第 42 卷 第 1 期 计 算 机 工 程 2016 年 1 月
Vol. 42 No. 1 Computer Engineering January 2016

·人工智能及识别技术· 文章编号：1000-3428(2016)01-0180-07 文献标识码：A 中图分类号：TP18

基于核主成分分析与小波变换的高质量微博提取

彭 敏[1,2]，傅 慧[1]，黄济民[1]，黄佳佳[1]，刘纪平[1,2]

（1. 武汉大学计算机学院，武汉 430000；2. 武汉大学深圳研究院，广东 深圳 518000）

摘 要： 在线社交媒体中存在大量的噪音和冗余信息，为对其进行过滤和筛选，获取高质量的信息，提出基于核主分析和小波变换的高质量微博提取框架，并设计一种基于多特征融合的高质量信息的提取算法，将信息特征转换到小波域以更好地捕获信号间的细节差异。利用最大期望算法度量各个特征的权值，进一步融合得到特征综合值。为降低噪声特征对信息质量提取的影响并提高算法运算速度，引入核主成分分析对特征进行变换。实验结果表明，该框架能够提取出更高质量的微博，并且大幅减少运算时间。

关键词： 信息提取；特征融合；小波变换；期望最大算法；核主成分分析

中文引用格式： 彭 敏，傅 慧，黄济民，等. 基于核主成分分析与小波变换的高质量微博提取[J]. 计算机工程，2016，42(1)：180-186.

英文引用格式： Peng Min，Fu Hui，Huang Jimin，et al. High Quality Microblog Extraction Based on Kernel Principal Component Analysis and Wavelet Transformation[J]. Computer Engineering，2016，42(1)：180-186.

High Quality Microblog Extraction Based on Kernel Principal Component Analysis and Wavelet Transformation

PENG Min[1,2]，FU Hui[1]，HUANG Jimin[1]，HUANG Jiajia[1]，LIU Jiping[1,2]

（1. School of Computer，Wuhan University，Wuhan 430000，China；
2. Institute of Shenzhen，Wuhan University，Shenzhen，Guangdong 518000，China）

【Abstract】 Massive social event relevant messages are generated in online social media，which makes the filtering and screening of them be a challenge. In order to obtain massages with high quality，a high quality information extraction framework based on Kernel Principal Component Analysis and Wavelet Transformation（KPCA-WT） is proposed. Based on multiple features fusion，the paper designs an algorithm to extract the microblogs of high quality，which transforms the features into wavelet domain to capture the details differences between the feature signals. The features' weights are evaluated by employing Expectation Maximization（EM） algorithm and fused further to get a comprehensive value of each message，in order to reduce the effect of noise features，and to speed up the operation，the features are transformed through KPCA. Experimental results show that the proposed framework can extract information with higher quality and greatly reduce the time consumption.

【Key words】 information extraction；feature fusion；wavelet transformation；Expectation Maximization（EM） algorithm；Kernel Principal Component Analysis（KPCA）

DOI： 10.3969/j.issn.1000-3428.2016.01.032

1 概述

微博（Microblog）作为 Web2.0 时代新生网络应用形式，以其用户数量基数大、状态信息更新频繁、信息传播迅速等特点，在近几年得到了迅猛发展。急剧增长的用户，参与到社交媒体的直接后果是信息爆炸。用户产生的微博内容丰富，从信息传播的数量和质量看，大量同质化、无用的信息导致信息泛滥。因此，在信息检索、评论挖掘等方面，对信息的过滤和筛选成为一个具有挑战性的问题。

在线社交网络领域，信息挖掘的研究主要有2 个方面：(1)检测垃圾信息[1]或垃圾评论者[2]，这有利

基金项目： 国家自然科学基金资助项目(61472291，61303115)；2013 年深圳知识创新计划基础研究基金资助项目。

作者简介： 彭 敏(1973－)，女，教授、博士后，主研方向为主成分分析、自然语言处理；傅 慧，硕士研究生；黄济民，本科生；黄佳佳，博士研究生；刘纪平，讲师。

收稿日期： 2014-11-17 **修回日期：** 2014-12-17 **E-mail：** pengm@whu.edu.cn

于排除错误信息;(2)鉴定高质量的信息,提高查询检索效率。本文着重于第 2 方面的研究。

目前,大部分鉴定高质量信息的研究侧重于信息提取[3-4]、信息摘要[5]或以其他方式呈现微博事件的内容[6]。传统的方法主要考虑信息的内容特征。这些方法主要基于 LDA(Latent Dirichlet Allocation)模型[7-9]、TF-IDF(Term Frequency-Inverse Document Frequency)模型[5]或者其他主题模型[10-12]来提取与主题相关的信息。比较有代表性的有:文献[13]通过几个特定领域的相似度和加权集成聚类,结合各种社交媒体的特征来定义事件;文献[14-15]基于多个特征提出了新的压缩采样的抽样方法;文献[16]通过时间序列和特征选择,提出了另一种高质量主题的发现方法;文献[17]运用支持向量机(Support Vector Machine,SVM)回归和文本的多个特征来有效地预测评论;文献[18]提出一个相互强化学习框架,同时预测信息内容质量和用户口碑。然而,本文研究的问题与之前的有较大的不同:(1)处理的文本非常短,每条微博消息都被限制在 140 个字符之内;(2)短文本的重要性(质量)不仅与内容有关,还与其他因素有关;(3)微博中重要(高质量)的信息只占平台信息中非常小的一部分,如何从大量低质量的微博文本信息集合中提取出高质量内容的文本更具有挑战性。

对于一个给定的微博事件集,本文综合考虑了微博短文本的多方面特征,基于核主成分分析设计了一个微博短文本的提取框架(KPCA-WT)。

2　问题定义

2.1　高质量微博集合

在线社交网络中的信息总是与社会事件相关联,所以可以将信息分为多个事件集。对于一个与某个具体事件相关的由海量微博组成的集合 Γ,该微博集合包含了社会结构、文本内容、发布者权威性等与事件相关的信息特征。由于不同的用户所发布的微博信息的质量有很大区别,因此集合 Γ 中存在大量的垃圾和冗余信息。本文尝试通过提取一个高质量的微博子集,来对事件进行高度概述。一个高质量的微博子集具有以下特点:

(1)是一个与事件内容高度相关,数据量远小于原始微博集合 Γ 的微博子集;

(2)体现整个事件的全过程或参与者的观点和行为;

(3)在参与者之间产生了广泛的关注;

(4)由有一定影响力的参与者发布;

(5)可读性强,冗余内容少。

本文的研究成果可以很好地应用于应急响应、病毒式营销、疾病的传播、社会管理、社会事件预报等方面。

2.2　特征分析

本文首先定义几组特征,以便从不同角度来表示微博消息的质量。这些特征包括内容特征、用户权威特征和微博行为特征。

(1)内容特征

用户在检索的时候一般希望获得质量高的短文本,具体的更倾向于内容相关度较高、长度较长的微博文本。

微博长度:由于微博消息的长度被限制在140 个字符以内,因此长的微博可能包含更多的信息。

内容的相关性:对于每条微博 i,都有一个 TF-IDF 值来衡量它与事件主题的相似度。TF-IDF $=\sum t_k(\text{if } t_k \text{ in } i)$,$t_k$ 是整个事件集中,第 k 个词的 $tf \times \log(idf+1)$ 值。

(2)用户权威特征

一个信息真实完整的用户可能发布更可靠的微博,而权威用户(特别是名人用户)更可能发布高质量的微博。

用户权威:如文献[19]可以通过利用粉丝得分(用户的粉丝数)来衡量一个用户的权威性。

(3)微博行为特征

微博有些特有的行为特点可以体现消息的行为特征。例如转发量和评论量,值越大说明微博内容越受到关注,所包含的信息也越丰富。

URL:对于第 i 条微博,w_k 是整个事件集中第 k 个URL 的出现频率,那么它的 URL 值 $=\sum w_k$ $(\text{if } w_k \text{in } i)$。

发布时间:更近的微博可能包含更新的信息,更有可能满足用户的需要。

基于微博数据所包含的不同特征,本文对高质量微博的提取任务作如下定义:给定一个事件的微博集 Γ,每条微博 $d_i \in \Gamma(i=1,2,\cdots,N)$ 有 L 个特征值 $F_l=(f_{i1},f_{i2},\cdots,f_{iL})$。通过核主成分分析,对特征进行降维,得到新的 K 维特征,则 N 条微博构造了一个 K 维特征矩阵 $\boldsymbol{F}=\{f_{ij},i=1,2,\cdots,N,j=1,2,\cdots,K\}$。接着量化并融合与这些微博特征相关联的权值,来定量评估微博的质量,并抽取打分较高、排序靠前的微博,构成一个高质量微博子集。本文提出的高质量微博提取框架(*KPCA-WT*)具体步骤如下:(1)对微博的多个特征进行无量纲化处理;(2)对特征进行核主成分分析,得到新的 K 维特征,构建特征矩阵 $\boldsymbol{F}=\{f_{ij},i=1,2,\cdots,N,j=1,2,\cdots,K\}$;(3)对每个特征向量 $\boldsymbol{F}_k(k=1,2,\cdots,K)$ 进行小波变换;(4)基于 EM 算法,对时间频率域的系数进行融合;(5)重构转换为信号的特征值,得到高质量微博。

3 基于核主成分分析的特征降维

微博消息的特征之间的关系是非线性的,运用传统的主成分分析,往往导致实验效果不理想。核主成分分析作为 PCA 方法的在处理非线性问题时的扩展,近年来得到了快速的发展。核主成分分析的目的是通过一个非线性变换 Φ,将输入空间映射到高维特征空间 F 中,使得高维特征空间线性可分,以便在此空间中提取主成分,并使它们的协方差结构满足单位矩阵[20-21]。基于此方法,不仅获得了良好的主成分提取性能,还能够保存数据中更多的高阶信息。

3.1 对原始数据的处理

原始数据一般包含了 2 个方面的重要信息:(1)各指标变异程度的差异信息,体现为各指标的变异系数(各指标方差与其均值之比);(2)各指标之间相互影响程度上的信息,体现为相关系数。但要对多组不同量纲、不同数量级的数据进行比较时,需要对它们先进行无量纲化处理[22-23]。传统主成分分析所采取的是“中心标准化”方法,不能准确反映原始数据所包含的全部信息。所以,需要对无量纲化方法进行改进。本文采用均值化方法,即:用各项指标的均值除以它们相应的原始数据。均值化处理不改变指标之间的相关系数,相关矩阵的全部信息都在相应的协方差矩阵中得到反映。

3.2 KPCA 具体步骤

KPCA 的具体步骤描述如下:

(1)获得数据集并进行无量纲化处理,得到输入样本 $X=[x_1,x_2,\cdots,x_l]$。

(2)选取适当的核函数,计算 $l\times l$ 维核矩阵 $\boldsymbol{K}=(k_{ij})_{l\times l}$,本文中选取高斯径向核函数。

(3)计算特征空间对映射数据进行中心化处理后的核矩阵 $\tilde{\boldsymbol{K}}$,如式(1)所示。

$$\tilde{\boldsymbol{K}}=\boldsymbol{K}-\boldsymbol{I}_l\boldsymbol{K}-\boldsymbol{K}\boldsymbol{I}_l+\boldsymbol{I}_l\boldsymbol{K}\boldsymbol{I}_l,(\boldsymbol{I}_l)_{ij}=1/l \tag{1}$$

(4)求解核矩阵 $\tilde{\boldsymbol{K}}$ 的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_l$ 和特征向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_l$。

(5)将特征值按降序排列,并调整与其对应的特征向量,并单位化特征向量。

(6)计算特征值的累积贡献率。一般取累计贡献率达 85% ~95% 的特征值 λ_1, $\lambda_2,\cdots,\lambda_m$ 所对应的第 1 个、第 2 个、…、第 $m(m\leqslant p)$ 个主成分,如式(2)所示。

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{2}$$

4 基于小波变换的 K 维特征融合

本文从全局范围内考虑特征的不同特点出发,基于小波变换提出了 K 维特征融合算法(K-dimensions Feature Fusion Based on Wavelet Transformation, KD-FF_WF),来提取高质量信息。给定输入信息的 K 维特征矩阵 $\boldsymbol{F}\in\mathbf{R}^{N\times K}$,将这些特征转换到小波域,获得它们的小波树 $T_k(k=1,2,\cdots,K)$。然后运用 EM 算法估计每个特征的贡献度来实现特征融合[24]。算法过程如图 1 所示。

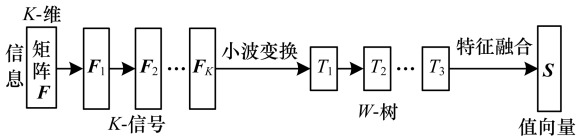

图 1 基于小波变换的 K 维特征融合算法

4.1 小波变换

本文对微博的特征进行均值化后,利用核主成分分析对特征进行降维,得到降维后的 K 维新特征的矩阵 $\boldsymbol{F}\in\mathbf{R}^{N\times K}$。每一维向量 $\boldsymbol{F}_k\in\mathbf{R}^{N\times 1}(k=1,2,\cdots,K)$ 代表了一个一维信号。通过小波变换将 K 维特征转换到时间-频率域[10],在实现多类特征融合的同时,捕获更加凸显的特征权值,降低计算复杂度。小波变换具有多尺度特性,体现了自适应分辨分析的思想,在低频时具有高的频谱分辨率,在高频时具有低的频谱分辨率,很适合用于捕获信号间的细节差异。小波变换的时间复杂度为线性时间 $O(N)$,远低于微博原特征空间的计算复杂度。小波函数能很好地描述数据的各种特性,如紧支性、消失矩、扩张性等[10]。小波基函数 h 可以采用 Haar 小波、Daubechies 小波或其他正交小波[10]。

4.2 小波域特征融合

根据实验可知,微博数据的特征向量可以相互影响,也可以共同体现出微博的一类特点,这里将全部特征向量都转换成小波域中的 L 个节点后,可以通过分别融合小波树 T_k 的第 $l(l=1,2,\cdots,L)$ 个节点的 K 个系数向量 $\boldsymbol{C}_{lk}(k=1,2,\cdots,K)$ 来融合需要共同进行分析的 K 维特征。本文采用 EM 算法来估计每个系数向量 $\boldsymbol{C}_{lk}$ 的贡献度。当分布包含隐性变量[24]时,EM 算法是参数学习的一种经典方法。具体过程如下:

对于每个节点集 $N_l=\{t_{kl}\mid k=1,2,\cdots,K\}$,需要确定系数矩阵 $\boldsymbol{C}_l$ 中的各维度 K_l 的贡献度 α_{lk}。设系数分布:$\boldsymbol{C}_l(\in\mathbf{R}^{nl\times K})$ 满足 K-分量的混合高斯分布。那么,基于 K-分量混合模型,融合系数 c_{li} 出现的概率,如式(3)所示。

$$f(c_{li})=\sum_{k=1}^{K}\alpha_{lk}f_{lk}(c_{lki}) \tag{3}$$

其中,$\alpha_{lk}\geqslant 0$,$\sum_{k=1}^{K}\alpha_{lk}=1$;$f_{lk}(c_{lki})$ 表示第 k 个分量分布的概率密度函数 $N(u_{lk},\sigma_{lk}^2)$;α_{lk} 表示节点集 N_l

中第 l 个节点的第 $k(k=1,2,\cdots,K)$ 个分量的贡献度。

设混合分布的参数向量为 $\boldsymbol{\Phi}_l=\{\boldsymbol{\Theta}_l;\boldsymbol{\Lambda}_l\}=\{\alpha_{l1},\alpha_{l2},\cdots,\alpha_{lK};\mu_{l1},\cdots,\mu_{lK},\sigma_{l1}^2,\cdots,\sigma_{lK}^2\}$;$\boldsymbol{\Lambda}_l$ 表示第 l 个节点集混合模型的分布参数向量。式(3)可表示为式(4):

$$f(c_{li}\mid\boldsymbol{\Phi}_l)=\sum_{k=1}^{K}\alpha_{lk}f_{lk}(c_{lki},u_{lk},\sigma_{lk}^2)=\sum_{k=1}^{K}\alpha_{lk}\frac{1}{\sqrt{2\pi\sigma_{lk}^2}}\mathrm{e}^{-\frac{(c_{lki}-u_{lk})^2}{2\sigma_{lk}^2}}\tag{4}$$

因此,所有的第 l 个节点的节点集 N_l 的似然函数可以表示为式(5):

$$L(\boldsymbol{\Phi}_l)=\prod_{i=1}^{n^l}f(c_{li}\mid\boldsymbol{\Phi}_{lK})=\prod_{i=1}^{n^l}\sum_{k=1}^{K}\alpha_{lk}\frac{1}{\sqrt{2\pi\sigma_{lk}^2}}\mathrm{e}^{-\frac{(c_{lki}-u_{lk})^2}{2\sigma_{lk}^2}}\tag{5}$$

运用最大似然函数来估计向量参数 $\boldsymbol{\Phi}_l$,由式(5)可以得到式(6):

$$l(\boldsymbol{\Phi}_l)=\ln(L(\boldsymbol{\Phi}_l))=\sum_{i=1}^{n^l}\ln\left(\sum_{k=1}^{K}\alpha_{lk}\frac{1}{\sqrt{2\pi\sigma_{lk}^2}}\mathrm{e}^{-\frac{(c_{lki}-u_{lk})^2}{2\sigma_{lk}^2}}\right)\tag{6}$$

本文通过 EM 算法的迭代过程获得 $l(\boldsymbol{\Phi}_l)$ 的最大值。算法 1 给出了 EM 算法的全过程。

算法 1　EM 估计

输入　节点集 N_l,系数矩阵 $\boldsymbol{C}_l=\{C_{lk};k=1,2,\cdots,K\}$。

Begin

(1)对参数向量 $\boldsymbol{\Phi}_l$ 初始化:

$\boldsymbol{\Phi}_l^{(0)}=\{\alpha_{l1}^{(0)},\alpha_{l2}^{(0)},\cdots,\alpha_{lK}^{(0)};\mu_{l1}^{(0)},\cdots,\mu_{lK}^{(0)},\sigma_{l1}^{2(0)},\cdots,\sigma_{lK}^{2(0)}\}$

(2)repeat

运用式(7)计算 $\boldsymbol{C}_l$ 中每个元素的后验概率 w_{lki}:

$$w_{lki}^{(s)}=\frac{\alpha_{lk}^{(s-1)}f_{lk}(c_{lki},u_{lk},\sigma_{lk}^2)}{\sum_{t=1}^{K}\alpha_{lt}^{(s-1)}f_{lt}(c_{lti},u_{lt},\sum_{lt}^{2})}\tag{7}$$

对式(7)中 $\boldsymbol{\Phi}_l$ 的每个参数做偏导可以估算出向量 $\boldsymbol{\Phi}_l^{(s)}$ 新的值。

(3)until $|l^{(s)}\boldsymbol{\Phi}_l-l^{(s-1)}\boldsymbol{\Phi}_l|\leqslant\varepsilon$。

输出　一组最佳参数 $\boldsymbol{\Theta}_l=\{\alpha_{l1},\alpha_{l2},\cdots,\alpha_{lK}\}$。

End

通过 EM 算法得到了最佳参数 $\boldsymbol{\Theta}_l=\{\alpha_{l1},\alpha_{l2},\cdots,\alpha_{lK}\}$,通过对 $\boldsymbol{\Theta}_l$ 的线性加权可将 K 维系数矩阵 $\boldsymbol{C}_l$ 融合成 1-维系数向量 $\boldsymbol{C}_{l*}$:

$$c_{li}^{*}=\sum_{k=1}^{K}\alpha_{lk}c_{lki}$$

将融合后的系数集 $C^{*}=\{Cl^{*}\mid l=1,2,\cdots,L\}$ 通过多尺度二维小波重构函数逆变换重构得到一个新的信号 S,S 中的每个元素代表一条信息的综合值。算法 2 描述了基于小波变换的 K 维特征融合的全过程。

算法 2　基于小波变换的 K 维特征融合

输入　K-维特征矩阵 $\boldsymbol{F}\in\mathbf{R}^{N\times K}$

Begin:

(1) fork = 1 to K do

T_k = wavelet transformation h(F_k);//小波变换

end

(2) for l = 1 to L do

Estimate fusion parameter Θ_l of the lth coefficient C_l with EM algorithm;//训练参数

Calculate the fused coefficient C_{l*};//系数融合

end

(3) S = Inverse wavelet transformation h(C^{*});//小波逆//变换

输出　新的信号 S

End

4.3　时间复杂度

由于微博的数量 N 相当大,因此 KPCA-WT 框架的可扩展性变得相当重要。框架 KPCA-WT 的时间复杂度和 KD-FF_WT 算法的复杂度近似,但是经过 KPCA,特征的减少可以减少 EM 算法的迭代次数,大大地减少了 KPCA-WT 框架的时间开销。

算法 KD-FF_TF 中每个维度小波变换的计算复杂度是 $O(N\times j)$,其中,j 是小波分解层数。估计每个节点集 N_l 的参数向量 $\boldsymbol{\Theta}_l$ 的计算复杂度是 $O(s^l\times K^3\times n^l)$,$s^l$ 是迭代次数,K 是微博特征个数,n^l 是小波树 T_k 第 l 个节点的系数个数。因为每个 $\boldsymbol{\Theta}_l$ 的计算是独立的,所以 EM 估计过程可以针对所有的节点集 $N_l(l=1,2,\cdots,L)$ 并行执行。EM 估计过程的计算复杂度即为 $\max(O(L\times s^l\times K^3\times n^l)$。那么 KD-FF_WT 算法的整体计算复杂度为 $O(N\times j)+\max(O(L\times s^l\times K^3\times n^l)$。然而,$\max(n^l)\approx\frac{N}{2}$,分解层数 $j\in[2,8]$,同时 $s^l\in[30,50]$,所以 $O(N\times j)$ 远小于 $\max(O(L\times s^l\times K^3\times n^l)$。因此,KD-FF_WT 算法的复杂度近似 $\max(O(L\times s^l\times K^3\times n^l)$。

KPCA 的时间复复杂度为 $O(m^3)$,m 是样本的个数,那么 KPCA-KD-FF_WT 框架的整体计算复杂度为 $O(m^3)+\max(O(L\times s^l\times K^3\times n^l)$。因此,它的计算复杂度也近似为 $\max(O(L\times s^l\times K^3\times n^l)$。所以,特征的减少可以减少时间开销。

5 实验结果与分析

5.1 实验数据

本文研究的对象主要是微博消息,采用国内热度最高的微博平台——新浪微博作为实验数据收集对象。通过新浪微博 API 接口,将 2013 年 1 月-5 月中 8 个热点事件的微博作为实验数据。每个热点事件的微博数和涉及的用户数如表 1 所示。

表 1 事件集的微博数和涉及的用户数

热点事件	微博数	用户数
两会召开	73 337	59 257
第一夫人	9 603	8 571
国五条颁布	26 519	20 072
H7N9	78 814	68 760
雅安地震	101 978	87 577
北京雾霾	32 613	30 201
黄浦江死猪	19 282	17 176
撒切尔逝世	20 421	18 670

本文选取的微博特征共 7 个:微博发布时间,微博的评论量,微博的转发量,微博发布者的粉丝数,微博分词后的分次数,微博中的 URL 值和微博内容的 TFIDF 值。对数据做预处理后,KPCA 选取高斯径向核函数,并选取累计贡献率大于 95% 的主成分。小波基函数 h 选择离散小波 Daubechies7。

本文将所提出的提取框架和未经降维的KD-FF_TF 算法以及几个经典的微博提取方法,在时间开销和微博内容冗余度方面进行了对比。实验参照算法包括:

(1) Most Recent Tweets[6,14] (MR)算法。对于一个给定的事件集,基于微博发布时间的逆序排列提取微博的样本;

(2) Most Tweeted URL-based tweets[14] (MTU)算法,统计了每个事件集中的所有 URL 出现次数,并计算每条微博的 URL 值,基于 URL 值的逆序排列提取微博样本;

(3)常用于微博摘要和提取的 TF-IDF 算法[3,6,25]。

5.2 数据集

5.2.1 时间开销

通过 KPCA,可提取累计贡献率大于 95% 的核主成分,最终特征维度降低了将近一半。表 2 中展示了 KPCA-WT 框架和 KD-FF_WT 框架的整体运行时间。KPCA-WT 框架的时间开销包括:KPCA 运行时间,小波分解与信息重构时间,EM 训练参数时间。可以看到,通过主成分分析,提高了运算效率,降低了时间开销。

表 2 算法的运行时间 s

热点事件	KPCA-WT 框架	KD-FF_WT 框架
两会召开	75.77	104.14
第一夫人	6.56	19.26
国五条颁布	24.97	46.40
H7N9	81.13	114.24
雅安地震	54.84	80.70
北京雾霾	32.66	58.55
黄浦江死猪	17.58	36.81
撒切尔逝世	19.13	36.80

5.2.2 信息量

微博内容是微博的一个核心特征,这里基于信息量计算来衡量微博中包含信息的多少,通过最终的综合值对微博进行排序,将前 N 条微博总的信息量设为 $H(N)$($H(N)=\sum_{i=1}^{N} si\lg(si)$)。在图 2 中展示 8 个事件集由 KPCA-WT 框架和 KD-FF_WT 算法所提取出的前 150 条微博的信息量。可以看到 KPCA-WT 框架所提取出的微博总的信息量更大。

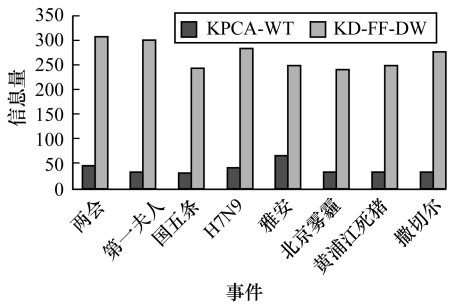

图 2 信息量提取结果

5.2.3 冗余内容减少的影响

本文通过计算不同方法提取的前 N 条微博的方差,来对比提取出微博内容的冗余度。前 N 条微博的方差表达式为(8):

$$\frac{1}{n-1}\sum_{i=1}^{N}(si-\overline{si}) \tag{8}$$

其中,s_i是第 i 条微博的内容得分。方差越大代表内容冗余度越小。

以事件集“国五条颁布”、“第一夫人”、“两会召开”、“雅安地震”为例,这里本文对比了 KPCA-WT 框架和 4 个实验参照算法(KD-FF_WT,MR,MTU,TFIDF)提取的前 N 条微博内容得分的方差,如图 3 所示。从图 3 可以发现,在一些事件集中,当选取更多特征时,未经过特征降维的 KD-FF_WT 算法所提取的微博内容冗余较大。这说明微博的多个特征中确实存在噪声和冗余,这使得有效特征的权重降低,影响了提取效果。

如图 3 所示,通过本文提出的 KPCA-WT 框架提取的高质量微博集,信息冗余最小,即这组微博集拥有最丰富的信息。

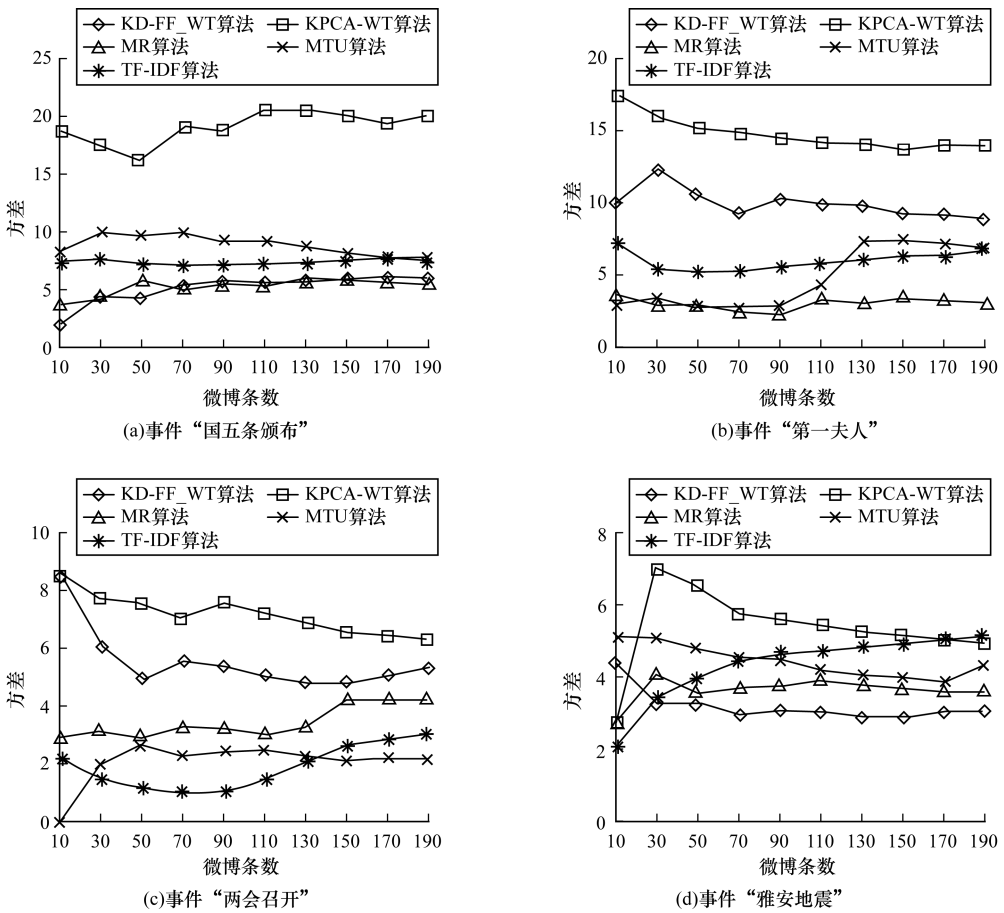

图 3　微博热门事件方差计算结果

5.2.4　微博内容

IK Analyzer 是一个开源的、基于 Java 语言开发的轻量级的中文分词工具包。本文通过中文分词器 IK Analyzer,对每个事件集内的所有微博进行分词并统计词频。同时,通过 KPCA-WT 框架,本文提取出了的每个事件集中排名前 100 的微博,对这100 条微博也进行分词和词频的统计。表 3 分别展示了微博分词后除去停用词的高频词汇。实验结果表明,基于本文 KPCA-WT 框架提取的前 100 条微博,能很好地表示整个事件集的主题内容。

表 3　微博事件集合中的高频词汇

主题	每个事件集合所有微博的高频词汇	每个事件集合前 100 条微博的高频词汇
两会召开	代表、中国、全国、提案、关注、问题、北京、委员	代表、全国、国家、委员、温家宝、奶粉、选举
第一夫人	走红、中国、头条新闻、出访、品牌、着装、百雀羚、山东	品牌、中国、出访、第一夫人、百雀羚、本土、习近平、服饰
国五条颁布	细则、出台、二手房、放假、政策、税、调控、楼市	细则、地产、二手房、税、出台、调控、楼市、北京
H7N9	禽流感、感染、病例、上海、死亡、患者、确诊、板蓝根	禽流感、上海、死亡、患者、板蓝根、感染、南京、病毒
雅安地震	灾区、四川、加油、祈福、捐款、爱心、庐山、希望	加油、祈福、爱心、灾区、爱、希望、救援、四川
北京雾霾	天气、北京、污染、空气、严重、口罩、空气质量、pm2.5	污染、空气、天气、上海、空气质量、环境、pm2.5、城市
黄浦江死猪	嘉兴、上海、打捞、漂浮、浙江、生猪、水质、耳标	上海、嘉兴、打捞、漂浮、殡葬、发现、处理、水质
撒切尔逝世	英国、铁娘子、去世、葬礼、前首相、中国、中风、87 岁	英国、铁娘子、去世、葬礼、中风、前首相、中国、伦敦

6　结束语

在社交媒体平台上,社会事件常常引起广泛的用户关注,一个热点事件往往会引起十万甚至上百万条的微博讨论。因此,为社交媒体上的用户提取数量较小、质量更高的热点事件的信息子集有着较高的社会意义和应用价值。本文从微博信息的多特征抽象和融合出发,提出了一个高质量的微博信息提取框架,并通过核主成分分析对微博多类特征进行降维,进一步提高微博提取速度和质量。基于一个包含 8 个热门事件的新浪微博实验数据集,本文从内容冗余、时间消耗这 2 个方面评估了所提出的

框架。实验结果表明,该框架提高了提取出的前 N 条微博总的信息量、减少了冗余内容并且有效地降低了时间消耗。通过词频统计可知,本文所提取出的微博能很好地代表整个事件集。

本文的主要贡献有:

(1)构建特征矩阵,并尝试通过小波变换,将微博信息的不同特征空间转换到一个统一的时域空间,以便凸显高质量微博信息的相关特征指标,并降低计算复杂度。

(2)在小波域内,基于 EM 算法评估各个特征的重要性程度,并结合重构算法,来对每条微博信息进行综合打分。

(3)基于特征变换来降低特征维度。EM 算法处理多维数据,维度增加时,其总迭代过程计算量很大,迭代速度相当慢,是次线性收敛速度。基于核主成分分析方法减少特征的维度,以便降低计算时间和噪声特征对提取效果的影响。

参考文献

[1] Jindal N, Liu B. Opinion Spam and Analysis[C]//Proceedings of the 2008 International Conference on Web Search and Data Mining. Los Angeles, USA: ACM Press, 2008: 219-230.

[2] Lim E P, Nguyen V A, Jindal N, et al. Detecting Product Review SpammersUsing Rating Behaviors[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada: ACM Press, 2010: 939-948.

[3] Becker H, Naaman M, Gravano L. Selecting Quality Twitter Content for Events[C]//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI Press, 2011: 442-445.

[4] Choudhury M D, Counts S, Czerwinski M. Find Me the Right Content! Diversity-based Sampling of Social Media Spaces for Topic-centric Search[C]//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI Press, 2011: 129-136.

[5] Sharifi B, Hutton M A, Kalita J K. Experiments in Microblog Summarization[C]//Proceedings of the 2nd International Conference on Social Computing. Minneapolis, USA: IEEE Press, 2010: 49-56.

[6] Ramage D, Dumais S, Liebling D. Characterizing Microblogs with Topic Models[C]//Proceedings of the International AAAI Conference on Weblogs and Social Media. Barcelona, Spain: AAAI Press, 2010: 130-137.

[7] Xia Wei, He Yanxiang, Tian Ye, et al. Feature Expansion for Microblogging Text Based on Latent Dirichlet Allocation with User Feature[C]//Proceedings of the 6th Joint International Technology and Artificial Intelligence Conference. Chongqing, China: [s. n.], 2011: 228-232.

[8] Titov I, McDonald R. ModelingOnline Reviews with Multi-grain Topic Models[C]//Proceedings of the 17th International Conference on World Wide Web. Beijing, China: [s. n.], 2008: 111-120.

[9] Li P, Jiang J, Wang Y. Generating Templates of Entity Summaries with an Entity-aspect Model and Pattern Mining[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010: 640-649.

[10] Daubechies I. TenLectures on Wavelets[M]. Philadelphia, USA: [s. n.], 1992: 213-222.

[11] Becker H, Naaman M, Gravano L. EventAdentification in Social Media[C]//Proceedings of ACM SIGMOD Workshop on the Web and Databases. Rhode Island, USA: ACM Press, 2009: 365-378.

[12] Sharifi B, Hutton M A, Kalita J. Summarizing Microblogs Automatically[C]//Proceedings of Human Language Technologies Conference. Los Angeles, USA: Association for Computational Linguistics, 2010: 685-688.

[13] Harabagiu S M, Hickl A. Relevance Modeling for Microblog Summarization[C]//Proceedings of the 5th International Conference on Weblogs and Social Media. Barcelona, Spain: AAAI Press, 2011: 514-517.

[14] Agichtein E, Castillo C, Donato D, et al. FindingHigh-quality Content in Social Media[C]//Proceedings of 2008 International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2008: 183-194.

[15] Lin Y R, Candan K S, Sundaram H, et al. SCENT: Scalable Compressed Monitoring of Evolving Multirelational Social Networks[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2011, 7(1): 29.

[16] Chen You, Cheng Xueqi, Yang Sen. Finding High Quality Threads in Web Forums[J]. Journal of Software, 2011, 22(8): 1785-1804.

[17] Kim S M, Pantel P, Chklovski T, et al. Automatically Assessing Review Helpfulness[C]//Proceedings of 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: Association for Computational Linguistics, 2006: 423-430.

[18] Bian Jiang, Liu Yandong, Zhou Ding, et al. Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain: ACM Press, 2009: 51-60.

[19] Duan Yajuan, Jiang Long, Qin Tao, et al. An Empirical Study on Learning to Rank of Tweets[C]//Proceedings of the 23rd International Conference on Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2010: 295-303.

[20] Schölkopf B, Smola A, Müller K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5): 1299-1319.

[21] Schölkopf B, Smola A, Müller K R. KernelPrincipal Component Analysis[C]//Proceedings of ICANN'97. Berlin, Germany: Springer, 1997: 583-588.

[22] 梅长林, 周家良. 实用统计方法[M]. 北京: 科学出版社, 2002.

[23] 徐雅静, 汪远征. 主成分分析应用方法的改进[J]. 数学的实践与认识, 2006, 36(6): 68-75.

[24] Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[25] Vosecky J, Leung K W T, Ng W. Searching for Quality Microblog Posts: Filtering and Ranking Based on Content Analysis and Implicit Links[C]//Proceedings of DSAA'12. Busan, Korea: Springer, 2012: 397-413.

编辑 索书志