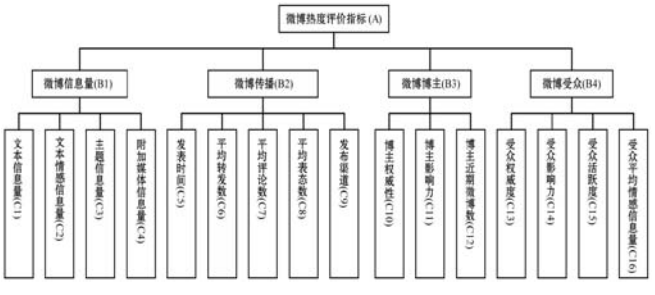

图 1 微博热度评价指标体系

2.1 微博信息量（B_1）

微博的吸引力主要来自于内容的信息量，据课题组对新浪热门微博数据集的分析统计，热门微博的文本平均长度达到了 117 个字，且长度与热门程度呈正相关态势，内容均含有图片或者视频，其中 46% 含有 URL 链接，61% 含有话题标签。同时情感词的增多可以使文本内容更活泼更容易引起共鸣，基于大连理工情感本体库进行统计，90% 以上的热门微博均包含情感词。基于上述统计，课题拟定了一级指标微博信息量（B_1），其特征选取依据主要为微博内容及其情感信息量的丰富程度。

文本信息量（C_1），文本即微博的核心内容，长度越大内容越丰富，才能完整清晰地传达事情的全貌，因此以微博文本的长度即字符数作为特征。文本情感信息量（C_2），情感词是文本情感分析的基础，是文本情感信息量的代表，把文本中的情感词词频作为分析指标。主题信息量（C_3），话题标签是微博内容的缩影，对微博传播有直接影响，因此把微博文本中标签的个数作为分析指标。附加媒体信息量（C_4），图片、视频、URL 链接均是对微博内容的补充，均有提高微博内容信息量及吸引力的能力，因此把图片、视频、链接的合计量作为分析指标。

2.2 微博传播（B_2）

一级指标微博传播（B_2）指微博的传播特征，主要包括时间、方式、转发数、评论数、表态数等，为了消除时间的累积效应，使其能够在实际情况中评估微博的传播能力，研究采用自微博发布时间起至被抓取的时间的统计数据均值作为指标，即：数量/每小时。

根据濮小燕等[10]研究，新浪微博在线人数和活跃度在每天的时间序列上呈现一定的规律，不同时间的活跃用户数会对热门微博的产生有直接的影响。课题组对热门微博发表的时间进行了统计，发现热门微博的发表时间集中于中午（午休）、傍晚（交通）、夜晚（文娱）三个时间段，即非工作时间段，用户活跃度较高，根据统计结果对每日 24 个小时进行切分，将发表时间指标（C_5）分为工作时段（7：00 － 12：00，14：00 － 18：00）、文娱时段（12：00 － 13：00，18：00 － 23：00）、睡眠时段（23：00 － 7：00）。

$$C_5=\begin{cases}1 & time\in(23{:}00,7{:}00)\\ 2 & time\in(7{:}00,12{:}00)\cup(14{:}00,18{:}00)\\ 3 & time\in(12{:}00,14{:}00)\cup(18{:}00,23{:}00)\end{cases}\quad(1)$$

同时以平均转发数（C_6），平均评论数（C_7），平均表态数（C_8）作为传播路径上的分析指标。相对手机

客户端而言,PC 端及第三方应用产生的微博发布步骤相对复杂,并且具有审核功能,具有更高的公信力,更容易产生社会舆情,因此将发布渠道(C_9),作为分析指标之一,计算方式如公式所示:

$$C_9 = \begin{cases} 1 & \text{from cellphone} \\ 2 & \text{from PC} \\ 3 & \text{from third-party} \end{cases} \quad (2)$$

2.3　微博博主(B_3)

自媒体是以人为核心,通过公众用户自我传播的。微博博主的影响力对微博的传播、热度有直接的影响。基于新浪微博采集的数据,博主的属性主要有认证、粉丝数、微博数等。其中认证代表着博主权威性指标(C_{10}),权威性越高则博主微博内容的可信度越高,越容易被受众接受并传播,已受到官方认证的博主权威性更高,量化计算方式如公式所示:

$$C_{10} = \begin{cases} 0 & \text{not verified} \\ 1 & \text{verified} \end{cases} \quad (3)$$

粉丝数指关注该博主的人数,粉丝越多,该博主的累积影响力越大,所发布的微博也容易被更多人阅读及转发,成为热门微博,因此将粉丝数作为博主影响力指标(C_{11})。相对粉丝数而言,博主的近期微博数(C_{12})不仅反映了博主的活跃度,也反映博主的近期影响力。活跃度较高的博主更容易受到注意,并且对粉丝有更强的影响力。

2.4　微博受众(B_4)

微博是自媒体网络社交平台,在自媒体平台上,人人皆可成为媒体,也可以称为“个人媒体”,意味着微博的受众,即微博信息的接受者和传播者均成为了舆情传播的重要环节。因此将微博受众(B_4)作为一级指标进行分析。

微博受众与博主分析指标类似,主要计算其在传播节点中的影响力,相对博主而言,受众具有较为庞大的数量,为了平衡微博受众间的数量及影响力差异,采用其平均数作为分析指标。

$$C_n = \frac{C}{N_{\text{audience}}} \quad (4)$$

式中:C 为指标 C_n 的统计数。

与博主指标类似,受众权威度(C_{13})来自于其受众博主的平均认证数,已认证的微博用户对自己的言论更慎重,对自己所参与转发、评论的微博也较为谨慎,其转发、评论的微博会具有较高的可信度。微博受众影响力(C_{14})来自于其微博受众用户的平均粉丝数量,反映传播路径上受众的影响力及水平。传播路径上较活跃的节点具有更高的影响力,受众活跃度(C_{15})则通过受众用户的平均微博数获取,平均微博数越多路径活跃度越高,其传播节点也越大,也对微博传播具有更深远的影响力。受众的平均情感信息量(C_{16}),根据文献[2]的研究,人们往往对于能够让自身产生情感的事件更关注,热门微博的回复评论当中,均呈现出大幅度的情感波动。因此当回复评论中出现大量情感波动时,则微博更容易成为热门微博。基于上述理论笔者基于研究小组提出的方法[11]计算回复及评论文本情感倾向,采用线性加和的方式计算总情感信息量。

3　热门微博特征选择及预测方法

微博热度评价体系中的多维特征能够覆盖微博本身及其影响力所涉及的各项因素。但各项评价指标之间有差异性也存在一定的关联性,其对微博热度评价的结果都存在正向或负向影响力,其影响力程度也呈不同水平,直接采用原始评价指标作为特征有时难以反映真实情况,增加了数据处理的难度和计算复杂度,容易对评价结果产生负面影响。

为了减少特征中的噪声,降低对热门微博预测的负面影响,使其能有效地应用于海量的热门微博识别中,本文采用因子分析法对指标进行降维处理,消除噪声指标的影响力,获取公共因子。

因子分析是一种能够将原始变量转化成几个综合变量的多元统计分析方法[12],其通过研究众多变量数据之间的信息关系,将相同本质的变量归入同一个综合变量,这几个综合变量被称作“因子”,其代表了多个原始变量的信息及结构,既实现了指标归总及特征降维,也有利于提高分类精确率及计算效率。

以公共因子作为特征向量,笔者拟采用支持向量机 SVM 训练热门微博预测模型。支持向量机[13]以统计学习理论的 VC 维理论和结构风险最小原理为基础,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中,是一种有监督学习模型,它在解决小样本、非线性及高维模式识别中表现出许多特有的优势,有较好的较好泛化性能。基于 SVM 的单条热门微博预测模型就是将微博能否成为热门微博的预测转换为一个二分类问题,即将单条微博分为热门微博及非热门微博。

本文以新浪微博作为研究对象,以 16 个热度评价指标作为输入,获取热门微博分类结果及其评价作为期望输出。研究框架如图 2 所示。

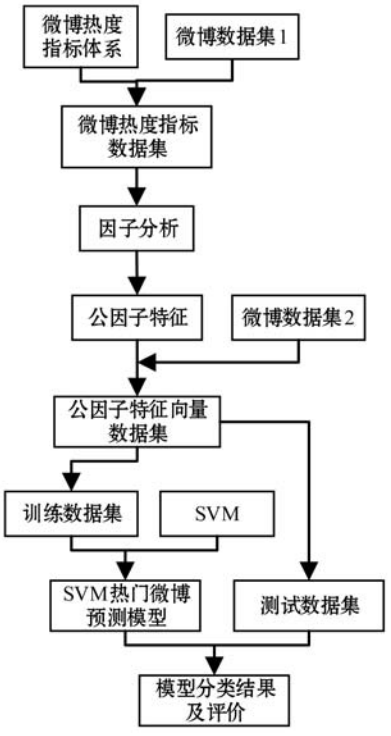

图 2 FA-SVM 热门微博研究框架

4 实验结果及分析

4.1 实验数据及工具

本文采集了新浪微博共 500 条,其中包含来自新浪微博热门排行榜的数据 150 条,其余 350 条为新浪微博热度值较高,但又未入榜单的微博数据。数据集中共包含 633 110 条微博用户数据以及 702 135 条评论数据。实验中所使用的情感词典来自于大连理工大学信息检索研究室(DUTIR)的情感词汇本体库,共 27 466 条情感词汇。

实验中的原始指标数据生成工具由课题小组基于 C#语言进行开发,同时使用 SPSS 20 进行数据标准化处理及因子分析研究。SVM 预测模型则采用基于 python 语言的 sklearn 包进行构建。

4.2 因子分析实验

本文采用 IBM SPSS 软件进行因子分析处理,其中因子提取方法为主成分分析法,因子旋转采用最大方差法。获取的 KMO 检验的结果为 0.717,根据 KMO 度量标准(KMO 值越接近于 1,意味着变量间的相关性越强。通常认为的度量标准是:0.6 ~ 0.9,这意味此时运用因子分析法是适当的。

从公因子方差表中可以看到,因子可提取到的原始变量信息成分最低为 0.58,其中博主影响力指标(C_{11})、博主的近期微博数(C_{12})、受众活跃度(C_{15})均未超过 0.6,即其在公因子中被提取的信息量均在 60% 以下;其中文本信息量(C_1)、平均评论数(C_7)、平均表态数(C_8)、博主权威性指标(C_{10})等变量提取的信息成分均超过 0.8,即公因子提取了该变量 80% 或以上信息量。平均信息提取量比例为 70.9%,说明了即将产生的几个主成分因子可提取到的原始变量信息的比例达到 70% 以上。

因子分析法共生成了 16 个公因子,基于因子分析理论,特征值大于 1 的因子才能解释所有数据的方差,因此提取前 6 个公因子作为分析对象,其贡献率如表 1 所示。

表 1 公因子累计贡献率

因子	F1	F2	F3	F4
特征值	3.362	2.178	2.03	1.631
贡献率/%	21.015	13.614	12.686	10.195
累计贡献率/%	21.015	34.629	47.315	57.510
因子	F5	F6	F7	F8 – F16
特征值	1.122	1.023	0.889	…
贡献率/%	7.011	6.395	5.559	…
累计贡献率/%	64.521	70.916	76.475	…

由表 1 中显示出前 6 个因子的累计方差贡献率已经达到 70.916%,说明了前 6 个公因子可以将原始指标中超过 70.9% 的信息保存了下来,其具有解释原始指标的评价能力,可以反映原始指标的大部分信息。最终确定主成分个数为 6 个,将原来的 16 个指标进行压缩后用 6 个因子特征来代替。

从成分得分系数矩阵可获知,每个公因子中,不同指标均有不同的权重,权重值范围为 – 1 至 1,代表着公因子中各指标所占重要程度。例如公因子 F1 的权重中 $C_6 - C_8$ 的权重均为负 0.6 以上,意味着在公因子 F1 中,这几个指标所提供的信息量极少,而 C_{12}、C_{14} 指标的权重则均超过了 0.5,意味着该指标在公因子 F1 中占比相对较高。而在其他公因子中,指标均呈现不同的权重。

从图 3 可以看出,在各公因子中,$C_5 - C_9$ 均占比较低,而 $C_{11} - C_{14}$ 平均占比较高。证明在公因子中,主要信息量及影响力来自于后者,在热度指标中,后者的重要程度更高。最终采用线性加权方法计算公因子特征值,如公式:

$$F_i = C_1 \times w_{i1} + C_2 \times w_{i2} + C_3 \times w_{i3} + \cdots + C_{16} \times w_{i16} \quad (5)$$

式中:F_i 是第 i 个公因子的特征值,w_{i1} 是第 i 个公因子中 C_1 的权重。实验以该 6 个公因子特征作为下一步

预测模型的输入特征向量。

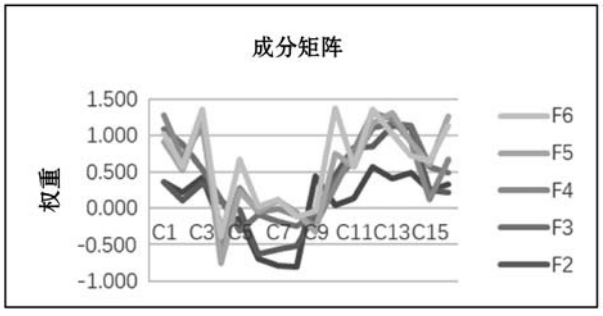

图 3　成分矩阵

4.3　基于 SVM 的热门微博预测模型实验

基于 SVM 的热门微博预测研究即将预测问题转为一个二分类问题,把微博分为热门或非热门类别,模型构建及评价流程如图 2 所示。实验因子分析的结果获取公因子特征数据集,以公因子作为输入向量,采用训练数据集结合 SVM 算法训练热门微博预测模型(FA-SVM),使用测试语料集来评价模型的效果。

为了有效地体现 FA 特征选择方法的有效性,实验同时以 16 个原始指标作为特征向量,采用 SVM 算法训练热门微博预测模型(SVM)作为对比。实验采用准确率、召回率对分类结果进行评价,得到结果如表 2 所示。

表 2　热门微博预测模型评价　　%

方法	总体准确率	热门微博召回率	非热门微博召回率
SVM	77	83.3	74
FA-SVM	84.6	90	82.3

从表 2 可以看出,本文提出的 FA-SVM 模型相对单纯 SVM 方法而言有效地提高了分类准确率及召回率,证明因子分析法能够有效地提取多个指标中的潜在信息,形成公因子特征,在降低特征维度的同时,能更准确地识别出单条热门微博。从召回率来看,FA-SVM 方法同时提高了热门微博的召回率及非热门微博的召回率,也意味着因子分析所提取的公因子特征中,不仅降低了特征维度,同时也消除了特征中的噪音,有效地提高了热门微博的识别能力。

经实验证明,FA-SVM 方法结合微博热度评价指标体系,能够获取热门微博的共性特征,并应用于热门微博预测研究领域。

5　结　语

单条微博是微博舆情的起点,热门微博预测研究有助于微博舆情监控研究。本文以新浪微博为研究对象,从微博内容、微博博主、微博传播、微博受众四个方面提出一套可量化的微博热度评价指标体系,采用因子分析法对指标进行分析,获取其公共因子,并以公共因子作为特征,结合 SVM 算法训练热门微博预测模型。实验表明该方法能有效地提取指标特征的共性因子,并提高热门微博的预测概率。

参　考　文　献

[1] Puvipadaw S, Murata T. Breaking news detectection and tracking in Twitter[C]//Proceedings of the 9th IEEE/WIC/ACM IntConf on Web Intelligence and Intelligent Agent Technology(WI-IAT'10) New York:ACM,2010:120-123.

[2] 杨亮,林原,林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报,2012,26(1):84-90.

[3] 吴青林,周天宏. 基于话题聚类及情感强度的中文微博舆情分析[J]. 情报理论与实践, 2016, 39(1):109-112.

[4] 叶成绪,杨萍,刘少鹏,等. 基于主题词的微博热点话题发现[J]. 计算机应用与软件, 2016,33(2):46-50.

[5] Unankard S, Li X, Sharaf M A. Location-based emerging event detection in social networks[M]//Web Technologies and Applications. Springer Berlin Heidelberg,2013.

[6] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. New York:ACM, 2010:261-270.

[7] Pal A, Counts S. Identifying topical authorities in microblogs[C]//Proceedings of the fourth ACM international conference on Web search and data mining. New York: ACM, 2011: 45-54.

[8] 郑志蕴,江国林,张行进,等. 基于多特征的热门微博预测算法研究[J]. 小型微型计算机系统,2017,38(3):494-498.

[9] 陈梦秋,周安民. 基于 SVM 的新浪热热微博预测[J]. 现代计算机,2017(9):23-27.

[10] 濮小燕. 基于多层结构的单条微博影响力研究[D]. 成都:电子科技大学,2015.

[11] Zhou J, Chen B, Lin Y. An Approach to Constructing Sentiment Collocation Dictionary for Chinese Short Text Based on Word2Vec[C]//International Symposium on Emerging Technologies for Education. Springer, Cham, 2017: 548-556.

[12] 张红兵,贾来喜,李潞. SPSS 宝典[M]. 北京:电子工业出版社, 2009.

[13] 邓乃扬,田英杰. 支持向量机——理论、算法与拓展[M]. 北京:科学出版社, 2009.