doi:10.3969/j.issn.1001-0505.2015.03.004

基于逻辑回归的中文在线评论有效性检测模型

吴含前[1] 朱云杰[1] 谢 珏[2]

([1]东南大学计算机科学与工程学院,南京 210018)
([2]东南大学-蒙纳士大学苏州联合研究生院,苏州 215123)

摘要: 为了实现电子商务和社交网络中文在线评论有效性的自动化检测,提出了一种单一主题环境下基于逻辑回归的垃圾评论检测模型.中文在线评论有效性的检测可以归结为分类问题,结合中文在线评论的特点提取了9个特征以构建分类模型;为获取核心特征主题的相关度,采用基于关联规则的评论名词模式优化了ICTCLAS中文分词系统的主题识别,进而利用交叉语言模型获取在线评论主题相关度.实验中采取了人为标定的1 000条评论作为样本,把支持向量机分类模型作为对比进行试验,利用数据挖掘工具Weka进行计算.结果表明,采用优化评论名词模式下基于逻辑回归的垃圾评论检测模型结果的准确率达到83.54%,比支持向量机分类模型计算得到的准确率高2.10%.

关键词: 在线评论有效性;逻辑回归;关联规则
中图分类号: P315.69 **文献标志码:** A **文章编号:** 1001-0505(2015)03-0433-05

Detection model of effectiveness of Chinese online reviews based on logistic regression

Wu Hanqian[1] Zhu Yunjie[1] Xie Jue[2]

([1]School of Computer Science and Engineering, Southeast University, Nanjing 210018, China)
([2]Southeast University-Monash University Joint Graduate School, Suzhou 215123, China)

Abstract: In order to realize automated detection of the effectiveness of Chinese online reviews in the context of e-commerce and social networks, a spam detection model based on logistic regression to solve single topic classification problem is proposed. The detection of effectiveness of Chinese online reviews can be regarded as a classification problem. According to the characteristics of Chinese online reviews, nine features are extracted to build the classification model. In order to extract the core feature-topic relevance, an association rule based review term mode is utilized to optimize the topics identification in ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System). The cross language model is then used to retrieve relevancy between online review topics. In the experiment, a sample of 1 000 human-labeled reviews is used, and the support vector machine (SVM) classification model is adopted as a comparison. The calculation results of the data mining tool Weka demonstrate that the accuracy rate of the proposed logistic regression classification model based on the optimized review term classification mode is 83.54%, which is 2.10% higher than that of the SVM classification model.

Key words: effectiveness of online review; logistic regression; association rule

电子商务领域中,在线评论对网购用户购买决策起着关键的影响作用.2013年中国网络购物市

收稿日期:2014-12-05. 作者简介:吴含前(1972—),男,博士,副教授,hanqian@seu.edu.cn.
基金项目:国家自然科学基金资助项目(60803057)、国家高技术研究发展计划(863计划)资助项目(2015AA015904).
引用本文:吴含前,朱云杰,谢珏.基于逻辑回归的中文在线评论有效性检测模型[J].东南大学学报:自然科学版,2015,45(3):433-437.
[doi:10.3969/j.issn.1001-0505.2015.03.004]

场研究报告指出:直至2013年12月,国内网购用户达到3.02×10^9人,37.5%的用户在购买不熟悉产品时主要考虑的是用户评价,其次为网站知名度和口碑[1].由于在线评论的好坏直接影响产品的销售[2-3],电子商务网站中出现了大量误导网购用户的、具有恶意目的的评论;同时,由于网购用户规模巨大,在线评论数量的爆炸式递增,增加了网购用户识别评论有效性的难度.因此,如何实现在线评论有效性的自动化识别成为了当前学术界和工业界的研究热点.

结合评论内容及其评论发布者,研究者们从以下2个方面对评论的有效性展开研究:① 评论者异常行为的检测[4-6],即通过研究评论者制造无效评论的方式和目的来发现无效评论者,从而识别无效评论;② 评论内容的检测[7-9],将评论有效性识别归结为基于监督学习的文本分类问题,通过构建分类模型识别无效评论.针对基于评论者异常行为的检测,通常采取的方法包括:① 建立无效评论者检测模型并对其打分[4],识别出无效评论制造者;② 采用关联规则[5]发现异常评论模式并识别产生无效性评论行为,发现无效评论制造者.由于网站十分重视对评论者信息资源的保护,在实际研究中难以完整获取评论者的行为信息,故对实际评论者行为检测的研究较为困难.基于评论内容有效性的检测是目前的研究重点,最初工作可以追溯到Jindal等[7]对亚马逊网站2.14×10^6位用户编写的5.8×10^6条英文评论中无效评论检测的研究,给出了无效评论的定义,从评论内容出发把无效评论划分为不真实评论、仅针对品牌的评论以及无关评论3种类型,通过重复评论的检测来识别不真实评论,并建立分类模型用于判别仅针对品牌的评论及无关评论.由于语言的差异性,这种基于英文评论的有效性检测结果难以适用于在线中文评论的处理.

本文研究了单一主题环境下中文在线评论有效性的检测问题.首先,结合中文评论特点,提取9个特征构建了分类模型;然后,针对ICTCLAS中文分词系统内置名称模式在单一主题中文评论环境下识别主题词准确度不高的问题,提出了一种具有更高精度的基于关联规则的评论名词模式,并采用交叉语言模型来判断评论名词与主题的相关度;最后,利用逻辑回归分类模型来检测中文在线评论的有效性.实验结果表明,该模型在中文在线评论的有效性检测中能够得到较高的检测准确率.

1 评论有效性检测分类模型

1.1 逻辑回归分类模型

评论有效性检测是一种典型的二值分类问题,通常利用分类模型进行研究.分类模型是通过对已知类别数据集进行学习,构造分类器来预测新数据的类别.数据集由特征值和类别组成,单条数据格式的表达式为$\{f_1,f_2,\cdots,f_n;y\}$,其中,f_j为特征值,y为类别.分类器可以采用逻辑回归分类模型或者支持向量机分类模型来构造.

逻辑回归分类模型可以描述为

$$h_\theta(x)=\frac{1}{1+e^{-\theta^{\mathrm{T}}x}} \tag{1}$$

式中,$h_\theta(x)$为预测值;x为分类模型特征向量;θ为特征向量系数.

逻辑回归分类模型是基于最大似然估计来计算对应特征向量系数的,即

$$P(y=1\mid x)=h_\theta(x) \tag{2}$$

$$P(y=0\mid x)=1-h_\theta(x) \tag{3}$$

由式(2)和(3)可得

$$P(y\mid x)=(h_\theta(x))^y(1-h_\theta(x))^{1-y} \tag{4}$$

最大似然估计为

$$L(\theta)=\prod_{i=1}^{m}P(y^{(i)}\mid x^{(i)})=\prod_{i=1}^{m}(h_\theta(x^{(i)}))^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}} \tag{5}$$

基于逻辑回归分类模型求解的关键是确定特征向量系数.因此,针对中文在线评论有效的检测需要结合中文评论内容的特点来提取相应的特征向量.

1.2 评论内容特征提取

Jindal等[7]在关于英文在线评论有效性的研究中,基于评论、评论者和评论对象提取了36个评论特征,其中包括了针对评论文本内容的7个特征,即评论正向情感词、评论负向情感词、评论与产品特性的相似度、品牌名个数、数字个数、大写个数和由大写构成的单词个数.由于语言的差异性,上述7个特征只有评论正向情感词和评论负向情感词适用于中文评论.评论具有主观性,应包含评论者的情绪;如果评论中没有情感特征,则为客观表述,应被判别为无效评论.针对中文在线评论,本文通过获取评论中的显式情感词[10]与中文情感词库[11]来判别评论的情感特征,从而获取评论正向情感度和评论负向情感度,即评论中包含赞扬产品

的形容词个数与贬低产品的形容词个数.

本文将评论主题相关度作为评论特征,以量化中文在线评论与评论主题之间的相关程度.

常规评论由评论对象和评论者态度构成,应具备一定的长度. 而现实的在线评论网站上,评论中往往只具备评论者态度而无评论对象. 无效评论制造者为吸引人们的注意力,往往会编写较长的评论. 因此,本文采用评论文本长度作为评论特征向量之一,评论文本长度即中文在线评论文本包含的字数.

基于对实际评论的观察和研究发现,中文文本注重采用整齐的、排比的句型,多采用短句,评论者在编写评论时,必定会合理使用标点符号. 而垃圾评论制造者在编写没有主题的评论时,为快速表达出自己的想法和意愿,会产生不使用或滥用标点符号的情况. 因此,本文采用评论标点数量和评论标点符号差异数量作为评论特征向量,其中评论标点数量是指评论中标点符号的总个数,评论标点符号差异数量是指评论中标点符号类型的个数.

同时,本文还引入了 Bhattarai 等[12] 检测博客空间中垃圾评论使用的 3 个特征向量: 评论词重复率(即中文在线评论中重复的中文字出现的比例) 、评论名词率(即评论词性标注之后名词所占的比例) 和评论句子数量(即在线评论文本中句子的个数) .

基于上述分析,针对中文在线垃圾评论检测,本文共提取了 9 个评论内容特征: 评论正向情感度、评论负向情感度、评论主题相关度、评论文本长度、评论标点数量、评论标点符号差异数量、评论词重复率、评论名词率以及评论句子数量. 其中,评论主题相关度的处理最为关键和复杂.

2　评论主题相关度

2.1　评论主题词的提取

评论主题词往往采用名词来表示. 目前,评论主题词获取的常用方式是通过中文分词系统对评论进行分词、词性标注处理,然后提取分词系统中内置名词模式标示的名词. ICTCLAS 是我国最具代表性的中文分词系统,其包含中文分词和词性标注的功能,分词准确率达到 98. 45%. ICTCLAS 考虑了文本的通用性,没有针对评论的特殊处理方式,因此很多评论主题词无法被 ICTCLAS 内置名词模式标示.

以电影《速度与激情 6》影评中的一条评论为例 “这个系列的任何一部,一点剧情都记不住. ” 该评论中出现的主题词包括 “系列”、“一部”和 “剧情”. 通过 ICTCLAS 处理之后,评论显示为: “这个/rz 系列/n 的/ude1 任何/rz 一/m 部/q,/wd 一点/m 剧情/n 都/d 记/v 不/d 住/vi. /wj”. 在这条评论中,评论主题词 “一部”没有被标示出. 考虑到相邻 2 个词性标注的组合关系可以归结为有序关联规则问题,因此,本文采用一种改进的 Apriori 算法以获取评论名词模式,从而在实际中提高中文评论主题词的获取精度.

本文采用关联规则[13] 来寻找具有最小支持度的评论名词模式. 通过关联规则可从大量数据中发现数据项之间的相关关系,其规则形式可以表示为

$$X \rightarrow Y$$

式中,X,Y 为数据集中的非空子集. 支持度是关联规则计算中的一个主要指标,即所计算的关联规则模式必须满足预先设置的最小支持度. 最小支持度计算数学表达式为

$$s = \frac{c}{n} \tag{6}$$

式中,c 为非空子集 X 和 Y 同时出现的次数; n 为数据集中记录总数.

评论名词模式是一种通过相邻词性标注组合关系得到的名词模式. 本文首先采用 ICTCLAS 中文分词系统对评论集进行词性标注,然后采用 Apriori 算法获取评论名词模式. 在 Apriori 算法中,主要采用以下步骤寻找关联规则: ① 生成任意 2 个非空子集(如 X 和 Y) 的并集,若其在所有数据集空间中出现的次数超过用户预先指定的值,则把该并集归类到频繁集中; ② 根据频繁集生成关联规则. 由于评论名词模式主要寻找相邻词性标注的组合关系,采用 Apriori 算法在频繁集生成过程中会产生大量的组合,这将导致 Apriori 算法性能下降. 为此,本文对 Apriori 算法进行了如下修改: ① 将数据集合中各词性标注实现有序排列; ② 频繁集由相邻词性标注构成. 由此便可有效降低频繁集生成规模,从而提高 Apriori 算法效率.

2.2　评论主题相关识别

获取评论主题词后,需要判断这些主题词与评论主题的相关度,可用 Zhai 等[14] 提出的交叉语言模型来判断名词与主题之间的关系. 该模型假定一个文档是由一个目标短语向量和一个资料库短语向量构成的,即

$$\boldsymbol{\theta}_1 = \alpha \boldsymbol{\theta}_{\text{corpus}} + \beta \boldsymbol{\theta}_{\text{query}} \tag{7}$$

式中,$\boldsymbol{\theta}_1$ 为从评论集合中获取的名词向量; $\boldsymbol{\theta}_{\text{corpus}}$ 为资料库名词向量; $\boldsymbol{\theta}_{\text{query}}$ 为与主题相关的名词向量;

α,β 分别为对应于 $\boldsymbol{\theta}_{\text{corpus}}$,$\boldsymbol{\theta}_{\text{query}}$ 的系数,且 $\alpha+\beta=1$.

Zhang 等[15] 利用时间复杂度为 $O(k\log(k))$ 的算法来获取交叉语言模型中的 $\boldsymbol{\theta}$. 交叉语言模型的简单表述为

$$\boldsymbol{r}=\alpha\boldsymbol{p}+\beta\boldsymbol{q} \tag{8}$$

式中,$\boldsymbol{r}$,$\boldsymbol{p}$ 和 $\boldsymbol{q}$ 均为多维向量.

为计算 $\boldsymbol{q}$,先假设 f_i,p_i 分别为 $\boldsymbol{r}$,$\boldsymbol{p}$ 中第 i 个词出现的频度. 计算步骤如下:

① 计算 p_i/f_i,按照从大到小的方式排列,结果为 $f_1/p_1>f_2/p_3>\cdots>f_k/p_k$.

② 寻找满足 $\dfrac{\dfrac{\alpha}{\beta}+\sum_{j=1}^{t}p_j}{\sum_{j=1}^{t}f_j}-\dfrac{p_t}{f_t}>0$ 和 $\dfrac{\dfrac{\alpha}{\beta}+\sum_{j=1}^{t+1}p_j}{\sum_{j=1}^{t+1}f_j}-\dfrac{p_{t+1}}{f_{t+1}}\leqslant 0$ 时的 t.

③ 计算得到

$$q_i=\begin{cases}\dfrac{f_i}{\lambda}-\dfrac{\alpha}{\beta}p_i & 1\leqslant i\leqslant t\\ 0 & i>t\end{cases}$$

其中,$\lambda=\dfrac{\sum_{i=1}^{t}f_i}{1+\dfrac{\alpha}{\beta}\sum_{i=1}^{t}p_i}$.

将计算结果大于 0 的 q_i 组成 $\boldsymbol{\theta}_{\text{query}}$,评论主题相关度可以通过计算单条评论中评论主题词在 $\boldsymbol{\theta}_{\text{query}}$ 中的个数获得.

3 实验结果与分析

本文以电影《速度与激情6》的影评为目标评论对象进行实验. 首先,选取1 000条影评,让5位研究生(编号为S1~S5)对其进行人工标示,将无效评论记为1,有效评论记为0;然后,将标示结果求和后取平均值,当平均值大于0.5时视为无效评论,反之则为有效评论. 5位研究生对1 000条影评做出的有效评论和无效评论统计数量结果见表1.

表1 评论样本人工标示的统计数量结果

标示值	S1	S2	S3	S4	S5
0	711	748	717	718	708
1	289	252	283	282	292

为了获取评论主题相关度和评论情感分析,在将评论文本转换为评论文本特征向量的过程中需要对评论进行预处理. 首先,构建中文情感词库,根据该词库来判断评论文本中形容词的词性. 然后,利用时间复杂度为 $O(k\log(k))$ 的算法来建立评论主题词库,从而获取评论主题相关度. 除评论主题相关度和评论情感分析外,其他特征向量值都能直接从评论文本中获取.

本文采用支持向量机分类模型作为对比,验证基于逻辑回归垃圾评论检测模型的有效性. 同时,为了验证评论名词模式在垃圾评论检测中的效果,将ICTCLAS中文分词系统的内置名词模式作为对比进行实验. 实验共分5次进行,每次对1 000条样本评论进行随机排列,计算时采用开源的数据挖掘工具Weka,并利用基于十折交叉验证法来获取垃圾评论检测模型的准确性. 给定样本评论的有效性检测准确率结果见表2.

表2 样本评论的有效性检测准确率 %

次数	内置名词模式		评论名词模式	
	基于逻辑回归	基于支持向量机	基于逻辑回归	基于支持向量机
1	82.7	80.9	83.6	81.5
2	83.3	80.7	84.1	81.2
3	83.0	80.7	83.5	81.3
4	83.1	80.6	83.0	80.9
5	82.8	81.6	83.5	82.3
均值	82.98	80.90	83.54	81.44

由表2可知,采用本文提出的评论名词模式较采用ICTCLAS中文分词系统的内置名词模式在计算垃圾评论检测模型时具有更高的准确率. 在4种情况的对比试验中,采用评论名词模式下基于逻辑回归的垃圾评论检测模型准确率(83.54%)最高,比支持向量机分类模型计算得到的准确率高2.10%.

4 结语

目前国外学术界关于在线评论有效性的研究对象大都采用英文评论,由于语言的差异性,相关英文在线评论有效性检测的研究成果难以推广到中文评论的有效性检测中. 本文研究了单一主题环境下中文在线评论有效性的检测问题. 结合中文评论特点,从评论文本内容中提取9个特征向量来构建逻辑回归分类模型;针对核心特征向量评论主题相关度的研究过程中,利用一种改进的Apriori算法来获取评论名词模式,从而提高了中文评论主题词的识别精度,并基于交叉语言模型计算评论名词与主题的相关度. 实验结果表明,基于逻辑回归的中文在线评论有效性检测模型在评论有效性检测中表现出较高的检测准确率.

参考文献(References)

[1] 中国互联网络信息中心. 2013年中国网络购物市场研究报告[EB/OL]. (2014-04-21)[2014-10-20]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/dzswbg/201404/

t20140421_46598. htm.

[2] Karkare V Y, Gupta S R. A survey on product evaluation using opinion mining [J]. *International Journal of Computer Science and Applications*, 2013, **6**(2): 306-312.

[3] Sheibani A A. Opinion mining and opinion spam: a literature review focusing on product reviews [C]//2012 *Sixth International Symposium on Telecommunications* (IST). Tehran, Iran, 2012: 1109-1113.

[4] Lim E P, Nguyen V A, Jindal N, et al. Detecting product review spammers using rating behaviors [C]// *Proceedings of the* 19*th ACM International Conference on Information and Knowledge Management*. New York, USA, 2010: 939-948.

[5] Jindal N, Liu B, Lim E P. Finding unusual review patterns using unexpected rules [C]//*Proceedings of the* 19*th ACM International Conference on Information and Knowledge Management*. New York, USA, 2010: 1549-1552.

[6] Mukherjee A, Kumar A, Liu B, et al. Spotting opinion spammers using behavioral footprints [C]//*Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA, 2013: 632-640.

[7] Jindal N, Liu B. Opinion spam and analysis [C]//*Proceedings of the* 2008 *International Conference on Web Search and Data Mining*. New York, USA, 2008: 219-230.

[8] Ott M, Cardie C, Hancock J T. Negative deceptive opinion spam [C]//*North American Chapter of the Association for Computational Linguistics-Human Language Technologies*. Atlanta, Georgia, 2013: 497-501.

[9] Lin Y, Zhu T, Wang X, et al. Towards online review spam detection [C]//*Proceedings of the Companion Publication of the* 23*rd International Conference on World Wide Web Companion*. New York, USA, 2014: 341-342.

[10] Liu B. Sentiment analysis and opinion mining [J]. *Synthesis Lectures on Human Language Technologies*, 2012, **5**(1): 1-167.

[11] **徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造** [J]. **情报学报**,2008,27(2):180-185.
Xu Linhong, Lin Hongfei, Pan Yu, et al. Constructing the affective lexicon ontology [J]. *Journal of the China Society for Scientific and Technical Information*, 2008, **27**(2): 180-185. (in Chinese)

[12] Bhattarai A, Rus V, Dasgupta D. Characterizing comment spam in the blogosphere through content analysis [C]//2009 *IEEE Symposium on Computational Intelligence in Cyber Security*. Nashville, TN, USA, 2009: 37-44.

[13] AL-Zawaidah F H, Jbara Y H, Abu-Zanona M A. An improved algorithm for mining association rules in large databases [J]. *World of Computer Science and Information Technology*, 2011, **1**(7): 311-316.

[14] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval [C]//*Proceedings of the Tenth International Conference on Information and Knowledge Management*. New York, USA, 2001: 403-410.

[15] Zhang Y, Xu W. Fast exact maximum likelihood estimation for mixture of language model [J]. *Information Processing & Management*, 2008, **44**(3): 1076-1085.