DOI:10.16515/j.cnki.32-1745/c.2019.01.013

在线评论文本特征表示方法研究

王倩倩,陈 康

(金陵科技学院人文学院,江苏 南京 210038)

摘 要:随着新兴技术与新的商业交易模式不断涌现,网络交易激增,网络交易评论也呈现出爆发式增长态势。针对大数据环境下网络评论文本空间高维的现象,提出借助商品标题和商品描述进行二重筛选的网络评论文本特征表示方法。该方法借助种子词而不是主题词典描述文本特征,降低了文档维度,减少了迭代次数,提高了在线评论文本分类的速度;同时,在文本映射时经过直接映射和间接映射二重筛选,减少了文本分类的疏漏,提高了文本分类的精度。

关键词:文本表示;种子词;词聚类;文本分类;降维;在线评论;文本特征

中图分类号:H08;TP391.1 **文献标识码**:A **文章编号**:1673-131X(2019)01-0056-05

Research on the Representation Method of the Text Features of Online Reviews

WANG Qian-qian, CHEN Kang

(Jinling Institute of Technology, Nanjing 210038, China)

Abstract: With the emergence of new technologies and new commercial transaction modes, the online transaction and the online transaction reviews has presented an unprecedented explosion growth. Aiming at the high dimensional phenomenon of the text space of online reviews in big data environment, this paper proposed a representation method, the text features of network reviews based on the double screening with the help of the title and the description of goods. The proposed method uses seed words to describe text features, does not need to use the theme dictionaries, thus lowers the dimension of the document, reduces the number of iterations, then improves the speed of the text classification of the online reviews; meanwhile, when mapping the text, namely direct mapping and indirect mapping, after the double screening, this method reduces the omissions in text categorization and improves the accuracy of text classification.

Key words: text representation; the seed word; word clustering; text categorization; dimension reduction; online reviews; text features

随着新兴技术与新的商业交易模式不断涌现,尤其是手机购物的兴起,便捷的购物方式使得网络交易呈现出爆发式增长态势。急速增长的网络交易量必然会产生大量的网络交易评论,这些交易评论对消费者的决策产生了重要影响。研究表明,购物时如信息不对称情况相对严重及羊群效应存在时,消费者的网购行为受网络评论的影响更大[1]。但是网络评论形成速度快、语言随意且多变,此外一些热门产品的评论数量巨大,这为获取有价值的信息带来了一定的困难。从大量的评论中找出有价值的关

收稿日期:2018-12-15

基金项目:金陵科技学院博士科研启动基金项目(jit-b-201622);江苏高校哲学社会科学研究基金项目(2017SJB0488)

作者简介:王倩倩(1985-),女,安徽淮南人,讲师,博士,主要从事数字出版、电子商务、数据挖掘研究。

键信息，帮助用户快速进行购物决策、减少用户对评价内容的参考成本，评论文本的特征表示就显得尤为重要。特征表示的好坏影响着分类器的分类精度和泛化性能，并直接影响着人们能否快速准确地获取自动摘要、辨析评论中的情感褒贬倾向等有用的信息[2]。

一、文本特征表示方法

目前，在评论文本信息处理时通常采用向量空间模型来描述文本向量。由于直接用分词和词频统计获取文本向量中的各个维，所以数据会很大[3]，因此需要找出最具代表性的文本特征，通过特征表示来对文本向量进行降维。

目前，对在线评论文本特征表示的研究主要集中在两个方面。一是通过构造各类评估函数，直接从原始特征中挑选出一些具有代表性的字、词或者词组、短语作为特征[4]，如信息增益法、互信息法、文档频率法等。但是由于词语本身存在同义、多义以及对短语和上下文的依赖，单纯将词语孤立地进行研究，破坏了文档中的相关关系和语义特征，导致这种提取存在较大的局限性[5]。二是采用映射或者变换的方法把原始特征变换成较少的新特征来对文本进行降维，如主成分分析法、潜在语义搜索法等。也有学者抽取《hownet 概念词典》中的概念作为特征来构成文本向量[6]。由于概念空间比词空间小，而且各分量之间相对独立，因此，概念特征比词特征更适合表示文本内容。但是概念词有限，不能涵盖网上出现的大量新词，尤其不适用在线评论这类发表自由、网络口语使用频繁的文本。鉴于在线评论灵活多变的特性，有学者提出了基于文本发现的 Web 表示方法，即用词和新词共同作为 Web 文本特征的表示项，从而提高了 Web 文本的表达能力[7]。但是新词依赖于原有的主题词典，召回率和准确率不高。

总而言之，网络产品评论具有句子较短、断句随意、用词口语化和语法标点符号使用不规范等特点，要从内容和形式自由度高、数据噪声大的评论信息中提取关键特征较为困难。因此，本研究在特征词选择算法的基础上，提出了将商品标题和商品描述作为训练集的文本特征表示方法，即使用可扩展的支持向量机（Lengthen Support Vector Machine，LSVM）方法。该方法不借助主题词典，先从商品标题和商品描述这些训练集中对词的贡献情况进行分析，测试集和训练集可能有不同的来源，是通过不同的途径取得的，换句话说，二者本来就是分开的。因此，本研究没有单独的测试集，而是通过词聚类生成表示某一主题的种子词，然后用种子词作为文本的特征项，最后得到评论文本的特征向量描述。具体研究思路如图 1 所示。

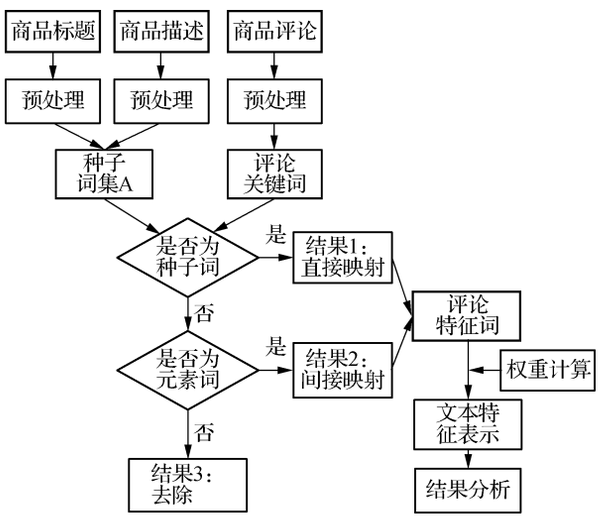

图 1　文本特征表示思路

二、训练集种子词选取

词聚类是从语义上通过词与词之间的距离来判断表达的意义是相同、相近还是不同。本文采用 k 模聚类[8]对范畴属性进行聚类，即使用一个较小的集合，用每个类的高频词来表示这个类，用种子词来代表其他近义词以表达某一主题的内容。例如：尺码（肥大、大码、宽松、紧身、长款）是以种子词“尺码”表示一个词类，词类中的任意一个词为该词类的一个元素，即“肥大”“大码”“宽松”“紧身”“长款”分别为种子词“尺码”表示的词类的一个元素，或者说是种子词“尺码”的一个元素。对训练集种子词的选取分两步进行：

一是从商品标题中选取种子词。因为商品标题需要尽可能多地涵盖商品属性，这样才更有可能被用户检索到，加之商品标题往往比较短，一般 30 字之内，因此从商品标题中选择特征词作为种子词，只要通过分词和去除停用词①就可以得到。如一件连衣裙的标题是“2017 秋冬装新款韩版时尚优雅气质大码显瘦收腰长袖打底连衣裙女”，我们从这条标题中得到的种子词为：秋冬装、新款、韩

① 停用词（Stop Words）是指在信息检索中为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉的某些字或词。

版、时尚、优雅、气质、大码、显瘦、收腰、长袖、打底、连衣裙、女。

二是从商品描述中选择特征词作为种子词。商品描述是卖家在商品详情页面写给买家看的有关商品特点、质地、款式等信息的陈述,包含着重要的商品特征内容。从商品描述中选取种子词可以减少对大量评论文本检索的时间。具体做法如下:分别对商品标题、商品描述进行文本预处理,分词,去除停用词,计算词频,设置一个阈值,选择频度大于阈值的词作为种子词,得到的两组种子词进行合并,取并集,得到最终的种子词集 $A(zc_1,zc_2,zc_3\cdots\cdots zc_k)$,$zc_1$、$zc_2$、$zc_3\cdots\cdots zc_k$ 为种子词。

三、文本特征的具体表示

所谓特征表示,就是从多个度量值集合中,按某一准则选取出供分类用的子集,以其作为降维的分类特征[9]。本研究在 SVM 模型的基础上,采用结合商品标题和描述的词聚类方法,将关键词对应到特征空间,对评论文本的特征向量进行描述。

(一)种子词文本映射

从商品标题和商品描述中得到种子词集 $A(zc_1,zc_2,zc_3\cdots\cdots zc_k)$,$zc_1$、$zc_2$、$zc_3\cdots\cdots zc_k$ 为种子词。对于评论文本 d 进行分词处理,去除停用词,选择频度大于阈值的词作为关键词,设为 c_1、$c_2\cdots\cdots c_h$。这些关键词可以映射到评论文本的特征空间,即:假设 c_1 在种子词集 A 中能找到,其与种子词 zc_1 相同,则直接将其映射为文本 d 的特征词;假设 c_2 不是种子词,但是种子词 zc_2 的元素词,则将 c_2 映射为种子词 zc_2,再映射为文本 d 的特征词 c_2';假设 c_3 在种子词集 A 中找不到,即 c_3 不是种子词也不是种子词的元素词,则将其去除(图 2)。

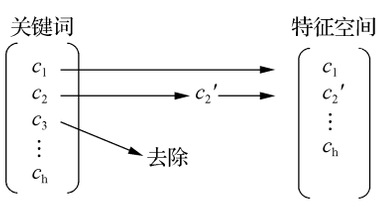

图 2 种子词文本映射示意

(二)权重计算

根据映射规则,将权重按照下面的公式进行计算,共三种情况:

第一,若 c_i 为种子词,设 $G(c_i)$ 为此类权重因子,$H(c_i)$ 为词 c_i 在文本 d 中所含的信息量。权重计算公式为

$$Q(c_i)=G(c_i)H(c_i)$$

第二,若 c_i 不是种子词,但是种子词 zc_i 的元素词,则用距离系数来计算权重。根据 WordNet 的语义相似度计算两个词之间的距离[10],用 $Dist(x_i,x_j)$ 表示两个词之间的距离,$Dist(x_i,x_j)$ 越小,表明两个词之间的相关度越大,故用 $1/Dist(x_i,x_j)$ 表示两个词之间的相关度,权重计算公式为

$$Q(c_i)=G(zc_i)H(c_i)/Dist(x_i,x_j)$$

第三,在实际中还可能出现两个或多个关键词生成相同特征词的情况。如 zc_i 是种子词,c_j 不是种子词但却是种子词 zc_i 的元素词,则这两个词映射的特征词都是同一个词 zc_i。这个特征词的权重为两个效应的叠加,权重计算需要依赖相似度进行计算,公式为

$$Q(c_i)=(1/Dist(x_i,x_j)+1)G(zc_i)H(c_i)$$

(三)文本特征表示

从以上两步得到种子词以及种子词所在类的元素词,可以表示成种子词的向量,用来表示文本特征。文本特征表示一般采用向量表示法,就是将文本表达为一个向量,看作是向量空间中的一个点,采用文本特征计算权重。由于文本特征表示包含种子词和元素词,因此,本研究中括号前的词为种子词,括号里面的词为元素词。根据以上距离,计算出权重,最终得到评论文本 d 的向量描述为

$$d=[(c_1,Q(c_1)],[c_2,Q(c_2)],[c_i,Q(c_i)],\cdots,[c_h,Q(c_h)](h\text{ 为正整数})$$

四、实证及讨论

淘宝网是中国最大的 C2C 电子商务网站,比其他网站有着更多的用户评论和更高的用户认可度,而且其商品评论必须是用户购买后才能评价,所以本文以淘宝网的网络评论为对象进行分析。从淘宝网选取销量较好的一款连衣裙获取评论语料库,规定字数大于或等于两个汉字(四个字节)的评论才算是有效评论,因此设置一个阈值:评论字数大于或等于两个汉字则进行提取,否则视为无效评论,最终得到评论 14 235 条,同时拷贝保存该商品的商品标题和商品描述文本。

(一)实证过程

1.种子词选取。先从商品标题“2017 秋冬装新款韩版时尚优雅气质大码显瘦收腰长袖打底连衣裙女”获取关键词作为一组种子词,“秋冬装、新

款、韩版、时尚、优雅、气质、大码、显瘦、收腰、长袖、打底、连衣裙、女”；再从商品描述中获得一组关键词为种子词，“长袖、面料、摸、舒服、垂感、厚、性价比、很高、大码、收腰、显瘦、秋冬、韩版、连衣裙”；最后将这两组种子词取并集，获得最终的训练集种子词集A(秋冬、韩版、大码、显瘦、收腰、长袖、连衣裙)。

2. 测试集文本预处理。对于该商品的网络评论，采用中科大 ICTLAS 分词系统对文本进行分词，去除语气助词、副词、介词、连词等无明确意义词，还有常见的“的”“在”等停用词，标注后的格式如表 1 所示。经过分词后的网络商品评论，去除停用词，选择频度大于阈值的词作为关键词。这些关键词可以采用本文定义的方式映射到评论文本的特征空间。

表 1　网络商品评论文本分词

评论 1	面料/n、摸/v、起来/vf、很/d、舒服/a、垂/v、感/vg、很/d、好/a、厚/a、很/d、适合/v、秋冬/t、穿/v、性/ng、价/n、比/p、很/d、高/a
评论 2	没有/v、色差/n、颜色/n、跟/p、图片/n、一样/uyy、的/udel、质量/n、也/d、很/d、好/a、大小/n、长度/n、刚好/d、我/rr、158/m、穿/v、M/x、码/v、很/d、好/a、看/v
评论 3	物流/n、很快/d、质量/n、也/d、挺/d、好/a、的/udel、裙子/n、很/d、漂亮/a、也/d、很/d、好/a、同事/n、都/d、说/v、好看/a、穿/v、起来/vf、也/d、显/v、瘦/a、非常/d、满意/v
评论 n	……

注释：“/”后的字母为对应词语的词性。

3. 种子词文本映射及权重计算。经过种子词提取，从商品标题和商品描述中得到 3 类种子词集合：尺码(0.18，肥大、大码、宽松、紧身、长款)；颜色(0.21，色差、色正、抬肤色、显黑)；价格(0.13，性价比、贵、便宜、划算、公道)。(括号前的词为种子词，括号里面的词为元素词，数字为种子词的权重因子)

从商品评论中获得了 5 个关键词：尺码(0.63)，颜色(0.77)，色差(0.43)，便宜(0.27)，收腰(0.21)。(括号中为权重因子，表示关键词的信息量)权重因子通过 TF-IDF 的统计方法，评估这 5 个关键词对于该商品评论语料库的重要程度[11]。根据公式得出颜色与色差的距离 $Dist(x_i,x_j)$ 为 1.34，便宜与价格的距离 $Dist(x_i,x_j)$ 为 1.65，并对这 5 个关键词进行映射，分别计算它们的权重。

第一，“尺码”是种子词，因此映射后的特征词仍为尺码，它的权重是

$$Q(c_i)=G(c_i)H(c_i)=0.18\times0.63=0.113\ 4$$

第二，“便宜”不是种子词，而是种子词“价格”的元素词，映射后的特征词为“价格”，它的权重为

$$Q(c_i)=G(c_i)H(c_i)/Dist(x_i,y_i)$$
$$=(0.13\times0.27/1.65)=0.021\ 3$$

第三，“颜色”是种子词，“色差”是元素词，它们都分别映射到种子词“颜色”上，因此特征词“颜色”的权重为

$$Q(c_i)=(1/Dist(x_i,y_i)+1)G(c_i)H(c_i)$$
$$=(1/1.34+1)\times0.21\times0.77=0.282\ 4$$

第四，“收腰”既不是种子词也不是元素词，将其删除。

最终评论文本 d 的向量描述为

$d=$[(尺码,0.113 4)、(颜色,0.282 4)、(便宜,0.021 3)]

(二)结果及分析

1. 性能对比分析。本文以 TP=(原始特征词数−结果特征词数)/原始特征词数×100% 作为衡量该方法优劣的标准。传统的方法中比较著名且效果较好的有 IG(信息增益法)和 CHI(卡方统计法)，本文将这两个方法与 LSVM 方法进行比较观察它们的优劣性，采用的指标是宏平均准确率 MP、宏平均召回率 MR 和宏平均 F_1 值 MF_1，实验结果如图 3、图 4 和图 5 所示。为方便绘图，将横坐标的特征维数缩小 1 000 倍，故横坐标 1—9 对应为 1 000—9 000 维。

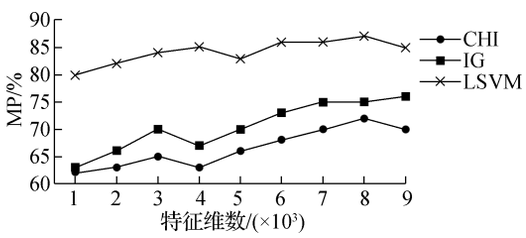

图 3　宏平均准确率 MP 比较结果

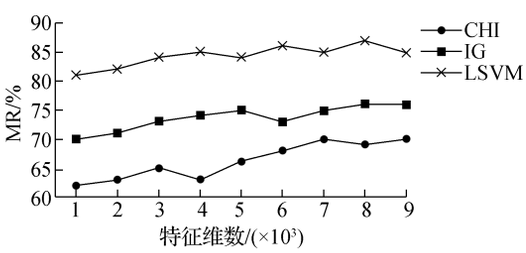

图 4　宏平均召回率 MR 比较结果

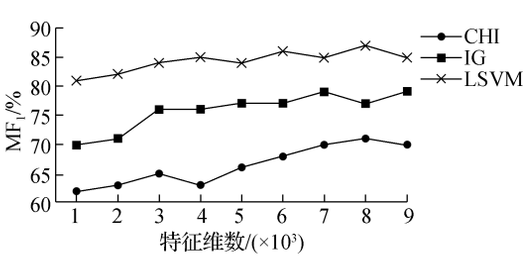

图 5　宏平均 F_1 值 MF_1 比较结果

从图中可以看出,上述几种评价函数均试图通过概率找出特征与主题类之间的联系,本文提出的LSVM方法比传统的IG和CHI分类算法准确率有所提高。IG方法只能考察特征对整个系统的贡献,而不能具体到某个类别上,这就使得其只适合用来做"全局"的特征选择,对判断文本类别贡献不大;而且其引入不必要的干扰项,使处理类分布和特征值分布高度不平衡的数据时选择精度下降。IG方法在高维空间上的效果比在低维空间好,在8 000维以上能够达到70%以上,可是在2 000维以下却不到70%,受维度影响较大。CHI方法随着维度的增加分类效果变好,但是总体较差。而LSVM方法在不同的维度下分类结果基本保持稳定,准确率MP值和召回率MR值均维持在80%以上。因为该方法将商品标题和商品描述考虑进来,通过训练集已经得到了比较稳定和准确的种子词作为特征词,同时,其在映射时,经过了"是否为种子词"和"是否为元素词"的两次映射判断,减少了文本分类的疏漏,提高了准确率。

2.效率对比分析。将LSVM方法与IG、GHI两种方法进行消耗时间的对比实验,选取前2 200维数据分成11组进行计时统计,每扫描200维统计一次,时间记录单位为毫秒,耗费时间如图6所示。

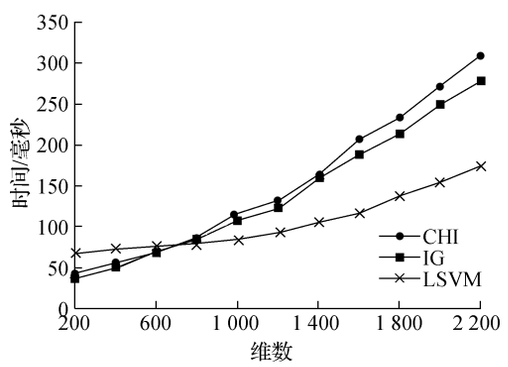

图6 消耗时间比较结果

从图6可以看出,LSVM方法消耗的时间较少,尽管在开始阶段由于需要扫描商品标题和商品描述,时间对比相差不大,但在600维及以后的评论挖掘时间上远远小于其他两种方法。这是由于简化了文本表示过程,使其在描述文本特征时不需要借助主题词典,减少了迭代次数,从而大大提高了文本分类的效率,缩短了信息挖掘时间。

五、结语

文本特征表示是文本分类的一项基础工作,直接关系到文本分类的效果。本文提出的基于商品标题和描述的在线评论文本特征提取方法(LSVM),能够利用商品属性之间的关联,快速找到文本中的同义词,并对它们进行聚类,从而简化了文本表示过程,提高了文本分类的效率。同时,LSVM方法在文本映射时经过两次判断,即直接映射和间接映射二重筛选,减少了文本分类的疏漏,提高了文本分类的精度。本研究成果还可以应用于大众点评网、饿了么、airbnb(爱彼迎)等其他行业的网络评论,并可对这些领域的评论进行舆情分析与检测,提高平台的市场决策能力,同时,平台还可以利用文本分类的结果对电商用户的购买决策进行个性化推荐。

参考文献:

[1] 游贵荣,吴为,钱沄涛.电子商务中垃圾评论检测的特征提取方法[J].现代图书情报技术,2014(10):93-100

[2] 李文慧,张英俊,潘理虎.多因素影响特征选择的短文本分类方法[J].计算机系统应用,2018(12):216-221

[3] Liu N,Zhang B Y,Yan J,et al. Text Representation: from Vector to Tensor[EB/OL]. (2012-08-10)[2018-11-12]. http://www.connex.lip6.fr/~gallinar/Enseignement/2009-Papiers-ARI/icdm2005-Liu-Tensors.pdf

[4] 孙水华,丁鹏,黄德根.利用句法短语改善统计机器翻译性能[J].中文信息学报,2015(2):95-102

[5] 马双刚.基于深度学习理论与方法的中文专利文本自动分类研究[D].镇江:江苏大学,2016

[6] 路永和,梁明辉.遗传算法在改进文本特征提取方法中的应用[J].现代图书情报技术,2014(4):48-57

[7] 吴春颖,王士同,蔡崇超.一种基于新词发现的Web文本表示方法[J].计算机应用,2008(3):764-767

[8] 蔡晓妍,戴冠中,杨黎斌.谱聚类算法综述[J].计算机科学,2008(7):14-18

[9] 苏丹,周明全,王学松,等.一种基于最少出现文档频的文本特征提取方法[J].计算机工程与应用,2012(10):164-166

[10] 张思琪.基于WordNet的语义相似度计算方法的研究与应用[D].北京:北京交通大学,2016

[11] 武永亮,赵书良,李长镜,等.基于TF-IDF和余弦相似度的文本分类方法[J].中文信息学报,2017(5):138-145

(责任编辑:李海霞)