基于K近邻的新话题热度预测算法

聂恩伦 陈 黎 王亚强 秦湘清 金 宇 于中华
(四川大学计算机学院 成都 610065)

摘 要 随着互联网的快速发展,网络舆情成为政府部门和企业以及社会大众关注的焦点,对网络舆情进行有效监管和正确引导是当前亟待解决的问题,话题热度预测是舆情监管和引导的基础。针对现有算法无法对新话题的热度进行有效预测的缺点,提出了一种基于K近邻的新话题热度预测算法。该算法利用与新话题相似的历史话题的点击数时间序列来对新话题的热度进行预测。实验结果表明,在允许相对误差分别低于10%、20%和30%的情况下,算法预测的前3天点击数的平均正确率分别为47.26%、61%和67.7%,点击数变化趋势平均正确率达到73.73%,这也说明了相似的话题在话题出现的初期具有近似的热度变化趋势。

关键词 热度预测,新话题,K-近邻算法,话题相似性,网络舆情
中图法分类号 TP391 **文献标识码** A

Algorithm for Prediction of New Topic's Hotness Using the K-nearest Neighbors

NIE En-lun CHEN Li WANG Ya-qiang QIN Xiang-qing JIN Yu YU Zhong-hua
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract With the rapid development of the Internet, the government, enterprises and public have paid more and more attentions on net-mediated public sentiment. How to effectively monitor and aright guide the public sentiment on the Internet has become an issue that should be coped urgently with. As a basis to solving the issue, it is necessary to have ability of predicting topic's hotness appearing on the Internet. As traditional algorithms could not predict aright new topic's hotness, a novel algorithm based on K-nearest neighbors(K-NN) was proposed in this paper. The algorithm predicts the hotness of new topics by using hotness times-series of their historical similar topics. The experimental results show that the average accuracies of the hotness prediction during the first 3 days are 47.26%, 61% and 67.7% respectively with the corresponding relative errors being less than 10%, 20% and 30%, and the average accuracy of the hotness trends within the first 3 days could be up to 73.73%. Meanwhile, the results also demonstrate that similar topics approximately have same hotness trends in their early developing stages.

Keywords Hotness prediction, New topic, KNN, Topic similarity, Net-mediated public sentiment

1 引言

互联网的快速发展和广泛应用是现代社会的一个鲜明特征,它是人类有史以来最大的信息资源库。其中,网络论坛作为开放的、自由发表言论的平台,是各种意见的集散地和思想交流的重要场所。对网络论坛上的言论信息进行分析、挖掘和利用,掌握或预测出网民所关注的热点话题,这样政府可以更好地对政策措施进行调整,企业可以及早更新产品和提升服务。因此话题热度预测成为当前Web数据挖掘领域的热点课题之一,引起了广泛的关注。

话题热度预测旨在挖掘话题的被关注度以及变化情况,为政府和企业决策提供支持。一个话题是指注册用户发表的一个主帖,主题是指由一系列相关的或者相似的话题组成的一个话题集合[5]。由于现有的热度预测算法多数都是依据话题本身已有的发展态势来预测其未来的发展情况,因此无法在话题刚出现时对其热度以及变化趋势进行预测。虽然对话题的长期预测具有重要的应用价值,但是由于网络舆情具有爆发周期短的特点,因此短期预测,尤其是针对新话题的热度预测更有意义。事实上,许多广为关注的舆情事件,从话题出现到广受关注只有短短几天的时间。例如,2010年7月17日发生的上海大众刘坚车祸事件,2010年7月19日关注度就达到了顶峰,然后开始下降,到7月23日就几乎没有人关注了[15]。基于上述原因,本文对新话题的热度预测问题进行了研究,提出了一种基于K近邻的新话题热度预测算法,该算法利用与新话题相似的历史话题的点击数时间序列来对新话题的热度进行预测。实验结果表明,内容相似的话题之间在话题初期具有相似的发展态势,可以利用相似话题的初期发展态势对新话题的初期热度进行有效的预测。

本文受高等学校博士学科点专项科研基金(20100181120029)资助。
聂恩伦(1986—),男,硕士生,主要研究方向为数据挖掘、计算语言学,E-mail:nieenlun@163.com;陈 黎 女,博士,讲师;王亚强 男,博士生;秦湘清 男,硕士生;金 宇 女,硕士生;于中华 男,博士,副教授,主要研究方向为数据挖掘与计算语言学,E-mail:yuzhonghua@scu.edu.cn(通信作者)。

2 相关工作

近年来，舆情研究在我国发展迅速[1-3]，热点话题挖掘与预测作为它的一项重要研究内容，也取得了丰硕的成果。为了挖掘用户关注的热点话题，文献[4]提出了一种基于文本聚类的算法，该算法通过对话题文本进行聚类，运用论坛话题的点击数和回复数对聚类的话题进行热度排名，从而找出热点话题。文献[5]提出了一种基于影响力扩散模型[6]计算论坛中主题影响力的算法，该算法能以35%的准确率发现当前的热点主题。文献[7]提出了一种话题追踪方法，即用话题聚类中心替代话题，在追踪过程中不断更新该中心，直到稳定。

上述研究的目的是发现当前的热点话题。在话题热度预测方面，文献[8]提出了小波分析和神经网络相结合的算法。该算法利用小波变换对帖子的原始点击数序列进行转换，得到低频和高频小波系数，再根据训练集(包括热帖和非热帖两类)的点击数序列选取出贡献度最高的若干系数作为类别的特征系数。对于未知类别的点击数序列，以特征系数作为神经网络的输入，输出的类别即是预测结果。文献[9]提出了基于小波多尺度分析的论坛话题热度预测算法。该算法通过对帖子的历史点击数时间序列进行小波分解与重构，来实现对未来点击数的预测。针对已有方法在长期预测方面的不足，文献[10]提出了一种预测事件长期发展趋势的方法，该方法首先通过周期分析和层次聚类为每类已发生的事件建立一个发展趋势模型；然后对待预测事件已有发展态势进行自适应缩放变换；最后应用最小二乘法从待预测事件所属类别的模型库中选取均方误差和最小的模型来预测该事件的发展趋势。虽然热点话题挖掘和预测方面已经取得了丰硕的成果，但是，现有的方法在预测话题热度时必须借鉴该话题已有的发展态势，无法对新话题发展初期的热度进行预测。

本文针对现有算法无法预测新话题热度这一问题，提出了一种基于K近邻的新话题热度预测算法，该算法利用与新话题相似的历史话题的点击数时间序列来对新话题的热度进行预测。实验结果表明，该算法对新话题发展初期的热度预测具有较高的精度。

3 基于K近邻的新话题热度和趋势预测算法

3.1 新话题热度和趋势预测基本思想

话题热度最常用的衡量因子是点击数[3]，热度预测本质上就是预测话题未来的点击数，趋势预测是对话题点击数变化情况的预测，如增加、降低等。本文算法的基本思想是：为了对新话题的热度和发展趋势进行预测，首先找到与该话题在内容上最相似的K个历史话题，然后根据这K个话题的热度和发展情况对新话题进行预测，即这种基于K-近邻[11]的预测算法认为内容相似的话题在热度和发展趋势上也应该具有相似性。本文第4节将根据实验结果对这种观点的正确性进行讨论。

3.2 话题相似性度量

为了找到与新话题在内容上最相似的K个历史话题，首先需要度量两个话题的内容相似性。为此，本文在对话题文本进行预处理的基础上，采用向量空间模型[13]来表示文本，将向量之间夹角的余弦值作为两个话题文本之间相似性的度量。

3.2.1 文本预处理

文本预处理的目的是过滤噪声，抽取出对表达文本主题重要的词作为文本特征，为后续文本向量化和度量相似性做准备。一般来说，文本主题主要是通过名词来表达的，因此，本文首先采用中科院分词系统[16]对话题文本进行分词和词性标注，然后抽取其中的名词(包括普通名词n、动词性名词vn、形容词性名词an)作为本文特征词。

3.2.2 文本向量化

给定历史话题文本集$D=\{d_1,d_2,d_3,\cdots,d_n\}$，$V=\{t_1,t_2,t_3,\cdots,t_{|V|}\}$是对$D$进行本文预处理后得到的所有不同的特征词构成的词表，其大小为$|V|$。按照向量空间模型的思想，任意话题文本d(d为历史话题或新话题)被表示为$|V|$维向量$\langle w_1,w_2,\cdots,w_{|V|}\rangle$，其中$w_i$为特征词$t_i$对于$d$的权重，度量方法为

$$w_i=tf_i\times idf_i \tag{1}$$

式中，tf_i为经过平滑处理的词项频率，定义为

$$tf_i=0.5+\frac{0.5+f_i}{\max\{f_1,f_2,\cdots,f_n\}} \tag{2}$$

式中，f_i表示特征词t_i在d中出现的次数。

idf_i表示经过平滑处理的逆文档频率，定义为

$$idf_i=\log\left(\frac{n-df_i+0.5}{df_i+0.5}\right) \tag{3}$$

式中，df_i表示包含特征词t_i的文本数。

3.2.3 话题相似度计算

本文将文档向量之间夹角的余弦值作为两个话题文档之间的相似性度量。假设d_i和d_k为任意两篇话题文档，经过预处理和向量化后，它们的特征向量分别为$W_i=\langle w_{i1},w_{i2},w_{i3},\cdots,w_{i|V|}\rangle$和$W_k=\langle w_{k1},w_{k2},w_{k3},\cdots,w_{k|V|}\rangle$，则$d_i$和$d_k$之间的相似度为

$$\begin{aligned}\mathrm{sim}(d_i,d_k)=Cosine(W_i,W_k)&=\frac{\langle W_i,W_k\rangle}{\|W_i\|\times\|W_k\|}\\&=\frac{\sum_{j=1}^{|V|}w_{ij}\times w_{kj}}{\sqrt{\sum_{j=1}^{|V|}w_{ij}^2}\times\sqrt{\sum_{j=1}^{|V|}w_{kj}^2}}\end{aligned} \tag{4}$$

对于任意的新话题，按照式(4)计算出它与每个历史话题的相似度后，取相似度最大的K个历史话题(即K个近邻)，根据这K个近邻话题的热度及其变化情况来预测新话题的热度及其变化趋势。

3.3 新话题点击数计算

对于任一新话题d，假设其K个近邻历史话题为$d_1,d_2,d_3,\cdots,d_k$，其中d_i的点击数时间序列[14]为$C_i=c_{i1}c_{i2}c_{i3}\cdots c_{it}$($c_{it}$为$d_i$第$t$天的点击数)，$K$个近邻的点击数时间序列分别为$C_1,C_2,C_3,\cdots,C_k$，$d$与$K$个近邻的相似度分别为$\mathrm{sim}_1,\mathrm{sim}_2,\mathrm{sim}_3,\cdots,\mathrm{sim}_k$(按照式(4)计算得到)，话题$d$第$t$天的点击数$c_t$预测为

$$c_t=\frac{\sum_{i=1}^{k}c_{it}\times\mathrm{sim}_i}{\sum_{i=1}^{k}\mathrm{sim}_i} \tag{5}$$

即对K个近邻历史话题的点击数根据它们与新话题的相似性进行加权平均，作为新话题点击数的预测值。

3.4 新话题热度变化趋势预测

利用K近邻算法得到新话题每一天的热度预测$c_1c_2c_3\cdots c_t$后，通过比较相邻两天点击数的大小关系，即可得到每一天

(从第二天开始)热度的变化趋势(上升、下降、持平),即任意一天的热度变化趋势 $t_i(i=2,3,\cdots,t)$ 预测为

$$t_i=\begin{cases}\text{上升}, & \text{如果 } c_i>c_{i-1}\\ \text{持平}, & \text{如果 } c_i=c_{i-1}\\ \text{下降}, & \text{如果 } c_i<c_{i-1}\end{cases} \tag{6}$$

4 实验

4.1 评价指标

本文采用相对误差和趋势预测准确率两个指标来对实验结果进行评价。

a)相对误差 *rela_err*

假设测试话题集为 D',d 是 D' 中的任一话题,它第 t 天的真实点击数为 c_t,预测出的点击数为 c_t',则该话题第 t 天热度预测的相对误差 $rela\_err_d$ 为

$$rela\_err_d=\frac{|c_t-c_t'|}{c_t} \tag{7}$$

假设算法对 D' 中话题的热度进行预测时,有 m 个话题第 t 天的热度预测有小于等于 p 的相对误差,则在允许相对误差低于 p 的情况下,算法对 D' 中所有话题第 t 天热度预测的正确率为 $m/|D'|$(其中 $|D'|$ 为测试话题的个数)。

b)趋势预测准确率

假设测试话题集为 D',d_r 是 D' 中的任一话题,d_r 的真实点击数时间序列为 $C_r=c_{r1}c_{r2}c_{r3}\cdots c_{rt}$,预测的点击数时间序列为 $C_o=c_{o1}c_{o2}c_{o3}\cdots c_{ot}$,$d_r$ 第 i 天到第 $i+1$ 天热度变化趋势预测正确与否记为 m_r,其定义为

$$m_r=\begin{cases}1, & (c_{ri}\leqslant c_{ri+1}\&\&c_{oi}\leqslant c_{oi+1})\,||\,(c_{ri}\geqslant c_{ri+1}\&\&c_{oi}\geqslant c_{oi+1})\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

则算法对 D' 中所有话题第 i 天到第 $i+1$ 天热度变化趋势预测的准确率 *per* 定义为

$$per=\frac{\sum_{r=1}^{|D'|}m_r}{|D'|} \tag{9}$$

式中,$|D'|$ 表示 D' 中测试话题的个数。

4.2 结果与分析

实验时选取中华论坛作为数据源,并针对中华论坛的中华山寨板块开发了数据获取器和网页净化器。实验数据涵盖了2011年6月10日至2011年7月20日共40天的帖子和这些帖子每天的点击数,近5000个话题。图1给出了不同 K 值($K=1\sim6$)下预测新话题前5天每天热度的相对误差,其中 p 表示相对误差,横坐标表示时间,纵坐标表示相对误差低于 p 情况下点击数预测的正确率。

从图1可以看出,在允许相对误差分别低于10%、20%和30%的情况下,前3天点击数预测的平均正确率分别达到了47.26%、61%和67.7%,而第4天分别下降到41.06%,53.79%和59.34%,第5天下降到34.45%,49.80%和52.77%。由此可见,随着时间的推移,正确率逐渐下降。一般来说,网络舆情爆发的周期都比较短,许多广为关注的舆情事件,从话题出现到广受关注只有短短几天时间。同时论坛中每天都有大量的新话题产生,民众一般会把注意力集中到新话题上,老话题被真正的关注者点击的可能性降低,而被漫游者点击的可能性增加,点击数的随机性变大,因此可预测性降低。但是从实验结果可以看出,对话题热度的短期预测,内容相似的话题可以提供重要的线索,内容相似的话题在话题初期具有类似的热度表现。

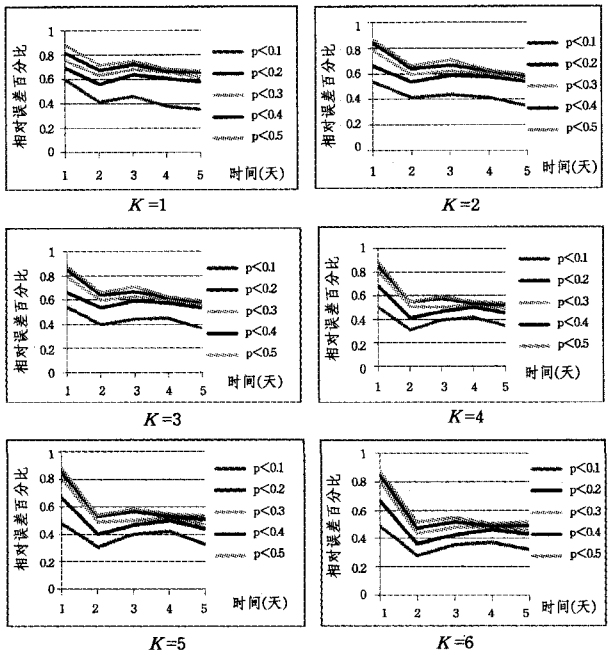

图1 不同 K 值点击数相对误差图

图2给出了不同 K 值下($K=1\sim6$)话题前5天热度变化趋势的预测情况,其中横坐标表示时间,纵坐标表示趋势预测准确率 *per*。从图2可以看到,对话题前3天热度变化趋势进行预测的准确率可以达到73.73%,说明相似的话题在发展初期具有近似的热度趋势。但是随着时间的推移,不同话题的热度变化趋势开始分化,基于内容相似的预测准确率也急剧下降。

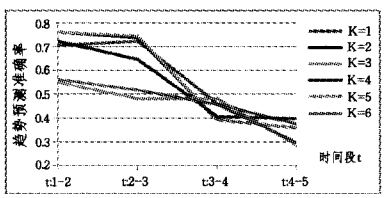

图2 K 从1到6的热度变化趋势预测准确率图

结束语 网络舆情是政府部门和企业以及社会大众所关注的焦点,而话题热度预测作为舆情监管和引导的基础,已经成为当前Web数据挖掘领域的热点课题之一。考虑到对话题进行短期预测,尤其是对话题发展初期进行预测具有重要的实际应用价值,而现有方法在这方面存在明显的不足,因此本文对新话题热度预测问题进行了研究,提出了基于K-近邻的新话题热度预测算法。实验验证了本文算法的有效性,结果表明该算法对新话题发展初期的热度预测具有较高的正确率,同时也证明了相似的话题在话题出现的初期具有近似的变化趋势。

参考文献

[1] 王来华.舆情研究概论-理论、方法和现实热点[M].天津:天津社会科学院出版社,2007

[2] 彭丹,许波,宋仙磊.基于网络评论的网络舆情研究[J].现代情报,2009,29(12):47

[3] 袁平波，俞能海，疏晓葵，等. 论坛帖子热度变化模型的研究[D]. 中国科技论文在线，2009
[4] 邱立坤，程葳，龙志祎，等. 面向 BBS 的话题挖掘初探[C]//全国第八届计算语言学联合学术会议(JSCL-2005). 北京：清华大学出版社，2005：401-407
[5] 高俊波，王晓峰，等. 一种新的主题影响力模型研究[J]. 计算机工程与应用，2007，43(25)：182-185
[6] Matsumura N，Ohsawa Y，Ishizuka M. Influence diffusion model in text-based communication[C]//Poster of the Eleventh International World Wide Web Conference. 2002
[7] 任晓东，张永奎，薛晓飞. 基于 K-Modes 聚类的自适应话题追踪技术[J]. 计算机工程，2009，35(9)：222-224
[8] 张虹，钟华，赵兵. 基于数据挖掘的网络论坛话题热度趋势预报[J]. 计算机工程与应用，2007，43(31)：159-161
[9] 张虹，赵兵，钟华. 基于小波多尺度的网络论坛话题热度趋势预测[J]. 计算机技术与发展，2009，19(4)：76-79
[10] 高辉，王沙沙，傅彦. Web 舆情的长期趋势预测方法[J]. 电子科技大学学报，2011，43(3)：440-445
[11] Mitchell T M. Machine Learning[M]. 曾华军，张银奎，等译. 北京：机械工业出版社，2009
[12] 刘里，何中市. 基于关键词语的文本特征选择及权重计算方案[J]. 计算机技术与发展，2009，19(4)：76-79
[13] Liu Bing . WEB DATA MINING[M]. 俞勇，薛贵荣，韩定一，等译. 北京：清华大学出版社，2009
[14] 程辉. 基于时间序列的网络舆情预测模型[J]. 网际网路技术学刊，2008，9(5)：16-17
[15] 大旗网. 口碑观察：上海大众刘坚车祸[DB/OL]. http://shehui. daqi. com/article/2951609. html，2011-08-20
[16] ICTCLAS 官网. ICTCLAS 汉语分词系统[DB/OL]. http://ictcla s. org/ictclas_about. htm，2011-08-20

(上接第 228 页)

都出现了高峰。但是随着时间的推进，两种情感都有所缓和，这主要是由于用户对该关注点整体关注度的降低。然而，随着'此次地震的起因是人祸而非天灾'的说法越来越受到网民的关注，网络上广泛掀起了对此次事故的批判声音。针对该事件的第二个关注点——核辐射相关内容，网民则一直表示出了很强的负面情感，不管是核辐射导致的海水和鱼虾蔬菜的污染，还是其导致的中国购盐热，都产生了很多的负面影响。而对于日本地震中所包含的第三个主题，即关于该次事故对中国本身的影响，以及国内新闻的报道、中国政府的应对措施等等，从图中曲线可以看出，国内民众的相关情绪逐渐得到稳定。

结束语 随着人们越来越多地通过网络平台发布自己对热点事件的观点和看法，针对热点事件主题的用户情感趋势挖掘能够快捷地反映网络用户在事件发生后的情感变化。而用户情感变化往往与事件本身发展的情况相关联，因此监测民众情感趋势能够对事件的发展状态进行实时评估，从而帮助决策部门及时制定相关措施，控制事件发展的不利影响。本文提出了一种面向微博客的热点事件情感分析方法，并实现了微博客背景下的情感趋势分析原型系统。突发事件的预测和异常事件的检测则是下一步工作的研究方向。

参 考 文 献

[1] 贺德方. 我国科技情报行业发展战略与发展路径的思考[J]. 情报学报，2007，26(4)：483-487
[2] 王飞跃. 开源情报与网络时代的国家安全[J]. 科学新闻杂志，2007，2(1)：3
[3] 王飞跃. 基于社会计算和平行系统的动态网民群体研究[J]. 上海理工大学学报，2011，33(1)：8-17
[4] 曾大军，王飞跃，曹志冬. 开源信息在突发事件应急管理中的应用[J]. 科技导报，2008，26(16)：27-33
[5] Savage N. Twitter as medium and message [J]. Commun ACM，2011，54(3)
[6] Jansen B J，Zhang M，Sobel K，et al. Twitter power：Tweets as electronic word of mouth [J]. Journal of the American Society for Information Science and Technology，2009，60(11)：2169-2188
[7] O'Connor B，Balasubramanyan R，Routledge B R，et al. From tweets to polls：Linking text sentiment to public opinion time series [C]//Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Washington，DC，2010：122-129
[8] Bollen J，Pepe A，Mao H. Modeling public mood and emotion：Twitter sentiment and socio-economic phenomena [J]. CoRR，2009，abs：0911-1583
[9] Thelwall M，Buckley K，Paltoglou G. Sentiment in Twitter events [J]. Journal of the American Society for Information Science and Technology，2011，62(2)：406-418
[10] Shen Y，Li S，Zheng L，Ren X，et al. Emotion mining research on micro-blog[C]// Proceedings of 1st IEEE Symposium on Web Society. Lanzhou，China，2009：71-75
[11] Song S，Li Q，Zheng N. A spatio-temporal framework for related topic search in micro-blogging [C]//Proceedings of the 2010 International Conference on Active Media Technology. Toronto，Canada，2010：63-73
[12] Sakaki T，Okazaki M，Matsuo Y. Earthquake Shakes Twitter Users：Real-time Event Detection by Social Sensors [C]//Proceedings of the 19th International World Wide Web Conference. Raleigh，NC(USA)，2010：851-860
[13] Singh V K，Gao M，Jain R. Situation Detection and Control using Spatio-temporal Analysis of Microblogs [C]// Proceedings of the 19th International World Wide Web Conference. Raleigh，NC(USA)，2010：1181-1182
[14] Vieweg S，Hughes A L，Starbird K，et al. Microblogging During Two Natural Hazards Events：What Twitter May Contribute to Situational Awareness [C]// Proceedings of the 28th International Conference on Human Factors in Computing Systems. Atlanta，Georgia，USA，2010：1079-1088
[15] 王素格，杨安娜. 基于混合语言信息的词语搭配倾向判别方法[J]. 中文信息学报，2010，24(3)：69-74
[16] 章成志，梁勇. 基于主题聚类的学科研究热点及其趋势监测方法[J]. 情报学报，2010，29(2)：342-349
[17] 夏天. 中文信息相似度计算理论与方法[M]. 郑州：河南科学技术出版社，2009
[18] Ku L，Liang Y，Chen H. Opinion extraction，summarization and tracking in news and blog Corpora[C]// Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. Stanford University，California，USA，2006