基于词向量模型的情感分析①

魏广顺[1,2], 吴开超[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院计算机网络信息中心, 北京 100190)

摘 要: 随着移动互联网的发展, 以商品评论等带有主观性的短文本信息急剧增加. 海量的文本信息使得人工管理越来越困难. 本文以商品评论为研究对象进行情感分析. 针对商品评论为短文本的特点, 本文在词向量的基础上提出了词向量叠加方法和加权词向量方法进行文本特征的提取, 从而更深层次的提取短文本特征. 在进行评论情感分析模型性能的比较中, 说明了本文所提方法的有效性. 基于情感分析技术可以解决人工难以胜任的海量商品评论的分类, 方便用户快速获取有效信息.

关键词: 情感分析; 加权词向量; 商品评论; 短文本

Sentiment Analysis Based on Word Vector Model

WEI Guang-Shun[1,2], WU Kai-Chao[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: With the development of Internet, text information, such as product review, increases rapidly. The mass text information makes it more difficult to make artificial management. Considering that product reviews are short text, this paper comes up with the method of word vector superposition and weighted word vector. In the result of sentiment analysis, the method is proved effective. Emotional analysis technology can solve the difficulty of artificial classification in the mass of product review, and help users to get information quickly.

Key words: emotion analysis; weighted word vector; product review; short text

随着互联网的快速发展, 推动了像淘宝、亚马逊、京东等电子商务网站的发展. 这些购物网站特别强调用户的参与, 为用户发表自己对商品的评价提供了在线评论机制. 这些评论不仅为厂家和商家提供了传统交易中难以获取的反馈信息, 而且影响着后续消费者的购买行为. 但是随着商品评价信息数量的快速增长, 使得人工判断这些杂乱无章的评论信息的主观情感倾向性越来越困难. 因此, 构建一个自动的商品评论文本的情感倾向性分类系统为消费者和商家提供在线评论的倾向性分析是很有必要的.

情感分析也称为观点挖掘、意见挖掘等, 是指通过分析文本中的统计和语义等信息, 挖掘出文本中所蕴含的情感倾向, 如消极、积极、中立等. 情感分析作为自然语言处理中的一个重要分支, 在越来越多的领域被应用, 如: 舆论监督、市场反馈、品牌营销、信息检索等. 按照处理文本的粒度可以将情感分析分为词语级、短语级、句子级、篇章级和多篇章级等[1]. 通过情感分析可以为用户决策提供依据. 情感分析方法主要分为基于语义的方法和基于机器学习的方法[2]. 本文主要研究商品评论的情感分析, 属于基于篇章级的研究. 通过情感分析方法构建情感分类系统, 可以实时的对用户评论进行分类, 为解决网上杂乱无章的文本信息提供了一种有效的方法.

商品评论作为一种用户反馈信息通常较短, 属于短文本. 短文本是指文本长度较短, 一般不超过 100 个字符[3]. 商品评论与其他文本信息(如新闻等)相比有其独有的特点: 表达不规范; 网络用语较多; 内容较短等. 商品评论作为一种消费者对购买物品的评价,

① 收稿时间:2016-06-21;收到修改稿时间:2016-08-18 [doi:10.15888/j.cnki.csa.005655]

带有明显的主观性情感倾向. 如“物流快, 服质优, 商品人性化定造, 设计美观大方, 尺码合适, 非常满意!”, 这条商品评论带有用户对所购买商品的的主观情感倾向. 这就为情感分析研究提供了可能性.

本文主要结构如下: 第一部分介绍情感分析领域的相关工作; 第二部分主要介绍基于词向量模型的文本特征提取方法; 第三部分实验结果对比分析; 最后对本文工作进行总结.

1 相关工作

情感分析方法主要分为两种: 基于语义的方法和基于机器学习的方法. 基于语义的方法主要是通过情感词典, 计算文本的情感值来进行确定文本的情感倾向[4]; 基于机器学习的方法主要是通过提取文本中的特征, 利用机器学习中的分类算法通过一定规模的样本训练来构建模型, 从而预测新文本的情感倾向[5,6].

基于语义的方法可以充分利用人工构建的情感词典, 这些情感词典中的词往往是那些能够明确反应人的情感倾向的词. 但是情感词典不可能包括所有的情感词, 而且网络用语多样化使得情感词典的构建难度增加. 基于机器学习的方法是通过机器学习算法学习给定训练集的特征来构建模型. 一般在文本分类中常见的机器学习算法包括决策树、KNN、Logistic 回归、支撑向量机(SVM)等. 在实际研究和实验中支撑向量机(SVM)被证明在情感分析中相对于其他方法具有更优的效果[7].

基于机器学的情感分析方法是一种统计学习方法, 需要对文本进行向量化, 从而更好的利用机器学习算法. 传统的向量空间模型(VSM)[8]是将文本看作一组词组成的序列, 通过有效的特征词选取方法如文档频率、信息增益(IG)、卡方统计等, 选取适当数量(N)的特征词. 这些特征词组成一个 N 维欧式空间, 每一篇文档被以向量[$W1$,$W2$,…,Wn]映射到这个 N 维空间中. 其中 Wi 表示文档中第 i 个特征词在空间的第 i 维坐标的权重值. 一般用 TF-IDF 作为权重. 一些传统的情感分析研究都是基于 VSM 模型进行的研究[9]. 向量空间模型一般维数在几千维甚至上万维, 对于商品评论这种短文本会形成特征的稀疏性问题, 即在文本向量化后会出现特别多权值为 0 的维度.

为了解决商品评论在向量空间模型中特征稀疏性问题, 本文提出了基于词向量模型[10]的特征提取方法. Bengio 等提出了一种神经网络语言模型 NNLM(Neural Network Language Model)用于预测在给定上下文的情况下生成当前词的概率[11]. 这个模型同时也成为了词向量模型的基础.

2 基于词向量模型的评论特征提取方法

传统的文本特征提取方法是基于向量空间模型的, 即将文本看作无序的词组成的序列. 这种向量空间模型存在数据稀疏、丢失词序信息的缺点. 为了解决向量空间模型的缺点, 出现了将词法和句法等一些复杂的文本特征加入到文本特征提取中的方法. 随着越来越多的特征加入, 使得基于机器学习的文本分析方法的性能得到极大的提升. 本文以词向量为基础, 将文本的特征表达引入到词向量空间. 并在词向量空间中对文本进行了多种方式的特征提取方法.

2.1 Word2vec 词向量模型

Word2vec 是 2013 年由 Google 开源的一款将词表示为实数向量的高质量工具, 是 Mikolov 等所提出的词向量模型的一种实现. Word2vec 是一种无监督学习工具, 它以未进行人工标记的语料作为训练集, 通过神经网络将词映射到一个 K 维欧式空间. 词向量在 K 维欧式空间上的特征同时反应了词之间的特征.

由于 Word2vec 学习的是语料中文本的语义关系, 这就要求用作训练的语料要充分大, 从而保证词向量的质量. 本文利用 Word2Vec 工具对 2 千万条商品评论进行训练, 最终得到一个 500MB 的词向量模型. 词向量在 K 维空间上的相似度, 同时反映了词在文本中的相似度关系. 可以通过计算词之间的相似度来说明此词向量模型的有效性.

表 1 “服务态度”最相似的 5 个词及其相似度

服务态度	相似度
服务	0.8988
态度	0.8965
服务质量	0.6776
服务业	0.5955
服务周到	0.554

由上表可以看出, 通过 2 千万条商品评论训练得到的词向量模型可以充分保证其词向量的质量.

2.2 词向量叠加文本向量化方法

词向量模型可以将每个词表示为一个 K 维的向量. 商品评论可以看作词序列化的表示, 一种简单的将商

品评论向量化的方法是将词向量进行拼接. 即将一个有 n 个不同词的商品评论表示为一个 $n\times K$ 维的向量. 这种方式的缺点是当 n 取值很大时, 会得到一个维度特别高的向量, 造成维度灾难; 每一条商品评论所含词的个数也不相同, 这会造成商品评论向量化之后维度的不一致.

为了解决词向量拼接方法的缺点, 本文首先提出将商品评论中词的词向量叠加来得到商品评论的向量化表示. 词向量叠加后会得到一个维数与词向量同维度的商品评论的实数化向量. 如评论“好吃, 便宜, 收银员态度很好, 总体来说是很好”, 分词后为[好吃, 便宜, 收银员, 态度, 很好, 总体, 来说, 是, 很好]. 每个词可以表示为一个 K 维的向量, 将“好吃”、“便宜”等这些词的词向量进行叠加, 得到一个 K 维的向量进行文本的向量化表达. 为了验证词向量叠加本文特征提取的有效性, 本文将其与传统的空间向量模型的文本情感分析效果进行了比较.

2.3 加权词向量文本向量化方法

TF-IDF 是在信息检索中的一个概念, 同时也被认为是信息检索领域最重要的发明[12]. 在搜索、分类等领域都有着广泛应用. TF 即 Term Frequency, 表示一个词在一篇文档中出现的频率. IDF 即 Inverse Document Frequency, 表示的是在文本集中多少篇文档包含该词, 是词的文档频率. TF-IDF 值为 TF 与 IDF 的乘积. 其既充分考虑了词在文档中的出现频率, 又充分考虑了词在整个文档集中的出现频率, 是一种对词在文本中的重要性比较综合的度量.

每个词在每条商品评论中的都有其重要性, 简单的将词向量相加将每个词在商品评论中的重要性视为相同, 丢失了词语重要性信息. 如评论“好吃, 便宜, 收银员态度很好, 总体来说是很好”, 分词后为[好吃, 便宜, 收银员, 态度, 很好, 总体, 来说, 是, 很好]. “很好”无论是在语义上还是在其权重上都对情感分析应该起到最重要的作用, 当采用简单的词向量相加时, 这种明显的特征词就会被视为与其他词一样来进行处理.

本文选取 TF-IDF 作为词在商品评论中的权重, 既充分考虑了词在当前商品评论中的重要性, 又充分考虑了该词在整个商品评论文档集中的重要性, 在对评论文本向量化的过程中保留了其重要性信息.

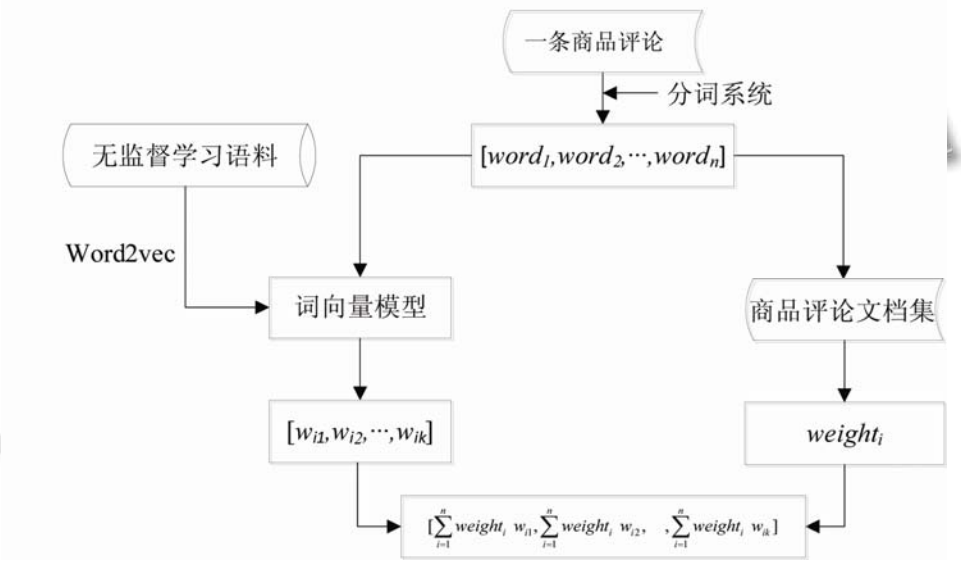

图 1　商品评论加权词向量流程图

为了充分利用商品评论中对情感分析起到更大作用词的信息, 本文进一步提出了一种加权词向量的方法. 此方法充分利用词在商品评论中的权重信息, 在将商品评论向量化的过程中, 将词在文档集中的 TF-IDF 值作为权重参与到向量化的过程中. 图 1 给出了将一条商品评论进行加权词向量表示的求法的流程图. $[word_1,word_2,\cdots,word_n]$表示一条商品评论分词后的结果. $W_i=[w_{i1},w_{i2},\cdots,w_{ik}]$表示 $word_i$ 在当前词向量模型中的向量化表示. $weight_i$ 表示 $word_i$ 在当前文档集中的 TF-IDF 值.

2.4 情感分析模型

支撑向量机[13]的基本思想是将输入空间中的输入转换到特征空间, 然后在特征空间中进行学习. 支撑向量机通过求解一个凸二次规划问题, 即:

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ s.t. y_i(w\cdot x_i+b)-1\geq 0, \quad i=1,2,L \quad N \end{cases}$$

在特征空间中找到一个最优分类超平面: w·x+b=0.

使得分类间隔最大化, 将样本分到不同的类别中.

SVM 算法解决的是二值分类问题, 当需要进行多值分类的研究时要通过构建多个二值分类的 SVM 模型以投票的方式进行解决.

在文本分类中, 有大量的分类算法, 如 KNN、Logistic 回归、决策树等. 但在大量的实验和研究中表明 SVM 在文本分类中较其他分类算法有更好的效果, 并且大量的文本分类研究都是以 SVM 为基础[14-16]. 本文以 SVM 算法构建文本分类器, 从而比较本文所提评论文本特征提取方法较传统空间向量模型的有效性. 图 2 为基于词向量模型的文本特征提取方法与线性支撑向量机算法构建分类器的流程图.

算法伪代码:

[1] 读取人工标注的商品评论

[2] 文本预处理, 分词、去除停用词等

[3] 商品评论初始化向量 *doc2vector*=[*0,0, …,0*]

[5] for $word_i$ in [$word_1$,$word_2$,…,$word_n$]

[6] if $word_i$ 在词向量模型中

[7] 取出 $word_i$ 的词向量 W_i

[8] 计算 $word_i$ 在文档集中的 tf_idf 值 $weight_i$

[9] $doc2vector= doc2vector+weight_i* W_i$

[10] SVM 算法进行模型训练得到情感分析模型

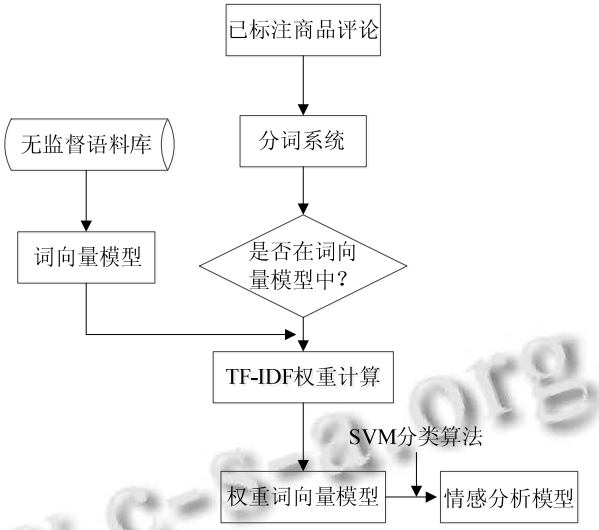

图 2 词向量模型评论情感分析流程图

3 实验结果比较

上节中介绍了两种文本特征提取方法:词向量叠加方法、加权词向量方法. 为了验证本文所提特征方法在情感分析上的有效性, 本文分别将两种特征提取方法与线性支撑向量机算法相结合, 构建情感分类系统, 并于传统的空间向量模型进行两个方面的比较: 计算量的分析和分类效果的比较.

3.1 计算量比较

本文使用 20000 条人工标注的商品评论作为实验数据. 评论的平均长度为 96 个字. 在试验中, 传统向量模型使用 8000 维特征词作为特征, 词向量模型维数为 300 维. 当对样本进行向量化后, 传统空间向量模型得到文件的大小约为 600MB, 而以词向量为基础的文本向量化方法得到的文件大小约为 25MB. 显然, 词向量模型在数据文本数据向量化后可以有效的减少数据的维数, 从而在小数据量的情况下加快分类器的训练速度. 模型训练时间, 以词向量为基础的文本向量化方法模型训练时间约为 17 秒, 而传统空间向量模型的模型训练时间约为 510 秒. 虽然在用大量无监督语料训练词向量时会耗费大量时间, 但是词向量的训

练是一次性的工作, 在模型训练和后期对新数据进行预测时一次性加载训练好的词向量模型即可. 因此本文比较分类器训练时间时不考虑词向量的训练时间.

3.2 分类效果比较

3.2.1 模型评估参数

在模型评估中采用 Precision、Recall 和 F1-Measure 作为模型的评价指标. 以下为各评价指标的说明.

表 2 混淆矩阵

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

TP 表示在样本中为正向, 被预测为正向的样本数;
FP 表示在样本中为负向, 被预测为正向的样本数;
FN 表示在样本中为正向, 被预测为负向的样本数;
TN 表示在样本中为负向, 被预测为负向的样本数.
两种类别的召回率分别为:

$$p\_recall = \frac{TP}{TP + FP}$$

$$n\_recall = \frac{TN}{TN + FN}$$

两种类别的准确率分别为:

$$g(x,\sigma) = e^{\frac{x^2}{2\sigma^2}} \quad g(x,y,\sigma) = e^{-\frac{(x^2+y^2)}{2\sigma^2}}$$

两种类别的 F 值分别为:

$$p\_F = \frac{2 * p\_recall * p\_precision}{p\_recall + p\_precision}$$

$$n\_F = \frac{2 * n\_recall * n\_precision}{n\_recall + n\_precision}$$

3.2.2 分类结果

本文使用 20000 条人工标注的商品评论作为实验数据. 其中 1 万条好评, 1 万条差评, 分别从好评和差评商品评论中取出 8000 条评论作为训练集, 2000 条评论作为测试集, 进行模型训练和模型评估. 模型评价指标采用 Precision、Recall 和 F1-Measure. 表 3 为各模型的评估结果.

表 3 实验结果

情感分析方法	Positive			Negative		
	Precision	**Recall**	**F 值**	**Precision**	**Recall**	**F 值**
传统空间向量模型**+SVM**	0.8143	0.8685	0.8405	0.8591	0.8020	0.8296
词向量叠加**+SVM**	0.8490	0.9025	0.8749	0.8959	0.8395	0.8668
权重词向量**+SVM**	0.8949	0.9325	0.9133	0.9295	0.8905	0.9096

4 结论

实验结果表明, 本文提出的以词向量为基础的文本向量化方法无论是在模型训练速度还是在分类效果都有更优的效果, 充分证明了本文所提方法的有效性. 商品评论是一种带有明显主观情感倾向的文本, 传统的向量空间模型在特征表示中丢失了大量统计和语义信息, 并且存在着特征稀疏性和高维度的缺点. 本文所提出的以词向量为基础进行文本向量化的方法, 通过词向量模型可以将向量控制在一个较小的维度并有效的解决了传统向量空间模型中的稀疏性问题; 通过权重可以保留词语在文本中的重要性信息.

参考文献

1 赵妍妍,秦兵等.文本情感分析.软件学报,2010,21(8):1834–1848.
2 张紫琼,等.互联网商品评论情感分析研究综述.管理科学学报,2010.
3 徐易.基于短文本的分类算法研究[硕士学位论文].上海:上海交通大学,2010.
4 林斌.基于语义技术的中文信息情感分析研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2006.
5 崔志刚.基于电商网站商品评论数据的用户情感分析[硕士学位论文].北京:北京交通大学,2014.
6 宋静静.中文短文本情感倾向性分析研究[硕士学位论文].重庆:重庆理工大学,2013.
7 张学工,等.关于统计学习理论与支撑向量机.自动化学报,2000.
8 Salton G, Wong A, Yang CS. On the specification of term values in automatic indexing. Journal of Documentation, 1973.
9 王素格.基于 Web 的评论文本情感分类问题研究[博士学位论文].上海:上海大学,2008.
10 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. NIPS, 2013.
11 Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. Meeting of the Association for Computational Linguistics. 2010.
12 吴军.数学之美.第 2 版.北京:人民邮电出版社,2014.
13 李航.统计学习方法.北京:清华大学出版社,2012.
14 叶志刚.SVM 在文本分类中的应用[硕士学位论文].哈尔滨:哈尔滨工程大学,2006.
15 伍岳.基于 SVM 的文本分类应用研究[硕士学位论文].成都:电子科技大学,2014.
16 张国梁,肖超峰.基于 SVM 新闻文本分类的研究.电子技术,2011.