垃圾评论自动过滤方法*

谭文堂,朱 洪,葛 斌,李芳芳,肖卫东
(国防科技大学 信息系统工程重点实验室,湖南 长沙 410073)

摘 要:针对互联网上存在的大量垃圾评论,提出一种基于电阻网络的垃圾评论检测方法,该方法用电阻距离来度量评论之间的上下文语义相似性,把整个评论数据表示成一个电阻网络,把垃圾评论当作该网络上的语义离群点来处理,根据网络节点对电阻网络平均电能消耗的影响,建立电离群因子来度量数据的离群程度,以此来识别垃圾评论。实验证明了该方法的有效性,在多个数据集上取得了较好的效果。

关键词:垃圾评论检测;电阻距离;电离群因子
中图分类号:TP316 **文献标志码**:A **文章编号**:1001-2486(2012)05-0153-05

Method of review spam detection

TAN Wentang, ZHU Hong, GE Bin, LI Fangfang, XIAO Weidong
(Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract: For detecting review spam in the Internet automatically, an method based on resistance network is proposed. By treating the distance between two reviews as a resistance, we represent the given dataset as a resistance network and the resistance distance between two nodes is a measure of the semantic distance between them. Spam reviews are semantic outliers in this network. An electrical outlier factor (EOF) for each review based on its influence on the power dissipated of the network was used to detect the spam reviews. Experimental results testified that EOF is suitable for detecting review spam, and is efficient and effective.

Key words: review spam detection; resistance distance; electrical outlier factor

伴随着 Web 2.0 应用的普及,互联网用户将网络作为自己表达主观性观点、态度、感觉、情绪的平台,在网络上发表大量有关人物、事件、产品等有重要价值的评价或意见,面向 Web 评论数据的数据挖掘应运而生,如:在线社会网络分析[1]、情感倾向性分析[2]等。但是互联网是一个开放的舆论平台,用户可以对任何帖子回复自己想说的内容,这导致评论数据当中存在着大量垃圾信息,严重干扰面向评论的数据挖掘工作。因此,如何识别并过滤这些垃圾评论具有重要的意义。Jindal 和 Liu 最先提出针对商品的垃圾评论概念[3-4],并把垃圾评论分为三类,第一类为错误或虚假的意见,这类评论有意地给某种商品好评,或者恶意诋毁某种商品;第二类是带偏见的评论,如对自己喜爱的品牌一律好评。第三类则是无关的评论,包括与主题无关的回复或广告等。

本文对第三类垃圾评论进行识别,但本文的工作不限于商品评论,还包括博客评论、论坛讨论。针对有监督方法泛化能力的不足,本文采用无监督方法对第三类垃圾评论进行识别,采用图的电阻距离刻画评论数据之间的相似性,在此基础上采用基尔霍夫指数即网络平均电能消耗的变化来度量数据的离群程度,建立电离群因子,根据电离群因子的局部变化和评论与主题的相关度,计算局部电离群因子,当局部电离群因子大于某个阈值时,把该评论当作垃圾评论。本文方法并不适用于实时过滤,因为在未知语境的情况下很难去定义垃圾评论。

1 相关工作

Jindal 等分别用逻辑斯蒂回归模型、支持向量机、决策树、朴素贝叶斯等分类器在手工标注好的垃圾评论集进行学习,建立统计模型来进行垃圾评论识别,达到了较高的识别率[3-5]。然而该方法需要大量的手工标注的垃圾评论,不同的商品评论数据的特征空间是不一样的,因此该方法只在特定领域或者商品有效,很难实现广泛应用。鉴于垃圾评论识别的难度,Lim 等转而识别发出

* **收稿日期**:2012-03-04
基金项目:国家自然科学基金资助项目(60903225);国防科技大学优秀研究生创新基金资助项目(S100502)
作者简介:谭文堂(1983—),男,贵州平塘人,博士研究生,E-mail:dean.tanw@gmail.com;
肖卫东(通信作者),男,教授,博士生导师,E-mail:wilsonshaw@vip.sina.com

垃圾评论的用户,即 Review Spammer[6],他们把观点总是与大部分人相左的人看作 Review Spammer,因此他们对用户的评论行为建模,把经常与大部分人意见相左的用户当作 Review Spammer。Mukherjee 等则认为 Review Spammer 一般都是团体活动,因此他们对 Spammer Group 进行识别[7]。另一种垃圾评论的定义是由 Mishne 等提出的 Comment spam[8],但是 Comment spam 是指在评论包含指向垃圾链接、页面的评论,并没有考虑没有含有链接的回复[9-12],该定义与本文所研究的垃圾评论具有较大的差异,在此不再赘述。

2 问题描述及分析

当前主要的垃圾评论识别方法都是有监督的方法,通过标注的垃圾评论进行训练,建立统计模型来识别垃圾评论。有监督的方法基于大量标注的样本,虽然能够得到较好的效果,但是由于评论的用户产生特性,垃圾评论的特征是动态变化的,有监督的方法难以广泛应用。因此本文采用无监督的方法自动识别第三类垃圾评论。本文遵循 Jindal 等关于垃圾评论的定义,但研究对象与其略有不同,本文认为识别前两类垃圾评论是很困难的,特别是评论对象不是商品而是公共事件时,即便人也很难识别哪些评论是垃圾评论。而 Lim 等把意见与大部分用户相左的用户当作 Review Spammer 显然也不科学。因此本文所研究的垃圾评论是指与评论对象如博客、新闻、主帖没有明显语义关联关系的评论。在本文中语义相似性是指评论在帖子的上下文语境中的相似性,而不是基于语义资源来计算评论之间的语义相似性,因为网络用语与语义资源的语义可能有较大出入且变化较快,如"破坏性试验"、"保护性拆除"等突然涌现的词语其含义与其本意有很大区别,因此上下文更能反映词在当前语境下的语义。垃圾评论涉及三个方面的问题,一是评论之间的语义相关性;二是网络习语的处理;三是垃圾评论的识别。

经过对评论数据进行统计分析我们发现,垃圾评论是评论中的少数,因为当一个帖子或博客中含有大量与其主题看似无关的评论时,这些评论与评论对象必有隐含的语义联系,如在一个有关官二代的帖子中含有大量"我爸是李刚"的评论,虽然该评论与帖子没有直接关系,但在网络语境下,二者具有密切的语义联系,这时就不能把这样的评论当作垃圾评论,因此本文把垃圾评论当作语义上的离群数据或者噪音来处理。针对第一个问题,本文首先用余弦公式计算评论之间的相似性,把评论之间的距离当作电阻,在此基础上建立电阻网络,采用电阻距离对评论之间的语义相关性进行度量。一般意义的文本相似性度量没有考虑数据的上下文,评论数据的特征很稀疏,因此对其进行简单的两两比较不能完整地体现数据间的相似性,如 A 与 B 相似,而 B 与 C 相似,虽然 A 与 C 之间没有共同特征,但如果考虑上下文,A 与 C 之间仍应具有一定程度的相似性。基于电阻网络,本文建立一种电离群因子来识别垃圾评论。针对网络习语,本文首先建立一个网络习语词库,对含有网络习语的评论,对其电离群因子进行惩罚,减去一个惩罚因子。以下先给出几个相关的定义。

2.1 电阻距离

Klein 在 1993 年首次提出图的电阻距离的概念[13],对于有限无向的连通图 $G=\{V,E\}$,$\boldsymbol{A}=[w_{ij}]$为图 G 的邻接矩阵,$w_{ij}=\frac{1}{r_{ij}}$,r_{ij}为邻接节点 i 和 j 间的电阻。图上的电阻距离定义如下:

定义 1 电网络中的任意两个节点 i,j 之间的电阻距离 r_{ij}是指在 i,j 之间施加一个单位电压之后,电路中电压与电流的比值,也叫节点 i,j 之间的有效电阻[13]。

Klein 根据基尔霍夫定律给出了电阻距离的计算方法:设 $\boldsymbol{L}$ 为图 G 的拉普拉斯矩阵,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$,$\boldsymbol{D}$ 为 $\boldsymbol{A}$ 的度矩阵,即 $\boldsymbol{D}$ 为对角阵且 $(\boldsymbol{D})_{ii}=\sum_{j=1}^{n}(\boldsymbol{A})_{ij}$,则

$$r_{ij}=l_{ii}^{+}+l_{jj}^{+}-l_{ij}^{+}-l_{ji}^{+}=l_{ii}^{+}+l_{jj}^{+}-2l_{ij}^{+} \quad (1)$$

电阻距离与图上随机游走时两个节点之间的平均往返时间 ECT(Exception Commute Time)[14] 是等价的,反映了两个节点之间的上下文相似性。在给定路径长度的情况下,节点间电阻距离与两点间最短路径数量成正比,两个节点之间最短路径越多,它们电阻距离就越小。

2.2 基尔霍夫指数

基于电阻距离,Klein 等提出了一种图的拓扑指标 —— 基尔霍夫指数(Kirchhoff index)[13],该指数与维纳指数相似[19],反映了图的拓扑特征。

定义 2 图 G 的基尔霍夫指数 $Kf(G)$ 定义为图 G 中所有点对之间电阻距离之和,即

$$Kf(G)=\sum_{i<j}r_{ij}=n\mathrm{Tr}(\mathrm{L}^{+}) \quad (2)$$

其中 $\mathrm{Tr}(\boldsymbol{L}^{+})$ 表示拉普拉斯矩阵 $\boldsymbol{L}$ 的 Moore-Penrose 逆矩阵($\boldsymbol{M}-\boldsymbol{P}$ 逆)$\boldsymbol{L}^{+}$ 的迹。基尔霍

夫指数反映了当随机地在电阻网络注入电流时,网络对电力的平均消耗,设一个电流 $J \in \mathbf{R}^n$ 被随机地注入为网络 G,且:$\mathrm{E}(J) = 0$;$\mathrm{E}(JJ^{\mathrm{T}}) = I - 11^{\mathrm{T}}/n$,E 表示数学期望,即 J 的期望为 0,协方差为 $I - 11^{\mathrm{T}}/n$。根据基尔霍夫电流定律可得:$1^{\mathrm{T}}J = 0$,电路消耗功率等于电流的平方与电阻的乘积,即功率 $P \sim J^{\mathrm{T}}L^{+}J$

$$\mathrm{E}(P) = \mathrm{E}(J^{\mathrm{T}}L^{+}J) = \mathrm{Tr}(\boldsymbol{L}^{+})\mathrm{E}(JJ^{\mathrm{T}}) = \mathrm{Tr}(\boldsymbol{L}^{+}) = \frac{1}{n}Kf(G) \tag{3}$$

由此可知基尔霍夫指数越小,单位时间内电阻网络消耗的电能就越少。在网络规模一定的情况下,网络连通性越好,节点间的电阻距离越小,则整个网络消耗的平均电能就少。

本文把评论数据之间的距离当作一个电阻,构建一个电阻网络。垃圾评论由于与大部分评论的距离较远,所处区域数据相对稀疏,因此垃圾评论到正常评论的电阻距离较大,因此由于垃圾评论的存在所导致的电能消耗就比较大。当去掉垃圾评论时,基尔霍夫指数的变化要比非垃圾评论大,因此本文根据电阻网络的基尔霍夫指数的变化提出一种电离群因子,以下给出电离群因子的定义:

定义 3　对于数据点 n,及其所在的连通图 G,假设去掉点 n 之后的图为 G_n,则电离群因子 EOF(Electrical Outlier Factor) 定义为图 G 与图 G_n 的基尔霍夫指数的比值,即

$$EOF_n = \frac{Kf(G)}{Kf(G_n)} \tag{4}$$

如果评论中含有网络习语,则

$$EOF_n = \frac{Kf(G)}{Kf(G_n)} - \min(EOF_n, \overline{EOF}) \tag{5}$$

其中 $\min(EOF_n, \overline{EOF})$ 为惩罚因子,$\overline{EOF}$ 为平均电离群因子。

定义 4　对于连通图 G 中的数据点 n,数据点 n 的电离群因子为 EOF_n,其 k 最近邻集合为 $N_k(n)$,则局部密度 LD(Local Density) 定义为

$$LD(n) = \frac{k}{\left(\sum_{o \in N_k(n)} \frac{EOF_n}{EOF_o}\right)} \tag{6}$$

由上式可以看出,局部密度 LD 完全符合人的直觉,如果一个评论的电离群因子越大,则 LD 越小,即该数据点所处的区域数据分布比较稀疏。根据局部密度 LD,本文给出局部电离群因子的定义。

定义 5　对于连通图 G 中的数据点 n,数据点 n 的电离群因子为 EOF_n,其 k 最近邻集合为 $N_k(n)$,则局部电离群因子 LEOF(Local Electrical Outlier Factor) 定义为

$$LEOF = \frac{k}{\left(\sum_{o \in N_k(n)} \frac{LD(o)}{LD(n)}\right)} \tag{7}$$

如果评论 n 所处的区域比较稀疏,与主题的距离较远,而其 k 最近邻所处区域相对较密集,则 $LEOF$ 越大,从电阻网络的角度来看,如果评论 n 消耗的电能比其 k 最近邻大很多,则其是离群点的概率越大。

3　基于 LEOF 的垃圾评论检测

3.1　电离群因子的计算

由于需要计算拉普拉斯矩阵 $\boldsymbol{L}$ 的 $M - \boldsymbol{P}$ 逆矩阵 $\boldsymbol{L}^{+}$,电离群因子的计算复杂性上界达到了 $O(n^4)$,因此,本文采用一种近似的方法,在计算节点 n 的 EOF_n,不是直接删除节点,而是把该节点置地,这样不改变图 G 的拓扑结构,此时 $Kf'(G_n)$ 相当于把节点 n 置地后随机地在其他 $n-1$ 个节点上注入随机电流 $J \in \mathbf{R}^{n-1}$ 时图 G 消耗的平均电能,因此 $EOF_n = \frac{Kf(G_n)}{Kf(G)}$。此时 G_n 的拉普拉斯谱矩阵 $\boldsymbol{L}_n$ 相当于 $\boldsymbol{L}$ 去掉第 n 行和第 n 列,这样就可以根据 Greville 迭代法的思想利用 $\boldsymbol{L}^{+}$ 来求 $\boldsymbol{L}_n^{+}$[17]。

定理 1　给定 $\boldsymbol{L}^{+}$,则 $\boldsymbol{L}^{+}$ 的每一个元素可由下式给出:

$$(\boldsymbol{L}_n^{+})_{ij} = (\boldsymbol{L}^{+})_{ij} - (\boldsymbol{L}^{+})_{in} - (\boldsymbol{L}^{+})_{nj} + (L^{+})_{nn} \tag{8}$$

证明　假设 M_{n-1} 是一个对称 $(n-1)\times(n-1)$ 矩阵,令 M_n 为 $(n-1)\times n$ 矩阵,且

$$M_n = [M_{n-1}\ \boldsymbol{a}] \tag{9}$$

即 M_n 为 M_{n-1} 加上一个列向量 $\boldsymbol{a}$ 所得;M_{nn} 为 $n \times n$ 矩阵,且

$$M_{nn} = \begin{bmatrix} M_n \\ \boldsymbol{a}'^{T} \end{bmatrix} \tag{10}$$

即 M_{nn} 为 M_{n-1} 加上一个列向量和一个行向量所得。

$$M_n^{+} = [M_{n-1}\ \boldsymbol{a}]^{+} = \begin{bmatrix} M_{n-1}^{+} - \boldsymbol{d}\boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix} \tag{11}$$

上式中:

$$\boldsymbol{d} = M_{n-1}^{+}\boldsymbol{a}$$

$$\boldsymbol{c} = \boldsymbol{a} - M_{n-1}\boldsymbol{d}$$

$$\boldsymbol{b}^{\mathrm{T}} = \begin{cases} \boldsymbol{c}^{+}, & \text{if } c \neq 0 \\ (1 + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{d})^{-1}\boldsymbol{d}^{\mathrm{T}}M_{n-1}^{+}, & \text{if } c = 0 \end{cases} \tag{12}$$

此处令 $M_{n-1} = \boldsymbol{L}_{n-1}$,则根据拉普拉斯矩阵的性质可得:$\boldsymbol{L}e = 0$,$\boldsymbol{a} = -\boldsymbol{L}_{n-1}e_{n-1}$,$\boldsymbol{d} = \boldsymbol{L}_{n-1}^{+}\boldsymbol{a} =$

$-\boldsymbol{L}_{n-1}^{+}\boldsymbol{L}_{n-1}e_{n-1} = -e_{n-1}$，于是

$$\boldsymbol{M}_n^{+} = [\boldsymbol{M}_{n-1}\ \boldsymbol{a}]^{+} = \begin{bmatrix} \boldsymbol{M}_{n-1}^{+} - \boldsymbol{d}\boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{M}_{n-1}^{+} + e_{n-1}\boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_{n-1}^{+} \\ 0^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{b}^{\mathrm{T}} \end{bmatrix} \quad (13)$$

由上式可得

$$(\boldsymbol{L}_{n-1}^{+})_{ij} = (\boldsymbol{M}_n^{+})_{ij} - (\boldsymbol{M}_n^{+})_{nj} \quad (14)$$

根据文献[16]的方法，同理可得

$$(\boldsymbol{M}_n^{+})_{ij} = (\boldsymbol{L}^{+})_{ij} - (\boldsymbol{L}^{+})_{in} \quad (15)$$

$$(\boldsymbol{L}_n^{+})_{ij} = (\boldsymbol{L}^{+})_{ij} - (\boldsymbol{L}^{+})_{in} - (\boldsymbol{L}^{+})_{nj} + (\boldsymbol{L}^{+})_{nn} \quad (16)$$

根据定理1只需要计算一次 $\boldsymbol{M}-\boldsymbol{P}$ 逆矩阵 $\boldsymbol{L}^{+}$，就可以根据 $\boldsymbol{L}^{+}$ 来求出 $Kf'(G_n)$：

$$Kf'(G_n) = (n-1)\mathrm{Tr}(\boldsymbol{L}_n^{+}) = (n-1)\sum_{i}^{n-1}(\boldsymbol{L}_n^{+})_{ii}$$

$$= (n-1)\sum_{i}^{n-1}((\boldsymbol{L}^{+})_{ii} + (\boldsymbol{L}^{+})_{nn}) \quad (17)$$

3.2 $\boldsymbol{M}-\boldsymbol{P}$ 逆矩阵 $\boldsymbol{L}^{+}$ 的计算

当数据规模较大时，直接计算 $\boldsymbol{M}-\boldsymbol{P}$ 逆矩阵 $\boldsymbol{L}^{+}$ 效率很低，当前有很多方法利用矩阵的稀疏性，采用迭代的方法来求解$\boldsymbol{M}-\boldsymbol{P}$逆[20-21]。本文采用文献[18]的基于稀疏矩阵 Cholesky 分解的方法。文献[16]利用该方法计算了150 000个节点的$\boldsymbol{M}-\boldsymbol{P}$逆，说明该方法能显著提高稀疏矩阵的 $\boldsymbol{M}-\boldsymbol{P}$ 逆的计算效率。

3.3 LEOF 算法描述

LEOF 算法的具体流程如下：

输入：评论数据集 S，最近邻个数 k。

输出：评论数据集 S 中的垃圾评论。

算法过程：

(1)利用余弦公式计算评论数据的相似性矩阵 $\boldsymbol{A}$ 及其拉普拉斯矩阵 $\boldsymbol{L}$；

(2)根据2.2中描述的方法计算数据集 S 的拉普拉斯矩阵的 $\boldsymbol{M}-\boldsymbol{P}$ 逆矩阵 $\boldsymbol{L}^{+}$；

(3)根据定理1和式(17)计算基尔霍夫指数 $Kf'(G_n)$；

(4)根据 *EOF* 的定义，计算 S 中每个数据点的 *EOF*；

(5)根据 *LEOF* 的定义，计算 S 中每个数据点的 *LEOF*；

(6)输出 *LEOF* 大于1的数据的评论数据。

相似性矩阵 $\boldsymbol{A}$ 的时间复杂度为 $O(n^2)$，计算 $Kf'(G_n)$ 的时间复杂度为$O(n^3)$，在数据比较稀疏的情况下略大于 $O(n^2)$，*LEOF* 的计算复杂度为 $O(n^3)$，因此算法计算复杂度为$O(2n^3+n^2)$。

4 实验结果及分析

本节将对本文算法进行综合评价，由于当前识别垃圾评论的主要方法是有监督的方法，与本文工作不具可比性，因此本文没有就方法与其他人的工作进行比较。实验在两种数据上进行，第一种数据利用已有的评论数据集，在其中加入一些与评论无关的数据构成垃圾评论，数据主要包括谭松波的评论数据集 ChnSentiCorp [22]，康奈尔大学的电影评论数据集 Movie Review[23]，Liu 等的产品评论数据集 Products Review[24]：第二种为来自互联网的真实数据，本文分别选取来自天涯论坛的三个比较大的讨论帖：T1，T2，T3，平均回帖数在1000条左右；另外选取太平洋电脑网的有关电脑评论的三个帖子C1，C2，C3。实验数据的组成见表1，实验效果采用准确率(Precision)和召回率(Recall)来评价：

$$Precision = \frac{Number\ of\ correct\ spam\ reviews}{|Number\ of\ spam\ reviews\ detected|}Recall$$

$$= \frac{Number\ of\ correct\ spam\ reviews}{|Number\ of\ all\ spam\ reviews|}$$

实验从以下两个方面验证算法的效率与稳定性。

(1)算法的性能分析

实验结果如表2所示，其中 *EOF* 方法按照数据中的垃圾评论比例取 *EOF* 的在前$(p+1)\%$的评论作为垃圾评论，而 *LEOF* 的参数 $k=8$；实验中 *EOF* 方法由于按照指定的比例提取垃圾评论，其准确率和召回率相差不是很大，而 *LEOF* 方法则具有一定的差距。实验结果证明无论是 *EOF* 方法还是 *LEOF* 方法在垃圾评论识别中，都具有较高的识别效率，说明电离群因子能抓住正常评论与垃圾评论之间的语义上的差别。

表1 实验数据说明

Tab. 1 Description of experimental data

数据名称	评论数量	垃圾评论数	垃圾评论比例 $p\%$
ChnSentiCorp1	2000	200	10%
ChnSentiCorp2	4000	400	10%
Movie Review	1000	100	10%
Products Review	1000	100	10%
T1	893	131	14.7%
T2	1027	240	23.4%
T3	1213	298	24.6%
C1	287	56	19.5%
C2	208	47	22.6%
C3	322	52	16.1%

表 2　实验结果

Tab. 2　Experimental result

数据集名称	*EOF*		*LEOF*	
	Precision	Recall	Precision	Recall
ChnSentiCorp1	0.92	0.93	0.96	0.97
ChnSentiCorp2	0.94	0.95	0.95	0.93
Movie Review	0.94	0.96	0.97	0.99
Product review	0.92	0.93	0.96	0.94
T1	0.87	0.89	0.93	0.92
T2	0.87	0.88	0.94	0.92
T3	0.82	0.83	0.90	0.89
C1	0.86	0.88	0.91	0.89
C2	0.88	0.90	0.92	0.93
C3	0.83	0.85	0.98	0.96

虽然 *EOF* 与 *LEOF* 在评论数据集上表现不相上下,但是在实际的论坛帖子的垃圾评论识别上,*LEOF* 方法占有明显优势,这是因为人工向评论数据集加上垃圾数据时,垃圾数据的分布是比较均匀的,即垃圾数据的局部差别不大,而在实际的论坛帖子里,垃圾评论的分布是不均匀的,局部密度变化比较大,*LEOF* 方法则刚好反映了局部密度的变化,因此在论坛帖子中,*LEOF* 方法表现要比 *EOF* 好。

(2)算法参数的影响

对 *LEOF* 算法的受参数 k 的影响进行分析,采用评论数据集 Movie Review 测试参数 k 对算法的影响,实验结果如图 1 所示。实验证明 *LEOF* 算法对参数 k 的影响较小,虽然在 k 较小和较大时,准确率发生了一点波动,但幅度并不大,说明 *LEOF* 方法的密度对于参数 k 不是很敏感,它真实地反映了评论中的语义离群特征。

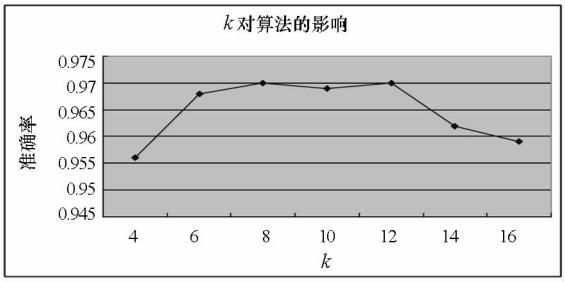

图 1　参数 k 对算法的影响

Fig. 1　Effection of parameter k

5　结论与下一步工作

本文研究了网络垃圾评论自动过滤问题,提出了一种基于网络平均电能消耗的垃圾评论检测方法。该方法建立评论之间的电阻网络,通过研究评论对于该网络的平均电能消耗的影响来检测垃圾评论,电能消耗越大,则离群因子越大。该方法考虑了评论之间的上下文相似性,对于高维稀疏的评论数据,具有较好的鲁棒性能。实验证明,算法在多个数据集上取得了很好的效果。下一步的工作中,将进一步降低算法的计算复杂性,并结合语义资源对垃圾评论进行实时过滤。

参考文献(References)

[1] Lin Y R, Sundaram H, Chi Y, et al. Discovery of blog communities based on mutual awareness [C]//Proceedings of Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.

[2] Liu B, Hu M Q, Cheng J S. Opinion observer: analyzing and comparing opinions on the web[C]//Proceedings of the 14th International Conference on World Wide Web, 2005: 342 - 351.

[3] Jindal N, Liu B. Analyzing and detecting review spam[C]// Proceedings of Seventh IEEE International Conference on Data Mining, 2007:547 - 552.

[4] Jindal N, Liu B. Review spam detection[C]//Proceedings of the 16th International Conference on World Wide Web, Banff, 2008:1 - 2.

[5] Jindal N, Liu B. Opinion spam and analysis [C] // Proceedings of the International Conference on Web Search and Web Data Mining, Palo Alto, 2008:1 - 11.

[6] Lim E P, Nguyen V A, Jindal N. Detecting product review spammers using rating behaviors[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, 2010: 1 - 10.

[7] Mukherjee A, Liu B, Wang J H, et al. Detecting group review spam[C]//Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 2011.

[8] Mishne G, Carmel D, Lempel R. Blocking blog spam with language model disagreement[C] //Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb, Chiba, Japan, 2005.

[9] Wong B, Locasto E M, Keromytis D A. PalProtect: A collaborative security approach to comment spam [C] // Proceedings of the IEEE Workshop on Information Assurance United States Military Academy, NY, 2006.

[10] Yang Y H, Zhao T J, Zheng D Q, et al. An efficient approach to comment spam identification [J]. Journal of Electronic (China), 2009, 26(5):644 - 651.

[11] Nagamalai D, Dhinakaran C B, Lee K J. Bayesian approach based comment spam defending tool[J]. International Journal of Network Security & Its Applications (IJNSA), 2010, 2 (4):267 - 280.

[12] Huang C R, Jiang Q C, Zhang Y. Detecting comment spam through content analysis [C] //Proceedings of the 11th International Conference on Web - Age Information Management, 2010.

[13] Klein D J, Randi M. Resistance distance[J]. Journal of Mathematical Chemistry, 1993(12):81 - 95.

[14] Chandra K, Prabhakar R, et al. The electrical resistance of a graph captures its commute and cover times [J]. Computational Complexity, 1997, 6:312 - 340.

(下转第 168 页)

$$+ C_5(p+t)\left(\int_{\sigma Q} |k(u-u_\Omega)|^{p+t}\mathrm{d}x\right)^{1/p}。$$

令 $s = p + t$ 并利用不等式 $(a^{1/p} + b^{1/p})^p \leqslant (a^{1/s} + b^{1/s})^s$，对任意 $a, b > 0$ 和 $p < s$，可得

$$\| u^+ \|_{\frac{sp}{n-p},Q} \leqslant C_6 \| u^+ \|_{s,\sigma Q} + C_7 \| k(u-u_\Omega) \|_{s,\sigma Q}。$$

同样的讨论对 u^- 一样成立。定理3证毕。

参考文献(References)

[1] Agarwal P R, Ding S, Nolder C A. Inequalities for di_erential forms[M]. Springer, 2009.

[2] Aronsson G, Lindqvist P. On p-harmonic functions in the plane and their stream functions[J]. J. Differential Equations, 1988,74:157 – 178.

[3] Ball J M. Convexity conditions and existence theorems in nonlinear elasticity[J]. Arch. Rational Mech. Anal., 1977, 63:337 – 403.

[4] Cao Z, Bao G, Xing Y, et al. Some Caccioppoli estimates for di _ erential forms [J]. J. Ineq. Appl., Article ID 734528, 2009.

[5] Ding S. Two-weight Caccioppoli inequalities for solutions of nonhomogeneous A – harmonic equations on Riemannian manifolds[J]. Proc. Amer. Math. Soc., 2004, 132: 2367 – 2375.

[6] Ding S. Local and global norm comparison theorems for solutions to the Nonhomogeneous A-harmonic equation[J]. J. Math. Anal. Appl, 2007,335:1274 – 1293.

[7] Giaquinta M, Soucek J. Caccioppoli's inequality and Legendre-Hadamard condition[J]. Math. Ann., 1985,270:105 – 107.

[8] Heinonen J, Kilpel ainen T, Martio O. Nonlinear potential theory of degenerate elliptic equations[M]. Dover Publications, Inc. Mineola, New york, 2006.

[9] Iwaniec T, Sbordone G. Weak minima of variational integrals [J]. J. Reine Angew. Math., 1994,454:143 – 161.

[10] Li X. On the strong Lp-Hodge decomposition over complete Riemannian manifold[J]. J. Fun. Appl., 2009, 257:3617 – 3646.

[11] Nolder C A. Hardy-Littlewood theorems for A-harmonic tensors[J]. Illinois J. Math., 1999,43:613 – 631.

[12] Nolder C A. Global integrability theorems for A-Harmonic tensors[J]. J. Math. Anal. Appl., 2000,247:236 – 247.

[13] Nolder C A. Conjugate harmonic functions and Cli _ ord algebras[J]. J. Math. Anal. Appl., 2005,302:137 – 142.

[14] Serrin J. Local behavior of solutions of quasi-linear equations [J]. Acta Math., 1964,111:247 – 302.

[15] Stro _ olini B. On weakly A-harmonic tensors[J]. Studia Math., 1995,114:289 – 301.

(上接第157页)

[15] Radl A, Von Luxburg U, Hein M. Getting lost in space: large sample analysis of the resistance distance[C] // Proceeding of 23rd Annual Conference on Neural Information Processing Systems(NIPS), 2010, Curran, RedHook, USA, 2010:1 – 9.

[16] Fouss F, Pirotte A. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on Knowledge and Data Engineering. 2007, 19(3):355 – 369.

[17] Ben-Israel A, Greville T. Generalized inverses: theory and applications[M]. seconded. Springer – Verlag, 2003.

[18] Herstein I, Winter D. Matrix theory and linear algebra[M]. Maxwell Macmillan International Editions, 1988, 440 – 441.

[19] Wiener H. Structural determination of paraffin boiling points [J]. Journal of the American Chemical Society,1947, 69:17 – 20.

[20] Greenbaum A. Iterative methods for solving linear systems [M]. Society for Industrial and Applied mathematics, 1997.

[21] Saad Y. Iterative methods for sparse linear systems[J]. Society for Industrial and Applied mathematics, 2000.

[22] Wu Q, Tan S B, Zhai H J, et al. SentiRank: cross – domain graph ranking for sentiment classification[C] //Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009: 309 – 314.

[23] Pang B, Lee L L, Vaithyanathan S. Thumbs up Sentiment classification using machine learning techniques[C] // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, 2002: 79 – 86.

[24] Hu M Q, Liu B. Mining opinion features in customer reviews[C]// Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI – 2004), 2004.