第 43 卷 第 6A 期
2016 年 6 月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于 word2vec 的互联网商品评论情感倾向研究

黄 仁 张 卫
(重庆大学计算机学院 重庆 400044)

摘 要 在电子商务蓬勃发展的网络环境下,产品的评论数据已成为企业提高商品质量和提升服务的重要数据源。这些评论中包含用户对产品各个方面的情感倾向,对其进行情感分析可以帮助商家了解产品的优缺点,也能为潜在消费者的购买决策提供数据支持。提出了基于组合神经网络的商品属性聚类及基于 word2vec 的商品评论情感分析新方法,通过 word2vec 计算语义相似度,建立情感词典,用构建的情感词典对测试文本进行情感分类。实验验证了该方法在互联网商品评论中的有效性和准确性。

关键词 word2vec,情感倾向,情感词典,情感分类
中图法分类号 TP391 **文献标识码** A

Study on Sentiment Analyzing of Internet Commodities Review Based on Word2vec

HUANG Ren ZHANG Wei
(Department of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract With the rapid development of e-commence under the network environment, product review has become an important data source for enterprises to improve quality and enhance service. The review comprises user's emotional tendency in all aspects of the product. Emotional analysis can not only help business to understand the advantages and disadvantages of the product, but also provide data support for the potential consumer's purchase decision. This paper presented a novel method to cluster commodity attribute based on combination neural network and computd sentiment of internet commodities review using word2vec. This essay computed the semantic similarity and built emotional dictionary based on word2vec, then used the emotional dictionary to obtain the emotional tendencies of the test texts. The effectiveness and accuracy of the method is validated through experiments.

Keywords Word2vec, Emotional tendency, Emotional directory, Emotional classification

随着互联网电子商务的蓬勃发展,越来越多的人青睐于网络购物。为了提高客户满意度,网络商家通常允许客户对他所购买的商品进行评价,导致商品评价的数剧迅速增长。分析隐藏在这些主观性评论文本中的情感倾向,不仅可以为潜在消费者提供网购指导,而且能帮助生产商和销售商通过反馈信息来改进产品、改善服务,提高竞争力。情感分析就是一种对该类信息进行分类的方法,又称为意见挖掘,是指通过自动分析某种商品评论的文本内容,发现消费者对该商品的褒贬态度和意见[1]。

1 文本情感分析技术

目前,文本的情感分析研究内容主要分为 3 个方面[2]:文本内容的主客观分类、文本的情感倾向性分类和文本的情感强度计算。其中,本文研究的重点是情感倾向性分类,它的主要研究内容是通过分析主观性文本中的情感词将文本情感分为正面或负面两类。它的研究思路可以归纳为以下两种:

1)基于语义的方法。通过统计和分析文本中情感词的褒贬性判断文本的情感倾向。采用的方法主要包括基于语料挖掘的方法[3,4]和基于情感词典[5]的方法。

2)基于机器学习的方法。传统的机器学习方法主要应用在文本主题分类,它将 k 近邻(k-Nearest Neighbor, KNN)、支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(Naive Bayesian, NB)、最大熵等机器学习方法应用于情感分类。本文提出的商品评论情感分析流程如图 1 所示。

图 1 本文情感分析流程

2 基于组合神经网络的商品属性聚类

众所周知,任何一个词语都有与其相对应的语言环境,即通常所说的上下文[6]。如果两个词语之间的上下文越相似,那么它们在应用环境和语义上也越相似。一般地,可以将词语左右各 d 个词语作为其上下文,或称开辟了一个长度为 $2d$ 的词语上下文窗口。

本文提出了面向商品评论文本的属性自动聚类算法——组合神经网络算法,该算法的主要思想如下:

黄 仁(1962—),男,教授,硕士生导师,CCF 会员,主要研究方向为图像处理、嵌入式应用技术;张 卫(1990—),男,硕士,主要研究方向为数据挖掘、自然语言处理,E-mail:xiaowei.hongye@163.com。

1)对评论文本分词时,对分词后的词语进行词性和词频的标注。

2)抽选出标注为名词的词语作为候选属性,构成候选属性集。

3)由于统计了分词之后的词语的词频,采用哈夫曼编码对每个词语进行编码,以方便计算机处理。

4)设置上下文窗口参数,抽取出候选属性集中名词 w 在文本中的上下文。将词语 w 作为第一层 BP 神经网络的输入,词语 w 的上下文作为第一层神经网络的输出,对 BP 神经网络进行训练。因为每个候选名词在评论文本中会出现多次,用 BP 神经网络对这些多次出现的上下文进行训练,最终得到一个最能代表词语 w 的上下文窗口向量。

5)将 BP 神经网络的输出作为 SOM 自组织神经网络的输入,进行 SOM 神网络的学习,最终 SOM 神经网络的输出即为属性相似的词语的聚类结果。

组合神经网络的结构模型如图 2 所示。

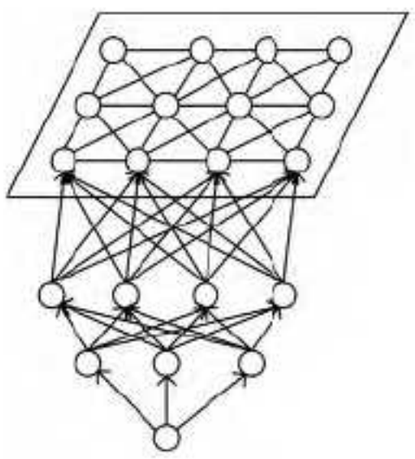

图 2 组合神经网络模型

3 word2vec 介绍

word2vec[7]是 Google 在 2013 年开源的一款将词表示为实数值向量的高效工具。通过训练可以把对文本内容的处理简化为 K 维向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度。word2vec 输出的词向量可以用来做很多 NLP 相关的研究,比如词聚类、找同近义词、词性分析等。

word2vec 生成词向量的基本思想来自于 Bengio 提出的 NNLM(Neural Network Language Model),其原理示意图如图 3 所示。

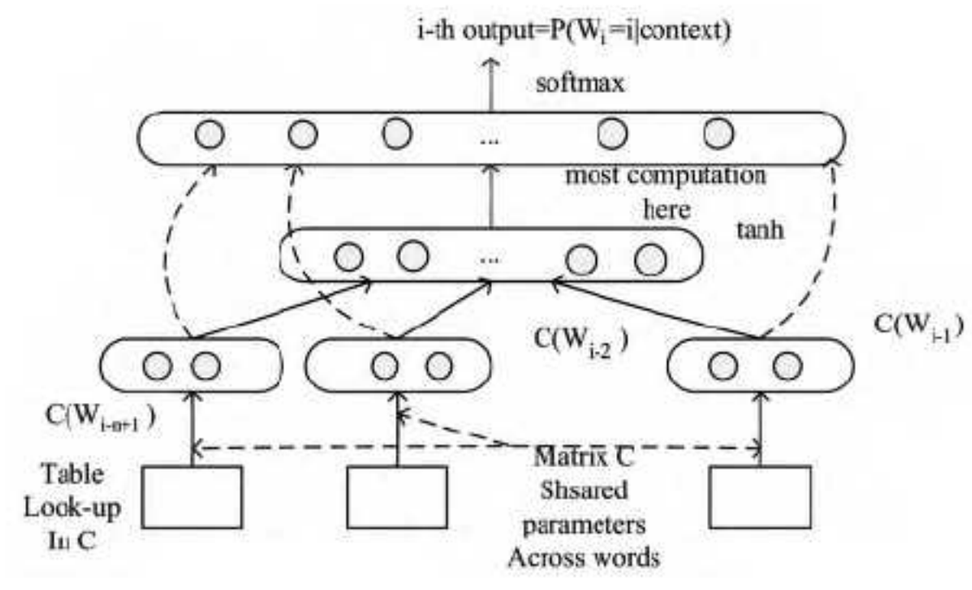

图 3 NNLM 原理模型图

图 3 中,每个输入词都被映射为一个向量,该映射用 C 表示,即 $C(W_{t-1})$ 为词语 W_{t-1} 的词向量。g 为一个前馈或递归的神经网络,其输出是一个向量,向量中的第 i 个元素表示概率 $P(W_t=i|W_1^{t-1})$,目标是学到一个好的模型:

$$f(w_t,w_{t-1},\cdots,w_{t-n+2},w_{t-n+1})=p(w_t|w_1^{t-1})$$

需要满足约束条件:

$$f(w_t,w_{t-1},\cdots,w_{t-n+2},w_{t-n+1})>0$$

$$\sum_{i=1}^{|V|}f(i,w_{t-1},\cdots,w_{t-n+2},w_{t-n+1})=1$$

word2vec 采用包含 CBOW(Continuous Bag-Of-Words)和 Skip-Gram 两种算法利用上下文信息来预测当前词的思想来生成词向量。将文本集作为输入,将每个词对应的生成向量作为输出。通过生成的词向量,可以计算与用户指定词语之间的距离(相似度)。比如用户指定输入"北京",将显示训练文本中与"北京"最接近的词语以及它们之间的余弦距离。利用生成的词向量,word2vec 采用 k-means 聚类算法可以实现大数据集的文本分类。

4 商品评论情感词典构建

随着互联网的发展,越来越多的网络用语出现在商品评论文本中,而且这些词汇不同于现有情感词典中的情感词,但却也表达了情感倾向。因此,有必要构建一个网络情感词典,以提高评论文本情感倾向的准确性。情感倾向判断包括两个部分:情感极性和情感强度[8]。情感极性表示情感是正向还是负向的,情感强度指的是相应的情感强烈程度。为了便于计算和分类,本文暂不考虑情感强度。用 EP 表示情感极性,EI 表示情感强度,EO 代表情感倾向,则一个情感词可以按如下表示:

$$EO(word)=(EP,EI),EP\in\{-1,1\} \tag{1}$$

集合 EP 中,-1 表示负向情感倾向,1 代表正向情感倾向。本文用基于语义的语义相似度计算方法来计算商品评论文本中的词语与情感词典中的词语的距离以判断语义倾向。利用 KNN 算法将词语的极性划分到与之最相近的 k 个词语中的大多数词语属于的类别当中。用 Pos 表示具有正向情感极性的词语的集合,Neg 表示具有负向情感极性的词语的集合,$SD(word1,word2)$ 表示 $word1$ 与 $word2$ 之间的语义距离:

$$EO(word)=\sum_{posword\in Pos}SD(word,posword)-\sum_{negworkd\in Neg}SD(word,negword) \tag{2}$$

我们将 0 作为分类正向极性和负向极性情感的阈值。EO(word)>0 时,word 的情感极性为正向,EO(word)<0 时,情感极性为负向。

在实验中,通过网络爬虫软件在某电子商务网站上爬取了 54368 篇商品评论的真实文本。将爬取的文本经过中文分词处理之后,用 50%的文本作为训练集训练 word2vec。经过 word2vec 训练之后,能得到每个词语的向量表示,计算两个向量的余弦值来表示两个词语的语义相似度距离,余弦值越大,表示两个词语的语义越相近。例如两个 n 维向量 $a(x_{11},x_{12},\cdots,x_{1n})$ 和 $b(x_{21},x_{22},\cdots,x_{2n})$,余弦值的计算公式如下:

$$\cos(\theta)=\frac{\sum_{k=1}^{n}x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n}x_{1k}^2}\sqrt{\sum_{k=1}^{n}x_{2k}^2}} \tag{3}$$

构建中文商品评论情感词典的步骤如下:

1)应用 50%的文本训练 word2vec,得到词语 $W(EP,EI)$

的向量表示并将其存入 vector.bin 文件中；

2)如果能在情感基准词典中找到词语 W，则可直接跳入步骤 5)标注 W 的情感极性；否则，跳入步骤 3)；

3)在 word2vec 中执行"./distance vector.bin"获取与词 W 最接近的 20 个词，在扩展情感词典中查找与 W 最接近的 20 个词语；

4)用式(1)计算 $SO-SD(W)$ 并判断 $SO(W).P$ 的极性值：$SO(W).P=SO$(与向量 W 的余弦值最大的词).P；

5)将 W 存入商品评论情感词典中。

5 商品评论情感分析

在一篇真实的商品评论文本中的句子不仅包括情感词，而且也包含了一些如程度副词、感叹词、否定词等影响判断情感词极性的因素。因此，判断商品评论文本的情感倾向需要借助构建好的商品评论情感词典、HowNet 中的否定词词典和程度副词词典[9]。商品评论文本绝大多数由较短的句子构成，我们将评论文本中的句子表示为 $S_1,S_2,\cdots,S_n$。每个评论句子中的情感词表示为 $W_1,W_2,\cdots,W_n$。句子 S_i 的情感极性表示为：

$$EO(S_i)=\sum_{i=1}^{n}EO(W_i) \tag{4}$$

其中，$EO(W_i)$ 表示词 W_i 的情感极性。如果情感词 W_i 前面的否定词的个数为奇数个时，则应该倒置 W_i 的情感极性。例如若 W_i 的情感极性为正向，则此时应该变成负向。

6 实验与分析

利用网络爬虫工具从 www.jd.com 网站上爬取 54368 篇关于某智能手机的评论文本。进行分词和去停用词以及词性标准处理之后，将标准为名词的词语作为候选属性集，利用组合神经网络进行聚类。随后将 50% 的文本用于训练 word2vec，剩下的 50% 的文本作为测试商品情感分类的测试集。评估文本分类的主要指标有准确率[10]、召回率及 F1 值，具体计算公式如下：

$$precision=\frac{a}{a+b}\times 100\% \tag{5}$$

$$recall=\frac{a}{a+c}\times 100\% \tag{6}$$

$$F_1=\frac{2\times precision\times recall}{precision+recall} \tag{7}$$

其中，a 表示分类器正确判断文档类别归属此类的数目；b 表示分类器错误判断文档属于该类但实际不属于的文档数目；c 表示分类器错误判断文档不属于该类但实际属于的文档数目。

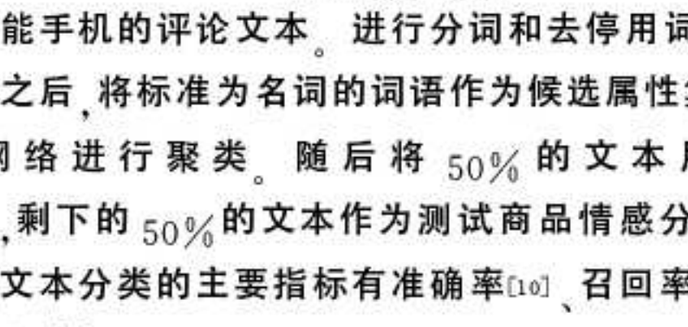

图 4 商品评论情感分析结果

表 1 基于 word2vec 的商品评论情感分析结果

	准确率	召回率	F1 值
正向极性	0.87	0.85	0.86
负向极性	0.86	0.83	0.84

从图 4 中可以看出，本文提出的基于组合神经网络的商品属性聚类方法的效果，从表 1 中能看出基于 word2vec 的商品评论情感分析针对正向和负向情感的情感词都有比较高的准确率和召回率，证实了本文提出的方法在判别商品评论文本情感倾向上的可行性。

结束语 本文提出了基于组合神经网络的商品属性聚类与基于 word2vec 和计算情感倾向相似度(SO-SD)的方法来进行中文商品评论情感分析。用真实商品评论文本训练 word2vec，再用测试文本集对所提模型进行测试。结果表明了本文提出的方法的有效性和准确性。由于本文提出的方法中没有考虑到中文比较性的句子的情感极性及情感强度的判断，因此，对文本中比较性的句子的情感倾向判断以及情感强度的研究是本文后续工作中的一个重点。

参 考 文 献

[1] 张紫琼，叶强，李一军. 互联网商品评论情感分析研究综述[J]. 管理科学，2010，13(6)：84-96

[2] 钟将，杨思源，孙启干. 基于文本分类的商品评价情感分析研究[J]. 计算机应用，2014，34(8)：2317-2321

[3] Kim S M，Hovy E. Determining the sentiment of opinions[C]// Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics，2004：1367-1373

[4] Fellbaum C. WordNet：An Electronic Lexical Database [M]. Bradford Book，1998

[5] Esuli A，Sebastiani F. Sentiwordnet：A publicly available lexical resource for opinion mining[C]// Proceedings of LREC. 2006(6)：417-422

[6] Yi J，Nasukawa T，Bunescu R，et al. Sentiment analyzer：Extracting sentiments about a given topic using natural language processing techniques[C]// Third IEEE International Conference on Data Mining，2003(ICDM 2003). IEEE，2003：427-434

[7] 唐小丽，白宇，张桂平，等. 一种面向聚类的文本建模方法[J]. 山西大学学报(自然科学版)，2014，37(4)：595-600

[8] Baumgarten M，Mulvenna M D，Rooney N，et al. Keyword-Based Sentiment Mining using Twitter[J]. International Journal of Ambient Computing and Intelligence (IJACI)，2013，5(2)：56-69

[9] Turney P D，Littman M L. Measuring praise and criticism：Inference of Semantic Orientation from Association[C]// Proceedings of ACM Transactions on Information Systems，2003. New York，USA，2003：315-346

[10] Yuan Cheng-xiang，Zhuang Yi，Li Hao-hong. Semantic Based Chinese Sentence Sentiment Analysis[C]//2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE，2011：2099-2103