情 报 学 报 ISSN 1000 - 0135
第 34 卷 第 5 期 493 - 500, 2015 年 5 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 34 No. 5, 493 - 500 May 2015

doi:10.3772/j.issn.1000-0135.2015.005.006

基于特征提取改进的在线评论有效性分类模型[1)]

夏火松 杨 培 熊 淦

(武汉纺织大学管理学院,武汉 430073)

摘要 随着国内电子商务的蓬勃发展,各大电商网站评论量飞速增长,如何从 Web 大量的商品评论中挖掘出价值信息并进行有效分类对消费者和生产厂商决策有重要的影响。传统分类方法能有效地抽取商品评论中的特征及观点,但对中文评论分类仍存在一些不足。为了进一步提高商品评论分类有效性,首先,综合前人研究提出一种基于评论长度的特征提取方法,提高分类准确率;然后,设计了评论样本自动标注方法,并构建评论的有效性分类模型,改善分类效率;最后,以京东商城上爬取的 1710 条商品评论为例,对提出的特征提取与自动标注方法进行验证。实验结果表明,根据该方法,评论分类准确率得到明显提高。

关键词 在线评论 有效性 文本分类 领域词典 自动标注

The Classification Model for Online Reviews' Effectiveness Based on Feature Extraction Improvement

Xia Huosong, Yang Pei and Xiong Gan

(*School of Management*, *Wuhan Textile University*, *Wuhan* 430073)

Abstract With development of e-commerce, the major e-commerce website product reviews increased rapidly. How to utilize the large number of reviews and classify it efficiently make an important impact on manufacturers' decisions. Traditional classification method can effectively extract the product opinions, but not for Chinese reviews. In order to improve the effectiveness of product reviews classification, firstly we present a kind of feature extraction method based on the length of feature to improve classification accuracy; secondly we design a comment sample automatic annotation methods and construct the classification model; and finally, take 1710 product reviews from Jingdong Mall and proposed this methods. The results show that this method could improve the classification accuracy significantly.

Keywords online review, effectiveness, text categorization, field dictionary, automatic marking

1 引 言

随着国内电子商务的蓬勃发展,越来越多的人倾向于通过电子商务网站购买各类产品。近几年,阿里巴巴、淘宝、京东、亚马逊等主要电子商务网站都积累了庞大的用户群体,而且为了提高消费者满意度与改善消费者购物体验,这些电商网站大都允许消费者对其购买的产品发表意见与建议。在此基础上,包含有客户体验与评价信息并对企业具有重要价值的商品评论大量产生。通过对评论的分析与研究,消费者可以有比较地制定购买计划,生产商亦可了解自己产品的优势和不足,把握用户的需求,改善产品与服务[1]。

收稿日期:2014 年 11 月 18 日

作者简介:**夏火松**,男,1964 年生,武汉纺织大学管理学院教授,博士,主要研究方向:知识管理、数据挖掘、物流信息管理和电子商务、DSS,E-mail:bxxhs@sina.com。**杨培**,男,1990 年生,武汉纺织大学硕士研究生,主要研究方向:数据挖掘、信息管理。**熊淦**,男,1992 年生,武汉纺织大学研究生,主要研究方向:信息管理、知识管理。

1) 基金项目:国家自然科学基金项目(71171153)"24 小时知识工厂的知识共享活动模型与服务支持系统研究"。

在线评论(Online Review)是一种重要的在线口碑(Online Word-of-Mouth)形式,它是消费者之间通过网络交流的所有关于产品和服务的具体特性、使用或提供商的信息[2]。利用在线评论可以帮助生产商做决策支持,但并非所有评论都有价值。一方面,由于网络的匿名性、非面对面的接触和沟通成本低廉等特性[3],部分评论者会不负责任地随意发表评论,评论的质量往往参差不齐;另一方面,大数据的兴起以及爆炸性增长的产品评论也对在线评论的研究可操作性提出了更高的要求,人工分析阅读筛选有效评论已不具备可行性。即使部分电子商务网站已建立评价有用性投票,但需要长期的数量积累,对于冷门商品与刚刚上架的商品,评论数量较少,评价有效性亟待度量。

2 相关研究工作

目前关于评论有效性的研究主要分两方面,对产品评论有效性的意见挖掘与对产品评论有用性的影响因素研究。从意见挖掘的角度,研究人员主要采用方面级的意见挖掘方法找到与商品相关的某个物理组成部分、功能或者性质[4],以提取出商品评价中被重点关注的商品相关信息。Miao[5]使用半监督的方法建立一个商品属性的抽取模型,并根据语义相似度提出了一种商品属性的聚类模型。Hu[6]提出了一个多重方程模型探索评论评级、情感与销量的关系,研究结果发现越是靠前以及越是容易理解的评论对销量的影响较大,而评级对销量没有直接影响,只能间接影响情绪。而从产品评论有效性影响因素角度,学者大多从客户的角度出发,探索商品评论对网络口碑的影响[7,8]。严建援等[2]通过中国大型B2C网站中221个有效样本的实证分析发现评论内容越客观,传达更多实物与网站描述是否相符或者包含更多产品特征的评论更有价值,而包含主观感受越多,则评论有用性更低。廖成林等[9]通过对淘宝网445条有效样本做实证分析,发现越是中差评,其包含的购买者对商品更深的了解,其给出的评论也更有价值。郝媛媛等[7]通过建立在线评论有用性影响因素模型,并做出分类预测证明评论的平均句子长度对评论的有用性具有正向影响。同时杨朝君和汪俊奎[10]也证实评论长度对评论有用性存在正向影响。

综上所述,虽然前人研究在实际应用中取得了很好的效果,但也存在一些不足,第一,学者研究主要集中在对评论的情感分析与观点挖掘及行为研究,而关于评论有效性与自动分类问题的研究较少[11,12];第二,对于文本评论的有效性分类而言,传统文本分类缺少一种准确率较高并且对人工需求并不大的特征提取方法;第三,样本集标注方面,传统文本分类为基于监督的分类模式,需要花费较多人力对样本集进行类别标注再供分类器训练用。

因此,我们从有效评论分类以及自动标注方面对传统文本分类方法的关键技术进行改进,首先,综合前人研究提出一种基于领域词典结合评论长度的特征提取方法,提高分类准确率;然后,设计了评论样本自动标注方法,并构建评论的有效性分类模型,改善分类效率。研究结果对于未来电子商务评论有效性自动分类方面有着重要的参考价值。

3 评论有效性分析关键技术

结合以上论述,我们对中文评论文本分类过程中的文本表示这一关键部分作深入研究,拟采用提出的基于领域词典结合评论长度的特征提取方法与语料类别自动标注两方面提高评论分类准确率与效率。

3.1 特征提取

特征提取原理是对候选特征集 T 中的特征,利用一种评价函数评价特征项的重要性,可以设定一定的阈值剔除掉重要性权值小于这个阈值的特征项从而达到特征提取的目的。虽然用向量空间模型可以把文本表示为向量形式,但是不可能将文本集中出现的所有词都作为特征项,因为文本集的规模越大,则词的个数越多,特征项的个数越多,计算机计算的难度也就越大。因此需要选择有限个比较能代表文档内容的词特征项,因此特征提取的重要性不言而喻。前面综述中提到,特征提取方法主要分基于领域词典的提取方法和基于统计的特征提取方法两类,基于领域词典的特征提取方法虽然精度高,但是需要耗费大量人力,并且每个领域的特征提取都需要对其领域建立特定的领域词典,通用性较差,而基于统计的特征提取方法虽然依赖计算机自动处理,处理速度极快,大大节省人力,但精度较差。因此,我们提出一种通用领域词与评论长度结合的特征提取方法,并与基于统计的特征提取方法之一的基于信息增益的特征提取方法作对比分析。

(1)传统基于信息增益的特征提取方法

基于统计的特征提取方法是通过一定的函数自

动计算特征的重要性而进行的特征提取,不需要建立词典,并且与特定领域无关,只与特征项重要性大小相关,同时,阈值设定高低会影响特征项的个数。基于统计的特征提取方法虽然精度较低,但因为不需要多少人工支持,能够快速计算,适用范围较广。

这里我们主要以基于统计的特征提取方法中基于信息增益的特征提取方法作代表性说明。信息增益是基于统计的特征提取方法之一。它的原理是计算整个文本集合在包含某个特征与不包含有该特征的信息量的差值,这个差值越大,则这个特征的信息增益越大,对文本集合的重要性越强。要计算信息增益,要先计算"熵"。对于一个 n 类问题,"熵"的计算公式如式(1)所示。这里 $P(C_i)$ 表示类别 C_i 出现的概率。再计算特征 t 的"条件熵",其计算公式如式(2)所示。

$$H(C) = -\sum_{i=1}^{n} P(C_i) \cdot \log_2 P(C_i) \quad (1)$$

$$H(C \mid T) = P(t)H(C \mid t) + P(\bar{t})H(C \mid \bar{t}) \quad (2)$$

其中,$P(t)$ 和 $P(\bar{t})$ 分别表示特征 t 在总文本中出现的概率与不出现的概率,而 $H(C \mid t)$ 和 $H(C \mid \bar{t})$ 分别表示在特征 t 出现的条件下文本的熵与特征 t 不出现的情况下文本的熵。其计算方法如式(3)、式(4)所示。

$$H(H \mid t) = -\sum_{i=1}^{n} P(C_i \mid t)\log_2 P(C_i \mid t) \quad (3)$$

$$H(H \mid \bar{t}) = -\sum_{i=1}^{n} P(C_i \mid \bar{t})\log_2 P(C_i \mid \bar{t}) \quad (4)$$

其中,$P(C_i \mid t)$ 和 $P(C_i \mid \bar{t})$ 分别表示在特征 t 存在的条件下类别 C_i 出现的概率与特征 t 不存在的条件下类别 C_i 出现的概率。有上述公式得到特征 t 的信息增益公式如式(5)所示。

$$IG(t) = H(C) - H(C \mid T) \quad (5)$$

对每个特征都可以用这个方法计算出其信息增益量,对于信息增益量小于"阈值"的特征项去掉该特征,可以根据不同的情况设定不同的阈值。

(2)基于领域词典结合评论长度的特征提取方法

基于领域词典的特征提取方法即基于特定的领域,人工建立一个备选的特征项集合。同时也可以通过语义相似度、建立同义词典等方式对于未登录词进行判断[6]。但是无论是否用同义词词典,都需要人工建立一个领域种子词库,根据这个种子词库来扩充领域词典的规模。我们采取利用较少的领域词结合评论长度共同作为特征项的特征提取方法,能一定程度解决建立领域词典需要大量人力以及无法跨领域的问题。即该提取方法当成一个特殊的"领域词典",通过它对评论集进行特征提取,具体步骤分为提取通用领域词与统计评论的含词量两步。

① 通用领域词

通用领域词是不同领域产品评论内容中共有的词汇,比如"物流"、"服务"等词,在所有网购产品的在线评论内容中常常出现,这些词在某条评论中是否出现也能够一定程度上反映该条评论是否有效。这里的通用领域词只有所有商品共有的特征,在维度上会比较低。通过人工采集两个评论集合中的商品特征词,发现除了商品特有的特征外,顾客还对商品的外观,做工,商品发货送货的速度,价格的合理与否,商品是否是正品,与描述是否享福,电子商务公司的服务(包括该产品的客服人员的服务,售前售后以及快递人员的态度等)等比较关心,这些关键词,在两种不同的商品评价中都常出现,可以作为有用评价的通用领域词。论文通过对这些词进行归类,参考李杰[12]对服装产品建立的评论要素概念模型,建立了三层树状概念模型如图 1 所示。

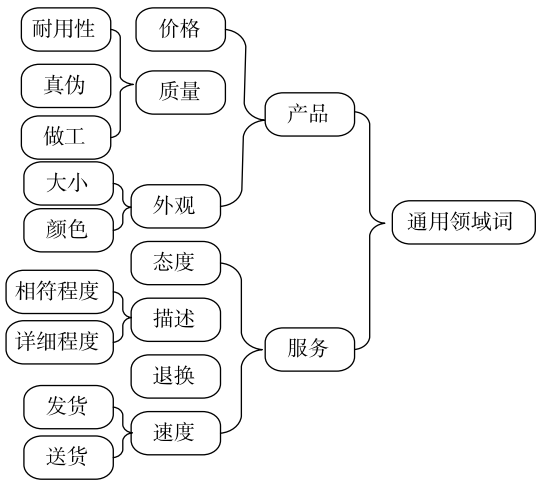

图 1　通用领域词概念模型

依据此概念模型,初步提取以下领域词见表 1。

表 1　初步领域词

产品	服务
包装盒	速度
条形码	态度
质量	素质
…	…
20	35

② 评论长度

评论的长度是影响一条评论是否有效的一个重要因素。在中文商品评论中,一条评论的长度越长,其有用的可能性就越大。单纯的依赖通用领域词做特征提取会损失大量可以代表产品特征的专用领域词特征,且很容易对这些评论的有效性进行误判,使得分类准确率降低。例如(1)"一买就降价 WiFi 受限,垃圾电脑,劝别买。"与(2)"垃圾电脑,收到货就一直出毛病,无线网连上,过一分钟不到就掉线,要买的朋友注意,千万别在这买。垃圾京东。"前者与后者所描述的主要意思相同,但是后者对于问题的描述更加详尽,对产品的认识也更加深刻。当然,一条很长的评论不一定有用,但是可以将评论长度作为衡量评论有用性的重要指标,研究将评论长度作为特征权重之一,使分类器覆盖更多的有效评价。这里考虑到一般的特征权重的值大于1的可能性较小,对评论长度作归一化处理,其公式如式(6)所示。

$$l = \frac{x - \min(d)}{\max(d) - \min(d)} \tag{6}$$

式中,l 为归一化前的评论长度,$\max(d)$、$\min(d)$ 分别表示评论所在的评论集的最大长度与最小长度,y 为转化后的评论长度。经过转化后的评论长度都不大于1。

综合以上,提出特征提取方法形式如式(7)所示。

$$T = T(l, t_1, t_2, \cdots, t_n) \tag{7}$$

式中,T 表示特征项集合,l 为评论长度,t_1 至 t_n 为通用领域词。

3.2 样本类别自动标注

有用评价的分类是一个二类分类问题,特征选择步骤与分类器训练都需要用到语料标注,在信息增益里计算"熵"时需要知道样本集中正类与负类的个数各是多少,而在训练分类器时,正类和负类都要分别标注为1和-1。通常做文本挖掘语料标注都要基于人工理解,需要花较多时间阅读语料样本集,因此研究基于上述特征提取方法提出一种语料类别自动标注方法。

第一,对于较长的评论,无论其是否含有领域词,都标注为有用。这种方法能够避免将评论较长,含有较多信息,却因为领域词典不全而导致被标注为无用的情况。第二,对于不长并且不太短的评论,若其含有领域词,则标注为有用。有些评论虽然含有领域词,但是其长度过短,无法带来商品较细节的信息,对这一类的评论依然将其标注为无用,而长度虽然不是特别长,却含有领域词的则应该标注为有用。该标注方法伪代码见表2。

评论长度可以通过代码编程统计,只要设定的阈值 A 与 B 合适,就能很好地对样本集进行自动标注,自动标注效果在后续实验中作详细讨论。

基于普通文本分类过程与以上改进,研究提出的评论有效性分类模型如图2所示。

4 评论有效性分类实验

研究主要采用文本分类方法,即按照预先定义的主题类别,通过分析文本的内容将文本集合中的每个文本分配到预先定义的类别中[13],并基于以上分类模型对传统文本分类进行改进。郑丽娟等[14]为了测试情感分类在句子与段落上的分类效果差异,对传统文本分类的过程进行了相应的描述,其分类流程大致如下:

表2 样本类别自动标注伪代码

Marking Method
D_i :样本集 D 中第 i 条评论; A、B:设定的有效评论长度的阈值; l_i :为评论的长度。

for each D_i in D {　if ($A \leqslant l_i \leqslant B$ 且 D_i 中存在一个或以上个 $t_k \in \{t_1 \rightarrow t_n\}$ 不为0)　或　($l_i > B$) {
评论 D_i 为有效评价,标注为1; }
else {
评论 D_i 为无效评价,标注为-1; }
return 标注后的样本集;

End

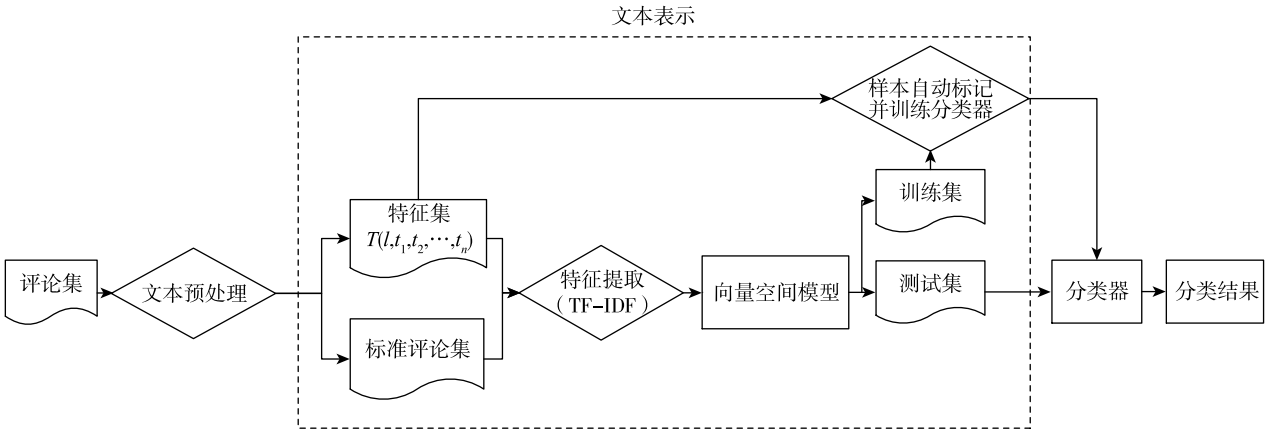

图 2　评论有效性分类模型

首先,文本预处理,通过文本预处理能够将不规则的文本转化为计算机较易处理的规范化的文本。

其次,文本表示,先进行特征提取与特征降维,提取文本中能够表示文本主要内容的关键词,去掉对文本内容没有太大影响的词汇,再利用提取出的词汇作为向量的维度,计算每一篇文本中每个维度的权重,将文本最终转化为数据形式。

最后,训练分类器并进行分类,利用数值形式的文本数据训练分类器,并运用该分类器将未知样本映射到已知样本,从而达到对未知样本进行分类的目的。

4.1　实验数据与预处理

阿里巴巴、淘宝、京东、亚马逊等主要电子商务网站都积累了庞大的用户评论数据。考虑到从代表性的负面评论中更能提取出较有价值的特征词汇,而且京东商城上已经对评论进行了好评、中评、差评的分类,我们从京东商城的左旋肉碱(减肥药)与联想笔记本电脑这两种完全不同的商品评论中提取了其负面评论各 1210 条与 500 条,进行下一步的对比分析。

对文本评论的预处理工作首先,去除空白、重复评论,避免重复评论对特征选择计算造成的干扰。然后,去停用词,筛除如“我”、“一方面”、“不然的话”等没有任何价值但是出现的频率较高的副词与人称代词,减轻文本表示计算量,研究主要采用“四川大学机器智能实验室停用词库”自动去除停用词。最后,分词并建立样本词集,将每一条评论切分为词集并保存于数组中便于计算机统计词频,计算特征权重。整个预处理过程采用 python 语言编程实现,分词过程采用 python 语言接口“结巴分词”作为分词软件。经过删除无用与重复的评论的预处理,得到两种商品的评论各 1050 条与 500 条。

4.2　文本表示(TF-IDF)

特征提取与特征降维方法采用上一章中经过改进的基于领域词典结合评论长度的特征提取方法,并以结果数据作实验组,与传统基于信息增益的特征提取方法的对照组实验作对比分析。

经过特征提取后文本表示主要采用向量空间模型(VSM)方法来实现,基本过程是将经过预处理后的文本生成一个 n 维特征项集合 $T = T(t_1, t_2, \cdots, t_n)$,其中 t_j 表示第 j 个特征项,再根据 T 把每个文本转化为一个由这些特征项的权值所构成的权值向量 $W_i = W_i(w_{i1}, w_{i2}, \cdots, w_{in})$,其中 w_{ij} 为文档集 D 中第 i 篇文档的第 j 个特征项的权值,这样,文档集 D 中的每一篇文档都可以表示成一个权值向量,便于计算。特征项权重常见的方法有:布尔函数,平方根函数,对数函数,TF 算法,逆文档频率(TF-IDF 算法)等。

这里我们主要采用逆文档频率(TF-IDF)计算每个文本的每维特征的权重,以特征权重作为特征向量的每维的值,从而一个文本集合即转为一个由特征权重构成的向量空间,计算公式(8)所示。

$$tf - idf = tf_{ij} \times idf_j = tf_{ij} \times \lg(d/df_j) \quad (8)$$

其中,tf_{ij} 为词项 t_j 在文档 D_i 中出现的概率,d 为总文本的个数,df_j 为出现词项 t_j 的文本个数。例如,“大小”是向量空间模型的一个特征词项,文档 D_i 中“大小”出现 2 次,且文档 D_i 的总词数为 50,那么“大小”的词频为 $tf_{ij} = 2/50 = 0.04$,总文本个数 d 为 1000 个,出现“大小”的文本数为 10 个,那么词

项“大小”的idf值为 $idf_j = \lg(1000/10) = 2$，则对于文档 D_i，其特征项“大小”的权值为 $tf_{ij} \times idf_j = 0.04 \times 2 = 0.08$。

4.3 文本分类器

将文本数据转为数值数据后即可用分类器进行分类。常用的文本分类器包括支持向量机、最大熵和朴素贝叶斯。而支持向量机算法应用到文本分类时取得了较好的效果，郑丽娟等[14]将支持向量机应用到文本情感分类得到了较高的准确率，因此研究采用支持向量机作为文本分类器。支持向量机是1995年Cortes和Vapnik[15]提出的，它在解决小样本分类问题上有优势，其分类原理是解一个最优化问题，对于一个样本集 $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \cdots, (\vec{x}_n, y_n)\}$，其中 $\vec{x}_i$ 是文档 D_i 经过逆文档频率计算后转化成的向量形式，$y = \{1, -1\}$ 是样本标签，即上面的样本标注结果，解公式(9)的优化问题。

$$\begin{cases} \min \dfrac{1}{2}\vec{w}^2 \\ s.t.\ y_i(<\vec{w}, \vec{x}_i> + b) \geqslant 1, i = 1, 2, \cdots, n \end{cases} \tag{9}$$

求出 $\vec{w}$，得到一个分类超平面 $f(\vec{x}) = \vec{w} \cdot \vec{x} + b$，使得一个未知样本 $\vec{x}$ 代入后，当 $f(\vec{x}) > 0$ 时，$\vec{x}$ 属于正类，$f(\vec{x}) < 0$ 时为负类。

5 实验评价与结果分析

根据以上有效性分类模型与实验流程，将用两种不同领域的评论集做对比试验，巡演分类器测试分类器性能，验证论文提到的特征降维方法与自动标注方法。这里定义准确率的概念，如公式(10)所示。

$$Acc = \frac{\text{正确分类的语料数}}{\text{总语料数}} \times 100\% \tag{10}$$

实验结果分析分两步展示，验证自动标注的可行性与验证评论有效性分类模型的可行性。

5.1 基于自动标注的结果分析

按照上面提出的标注方法，设定两个阈值A和B，并且确定如下规则：第一，对于某条评论，若其总词数大于A，无论其是否含有领域词，都认为该条评论有用。第二，若其总词数不大于A，但是大于B，且含有领域词，则也认为该条评论有用。不满足这两个条件的，认为该条评论无用。对比自动标注与人工标注结果，得到自动标注准确率。由于同时考虑A与B难以测试A、B对评论标注准确率的影响，这里设定B固定为0，先考虑A变化时，自动标准准确率的变化，结果见表3。

由表3可以看出，对左旋肉碱评论集而言，A为70时，准确率达到最大；对联想而言，A为10时准确率达到最大。这里，设定B为0会将很多含有领域词但不是有用的评论误判为有用，因此，这里进一步固定左旋肉碱的A为70，联想的A为10，改变B的值，测试两个评论集的准确率，测试结果如表4所示。

结果表明，自动标注方法的准确率受到长度的影响较大，对左旋肉碱而言，A、B分别为70词和20词时标准准确率达到最高，对联想而言A、B分别为10词与3词的时候准确率达到最高。

表3 设定长度阈值A准确率

评论集		长度阈值A(词数)(这里固定B为0)							
		0	10	20	30	40	50	60	70
准确率(%)	减肥药减肥药	84	71.9	81	82.8	83.7	83.9	84.2	84.3
	联想	57.8	70.2	70	66.2	62.8	61.2	59.4	58.8

表4 设定长度阈值B准确率

评论集	长度阈值B						
减肥药	准确率(%)	10	20	30	40	50	60
		85.8	86.6	86.1	85	84.5	84
联想	准确率(%)	1	2	3	4	5	6
		70.2	70.8	71.4	71.4	70.8	71

5.2 分类器性能分析

为了更好地验证评论有效性分类模型,将基于信息增益的特征提取方法训练普通的分类器与基于有效性分类模型训练出的分类器性能进行对比。

从基于信息增益的特征提取方法及人工标签的常规分类方法上,由于联想笔记本的特征信息增益值普遍大于左旋肉碱的特征信息增益值,因此对于左旋肉碱评论集选择特征信息增益值大于 0.374 的特征作为特征选择阈值,对联想笔记本电脑评论选择权重大于 0.775 的特征作为特征选择阈值。根据该传统文本分类方法,实验得到左旋肉碱与联想的分类准确率分别为 82% 与 63% 。

从基于研究提出的结合通用领域词典与文档长度的特征提取方法及自动标注标签的分类上,以前文所述自动标注结果作为标签,训练分类器,对比自动标注标签求准确率,最后得到左旋肉碱与联想的准确率分别为 91.05% 与 76% 。实验的对比结果见表 5。

表 5 实验对比结果

评论集		减肥药	联想
准确率(%)	基于人工标签的分类	82	63
	基于自动标注的分类	91.05	76

由表 5 结果可以发现,基于人工标签的分类器准确率明显低于基于自动标签的分类器准确率,而联想评论集在分类准确率上普遍弱于左旋肉碱,可能存在以下原因,第一,联想评论集里有不少网络水军,不停的夸赞其他品牌的笔记本优秀,带来了大量的长评论,是无用评论被自动标注结果误判为有用;第二,人们对电子产品的了解程度较高,部分顾客评论含有较多专有领域词,其评论不长但很有深度,虽是有用评论,但由于通用领域词典不含专有领域词,而被误标注为无用评论。

6 结 论

研究提出了一种基于通用领域词与评论长度的特征提取方法,然后以此为基础设计出一种样本自动标注方法,并构建领域词库与在线评论有效性分类模型,最后以实际商品评论数据对提出的特征提取方法与自动标注方法的性能进行了验证,实现了预期的效果。从理论研究的角度,研究不仅在技术上改进了普通领域词典建立耗时长的问题,而且结合评论长度解决了领域词典难以跨领域的问题;其次,提出的样本自动标注方法,对基于监督的数据挖掘方法中人工标注训练集类别耗时较长的问题得到了较好的解决。从管理实践的角度,基于领域词典的有效评论分类方法,能够帮助企业更好的筛选出有价值的评论,从而为电商企业改善服务提供可信的评论分类模型,也为网络舆情的分析提供快速的分类模型。

但研究也存在一些局限性与不足,这也是未来继续研究的方向。第一,通用词典规模有限,因此在后续研究中可以扩充通用词典的规模,尽量减少因为词典涵盖面太窄而漏掉较多有用评论的问题;第二,研究结论虽然得出评论长度阈值对分类准确率有明显影响,但并未找到一个普适性的评论长度阈值做样本自动标注,省去人工测试最合适的样本长度阈值标注仍然须要人工标注进行对比的问题,这也是未来深入研究的方向;第三,在今后的研究中还可以考虑更多有用评论的特征,而不仅限于评论长度与领域词,从而进一步增强特征提取方法与自动标注方法的准确率。

参 考 文 献

[1] 李俊. 面向产品评论的意见挖掘研究综述[J]. 现代计算机, 2013(5): 11-16.

[2] 严建援, 张丽, 张蕾. 电子商务中在线评论内容对评论有用性影响的实证研究[J]. 情报科学, 2012, 30(5): 713-716.

[3] Harrison-Walker L J. The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents [J]. Journal of Service Research, 2001, 4(1): 60-75.

[4] 吕品, 钟珞, 蔡敦波,等. 基于 CRF 的中文评论有效性挖掘产品特征[J]. 计算机工程与科学, 2014, 36(2): 359-366.

[5] Miao Q, Li Q, Zeng D, et al. Entity attribute discovery and clustering from online reviews [J]. Frontiers of Computer Science, 2014, 8(2): 279-288.

[6] Hu N, Koh N S, Reddy S K. Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales [J]. Decision Support Systems, 2014, 57: 42-53.

[7] 郝媛媛, 叶强, 李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报, 2010, 13(8): 78-88.

[8] 闫强, 孟跃. 在线评论的感知有用性影响因素——基

于在线影评的实证研究[J]. 中国管理科学, 2013(6): 126-131.
[9] 廖成林, 蔡春江, 李忆. 电子商务中在线评论有用性影响因素实证研究[J]. 软科学, 2013, 27(5): 46-50.
[10] 杨朝君, 汪俊奎. 商品在线评论有用性——基于品牌的调节作用分析[J]. 现代情报, 2014, 34(1): 123-127.
[11] 郑丽娟, 王洪伟, 郭恺强. 中文网络评论的情感分类: 句子与段落的比较研究[J]. 情报学报, 2013, 32(4): 376-384.
[12] 郁亚辉. 产品评论特征及观点抽取研究[J]. 情报学报, 2014, 33(3): 326-336.
[13] 龚静, 胡平霞, 胡灿. 用于文本分类的特征项权重算法改进[J]. 计算机技术与发展, 2014(9):128-132.
[14] 郑丽娟, 王洪伟, 郭恺强. 中文网络评论的情感分类:句子与段落的比较研究[J]. 情报学报, 2013, (4):376-384.
[15] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

(责任编辑 马 兰)

第二十一届中国竞争情报年会征文通知

由中国科技情报学会竞争情报分会(SCIC)主办的“中国竞争情报年会”是情报和信息领域分享学术研究成果、交流竞争情报实践的盛会,已成为业界品牌,吸引了情报和信息界、咨询界及企业界的专家学者和实践者的积极参与,并引起了社会和媒体的广泛关注。2015年度第二十一届年会将于2015年11月4-6日在西安举办,内容包括:大会报告、多场专题报告、互动论坛、学术论坛和成果展示。大会期间,我们将组织专家对第二十一届年会投稿论文进行评选,设立一等奖、二等奖和三等奖。会议期间将设论文宣讲论坛,举行获奖论文颁奖仪式,出版论文集。欢迎大家围绕当前国家战略与经济热点,从理论角度探讨竞争情报研究与工作的战略定位和升级转型问题,探索情报服务智库化发展的新路子,为情报工作可持续化发展提供理论支撑。可围绕(1)竞争情报理论与方法;(2)战略情报与竞争战略;(3)信息资源与搜集方法研究;(4)竞争情报分析方法;(5)竞争情报系统研究;(6)竞争情报组织模式和激励机制研究;(7)竞争情报教育与能力培养;(8)竞争情报案例研究;(9)国家安全与商业秘密保护;(10)竞争情报趋势研究等其他有关情报、商业情报、竞争情报等国内外竞争情报的发展、自身的研究与实践成果积极撰写稿件。论文截稿日期:2015年9月15日。

1. 来稿请发至:scic@onet.com.cn(主题为“二十一届年会征文”)
   联 系 人:刘玉、殷锦红
   联系电话:(010)68961820
   传 真:(010)68962474
2. 论文录用函与会议邀请函,将于2015年10月15日开始通知作者。
3. 论文格式与投稿详情以及往届年会及本届年会筹备请参阅分会网站。(http://www.scic.org.cn)