软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(3):506−527 [doi: 10.13328/j.cnki.jos.004517]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

用户评论的质量检测与控制研究综述*

林煜明[1,2], 王晓玲[1], 朱 涛[1], 周傲英[1]

[1](上海市高可信计算重点实验室(华东师范大学 软件学院),上海 200062)

[2](广西可信软件重点实验室(桂林电子科技大学),广西 桂林 541004)

通讯作者: 王晓玲, E-mail: xlwang@sei.ecnu.edu.cn

摘 要: 随着网络技术的发展,越来越多用户生成的内容(user-generated content)出现在网络应用中,其中,用户评论富含用户的观点,它们在网络环境中充当越来越重要的角色.据美国 Cone 公司 2011 年的调查报告,64%的用户在购买行为之前会参考已有的用户评论.因此,为用户提供准确、简洁和真实的评论是一个迫切且重要的任务.主要围绕评论质量评估、评论总结和垃圾评论检测这 3 个方面综述了国际上评论质量检测与控制的研究内容、技术和方法的研究进展.在此基础上,展望该领域的发展给出了可能的研究方向.

关键词: 用户评论;质量评估;评论总结;垃圾评论检测

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 林煜明,王晓玲,朱涛,周傲英.用户评论的质量检测与控制研究综述.软件学报,2014,25(3):506−527. http://www.jos.org.cn/1000-9825/4517.htm

英文引用格式: Lin YM, Wang XL, Zhu T, Zhou AY. Survey on quality evaluation and control of online reviews. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):506−527 (in Chinese). http://www.jos.org.cn/1000-9825/4517.htm

Survey on Quality Evaluation and Control of Online Reviews

LIN Yu-Ming[1,2], WANG Xiao-Ling[1], ZHU Tao[1], ZHOU Ao-Ying[1]

[1](Shanghai Key Laboratory of Trustworthy Computing (Software Engineering Institute, East China Normal University), Shanghai 200062, China)

[2](Guangxi Key Laboratory of Trusted Software (Guilin University of Electronic Technology), Guilin 541004, China)

Corresponding author: WANG Xiao-Ling, E-mail: xlwang@sei.ecnu.edu.cn

Abstract: With the development of Web2.0, more and more user-generated content (UGC) occur in Web applications. These contents, especially reviews, are opinion-rich and play important roles in e-commerce. According to the Cone's survey published in 2011, 64% of users like to read the related reviews on goods before they make purchase decisions. It is vital to provide users the accurate, succinct and true reviews. This work surveys the research progresses on quality evaluation and control of reviews focusing on the prediction, summarization and spam review detection. At the end, some potential research topics are pointed out based on these analyses.

Key words: review quality; evaluation; summarization; review spam detection

Web 2.0 技术的发展与普及,使网络用户的沟通方式发生了明显变化,越来越多的用户喜欢通过网络论坛、博客、微博、社交网站等网络平台浏览、发布和转发消息,以此与其他用户进行交流,分享各自的体验和观点.同时,网络社交技术的发展和用户交流模式的转变,对电子商务领域也产生了巨大的影响:一方面,良好的网络环境为电子商务的开展提供了更广阔的发展空间和更便利的途径.根据 2013 年 7 月中国互联网络信息中心

* 基金项目: 国家自然科学基金(61170085, 61033007); 国家重点基础研究发展计划(973)(2010CB328106); 教育部新世纪优秀人才计划(NCET-10-0388); 广西自然科学基金(2013GXNSFBA019267); 广西可信软件重点实验室课题(kx201314)

收稿时间: 2013-01-22; 修改时间: 2013-10-11; 定稿时间: 2013-10-31; jos 在线出版时间: 2013-11-28

CNKI 网络优先出版: 2013-11-28 14:40, http://www.cnki.net/kcms/detail/11.2560.TP.20131128.1440.002.html

© 中国科学院软件研究所 http://www.jos.org.cn

(China Internet Network Information Center,简称 CNNIC)发布的第 32 次中国互联网络发展状况统计调查报告(http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201307/t20130717_40664.htm),截至 2013 年 6 月底,国内网络购物用户的规模已经达到 2.7 亿,较 2012 年底增长了 11.9%;另一方面,由于用户产生的内容富含个人的观点,对于顾客和企业等多方面都具有重大的参考价值.潜在客户通过阅读商品的评论了解产品或服务的性能、质量和用户体验等信息,可以辅助他们判定目标商品是否满足其需求,以便做出正确的购买决策.2011 年,美国 Cone 公司的调查(http://www.conecomm.com/contentmgr/showdetails.php/id/4008)指出:64%的用户会通过阅读商品的相关评论了解商品信息,87%的用户在阅读了肯定的评论后做出了购买的决定,而 80%的用户在阅读了否定的评论后放弃了购买的意向.从生产商或商家的角度看,评论作为网络上口碑(word of mouth,简称 WOM)传递的主要途径之一,可以辅助他们了解产品或服务存在的不足,由此进行针对性的改进,提升用户的满意度、企业的服务质量和品牌形象,从而提高产品的销量.国内外的大量研究表明,用户评论对商品的在线销售具有重大的影响作用[1−6].

用户生成的评论具有数据量大、针对性强、主观性高、规范性低、更新快等特点,因此,向评论的阅读者和下一级的服务应用提供快速、准确、简洁的评论具有重要的应用价值和研究意义.从评论的输入开始到向用户显示评论这一过程,可以通过不同的手段和技术控制评论的质量,主要包括 5 种途径:输入约束或政策激励、评论质量评估、垃圾评论检测、评论总结和评论排序显示.图 1 中自下而上描述了这个过程中评论质量控制的技术框架,其中,虚线部分指对应的内容主要处于研究的初始阶段,在实际应用中还较少.由评论人员按照预先定义的约束发表评论后,可对评论集中的垃圾评论过滤,然后按特定的排序算法进行排序显示,也可以对过滤后的评论集进行评论质量评估和评论总结,再进行排序显示.评论质量的检测和评论归纳可在用户生成的评论上直接进行,也可以在垃圾评论检测之后完成.经过处理后的评论最后直接向潜在顾客、商家或制造商显示,以便他们针对各自的需求了解商品的信息和顾客的观点,还可为其他基于评论的服务应用提供高质量的分析数据,例如推荐系统、电子商务/政务的智能系统等.

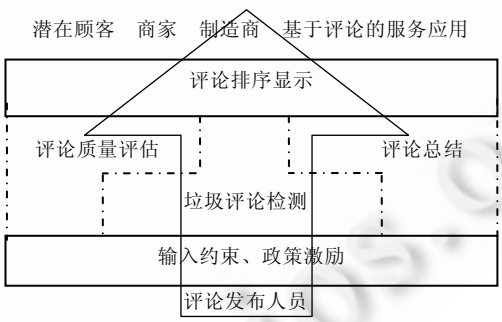


Fig.1 The framework of review quality evaluation and control

图 1 评论质量评估与控制的框架图

目前,大多数电商网站对于评论的处理还处于初级阶段,通常只是采用约束或激励控制用户产生评论的质量,以及允许用户选择评论的排序显示方法,这些排序方法虽然简单和直观,但效果往往不够理想.例如:按时间排序可能会导致用户错失很多高质量的评论以及用户容易受到虚假评论的误导;按评论有用性得票率排序会存在得票不公平性的问题等.对于用户评论质量的检测与控制技术,国内外的研究领域主要集中在评论质量评估[5,7−18]、评论总结[19−30]和垃圾评论检测[31−43]这 3 个部分:

(1) 评论质量(review quality)也称为评论的有用性,其评估主要是在统一的范围内量化评论的质量或衡量评论对用户的有用程度;

(2) 评论总结(review summarization)是对评论中包含的信息进行归纳和汇总,由于评论数据量大的原因,通常用户无法阅读所有评论,经过评论总结以简洁的方式向用户展示原评论集中的内容,方便用户

© 中国科学院软件研究所 http://www.jos.org.cn

掌握所有评论中包含的主要信息;

(3) 垃圾评论(review spam)检测是识别出商家或用户恶意发布的虚假评论或垃圾评论的发布者(spammer),尽可能降低这些评论(人)对其他用户或商家造成的不良影响.值得注意的是,用户书写的评论质量受作者的文化背景、当时的情绪等多方面因素的影响.本文中,垃圾评论是指并非反应评论人自身对商品体验的评论,因此,质量低的评论不一定就是垃圾评论;反之,垃圾评论的发布者通常尽量保证评论的质量,以此提高垃圾评论的影响力.

本文以评论质量检测与控制的框架为主线,综述了评论质量检测与控制在评论质量评估、评论总结和垃圾评论检测等方面的研究进展.本文第 1 节介绍预备知识.第 2 节~第 4 节分别介绍评论质量检测与控制包括的主要内容及其分析对象、采用的方法和技术、评价体系.第 5 节介绍目前在这一研究领域中常用的数据集.第 6 节总结全文,并对该领域在研究和应用上的发展趋势进行展望.

1 预备知识

网络上用户评论的数量非常巨大,采用人工的方法对评论逐条进行质量的评估分析不仅需要耗费大量的时间和费用,而且容易出错.目前,机器学习的方法是进行这类任务的主要途径之一.本节围绕用户评论的质量评估与控制问题中的垃圾评论检测、评论质量评估和评论总结等方面共有的基础知识进行介绍.

使用机器学习的方法来解决用户评论的质量评估和控制问题时,通常采用词袋(bag of words)的框架来描述评论.在此框架中,用户评论被表示为向量的形式,如图 2 所示,其中,每个评论被表示为一个 m 维向量,每个分量表示某个特征在评论中对应的值,它由特征函数 $\varphi:F'\to\mathbb{R}$ 确定,其中,$F'=\{f_1,f_2,\ldots,f_m\}$ 为预先确定的特征集.例如在进行评论的质量评估时,文献[14]认为,评论的长度可以从某个方面上反映评论的质量,因为评论长度越长则包含更多信息量的概率就越大,因此其质量就可能越高.在这种情况下,某个特征 $i(1\leqslant i\leqslant m)$ 可以用来表示一个评论的长度,而特征函数就是从评论到长度的一个映射.

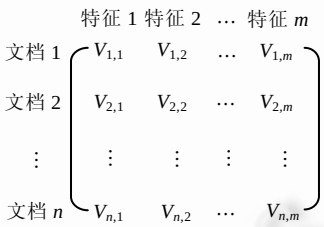


Fig.2 The diagram of review vectorization

图 2 评论向量化的示意图

评论进行向量化的时候,确定特征集是关键的步骤,这包含两个因素:特征类型和特征函数.特征类型是指采用什么样的特征,它进一步可分为通用型的特征和特定型的特征.通用型的特征类似一元词项(unigram)和二元词项(bigram)等,这类特征在评论进行向量化的时候需要通过预先设定特征函数来确定它们的特征值,例如 TF-IDF(term frequency-inverse document frequency)、特征在文档中是否出现等.特定型的特征主要是指通过对当前问题的分析而设定的特征.例如评论质量评估时评论的长度、在垃圾评论检测时评论的评分与平均评分的偏差、在评论总结时句子在评论中的位置,等等.常用的特征选择方法包括:

(1) 文档频率阈值

计算每个词项(term)在文档中出现的次数,将超过阈值的词项作为特征.这种方法的一个基本假设是:稀疏的词项不能为分类提供有用的信息,而且对整体的分类性能不产生影响.在此假设下,文档频率阈值的方法能够降低特征空间的维度,并且如果这些稀疏的词项恰好是噪音的时候,那么也可能提高分类的准确度.

© 中国科学院软件研究所 http://www.jos.org.cn

(2) 信息增益(information gain)

该方法主要是通过评估词项能够给分类带来多少的信息量[44],带来的信息量越大,说明该词项越重要.假设集合$\{C_1,C_2,\ldots,C_k\}$表示类型的集合,词项 t 的信息增益定义为

$$G(t)=-\sum_{i=1}^{k}\Pr(C_i)\log\Pr(C_i)+\Pr(t)\sum_{i=1}^{k}\Pr(C_i\mid t)\log\Pr(C_i\mid t)+\Pr(\bar{t})\sum_{i=1}^{k}\Pr(C_i\mid\bar{t})\log\Pr(C_i\mid\bar{t}),$$

其中,$\Pr(X)$为 X 出现的概率.

(3) 互信息(mutual information)

互信息可用来度量词项与类型标签间关系的紧密程度,从而确定哪些词项可以作为特征,如文献[45,46].例如,将评论的质量评估作为一个二分类问题(质量高/质量低)的时候,可通过下式计算每个词项 t 和两种类型的关系:

$$I(t,l)=\log\frac{\Pr(t\wedge l)}{\Pr(t)\times\Pr(l)}\approx\log\frac{A\times N}{(A+C)\times(A+B)},$$

其中,$l\in\{$质量高,质量低$\}$,A 为 t 和 l 共同出现的次数,B 为 t 出现但 l 不出现的次数,C 为 l 出现但 t 不出现的次数,N 为评论的总数.最后,通过词项和标签的平均互信息或者最大互信息作为它的分值:

$$I_{avg}(t)=\Pr(l_1)I(t,l_1)+\Pr(l_2)I(t,l_2),I_{\max}(t)=\max\{I(t,l_1),I(t,l_2)\}.$$

(4) χ^2 统计(χ^2 statistic)

对于分类问题,这种方法主要考察词项 t 和类型 l 间的独立性[47].若独立,则说明 t 对文档是否划分为类型 l 缺乏判断力.对于上述方法(3)的例子,词项 t 和类型 l 的 χ^2 值可通过下式计算:

$$\chi^2(t,l)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)},$$

其中,N 为评论总数,A 为 t 和 l 共同出现的次数,B 为 t 出现但 l 不出现的次数,C 为 l 出现但 t 不出现的次数,D 为不包含 t 且不属于类型 l 的评论数.

词项 t 的最终 χ^2 值也可以通过类似方法(3)中的平均 χ^2 值或最大 χ^2 值来确定.

(5) 词项强度(term strength)

文献[48]采用这种方法通过词项 t 在紧密相关的文档中出现的概率来衡量它的重要性.假设 d_i 和 d_j 是任意一对相关的文档,则词项 t 的强度定义为

$$S(t)=\Pr(t\in d_i\mid t\in d_j),$$

其中,紧密相关的文档通常是指在将文档聚类之后属于同一个类中的文档.

对于上述 5 种特征选择方法,其中,信息增益、互信息和 χ^2 统计需要使用到文档类型的信息.文献[47]针对文本分类问题,在 kNN(k nearest neighbor)和 LLSF(linear least squarest fit)两种分类方法上对比了上述 5 种不同特征选择方法的效果,在他们的数据集中,χ^2 统计和信息增益效果最佳,词项频率阈值和前两者性能上大体相当,词项强度效果稍差,而互信息的最差.

在评论质量的评估与控制研究中,通常更偏向于特定型的特征,因为这样的特征直观上与具体的任务具有更直观的关系,但是通用型的特征可以作为一个有效的补充.在确定了采用的特征基础上对评论进行向量化之后,很多问题都可以通过回归、分类等机器学习的方法来解决.例如:预测一个评估的质量时,可以通过回归模型得到该评论的质量分值,也可以预测评论是属于“质量高 C_h”的类型还是属于“质量低 C_l”的类型.这些相关的技术和模型,我们在后面的内容中将进行详细的分析和讨论.

2 评论质量检测与控制的内容

为用户提供简洁、全面和准确的评论,可以从输入约束或政策激励、评论质量评估、评论总结、垃圾评论检测以及评论排序显示几个方面入手.输入约束或政策激励直接与评论人交互,主要通过引导和约束他们的行为控制评论的质量;评论质量评估和评论总结可在垃圾评论检测的基础上实施;经过评论排序后的结果直接向

© 中国科学院软件研究所 http://www.jos.org.cn

评论阅读者或其他基于评论的服务应用提供.

2.1 输入约束或政策激励

输入约束或政策激励作用在用户评论的生成阶段,提供评论功能的网站通过限制用户的一些行为或者影响用户的心理,使用户更倾向于发表真实的、高质量的评论.

很多电子商务网站只允许用户对已购买商品发表评论,这从某种程度上保证评论是用户对相应商品的真实体验,另一方面,可以提高恶意发布垃圾评论的成本,因此,在一定程度上能够减少低质量评论和垃圾评论的数量;但是,也可能限制了潜在的高质量评论的发表.大多数评论网站对用户评论中包含的字数进行限制,字数过少不能全面描述商品的特征和表达用户的观点,字数过多则产生冗余或无用的信息,给阅读和书写带来不便.

大部分网站允许评论的阅读者对某评论对其是否有用进行投票,将有用性得票率(有用性票数/总的投票数)作为衡量评论用户贡献的主要标准之一,对于贡献大的评论用户给予一定物质或精神上的奖励,例如显示前 k 个有用性投票率最高的评论用户,给予不同贡献的用户授予不同等级的名誉称号或物质性的奖励,在对应商品所有评论的前端强调该商品的高质量评论等都可以给予用户心理上的暗示和激励.用户还可以针对垃圾评论向网站管理方进行举报,以及可对已有评论发表自己的观点.有用性投票和举报这种聚集了集体智慧的结果不仅有助于网站维护评论的质量,还可在研究中作为潜在的验证基准(ground truth).此外,向用户提供评论书写的注意事项来指导他们发表高质量的评论.表 1 给出了国内外一些著名的电子商务网站在输入约束与激励方面采用的措施.

Table 1 The ways of review quality control on some well-known e-commerce sites

表 1 部分著名电子商务网站常用的评论质量控制方法

网站	购买评论	长度要求	有用性投票	举报评论	用户荣誉	书写指南
Amazon	√	长度≥20 单词	√	√		√
eBay		100≤长度≤3500 字符	√	√	√	√
TripAdvisor		长度≥200 字符	√	√	√	
IMDB		10 行≤长度≤1000 个单词	√	√	√	√
淘宝	√	0≤长度≤500 字	√		√	√
大众点评网		50 字≤长度≤2000 字	√	√	√	√
豆瓣		0≤长度≤140 字	√			√
当当网	√	0≤长度≤3000 字	√		√	√

表 1 中的第 1 列从上到下网站的网址分别为 www.amazon.com, www.ebay.com, www.tripadvisor.com, www.imdb.com, www.taobao.com, www.dianping.com, www.douban.com, www.dangdang.com.除 IMDB 外,这些网站的访问时间为 2012 年 12 月.IMDB 的访问时间为 2013 年 12 月.

2.2 评论质量评估

评论质量是一个主观的概念,不同的人对评论质量高低的衡量标准也不一样.文献[7]针对书的评论展开研究,认为好的(有用的)书评应当就书的主题/情节/写作风格/写作背景等方面向潜在读者提供充分的信息,以便他们做出适当的阅读或购买决定.本文不局限于书的评论而是围绕各种商品的评论,因此将高质量的/有用的/好的评论定义为:能具体描述商品的特征/性能等信息,辅助潜在用户做出适当决策的评论.

评论质量评估的目的是在一定范围内量化评论的质量或者根据质量对评论进行分类,识别出高质量的评论.在此基础上,可以对评论进行过滤、排序等操作.根据用户对评论的有用性投票进行统计,是目前大多数网站衡量评论质量的主要途径,但这种评估方法存在得票不公平的情况,这种不公平可能来源于用户,也可能来源于评论本身.若一个用户对某品牌具有偏见,则这种偏见就可能反映在用户对相关评论的投票中,因此在这种情况中,以用户的投票衡量评论的质量时,应当对有偏用户的投票给予较低的权值,从而降低这类用户的影响力[49].对于来自评论的不公平,文献[8]将其归纳为 3 类:

(1) 用户更倾向于投有用的票;

(2) 得到有用性投票多的评论容易得到更多的票;

(3) 发布早的评论容易得到更多的有用性投票.

© 中国科学院软件研究所 http://www.jos.org.cn

评论的质量受多种因素(特征)的影响[9−16],根据这些因素,可对评论的质量进行打分或排序.从评论内容的角度看,影响评论质量的因素主要包括:

- 语法特征:主要考虑评论中各种词性(part-of-speech)的词的数量或比例、是否包含情态动词、感叹词、比较级(最高级)的形容词或副词、疑问词等;
- 语义特征:表示主观性或客观性词、句子的数量、评论中包含的肯定情感/否定情感的词或句子、评论涉及的商品特征数等;
- 评论的元特征:评论中用户的评分、评论对应商品的平均评分、用户评分与该商品平均评分间的差值、评论的发表时间、有用性得票率等;
- 文本的统计特征:评论中包含词、句子、段落的数量、大小写字母的比例、超链接的数量等;
- 可读性:可通过拼写错误数、评论的长度、句子的平均长度等来度量,也可以借助于各种可读性指标[50],例如 ARI 指标(automated readability index)[51]、SMOG 分值(simple measure of gobbledygook score)[52]、Gunning-Fog 指标[53]等;
- 相似性特征:这类特征主要比较当前评论与其他评论之间的相似程度,常用的度量指标有 cosine 相似度、KL(Kullback-Leibler divergence)距离等.

从评论人的角度看,影响评论质量的因素主要有:

- 平均评分.该评论人所有评论的平均评分;
- 评论人是否登记了真实姓名、是否有 Top 徽章、评论数量、所有评论人的平均评论数量与标准差;
- 平均有用性得票率及其与所有评论人平均得票率的差值.该评论人所有评论的平均有用性得票率,该值与所有评论人平均得票率的差值可以反映该评论人的评分习惯;
- 社交特征.如果评论人之间有社交网络关系,那么可以通过图对评论人之间的关系进行建模,此类特征可包括评论人结点的入度、出度和 PageRank 值[54]等;
- 评论人的经验知识.商品的评论通常受评论人的经验阅历影响,对不同商品的认识度也有差异.例如,相对于《美国思潮》这类电影,喜欢科幻小说的人可能更喜欢/擅长写《矩阵革命》和《星球大战》这类电影的评论.因此,在设计评论质量的预测模型时应当考虑这样的因素.

对于不同类型的商品,影响评论质量的因素也有所不同.Nelson 将商品分为两类:体验性商品(experiment goods)和搜索商品(search goods)[55].体验性商品是指必须经过试用或购买之后才能获取质量信息的商品,例如电影、音乐等.搜索商品指顾客在购买之前就可以收集到关于质量信息的商品.文献[17]指出:对于体验性商品,中等评分的评论相对评分在两个极端的评论更有用;评论的深度对搜索商品的评论比对体验性商品的评论影响更大.

在这些因素的基础上,文献[8,13,56,57]采用分类的方法预测评论质量,他们都将评论质量的预测作为二分类问题,即预测评论的质量高/低(或者有用/无用、推荐/不推荐).分为两类的主要优点在于:可以获得更准确的标注数据和对比基准(ground truth),因为随着评论质量类别的增加,相邻两类间的界线变得越来越模糊,不同的用户对质量也有不同的评价标准,在这种情况下,预测的效果不能有效地评估;另一方面,划分粒度过粗会导致在实际对评论应用的精准度下降,可能难以满足用户的需求.上述 4 篇文献中:文献[8,56,57]采用的是硬分类,即硬性确定评论的质量是高还是低;但文献[13]采用的是带有置信度的分类,即给出属于每个类的概率,通过这种途径的好处在于,用户可以对同一类的评论进行排序,根据具体的需求选取不同数量的评论,例如向读者推荐前 k 个质量高的评论.

文献[9−12,14−16]使用回归的方法评估评论的质量,这种方法的优势在于:能够直观地反映评论质量变化和各种特征变化之间的关系,而且由于回归的方法输出连续型的质量分值,用户一方面可以根据该值对评论的质量排序,另一方面通过设置阈值的方式也可以实现分类的效果,如将质量分值大于 0.5 的作为有用的评论[11].此外,还可以使用多重回归模型对评论的质量进行建模.文献[11]中,在最上层通过评论人的专业知识、写作风格和发表时间这 3 个因素的线性回归模型预测评论质量,而对于前两种因素,分别通过多重径向基函数(radial

© 中国科学院软件研究所 http://www.jos.org.cn

basis function,简称 RBF)进行拟合,使用指数回归模型对评论质量与发表时间之间的关系进行建模.这样的方法通过多重不同类型的函数对不规则的目标函数曲线具有很好的拟合效果.回归方法的主要缺点是:确定哪些因素会影响评论质量以及这些因素的建模需要预先进行推测,但由于实际中这些因素具有多样性和不可测性,在一定程度上限制了回归方法的使用.

除了分类和回归的方法之外,文献[7]则将待评估评论与一个虚拟最优评论之间的相似度作为质量的衡量指标,其中,虚拟的最优评论在一个外部通用文集的基础上结合特定产品评论集中每个词项的词频确定.这种无监督的方法避免了人工标注的过程,避免引入人在标注时的主观性而引起的标注不一致问题.文献[10]首先在文本特征的基础上使用线性回归预测评论的质量,然后进一步利用评论人的社交网络信息来提高评论质量的评估准确度.第 3 节将进一步分析评论质量预测中使用的技术和模型.

2.3 评论总结

一些商品特别是热门商品通常评论的数量众多,用户无法阅读所有的评论.评论总结是为方便用户有效地获取商品信息,对当前的所有评论进行概述而形成的(新的)评论或评论集.将用户评论总结后再向用户显示,可减少用户的评论阅读量而又不丢失评论中包含的重要信息.

用户评论是一种短文本,传统的文本总结技术主要分两类:抽取总结(extractive summarization)和抽象总结(abstractive summarization)[58].抽取总结是通过在原文本中选取一些重要的句子、段落等拼接形成一个较短的文本,如文献[59,60].这种方法虽然简单直接,但是对于多文档总结时可能出现冗余度过大或偏向部分文档的现象.抽象总结需要从文档收集到的信息中产生新的句子进行总结.这种技术的难度较大,目前采用的方法是预先定义一个总结的模板,然后将原文本中最重要的信息抽取和统计填充到模板中生成一个新的文本,如文献[61,62].

评论的观点分析(观点识别、情感分类)[63−69]可看作为一种粗粒度的评论总结,它的目标在于识别评论的总体观点倾向,例如是肯定的(positive)还是否定的(negative).Pang 等人[63]将观点识别问题看作二分类问题,并首次使用机器学习的方法识别 IMDB 上电影评论的观点,其中,unigram、特征出现与否的信息和 SVM(support vector machine)的搭配达到最高的分类准确度.这种有监督的学习方法的局限在于:需要费时地人工标注数据,跨领域的观点识别[64−66]将一个领域中标注的评论知识应用到另一个领域中的学习中,从而降低了人工标注数据的工作量.此外,Dasgupta 等人[67]和 Turney 等人[68]分别采用半监督和无监督的学习方法识别评论的观点:前者首先通过谱聚类(spectral clustering)的方法标注观点明确的评论,然后,采用一个主动学习器挑选少量观点不明确的评论由人标注,最后,将所有标注样本训练多个分类器,并以此对未标注的样本进行预测;后者在一些观点鲜明的种子词基础上,通过一个外部的搜索引擎计算评论中的观点词与种子词的点互信息(pointwise mutual information,简称 PMI)值评估观点词和评论的观点类型,这种方法虽然不需要标注样本,但其效果依赖于种子词以及外部的工具,在一定程度上影响了它的效果和使用.文献[69]在不同的数据集上比较了传统的 tf*idf 及其变体在观点分析中的效果,而 Delta tfidf[70]和 tf*MI[71]是专门针对观点分析提出的两种特征加权的机制,两者的特征函数中都整合了词项的情感信息,但是 tf*MI 进一步考虑了词项在文档中的分布情况,因此在他们的实验中具有更好的效果.文献[72]中对观点分析进行了较全面的综述.

评论总体的观点倾向从评论的粒度上反映用户的观点,不能够描述评论中商品的特征,而潜在顾客通常希望能够了解商品的主要特征.因此,围绕着商品特征(如尺寸、电池、重量等)进行的总结更符合用户的要求.基于商品特征的评论总结系统输入某种商品的所有评论,输出该商品每个特征的汇总评分以及支持评分的文本信息.这类系统的实现通常包括 3 个步骤:

(1) 抽取评论中提及的商品特征;

(2) 识别评论人对商品特征的观点;

(3) 对商品特征的观点进行汇总和可视化展示.

文献[20,24]通过频繁项挖掘的方法从评论中抽取候选的商品特征,继而通过剪枝删除冗余的选项.这种方法能够找出经常在评论中出现的产品特征,但会遗失非频繁的产品特征.对于这类非热门讨论的特征,文献[73]

© 中国科学院软件研究所 http://www.jos.org.cn

使用含有观点词和评价对象的模板去挖掘产品的特征.这种方法的主要优点是具有针对性,可以根据具体的应用或问题制定对应的模板或规则.但是基于模板或规则的方法可扩展性较差,而且人工耗费较大.文献[25]通过评估部分关系指示词(例如,of scanner)和与该指示词相关的名词短语间的点互信息值来获取产品特征,这种方法的产品特征识别效果甚至超过了基于模板和规则的方法,但难点在于部分关系指示词的获取.基于这些识别出的产品特征,文献[20]抽取出描述特征的形容词,利用 WordNet[74]判断这些形容词是种子词(一些情感明确的形容词,例如 perfect,poor 等)的同义词还是反义词,以此确定用户对该特征的观点;最后,针对商品特征根据持肯定观点的用户数和持否定观点的用户数进行用户的观点汇总.文献[26]采用分类的方法预测所有评论对每种商品特征的评分,抽取一些代表性的短语辅助用户理解对每种特征的评分.这些基于产品特征的总结方法,可以通过图形化的方式对产品的每种特征进行描述,具有表述更直观的优点.

基于商品特征的评论总结,通过在商品每个特征上的评分概述用户的观点,这属于抽象总结的范畴.从抽取总结的角度出发,文献[27]从评论集中选取 k 个使目标函数值最大的句子形成一个总结,其中,目标函数从总结的情感强度、评分差异以及商品特征的覆盖率这 3 个方面考虑.对于这种从评论中选择若干个句子作为总结的方法,它们的缺点在于总结的结果缺乏人为书写评论的直观和叙述性结构.因此,文献[21]转而采用从评论集中选出一组满足目标函数的评论作为总结,它通过要求选择作为总结的评论从每种观点类型上能覆盖尽可能多的产品特征,从而将问题转变成图的最大覆盖问题;文献[28]在文献[21]的基础上进一步考虑了每类观点在总结评论集中的比例问题,使得选择出来作为总结的评论集在观点类型上的分布更接近源评论集.

2.4 垃圾评论检测

垃圾评论是指在某些利益的驱动下,一些商家或用户恶意发布的虚假评论,以此误导潜在客户.反垃圾评论技术的目标在于识别那些误导消费者的与事实不符的评论或者这类评论的发布者.垃圾评论的发布者为了隐藏自己的身份并能够达到误导用户的目的,通常会使自己的评论写得与正常评论一样.因此很多情况下,正常的用户也无法识别一个评论是不是垃圾评论,从而造成标注数据不足和难以评估检测效果的困境,这也是垃圾评论检测研究面临的挑战之一.

垃圾评论的出现可以是主动的也可以是被动的,例如,某商家/品牌一方面为了提高自家商品在网络上的声望,亲自/雇人为自己发布虚假的肯定评论;另一方面,又可以向竞争对手发布否定评论以降低对手的声望而从中获益.这些虚假的评论严重影响了网络在线市场中的正常竞争,而且损害了消费者和商家的权益.因此,在向潜在消费者显示评论时应当将这类评论过滤.从检测的对象看,反垃圾评论技术的检测对象主要包括如下 3 类.

(1) 检测垃圾评论

目前,大多数的研究[32−38]主要集中在垃圾评论的识别技术,文献[32]将垃圾评论概括为 3 种:① 虚假的肯定(否定)的评论;② 只是谈论商品的牌子而不关注商品本身的评论;③ 不包含任何观点的评论(例如广告).这 3 种垃圾评论中:第①种的危害性是最大的,也是最难识别的;后两种垃圾评论相对而言危害性较小而且较容易识别.反垃圾评论技术通常分析的特征包括:

- 评论的特征.这类特征又可细分为:
  - 文本特征,例如品牌名字出现次数的频率、数字的频率、大写字母的频率、第一人称或第二人称的出现频率、该评论与其他评论(或商家提供的商品描述)的相似度等;
  - 情感特征,包括主观或客观词(句)的比例、肯定或否定词(句)的比例等;
  - 元数据特征,例如评论长度、评论的评分、得到的投票总数、有用性的得票率、是否为该商品第一个或唯一的评论、评论发布的时间等;
- 评论人的特征.这类特征描述评论人的个人信息,例如书写的评论数、是否包含真实姓名(个人主页或个人描述)、在评论站点上的排名;行为特征包括该评论人发布的评论中是第 1 个评论的比例、评论人的平均评分、评分的标准差、是否总是发布同类观点(肯定或否定)的评论、发表第 1 个评论次数的比例、不同品牌的评论数之间的差异、发表不同品牌的商品评论中的评分差异等;
- 商品的特征.这类特征包括该商品的评论数、平均评分、该商品评分的标准差、商品的价格、商品的

© 中国科学院软件研究所 http://www.jos.org.cn

销售排名、商品(或品牌)在一个评论中被提到的次数等.

(2) 检测垃圾评论的发布者

相对于垃圾评论的检测,垃圾评论发布者的识别更加困难,因为这些发布者可能发布垃圾评论的同时也发布正常的评论.对这类发布者的识别可通过对用户行为的建模来识别.文献[40]定义了 4 种垃圾评论的发布行为:针对商品的行为、针对商品组的行为、一般的评分偏差和早期的评分偏差.这 4 种行为的识别通过评论的文本内容和(或)评分的差异性来捕获.文献[39]在社交平台上注册一些特定的帐号来收集垃圾评论发布者的个人信息,在此基础上,训练分类器以此识别垃圾评论的发布者.

(3) 检测垃圾评论的发布团体

垃圾评论的发布团体是指针对目标商品一起写虚假评论来提升或降低目标商品声望的一组评论人[41].文献[42]中,首先通过频繁模式挖掘找出候选的团队,然后在定义的 8 个指标(团体内工作的时间窗口、团体内评分的差异度、团体内评论的相似度、成员自身的评论相似度、团体内最早的评论发布时间、成员数与商品评论总数的平均比例、成员规模、团体的攻击目标数支持度)基础上对候选垃圾评论发布团队进行排序.

2.5 评论排序显示

通常,用户无法浏览所有的评论信息,评论的排序非常重要,按照不同的条件将评论排序显示,可以满足用户的不同需要.排序的对象可以是经过评论质量评估、垃圾评论检测或评论总结后的结果,也可以是原始输入的用户评论.目前,大部分网站的评论都是直接在原始输入的用户评论上进行排序显示;淘宝可以按照“推荐”对评论排序,但无法了解其具体的推荐算法.表 2 给出了一些主流电子商务网站的排序策略和显示内容.

Table 2 The sort modes and display contents on some well-known e-commerce sites

表 2 部分主要电子商务网站常用的评论排序方法和显示内容

	按时间排序	按有用性投票排序	其他排序方式
Amazon	√	√	
eBay	√	√	相关度
TripAdvisor	√		评分
IMDB	√	√	用户等级、评分
淘宝	√	√	用户等级、评分、推荐
大众点评网	√	√	用户等级
豆瓣	√	√	
当当网	√	√	评论的回复数

3 评论质量评估与控制的模型和技术

评论质量的评估与控制技术研究主要围绕在评论的质量评估、评论总结以及垃圾评论检测 3 个方面,涵盖了机器学习、文本挖掘、观点分析和自然语言处理等多领域.目前,在这 3 个方向的研究中所采用的方法或模型如图 3 所示.

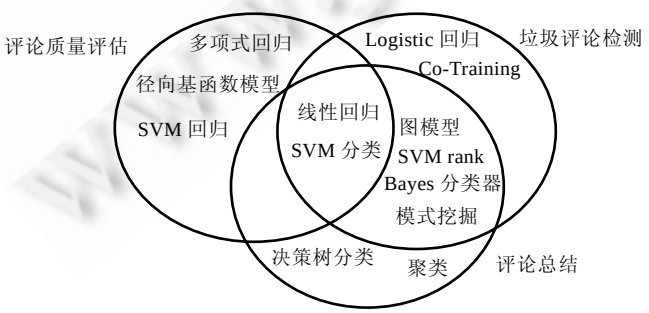


Fig.3 The techniques and models for review quality evaluation and control

图 3 评论质量评估与控制中使用的方法和模型

© 中国科学院软件研究所 http://www.jos.org.cn

为方便叙述,若无特殊说明,则标注的评论训练集 $S=\{(X_1,y_1),(X_2,y_2),\ldots,(X_n,y_n)\}$,其中,$X_i\in R^m(1\leqslant i\leqslant n)$表示根据预先确定的特征对评论进行向量化后的结果,$y_i$ 表示输出值.

3.1 回归分析

回归是指学习一个目标函数将每个 X_i 映射到一个连续的值 y,它可以用于预测输入变量和输出变量之间的关系,特别是当输入变量的值发生变化时,输出变量的值随之发生的变化[75].回归问题的学习等价于函数拟合:在标注数据集上找出一条函数曲线,使其能够以最小的误差来拟合训练的样本,并能很好地预测为标注数据.该模型的一般形式为

$$y=f(X)+\varepsilon,$$

其中,$f(X)$称为回归函数,ε称为随机误差,一般要求 $E(\varepsilon)=0$.该误差通常使用绝对误差 $\sum_{i=1}^{n}|y_i-f(X_i)|$或者平方误差 $\sum_{i=1}^{n}(y_i-f(X_i))^2$.

3.1.1 线性回归模型

线性回归分析中,假定分析对象可以表示为一些影响因素的线性函数.由于考察目标通常受多种因素(特征)的影响,因此最常使用的是多元线性回归模型,其形式为

$$f(X)=\beta_0+\sum_{i=1}^{p}\beta_i x_i+\varepsilon,$$

其中,β_0 为回归常数,β_i 为回归系数,ε为随机误差.

线性回归模型的参数估计常采用最小二乘法、最大似然估计等方法[76].

评论的质量可以看成是质量影响因素的线性组合(如文献[10,17])或者可转化为多元线性回归的模型(如文献[14,15,77]).Wang 和 Liu[29]针对抽取总结问题,通过句子与文本的相似度、句子与讨论主题的相关性、句子的情感以及长度的线性组合评估一个句子作为评论总结的分值.但是在实际应用中,影响因素和输出值之间的线性关系假设在很多时候不能满足.

3.1.2 多项式回归模型

当评论的质量与影响因素间的关系是非线性时,可采用多项式回归分析来预测评论的质量[17],多项式回归的最大优点是可通过增加影响因素的高次项对测试点进行逼近.多项式回归模型的形式为

$$f(X)=\beta_0+\sum_{i=1}^{p}\sum_{j=1}^{q}\beta_j x_i^j+\varepsilon.$$

从参数的角度看,多项式回归函数与未知的参数β为线性关系,因此可将 x_i^j 看作不同的独立变量而采用线性回归中的参数估计技术决定未知的参数.多项式回归的缺点在于:当自变量个数较多或自变量的幂较高时,计算量迅速增加;此外,回归系数间存在相关性,当删除一个变量的时候,必须重新计算出回归系数.

3.1.3 径向基函数模型

径向基函数(radial basis function,简称 RBF)是一个实值函数,它的值仅依赖于输入向量 X 和中心点μ之间的距离,一般形式为

$$RBF\phi(X|\mu,\Sigma)=f((X-\mu)\Sigma^{-1}(X-\mu)),$$

其中,f 是基函数,Σ为衡量各维度重要性程度的正定矩阵,$(X-\mu)\Sigma^{-1}(X-\mu)$为 X 到中心点μ之间的距离.

文献[11]采用多重高斯径向基函数对评论人的专业技能和写作风格分别进行建模,使用指数模型描述时间因素的影响,在这 3 个因素的基础上,设计评论质量的预测模型:

$$\hat{H}=p\sum_{i=1}^{k_1}u_i\phi(x\,|\,\mu_i,\sigma_i)+q\sum_{i=1}^{k_2}v_i\psi(y\,|\,\gamma_i,\xi_i)+r\cdot e^{-\beta(t-t_0)+d},$$

其中,$\hat{H}$ 是评论的有用性分值,p,q 和 r 分别是 3 种因素的权值,x 是专业技能的特征向量,y 是写作风格的特征向量,k_1 和 k_2 是对专业技能和写作风格建模的 RBF 网络的中心数,u_i,μ_i 和 σ_i 分别是专业技能模型中第 i 个 RBF 的

© 中国科学院软件研究所 http://www.jos.org.cn

权值、中心和跨距(spread),v_i,γ_i和ξ_i分别是写作风格模型中第 i 个 RBF 的权值、中心和跨距,t 是评论的发布时间,t_0 是评论的对象(例如商品)的发布时间,β控制有用性分值随时间的退化率.实验表明:3 种因素中,时间因素最具有预测力,评论人的专业技能次之,而写作风格的预测能力最差.

3.1.4 Logistic 回归

Logistic 回归为概率型回归模型,是研究分类观察结果与影响因素之间关系的一种多变量分析方法.假设离散型随机变量 Y 的取值集合是$\{1,2,\ldots,K\}$,影响因素向量 $X=(x_1,x_2,\ldots,x_n,1)$,影响因素的权值向量 $W=(w_1,\ldots,w_n,b)$,b 为截距,那么 Logistic 回归模型为

$$\left.\begin{aligned} P(Y=k\mid X)&=\frac{\exp(W_k^T X)}{1+\sum_{k=1}^{K-1}\exp(W_k^T X)}\\ P(Y=K\mid X)&=\frac{1}{1+\sum_{k=1}^{K-1}\exp(W_k^T X)} \end{aligned}\right\},\ k=1,\ldots,K-1.$$

权值向量 W 中的参数 w_i 和 b 常通过最大似然估计来确定.

Logistic 回归模型输出的概率可以作为评论的权值,在垃圾评论检测中,这个权值越高,表明该评论是垃圾的概率越大[32].因此在后续的处理中,只需要降低它们的影响而不需要删除.Logistic 回归模型中,影响因素变量可以是连续型变量,也可以是离散变量,而且该模型的输出是概率值,这使得它更符合实际应用;但它对训练样本量有一定的要求,如果训练样本过少时,拟合的函数会显得不稳定,系数或标准误差的估计可能会出现一些与预计相差很大的值,使得函数变得无法解释.

3.1.5 ε-SVR(ε-支持向量回归)

假设给定一个标注的评论集$\{(x_1,y_1),\ldots,(x_n,y_n)\}$,$x_i\in R^m$是向量化的评论,$y_i\in R$是对应的目标输出(评论对应的分值),如果$\varepsilon$范围内的偏差都是允许的,那么支持向量回归问题的标准形式就是:

$$\min_{w,b,\zeta,\zeta^*}\frac{1}{2}w^T w+C\sum_{i=1}^{n}(\zeta_i+\zeta^*),$$

$$\text{约束条件:}\begin{cases} w^T\phi(x_i)+b-y_i\leqslant\varepsilon+\zeta_i^*\\ y_i-w^T\phi(x_i)-b\leqslant\varepsilon+\zeta_i^*,\\ \zeta_i,\zeta_i^*\geqslant 0,i=1,\ldots,n \end{cases}$$

其中,w 是权重向量,b 是截距,ζ和ζ^*是松弛变量,C 和ε是给定的参数.

文献[9]在 5 种特征(结构化特征、词法特征、语法特征、语义特征和元数据特征)的基础上采用 SVM 回归预测评论的有用性分值,并且使用 SVM^{light} 对比了线性、多项式和径向基函数这 3 种不同的核函数在评论质量预测上的效果.其中,径向基函数的性能最佳.文献[12]则在语法相似性特征、词法特征和主观性特征的基础上对比了简单的线性回归和 SVM 回归,实验表明,SVM 回归效果更为显著.这主要归因于 SVM 回归模型可采用具有非线性能力的核函数,以及在同样的数据集训练上异常点对它的影响更小.

3.2 分 类

分类的任务是在标注样本集上学习得到一个目标函数 f,该目标函数能够将每个样本向量映射到预先定义的类型标签.与回归问题不同的是:回归模型中输出变量的值是连续型的,而分类模型输出离散型的值.例如,类型 C_h 表示质量高的类型,C_l 表示质量低的类型.

(1) 朴素贝叶斯(Naïve Bayes)分类

朴素贝叶斯分类的假设所有特征都是独立的,它的基本分类规则定义为

$$f(X)=\arg\max_{Y=\{C_h,C_l\}}\left(P(Y)\prod_{j:x_j=1}P(x_j=1\mid Y)\right),$$

其中,x_j表示评论样本 X 的一个分量(质量影响因素),$P(Y)$和$P(x_j|Y)$是在训练数据上计算出来的概率.但通常,特征

© 中国科学院软件研究所 http://www.jos.org.cn

的独立性假设在很多情况下不能保证,所以这种方法往往只是作为实验的对比基线(baseline)[34,35].

(2) 决策树(decision tree)分类

决策树分类模型是一种描述对实例进行分类的有向树型结构,内部结点表示一个特征或属性,叶子结点表示一个类[75].进行分类时,从跟结点开始,对实例的某一特征进行测试,根据测试的结果将实例分配到其子节点,再根据当前结点对应特征的取值分配下一个结点.如此递归,直到到达叶子结点就可得到该实例对应的类型.

决策树学习的本质是从训练集中归纳出一组分类规则,通常包括特征选取、决策树生成和剪枝这 3 个步骤,常用的学习算法有 ID3[78],C4.5[79]和 CART[80].决策树的优点在于易于理解和实现,而且能够有效地处理大规模的评论数据;但决策树在分类类别的增多时,可能导致错误率提高较快.

(3) 支持向量机(support vector machine,简称 SVM)分类

支持向量机[81]的原理是:将低维空间中不可分的点映射到高维空间,使它们成为线性可分的;然后,寻找一个使两个不同类的数据点间隔最大的超平面.对于一个评论实例 X,采用 SVM 分类时使用下面的函数给每个评论赋予一个值:

$$f(x)=\mathrm{sgn}\{w^TX+b\},$$

其中,w 是权值向量,b 是截距.

在评论质量的检测与控制中,当将目标问题看作分类问题时,SVM 分类由于它通过核函数的选择可以选择线性或非线性的划分方法,并且对异常点具有较高的抵抗力(resistant),因而在这类问题中,SVM 是采用最多的一种方法[8,13,18,35,56].此外,SVM 从理论上可以得到全局最优解,且它的最终决策函数只由少数的支持向量确定,而计算的复杂性只取决于这些支持向量的数量,因此在少量训练样本的情况下也能具有较好的泛化能力.这对评论质量评估与控制应用中很重要,例如在垃圾评论的检测中,由于垃圾评论所占的比例很小,如果要求收集大量垃圾评论作为训练样本,这将需要花费较多的时间,而在这段时间内,垃圾评论正在危害着正常用户的利益.另一方面,SVM 也具有自身的不足之处:由于 SVM 借助二次规划来求解支持向量,这使得它对大规模训练样本是需要耗费大量的时间和空间;传统的 SVM 只是针对二分类问题,对于多分类问题,目前主要借助多个支持向量机来实现.

SVM Rank[82]通过将排序问题转化为分类问题来间接实现对象的排序.文献[41,42]使用 SVM Rank 对候选的垃圾评论发布团队进行排序.文献[27]中定义了不同的句子打分函数,然后使用 SVM Rank 对评论中的句子进行排序,然后选择前 k 个句子作为评论的总结.文献[83]使用随机森林(random forest)分析评论可读性对评论有用性分类的影响.

除了回归和分类外,文献[34]从评论的相关特征和评论人的相关特征两个角度训练分类器,采用协同训练(co-training)的方法识别垃圾评论.文献[21,28]则从评论集中选出一组评论作为总结,前者要求选择的评论能够从不同的观点覆盖到尽可能多的商品特征,从而将评论的选取转化为图的最大覆盖问题;后者首先根据评论集生成一个平均的目标观点向量Γ,要求选取 k 个评论,它们所形成的平均观点向量π与Γ的距离最小. 文献[31]构建一个异质图(heterogeneous graph)捕获评论人、评论和商家之间的关系,并提出一个迭代的计算模型来评估评论人的可信度(trustiness)、评论的公正度(honesty)和商家的可靠度(reliability),以此识别可疑的评论和评论人.模式识别的方法可以用于检测突发性的垃圾评论发布行为[36]、异常的评论模式[37]、垃圾评论的发布团队[41,42]以及商品特征的识别[19,20,30].文献[26]对比了 k-mean、结构化的 PLSA(probabilistic latent semantic analysis)和无结构的 PLSA 这 3 种方法在商品特征聚类上的效果,在商品特征覆盖率上,3 种方法的差异不大(都在 80%左右),但结构化的 PLSA 在类的内聚性上最优.

4 评价体系

评论的质量检测和控制技术的评价本质上是十分困难的,主要是因为:

(1) 评论质量是一个主观性的概念,不同用户的关注点也不完全一样;

(2) 缺乏验证的基准,例如在垃圾评论检测中,一方面难以确定哪些评论是真正的垃圾评论,哪些评论人

© 中国科学院软件研究所 http://www.jos.org.cn

是真正的垃圾评论发布者;另一方面,这些垃圾评论(人)表面上与正常评论(人)一样,即使人工也难以识别.评论的质量检测和控制技术中,确定验证基准通常采用的方法有:

- 人工标注[7,8,19,20,24,26,27,29−31,34−36,40−42];
- 利用评论的有用性投票[5,9,11,12,15,21];
- 基于副本检测的方法[32,33];
- 利用外部的评论资源[31](例如一些权威的评论网站)等.

(3) 不同的模型和技术在不同的数据集上的效果不同.

由于目前在评论质量方面的研究还没有标准的数据集,在不同的数据集(评论网站)上的相关影响因素也不一样,例如评论包含的信息、打分的稀疏性、打分的形式和尺度以及数据集的其他特性.针对评论质量的评估与控制中的不同问题和方法,常用的评价指标有:预测准确度、分类准确度和排序准确度.

4.1 预测准确度

预测准确度主要考察预测值与验证基准之间的差异程度.当采用回归分析或者打分函数的方法时,预测准确度是一个重要的衡量指标.预测准确度的基本方法之一就是平均平方误差(mean squared error,简称 MSE):

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2,$$

其中,n 为预测对象的数量,Y_i 为真实值,$\hat{Y}_i$ 为预测值.平均平方误差比较直观且计算简单,可以评价数据的变化程度,MSE 的值越小,说明预测模型的预测结果具有更好的准确度.

文献[10]以评论的平均评分作为评论质量的真实值,而文献[11,12]以评论的有用性投票比例作为真实值,三者都通过对比预测值与真实值间的平均平方误差验证评论质量预测的准确度.

与平均平方误差相对应的还有均方根误差(root mean square error,简称 RMSE)和标准平均绝对误差(normalized mean absolute error,简称 NMAE)等[84].

4.2 分类准确度

如果评论的质量预测值为预先定义的若干个离散值时,此时更适合采用分类准确度对系统进行评估.例如:在评论质量检测时,可将问题转化为预测评论是属于质量高(或有用的)的类型还是质量低(或无用的)的类型[8,13,56];在垃圾评论(人)识别中,分类的类型可以是“垃圾评论(人)”和“正常评论(人)”[34,35];评论总结中,可以根据一个句子是否被选取作为评论总结的内容进行分类[85].

对于分类任务,查准率(precision)、查全率(recall)和 F-score 是最常使用的系统评估指标(http://en.wikipedia.org/wiki/Recall_(information_retrieval)),定义为

$$precision=\frac{tp}{tp+fp},\ recall=\frac{tp}{tp+fn},\ F_\beta=(1+\beta^2)\frac{precision\cdot recall}{(\beta^2\cdot precision)+recall},$$

其中,tp 为正样本被正确分为正样本的个数,fp 为负样本被误分为正样本的个数,fn 为正样本被误分为负样本的个数.F-score 将查准率和查全率统一起来综合考虑,正实数β调整两者的权重,β越大,意味着查全率的权重越高.当β=1 时,F_1 就是查准率和查全率的调和平均.此外,通常还使用准确率(accuracy)来评估分类的效果:

$$accuracy=\frac{tp+tn}{tp+tn+fp+fn},$$

其中,tn 为负样本被正确分为负样本的个数.

为了避免分类算法对数据的过分特化而导致分类结果的过分乐观估计,常采用 k-折交叉验证(k-fold cross-validation)的方法.在评论质量的分析中,根据具体的问题和目标上述的计算公式可作适当的调整.例如,文献[20]在评估“特征-观点”对的挖掘策略时,将查准率和查全率定义为

$$precision_{feature\text{-}opinion}=\frac{N_{correct}}{N_{mining}},\ recall_{feature\text{-}opinion}=\frac{N_{correct}}{N_{all}},$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中,$N_{correct}$指正确识别出的“特征-观点”对,N_{mining}指识别出的所有“特征-观点”对,N_{all}指评论中所有的“特征-观点”对.

2007 年,文本检索会议(text retrieval conference,简称 TREC)的垃圾检测针对垃圾邮件进行检测[86],它的一些统计性的评测标准也可以应用在垃圾评论的检测中(如文献[43]):

$$hm\% = \frac{b}{b+d},\ sm\% = \frac{c}{a+c},\ lam\% = \log it^{-1}\left(\frac{\log it(hm) + \log it(sm)}{2}\right),$$

其中,$\log it(x) = \log\left(\frac{x}{1-x}\right)$,$a$ 为被正确识别的垃圾评论数量,b 是正常评论被误分为垃圾评论的数量,c 为垃圾评论被误分为正常评论的数量,d 为被正确识别的正常评论数量.hm%和 sm%分别衡量的是正常评论和垃圾评论的误分率,lam%是正常评论和垃圾评论被误分概率的几何平均.

4.3 排序准确度

不同于分类准确度,排序准确度适用于需要给用户提供一个排序列表的评论质量评估与控制系统.这类系统不需要对评论进行硬性的划分,只需确定评论间的相对位置关系,例如,按照评论质量高低排序,或者返回根据评论属于垃圾评论可能性高低进行排序的列表.

在评论质量的评估与控制技术中,常用的排序技术评估指标包括 *precision@k*,*DCG*(discounted cumulative gain)和 *nDCG*,其中,*precision@k* 用于测量在返回列表的前 k 个结果的准确度,假设前 k 个结果中有 c_k 个准确的结果,那么,

$$precision@k = \frac{c_k}{k}.$$

文献[42]使用 *precision@k* 反映前 k 个评论中很可能是垃圾评论所占的比例,但是 *precision@k* 并没有体现前 k 个结果中相互的位置关系.当评价不同的排序算法时,通常可用 *DCG* 和 *nDCG* 来衡量算法的前 k 个结果的排序效果:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

其中,rel_i为第 i 个结果的相关度分值.文献[39]将 rel_i定义为评论人被认为是垃圾评论发布者得到的投票数;而文献[42]中则定义为在不同排序算法下,评论人团队作为垃圾评论团队的得分值.由于不同的排序算法会产生不同的结果集,因此可以通过引入一个理想的排序结果进行规范化:

$$nDCG_k = \frac{DCG_k}{IDCG_k},$$

其中,$IDCG_k$ 是理想排序的 DCG_k 结果.此外,文献[26]使用排序损失(ranking loss)[87]度量理想排序和预测排序之间的平均距离.

4.4 准确度之外的指标

在评论的质量检测和控制研究中,一些文献[8,13,14,32,33]通过 ROC 曲线(receiver operating characteristic curve)和 AUC 曲线(area under ROC curve)评估预测或分类模型的效果.ROC 空间将真阳性率(true positive ratio)和假阳性率(false positive ratio)定义为坐标的 x 轴和 y 轴,它描述了真阳性和假阳性之间的博弈.该曲线越靠左上方,说明模型的效果越好.AUC 曲线指 ROC 曲线下的面积,因此对于两条相互交叉的 ROC 曲线,AUC 能够更直观地描述模型的效果.

lift 曲线通常适用于评估在类型分布具有高度偏差的数据集上训练的模型,而由于垃圾评论在整个评论集中所占的比重很小,因此文献[33]中还采用了 lift 曲线可视化所提出的垃圾评论检测方法的性能.

由于评论的质量高低/评论总结的好坏因人而异,以及真正的垃圾评论和垃圾评论发布者的信息难以获取,因此,人工评估是一种重要的评估手段[7,21,27].人工评估的参与者可以是志愿者,也可以通过 AMT(Amazon mechanical turk)完成.对于多人参与的评估结果,通常采用 Fleiss' Kappa 值[88]度量参与者评估的一致性程度.

© 中国科学院软件研究所 http://www.jos.org.cn

文献[29]使用 ROUGE(recall-oriented understudy for gisting evaluation)衡量评论总结的效果.相关系数也常使用在质量预测和排序性能的评测中.Spearman 相关系数和 Kendal's Tau 相关系数可用于评估预测的排序和理想排序之间的相关性[9,26],而 Pearson 相关系数在文献[12]中用于度量评论的有用性分值与理想分值之间的相关性.

5 数据集

评论质量评估与控制研究领域中,目前缺乏标准的数据集,大量的研究[7−9,12,20,21,32,39,41,42]采用 Amazon 上的评论数据,主要原因在于 Amazon 上的商品种类丰富、评论数据量大且评论信息全面.这些评论数据可通过 Amazon 的 API 接口获取,伊利诺斯大学芝加哥分校(UIC)的 Liu 发布了他们从 2006 年 6 月开始收集的 Amazon 上的评论数据(http://www.cs.uic.edu/~liub/FBS/sentiment-analysis. html),主要包括 4 类商品:书籍、音乐、DVD/VHS 和制造类商品.其中,制造类商品指类似计算机、电子商品等工业制造商品.表 3 给出了该数据集的一些统计性信息.

Table 3 Some statistical information on Amazon review dataset released by Liu[33]

表 3 Liu 发布的 Amazon 评论数据的统计信息[33]

商品类型	评论数	被评论商品数	评论人数	总的商品数
所有商品	5 838 032	1 195 133	2 146 048	6 272 502
书籍	2 493 087	637 120	1 076 746	1 185 467
音乐	1 327 456	221 432	503 884	888 327
DVD/VHS	633 678	60 292	250 693	157 245
制造类商品	228 422	36 692	165 608	901 913

该数据集中每个评论由 8 个部分组成:⟨Product ID⟩,⟨Reviewer ID⟩,⟨Date⟩,⟨Number of Helpful Feedbacks⟩,⟨Number of Feedbacks⟩,⟨Rating⟩,⟨Review Title⟩,⟨Review Body⟩.

此外,该数据集中还包括商品的相关信息,如商品编号、商品名、销售价格、商品描述等,以及 amazon 用户的概要信息,如用户 ID、用户名、评论数和用户自述等.

另一个关于 Amazon 上的评论数据集由宾夕法尼亚大学的 Blitzer 在 2007 年 8 月发布,它是 25 种类型商品的评论数据集(http://www.cs.jhu.edu/~mdredze/datasets/sentiment/该数据集的最后更新时间为 2009 年 3 月 23 日).该数据集中的数据以 XML(extensible markup language)格式存储,每条评论包含了以下信息:评论 ID、商品 ID、商品名、商品类型、有用性投票信息、评分、评论的题目、评论发表的时间、评论人的名字、评论人的位置和评论内容.每种商品的评论包含了标注数据和未标注数据,其中,四星和五星的评论标注为肯定评论,一星和二星评论标注为否定评论.但并非每种商品的评论标注都是平衡的,例如,办公用品的评论只包含了 64 个标注为否定的评论和 367 个标注为肯定的评论,而书籍的评论中却包含了 1 000 个标注为肯定的评论、1 000 个标注为否定的评论以及 973 194 个未标注的评论.该数据集的一些统计信息见表 4.

Table 4 Some statistical information on Amazon review dataset released by Blizer

表 4 Blitzer 发布的 Amazon 评论数据集的统计信息

商品种类数	评论总数	肯定的评论数	否定的评论数	未标注的评论数
25	1 422 530	21 972	16 576	1 383 982

TripAdvisor 是全球最大的旅游评论网站,拥有超过 1 000 万的注册会员及 2 500 万条评论,这些评论涉及到旅馆、餐饮、航空公司等范围.文献[13,35]采用 TripAdvisor 上的评论作为实验的数据集,前者根据评论的质量向用户推荐旅馆,后者则致力于虚假旅馆评论的自动检测.Myle 等人[35]从 TripAdvisor 上针对 20 个 Chicago 地区的著名旅馆收集了 6 977 个评论,删除了所有非 5 星级、非英语、长度小于 150 个字符和那些只发过一次评论的评论人所发布的评论,将剩下的 2 124 个评论作为可信评论,从中选取 400 个与通过 AMT 获取的 400 个虚假的评论构成一个平衡的数据集(http://www.cs.cornell.edu/~myleott/).

© 中国科学院软件研究所 http://www.jos.org.cn

康奈尔大学的 Pang 在 2004 年发布了一个从 IMDB 获取的电影评论数据集(http://www.cs.cornell.edu/people/pabo/movie-review-data),其中包含了 320 个作者(每个作者最多 20 个评论)书写的电影评论,这些评论通过一个基于用户评分的极性分类器进行标注,最后得到了 1 000 个肯定和 1 000 个否定的电影评论.该极性分类器主要是通过一些明确的规则在用户的评分上确定评论的极性.斯坦福大学大学的 Mass 等人发布了一个更大的 IMDB 评论数据集(http://www.andrew-maas.net/data/sentiment),该数据集中包含了 25 000 个肯定的评论和 25 000 个否定的评论,还包括了 50 000 个无标注的评论,其中,在 IMDB 的评分系统中,他们将 7 分以上的评论作为肯定评论,而将 4 分以下的评论作为否定评论,同时还提供了每条评论的 URL 供用户参考.

此外,国外著名的电子商务网站例如 eBay,IMDB,resellerratings(http://www.resellerratings.com/)等网站上的评论都被用于评论质量检测与控制的研究[20,26,31];国内的淘宝、当当网、大众点评网和豆瓣等都是潜在的数据源.表 5 概述了这些数据源上可用的评论信息.

Table 5 The available review information on some well-known e-commerce sites

表 5 国内外部分知名电子商务网站上可用的评论信息

	Amazon	eBay	TripAdvisor	IMDB	淘宝	大众点评	豆瓣	当当网
商品的平均评分	√	√	√	√	√	√	√	√
评分分布	√	√	√	√	√	√	√	√
推荐评论	√	√		√	√			
评论题目	√	√	√	√			√	√
评论内容	√	√	√	√	√	√	√	√
评论的评分	√	√	√	√		√	√	√
评论的时间	√	√	√	√	√	√	√	√
有用性投票	√	√	√	√	√	√	√	√
评论人的个人信息	√	√	√	√	√	√	√	
其他用户对评论的点评	√					√	√	√
对商品的特征评分		√	√		√	√		

6 总结与展望

近年来,用户生成的评论数据作为用户观点的重要载体越来越受到重视,这些评论具有数量大、噪音多、更新快、主观性高和针对性强等特点.如何快速地向用户提供准确的、简洁的和真实的评论,是评论质量检测与控制的主要目标,它主要的研究内容可以归结为 3 个部分:评论质量评估、评论总结和垃圾评论的检测.评论质量评估主要围绕量化评论质量,确定评论的有用性展开;评论总结的目标在于给用户提供全面、简洁的评论;垃圾评论的检测是为了识别虚假的评论或这类评论的发布人,降低这类评论和评论人对用户的影响.本文介绍了最近几年来国际上在这个领域中研究的动态,综述了 3 个部分研究内容中使用到的方法和技术,以及常用的评估指标和数据集.

评论质量检测与控制技术是一个新兴的、富有挑战的前沿性研究领域,该领域的研究还处于初始阶段,虽然目前已经取得了一定的成果,但仍面临着很多的困难和挑战:

(1) 各种网站上的评论格式多样化、评论信息的不一致性,严重影响了各种评估和控制方法的适用性;

(2) 评论质量的高低因人而异,个性化的评论质量评估是一个值得深入研究的问题;

(3) 验证的基准难以获取,以至评估的标准无法统一;

(4) 评论的语义理解是一个复杂和困难的问题,而网络语言的随意性进一步提高了该问题的复杂度;

(5) 垃圾评论发布者的行为难以捕捉,评论人同时发布垃圾评论和正常评论,这种混杂型的身份使垃圾评论发布者的识别变得更加复杂.

评论质量检测与控制涉及到机器学习、信息检索、自然语言处理、数据挖掘、数据管理等多个学科领域的知识,从技术上看,未来的工作可以围绕以下几点展开:

(1) 降低评论的冗余信息,需要进行大规模的评论副本(近似副本)检测,特别是语义近似性的检测;

(2) 海量评论数据的清洗和信息融合、评论数据的高效存储和管理;

© 中国科学院软件研究所 http://www.jos.org.cn

(3) 实时性的垃圾评论(人)检测可尽早消除垃圾评论带来的负面影响,维护良好的网络环境;

(4) 用户评论数量大且更新快,检测模型需要高的可扩展性和灵活性;

(5) 产品特征的提取、分类和整合为评论质量检测与控制提供定义规范的结构化信息;

(6) 用户社交网络的建立,为评论观点的挖掘和传播提供了新的途径.

用户评论是具有针对性的观点表达方式,对于很多的现实应用具有重大的价值.随着评论分析和挖掘技术的提高和完善,它的应用将进一步推广:

(1) 用户评论反映了用户的需求和喜好,将其应用于个性化的评论和商品推荐中,可以提高推荐的准确性、针对性和智能性;

(2) 用户的位置信息和产品的评论相结合,有助于提高基于位置的服务质量;

(3) 其他媒体上(如微博、论坛等)评论的质量分析和观点挖掘,有助于企业和政府部门进行舆情分析和监控,提高突发性事件的应对能力,快速准确地制定决策;

(4) 通过高质量的评论预测用户的网络活跃度及需求,在此基础上进行广告投放,可使用户和商家双方受益.

作为新的信息来源,用户评论对潜在顾客、商家、企业和政府部门等的情报分析具有重要的应用价值.但在实际的应用中,这些用户评论信息的动态性、松散性、海量性和质量多样性等对现有的技术提出了新的挑战,有很多新的理论、方法和应用还需进一步的探索和研究.

References:

[1] Chakravarty A, Liu Y, Mazumdar T. The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation. Journal of Interactive Marketing, 2010,24:185−197. [doi: 10.1016/j.intmar.2010.04.001]

[2] Duan WJ, Gu B, Whinston AB. The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. Journal of Retailing, 2008,84(2):233−242. [doi: 10.1016/j.jretai.2008.04.005]

[3] Hao YY, Zou P, Li YJ, Ye Q. An empirical study on the impact of online reviews sentimental orientation on sale based on movie panel data. Management Review, 2009,21(10):95−103 (in Chinese with English abstract).

[4] Forman C, Ghose A, Wiesenfeld B. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronics markets. Information Systems Research, 2008,19(3):291−313. [doi: 10.1287/isre.1080.0193]

[5] Yang M, Qi W, Yan XB, Li YJ. Utility analysis for online product review. Journal of Management Sciences in China, 2012,15(5): 65−75 (in Chinese with English abstract).

[6] Yu XH, Liu Y, Huang JX, An AJ. Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Trans. on Knowledge and Data Engineering, 2012,24(4):720−734. [doi: 10.1109/TKDE.2010.269]

[7] Tsur O, Rappoport A. REVRANK: A fully unsupervised algorithm for selecting the most helpful book reviews. In: Adar E, Hurst M, Finin T, Glance NS, Nicolov N, Tseng BL, eds. Proc. of the 3rd Int'l Conf. on Weblogs and Social Media. Palo Alto: AAAI Press, 2009. 154−161.

[8] Liu JJ, Cao YB, Lin CY, Huang YL, Zhou M. Low-Quality product review detection in opinion summarization. In: Eisner J, ed. Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2007. 334−342.

[9] Kim SM, Pantel P, Chklovski T, Pennacchiotti M. Automatically assessing review helpfulness. In: Jurafsky D, Gaussier É, eds. Proc. of the 2006 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2006. 423−430.

[10] Lu Y, Tsaparas P, Ntoulas A, Polanyi L. Exploiting social context for review quality prediction. In: Rappa M, Jones P, Freire J, Chakrabarti S, eds. Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 691−700. [doi: 10.1145/1772690.1772761]

[11] Liu Y, Huang XJ, An AJ, Yu XH. Modeling and predicting the helpfulness of online reviews. In: Giannotti F, Gunopulos D, Turini F, Zaniolo C, Ramakrishnan N, Wu XD, eds. Proc. of the 8th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2008. 443−452. [doi: 10.1109/ICDM.2008.94]

© 中国科学院软件研究所 http://www.jos.org.cn

[12] Zhang Z, Varadarajan B. Utility scoring of product reviews. In: Yu PS, Tsotras VJ, Fox EA, Liu B, eds. Proc. of the 2006 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2006. 51−57. [doi: 10.1145/1183614.1183626]

[13] O'Mahony MP, Smyth B. Learning to recommend helpful hotel reviews. In: Bergman LD, Tuzhilin A, Burke RD, Felfernig A, Schmidt-Thieme L, eds. Proc. of the 2009 ACM Conf. on Recommender Systems. New York: ACM Press, 2009. 305−308. [doi: 10.1145/1639714.1639774]

[14] Ghose A, Ipeirotis PG. Designing novel review ranking systems: Predicting the usefulness and impact of reviews. In: Gini ML, Kauffman RJ, Sarppo D, Dellarocas C, Dignum F, eds. Proc. of the 9th Int'l Conf. on Electronic Commerce: The Wireless World of Electronic Commerce. New York: ACM Press, 2007. 303−310. [doi: 10.1145/1282100.1282158]

[15] Otterbacher J. "Helpfulness" in online communities: A measure of message quality. In: Olsen Jr. DR, Arthur RB, Hinckley K, Morris MR, Hudson SE, Greenberg S, eds. Proc. of the 27th Int'l Conf. on Human Factors in Computing Systems. New York: ACM Press, 2009. 955−964. [doi: 10.1145/1518701.1518848]

[16] Danescu-Niculescu-Mizil C, Kossinets G, Kleinberg J, Lee LL. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In: Quemada J, León G, Maarek YS, Nejdl W, eds. Proc. of the 18th Int'l Conf. on World Wide Web. New York: ACM Press, 2009. 141−150. [doi: 10.1145/1526709.1526729]

[17] Mudambi SM, Schuff D. What makes a helpful online review? A study of customer reviews on amazon.com. MIS Quarterly, 2010, 34(1):185−200.

[18] Weimer M, Gurevych I. Predicting the perceived quality of Web forum posts. In: Angelova G, Bontcheva K, Mitkov R, Nicolov N, eds. Recent Advances in Natural Language Processing (RANLP 2007). Stroudsburg: John Benjamins Publishing Company, 2007. 1−6.

[19] Zhuang L, Jing F, Zhu XY. Movie review mining and summarization. In: Yu PS, Tsotras VJ, Fox EA, Liu B, eds. Proc. of the 2006 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2006. 43−50. [doi: 10.1145/1183614.1183625]

[20] Hu MQ, Liu B. Mining and summarizing customer reviews. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 168−177. [doi: 10.1145/1014052.1014073]

[21] Tsaparas P, Ntoulas A, Terzi E. Selecting a comprehensive set of reviews. In: Apté C, Ghosh J, Smyth P, eds. Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2011. 168−176. [doi: 10.1145/2020408.2020440]

[22] Lauw HW, Lim EP, Wang K. Summarizing review scores of "unequal" reviewers. In: Apte C, Liu B, Parthasarathy S, Skillicorn D, eds. Proc. of the 7th SIAM Int'l Conf. on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics Publications, 2007. 539−544.

[23] Hu MQ, Liu B. Opinion extraction and summarization on the Web. In: Cohn A, ed. Proc. of the 21st National Conf. on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conf. Palo Alto: AAAI Press, 2006. 1621−1624.

[24] Zhan JM, Loh HT, Liu Y. Gather customer concerns from online product reviews—A text summarization approach. Expert Systems with Applications, 2009,36:2170−2115. [doi: 10.1016/j.eswa.2007.12.039]

[25] Popescu AM, Etzioni O. Extracting product features and opinions from reviews. In: Mooney RJ, Brew C, Chien LF, Sinica A, Kirchhoff K, eds. Proc. of the Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2005. 339−346. [doi: 10.3115/1220575.1220618]

[26] Lu Y, Zhai CX, Sundaresan N. Rated aspect summarization of short comments. In: Quemada J, León G, Maarek YS, Nejdl W, eds. Proc. of the 18th Int'l Conf. on World Wide Web. New York: ACM Press, 2009. 131−140. [doi: 10.1145/1526709.1526728]

[27] Lerman K, Blair-Goldensohn S, McDonald RT. Sentiment summarization: Evaluating and learning user preferences. In: Lascarides A, Gardent C, Nivre J, eds. Proc. of the 12th Conf. of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2009. 514−522.

© 中国科学院软件研究所 http://www.jos.org.cn

[28] Lappas T, Crovella M, Terzi E. Selecting a characteristic set of reviews. In: Yang Q, Agarwal D, Pei J, eds. Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 832−840. [doi: 10.1145/2339530.2339663]

[29] Wang D, Liu Y. A pilot study of opinion summarization in conversations. In: Lin DK, Matsumoto Y, Mihalcea R, eds. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011. 331−339.

[30] Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis GA, Reynar J. Building a sentiment summarizer for local service reviews. In: Nakagawa H, Torisawa K, Kitsuregawa M, eds. Proc. of the WWW Workshop on NLP Challenges in the Information Explosion Era. New York: ACM Press, 2008. 21−30.

[31] Wang G, Xie SH, Liu B, Yu PS. Identify online store review spammers via social review graph. ACM Trans. on Intelligent Systems and Technology, 2012,3(4):Article 61, 21. [doi: 10.1145/2337542.2337546]

[32] Jindal N, Liu B. Analyzing and detecting review spam. In: Ramakrlshnan N, Zaïane OR, Shi Y, Ciifton CW, Wu XD, eds. Proc. of the 7th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2007. 547−552. [doi: 10.1109/ICDM.2007.68]

[33] Jindal N, Liu B. Opinion spam and analysis. In: Najork M, Broder AZ, Chakrabarti S, eds. Proc. of the Int'l Conf. on Web Search and Web Data Mining. New York: ACM Press. 2008. 219−230. [doi: 10.1145/1341531.1341560]

[34] Li FT, Huang ML, Yang Y, Zhu XY. Learning to identify review spam. In: Walsh T, ed. Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2011. 2488−2493. [doi: 10.5591/978-1-57735-516-8/IJCAI11-414]

[35] Ott M, Choi YJ, Cardie C, Hancock JT. Finding deceptive opinion spam by any stretch of the imagination. In: Lin DK, Matsumoto Y, Mihalcea R, eds. Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011. 309−319.

[36] Xie SH, Wang G, Lin SY, Yu PS. Review spam detection via temporal pattern discovery. In: Yang Q, Agarwal D, Pei J, eds. Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 823−831. [doi: 10.1145/2339530.2339662]

[37] Jindal N, Liu B, Lim EP. Finding unusual review patterns using unexpected rules. In: Huang J, Koudas N, Jones GJF, Wu XD, Collins-Thompson K, An AJ, eds. Proc. of the 19th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 1549−1552. [doi: 10.1145/1871437.1871669]

[38] Lappas T. Fake reviews: The malicious perspective. In: Bouma G, Ittoo A, Métais E, Wortmann H, eds. Proc. of the Natural Language Processing and Information Systems—17th Int'l Conf. on Applications of Natural Language to Information Systems. Heidelberg: Springer-Verlag, 2012. 23−34. [doi: 10.1007/978-3-642-31178-9_3]

[39] Lee K, Caverlee J, Webb S. Uncovering social spammers: Social honeypots + machine learning. In: Crestani F, Marchand-Maillet S, Chen HH, Efthimiadis EN, Savoy J, eds. Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2010. 435−442. [doi: 10.1145/1835449.1835522]

[40] Lim EP, Nguyen VA, Jindal N, Liu B, Lauw HW. Detecting product review spammers using rating behaviors. In: Huang J, Koudas N, Jones GJF, Wu XD, Collins-Thompson K, An AJ, eds. Proc. of the 19th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 939−948. [doi: 10.1145/1871437.1871557]

[41] Mukherjee A, Liu B, Wang JH, Glance NS, Jindal N. Detecting group review spam. In: Srinivasan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R, eds. Proc. of the 20th Int'l Conf. on World Wide Web (Companion Volume). New York: ACM Press, 2011. 93−94. [doi: 10.1145/1963192.1963240]

[42] Mukherjee A, Liu B, Glance NS. Spotting fake reviewer groups in consumer reviews. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S, eds. Proc. of the 21st World Wide Web Conf. 2012. New York: ACM Press, 2012. 191−200. [doi: 10.1145/2187836.2187863]

[43] Lai CL, Xu KQ, Lau RYK, Li Y, Jing L. Toward a language modeling approach for consumer review spam detection. In: Lau F, Chung JY, Shah N, eds. Proc. of the 2010 IEEE Int'l Conf. on e-Business Engineering. Washington: IEEE Computer Society, 2010. 1−8. [doi: 10.1109/ICEBE.2010.47]

[44] Lewis DD, Ringuette M. A comparison of two learning algorithms for text categorization. In: Pavlidis T, ed. Proc. of the 3rd Annual Symp. on Document Analysis and Information Retrieval. Las Vegas: University of Nevada, 1994. 81−93.

© 中国科学院软件研究所 http://www.jos.org.cn

[45] Schütze H, Hull DA, Pedersen JO. A comparison of classifiers and document representations for the routing problem. In: Fox EA, Ingwersen P, Fidel R, eds. Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1995. 229−237. [doi: 10.1145/215206.215365]

[46] Wiener E, Pedersen JO, Weigend AS. A neural network approach to topic spotting. In: Proc. of the 4th Annual Symp. on Document Analysis and Information Retrieval. Las Vegas: University of Nevada, 1995. 317−332.

[47] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Fisher DH, ed. Proc. of the 14th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 412−420.

[48] Yang Y. Noise reduction in a statistical approach to text categorization. In: Fox EA, Ingwersen P, Fidel R, eds. Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1995. 256−263. [doi: 10.1145/215206.215367]

[49] Mishra A, Rastogi R. Semi-Ssupervised correction of biased comment ratings. In: Mille A, Gandon FL, Misselis J, Rabinovich M, Staab S, eds. Proc. of the 21st World Wide Web Conf. New York: ACM Press, 2012. 181−190. [doi: 10.1145/2187836.2187862]

[50] Klare GR. Assessing readability. Reading Research Quarterly, 1974,10(1):62−102. [doi: 10.2307/747086]

[51] McCallum DR, Peterson JL. Computer-Based readability indexes. In: Burns WJ, Ward DL, eds. Proc. of the ACM'82 Conf. New York: ACM Press, 1982. 44−48. [doi: 10.1145/800174.809754]

[52] Mc Laughlin GH. SMOG grading—A new readability formula. Journal of Reading, 1969,12(8):639−646.

[53] Gunnning, R. The fog index after twenty years. Journal of Business Communication, 1969,6(3):3−13. [doi: 10.1177/ 00219436690 0600202]

[54] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998,30: 107−117. [doi: 10.1016/S0169-7552(98)00110-X]

[55] Nelson P. Information and consumer behavior. Journal of Political Economy, 1970,78(2):311−329. [doi: 10.1086/259630]

[56] Weimer M, Gurevych I, Mühlhäuser M. Automatically assessing the post quality in online discussions on software. In: Carroll JA, van den Bosch A, Zaenen A, eds. Proc, of the 45th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2007. 125−128.

[57] Siersdorfer S, Chelaru S, Nejdl W, Pedro JS. How useful are your comments? Analyzing and predicting youtube comments and comment ratings. In: Rappa M, Jones P, Freire J, Chakrabarti S, eds. Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 891−900. [doi: 10.1145/1772690.1772781]

[58] Gupta V. A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence, 2010,2(3): 258−268.

[59] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence, 2009,1(1):60−76. [doi: 10.4304/jetwi.1.1.60-76]

[60] Chen F, Han KS, Chen GL. An approach to sentence-selection-based text summarization. In: Yuan BZ, Tang XF, eds. Proc. of the 2002 IEEE Region 10 Conf. on Computers, Communications, Control and Power Engineering. Washington: IEEE Computer Society Press, 2002. 489−493. [doi: 10.1109/TENCON.2002.1181320]

[61] Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004,22(1):457−479.

[62] Hahn U, Romacker M. Text understanding for knowledge base generation in the SYNDIKATE system. In: Bench-Capon T, Soda G, Tjoa AM, eds. Proc. of the10th Int'l Conf. on Database and Expert Systems Applications (DEXA'99). Heidelberg: Springer-Verlag, 1999. 135−145. [doi: 10.1007/3-540-48309-8_12]

[63] Pang B, Lee LL, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Hajic H, Matsumoto Y, eds. Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002. 79−86. [doi: 10.3115/1118693.1118704]

[64] Pan SJ, Ni XC, Sun JT, Yang Q, Chen Z. Cross-Domain sentiment classification via spectral feature alignment. In: Rappa M, Jones P, Freire J, Chakrabarti S, eds. Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 751−760. [doi: 10.1145/1772690.1772767]

© 中国科学院软件研究所 http://www.jos.org.cn

[65] Wu Q, Tan SB, Cheng XQ, Duan MY. MIEA: A mutual iterative enhancement approach for cross-domain sentiment classification. In: Huang CR, Jurafsky D, eds. Proc. of the 23rd Int'l Conf. on Computational Linguistics, Posters Volume. Beijing: Chinese Information Processing Society of China, 2010. 1327−1335.

[66] Tan SB, Wang YF, Wu GW, Cheng XQ. Using unlabeled data to handle domain-transfer problem of semantic detection. In: Wainwright RL, Haddad H, eds. Proc. of the 2008 ACM Symp. on Applied Computing. New York: ACM Press, 2008. 896−903. [doi: 10.1145/1363686.1363893]

[67] Dasgupta S, Ng V. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In: Su KY, Su J, Wiebe J, eds. Proc. of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009. 701−709.

[68] Peter D. Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Burelle S, Somesfalean S, eds. Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2009. 417−424.

[69] Paltoglou G, Thelwall M. A study of information retrieval weighting schemes for sentiment analysis. In: Hajic J, Carberry S, Clark S, eds. Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010. 1386−1395.

[70] Martineau J, Finin T, Joshi A, Patel S. Improving binary classification on text problems using differential word features. In: Cheung DWL, Song IY, Chu WW, Hu XH, Lin JJ, eds. Proc. of the 18th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2009. 2019−2024. [doi: 10.1145/1645953.1646291]

[71] Lin YM, Zhang JW, Wang XL, Zhou AY. An information theoretic approach to sentiment polarity classification. In: Castillo C, Gyongyi Z, Jatowt A, Tanaka K, eds. Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. NewYork: ACM Press, 2012. 35−40. [doi: 10.1145/2184305.2184313]

[72] Pang B, Lee LL. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008,2(1-2):1−135. [doi: 10.1561/1500000001]

[73] Hu M, Liu B. Mining opinion features in customer reviews. In: McGuinness DL, Ferguson G, eds. Proc. of the 19th National Conf. on Artificial Intelligence, 16th Conf. on Innovative Applications of Artificial Intelligence. MIT Press, 2004. 755−760.

[74] Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ. Introduction to WordNet: An on-line lexical database. Int'l Journal of Lexicography, 1990,3(4):235−244. [doi: 10.1093/ijl/3.4.235]

[75] Li H. Statistical Learning Methods. Beijing: Tsinghua University Press, 2012 (in Chinese).

[76] Casella G, Berger RL. Statistical Inference. 2nd ed., Beijing: China Machine Press, 2010.

[77] Ghose A, Ipeirotis PG. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Trans. on Knowledge and Data Engineering, 2011,23(10):1498−1512. [doi: 10.1109/TKDE.2010.188]

[78] Quinlan JR. Induction of decision trees. Machine Learning, 1986,1(1):81−106. [doi: 10.1023/A:1022643204877]

[79] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers, 1988.

[80] Lon WY. Classification and regression trees. WIREs Data Mining and Knowledge Discovery, 2011,1(1):14−23. [doi: 10.1002/widm.8]

[81] Cortes C, Vapnik V. Support-Vector networks. Machine Learning, 1995,20(3):273−297. [doi: 10.1007/BF00994018]

[82] Joachims T. Optimizing search engines using clickthrough data. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining New York: ACM Press. 2002. 133−142. [doi: 10.1145/775047.775067]

[83] O'Mahony MP, Smyth B. Using readability test to predict helpful product reviews. In: Pasi G, Liu TY, Raghavan P, eds. Proc. of the Recherched Information Assistée par Ordinateur (RIAO 2010), the 9th Int'l Conf. on Adaptivity, Personalization and Fusion of Heterogeneous Information. Paris: CID, 2010. 164−167.

[84] Lehmann EL, Casella G. Theory of Point Estimation. 2nd ed., Heidelberg: Springer-Verlag, 1998.

[85] Neto JL, Freitas AA, Kaestner CAA. Automatic text summarization using a machine learning approach. In Bittencourt G, Tamalho G, eds. Advances in Artificial Intelligence, 16th Brazilian Symp. on Artificial Intelligence (SBIA 2002). Heidelberg: Springer-Verlag, 2002. 205−215. [doi: 10.1007/3-540-36127-8_20]

© 中国科学院软件研究所 http://www.jos.org.cn

[86] Cormack G. TREC 2007 spam track overview. In: Voorhees EM, Buckland LP, eds. Proc. of the 16th Text REtrieval Conf. (TREC 2007). Gaithersburg: National Institute of Standards and Technology, 2007. 1−8.

[87] Snyder B, Barzilay R. Multiple aspect ranking using the good grief algorithm. In: Sidner CL, Schultz T, Stone M, Zhai CX, eds. Proc. of the Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2007. 300−307.

[88] Fleiss JL, Lecin B, Paik MC. Staticstical Methods for Rates and Proportions. 3rd ed., Hoboken: John Wiley & Sons Inc., 2003.

附中文参考文献:

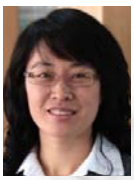
[3] 郝媛媛,邹鹏,李一军,叶强.基于电影面板数据的在线评论情感倾向对销售收入影响的实例验证研究.管理评论,2009,21(10):95−103.

[5] 杨铭,祁巍,闫相斌,李一军.在线商品评论的效用分析研究.管理科学学报,2012,15(5):65−75.

[75] 李航.统计学习方法.北京:清华大学出版社,2012.

林煜明(1978－),男,广西合浦人,博士生,主要研究领域为Web数据挖掘,观点分析.
E-mail: ymlinbh@gmail.com

朱涛(1989－),男,博士生,主要研究领域为Web服务计算.
E-mail: infozt@163.com



王晓玲(1975－),女,博士,教授,博士生导师,主要研究领域为数据管理技术,Web服务计算.
E-mail: xlwang@sei.ecnu.edu.cn

周傲英(1965－),男,博士,教授,博士生导师,CCF高级会员,主要研究领域为海量数据管理,Web语义搜索与挖掘,数据密集型计算,Web服务计算.
E-mail: ayzhou@sei.ecnu.edu.cn

© 中国科学院软件研究所 http://www.jos.org.cn