基于特征扩展与深度学习的短文本情感判定方法

杜永萍 陈守钦 赵晓铮
(北京工业大学计算机学院 北京 100124)

摘 要 针对中文短文本信息量少、特征稀疏等特点,面向微博短文本进行情感分类研究,为了更好地提取短文本情感特征,从评论转发等上下文内容中挖掘具有语义递进关系的语料对原文本进行扩展,并抽取具有潜在感情色彩的特征词,采用Word2vec计算词语相似度以进行候选特征词扩展,最后引入深度信念网络(Deep Belief Network,DBN)对候选特征词进行深度自适应学习。在COAE(Chinese Opinion Analysis Evaluation)2015任务评测数据集上的实验表明,该方法能够有效地缓解短文本特征稀疏问题,并且能够较为准确地挖掘情感特征,提高情感分类的准确率。

关键词 情感挖掘,短文本,特征扩展,深度信念网络
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.051

Method of Short Text Opinion Recognition Based on Feature Extension and Deep Learning

DU Yong-ping CHEN Shou-qin ZHAO Xiao-zheng
(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract This paper put forward the opinion recognition method on microblog short text, which contains a small amount of information, and the feature is sparse. The review and repost information of microblog were used to reconstruct the original microblog text. The tool of Word2vec was adopted to cluster the similar sentiment word for feature extension. And also the feature was learned by deep belief network, which achieves the high-quality sentiment feature. The experimental result on the data of COAE (Chinese opinion analysis evaluation) 2015 denotes that our method alleviates the problem of feature sparseness and also more effective sentimental features are mined. The system performance is improved with the precision of 64.1%。

Keywords Opinion mining, Short text, Feature extension, Deep belief network

1 引言

近年来,随着互联网技术的高速发展,Facebook、Twitter、微博、微信等随之兴起,其由于具有简单易用、传播迅速等特点,逐渐成为人们日常生活中不可或缺的交流媒介。因此针对此类短文本的情感分析对于产品口碑跟踪以及社会舆情的监测等都具有重要的意义。如何弥补现有短文本情感分析方法的不足及提高情感分析的准确度,成为近年来工业界及学术界关注与研究的重点。

不同于普通文本的情感分类,由于短文本表达简洁、携带特征少,传统的基于规则的方式不能较好地利用语义及特征对其进行情感分类。此外,汉语表达形式及语法的多样性也增加了中文情感分析的难度。张成功等[1]通过深入研究中文语言表达的特点,构建了以基础词典、领域词典为主的中文极性词典,提出了一种基于极性词典的情感倾向分析方法;Pang等[2]针对电影评论数据利用朴素贝叶斯分类器(Naive Bayes)、支持向量机分类器(Support Vector Machines,SVM)等方法实现了对篇章级别文本的情感分类。

随着研究的深入,众多方法被提出。孙艳等[3]通过在LDA主题模型中融入情感模型,实现无监督的主题情感词的抽取,进一步完成了对文本的情感分类。Mei Qiaozhu等[4]通过构建主题情感混合模型(Topic-Sentiment Mixture,TSM)获得作者关于混合主题不同方面(Facets)的情感倾向。杨震等[5]针对短文本特征稀疏及上下文缺失的问题,提出了基于特殊上下文关系的文本情感极性判别方法。王蒙等[6]结合全局特征改进现有的局部特征抽取方法,提出基于伪相关反馈的短文本扩展与分类方法,有效解决了短文本特征稀疏的问题。Turney[7]提出一种无监督的学习算法,通过计算文中包含的形容词与副词等的平均语义指向的方法判定情感倾向。何天翔等[8]利用大规模语料库及同义词集合构建情感词网,针对短文本特征稀疏、信息量少等问题,提出了结合情感词网的中文短文本情感分类方法。贺飞艳等[9]结合TF-IDF方法与方差统计方法,构建细粒度情感分析与判断流程,提出一种实现多分类特征抽取的计算方法。夏梦南等[10]针对微博短文本存在的口语化、简洁化等社交网络特征,充分利用句法依存关系以及条件随机场(Conditional Random Fields,CRFs)抽

到稿日期:2016-09-23 返修日期:2017-03-12 本文受国家科技支撑计划子课题(2013BAH21B02-01),北京市自然科学基金资助项目(4153058),上海市智能信息处理重点实验室开放基金(IIPL-2014-004)资助。

杜永萍(1977-),女,博士,副教授,主要研究方向为信息检索、自然语言处理,E-mail:ypdu@bjut.edu.cn;**陈守钦**(1991-),男,硕士,主要研究方向为情感分析、自然语言处理,E-mail:garychenqin@emails.bjut.edu.cn;**赵晓铮**(1994-),女,硕士,主要研究方向为情感分析、自然语言处理,E-mail:s201607008@emails.bjut.edu.cn。

取候选评价对象，并在基于机器学习的微博情感分类方法的基础上结合情感分析词典来提高分类的准确度。何炎祥等[11]针对微博中商品评论文本短小、结构多样等特征，在仅使用现有微博级情感标注的条件下，提出了一种基于层叠条件随机场模型进行情感分类的方法。Rao Yanghui[12]等利用潜在主题、情感标签以及读者的情感评分生成主题特征，提出一种基于主题最大熵模型(Topic-level Maximum Entropy，TME)的情感分类方法。Odbal[13]等利用依存句法的词语搭配特征和基于组合语义的深度特征，提出一种以短语为主要线索的半马尔科夫条件随机场情感分类模型。Wang Xuren[14]等通过潜在语义分析算法(Latent Semantic Analysis Algorithm)的研究，提出了一种改进的潜在语义算法，提高了文本情感分类的准确度。张佳明等[15]利用 BTM 模型挖掘隐含主题，并利用基本情感词典分析隐含主题的情感分布，从而实现文本的情感分类。苏艳等[16]利用动态随机特征子空间方法生成多个随机特征子空间，并利用协同过滤在每个子空间筛选置信度较高的未标注样本，并利用该样本反向更新训练模型，提高了情感分类的准确度。

本文基于现有短文本特征提取及情感分类方法的研究，首先挖掘具有语义递进关系的转发评论语料对原文本进行扩展，并抽取名词、动词、形容词、副词、表情符号等具有潜在感情色彩词性的词语作为候选特征词，然后利用同义词林及 Word2vec 训练同义词相似度计算模型以对候选特征词进行相似词扩展，最后通过深度信念网络(DBN)对候选特征进行深度自适应学习。

本文第 2 节介绍情感分析的组织结构；第 3 节介绍词典资源的自动构建及基于上下文语义关联词的文本扩展方法；第 4 节结合 Word2vec 及 DBN 深度信念网络介绍基于词语级别的特征扩展与深度特征学习的选取过程；第 5 节给出实验并对结果进行分析；最后总结全文。

2 面向微博短文本的情感分析结构

针对微博内容表达形式多样、语言简洁、特征稀疏等特点，本文利用转发评论等上下文信息及同义词林在篇章与词语级别上进行特征扩展，同时挖掘潜在感情色彩特征词与 DBN 深度特征自适应学习都将有助于抽取更高质量的特征词。情感分析的总体结构如图 1 所示。

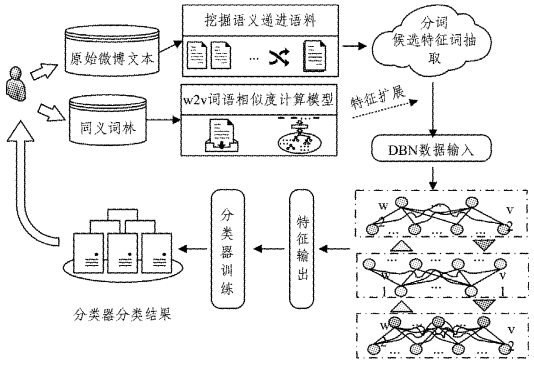

图 1 微博情感分析结构

3 词典资源构建及上下文关联重构

情感词典资源在情感分析任务中起着至关重要的作用，本文利用基本情感词典实现自动扩展，并根据微博文本特点构建了表情词典及语义递进关联词词典。

3.1 情感词典的构建及扩展

本文对台湾大学、大连理工大学、中国知网(HowNet)等提供的词典取并集后构建了基本情感词典。该词典虽然能够涵盖大部分通用领域，但是某些未出现在基本情感词典中的词语在特定领域能够表达明确的情感倾向，因此识别该类词语并对基本词典进行扩展具有积极的意义。

利用分词工具对文本进行分词并标注词性，去除停用词和过滤基本情感词典中已知情感倾向的词语；然后利用知网(HowNet)提供的情感倾向评价性词语并借助搜索引擎计算每个候选领域情感词与评价性词语组合的平均互信息值(PMI)；最后计算积极互信息(PMI_POS)与消极互信息(PMI_NEG)的差值，通过观察结果调整阈值以最终确定该词是否具有情感倾向，其中互信息的计算公式如式(1)所示：

$$PMI(a,b)=\frac{p(a,b)}{p(a)p(b)} \tag{1}$$

其中，$p(a)=A/N$，$p(b)=B/N$，$p(a,b)=C/N$，N 为所有相关检索结果的集合，A 为 a 结果出现的次数，B 为 b 结果出现的次数，C 为 a，b 结果共同出现的次数。

3.2 表情词典及递进关联词典的构建

鉴于表情符号对情感倾向的判定具有非常重要的作用，收集整理微博表情符号并对其进行情感标注，将具有情感倾向的表情符号加入基本情感词典中。表情符号示例如表 1 所列。

表 1 表情词典样例

编号	情感倾向	示例情感词
1	Positive	[嘻嘻]；[哈哈]；[互粉]；[礼物]；[给力]
2	Negative	[阴险]；[怒骂]；[抓狂]；[晕]；[弱]

本文从评论转发等上下文内容中挖掘具有语义递进关系的语料对原文本进行扩展，主要利用递进关联词确定该语料是否与原文本具有相同的情感倾向。递进关联词样例如“确实是”“赞”“非常认同”“说得对”等。

3.3 基于上下文语义关联词的文本扩展

为了缓解短文本的特征稀疏问题，本文将转发评论等上下文内容以原始微博内容为根节点构建关联树。利用递进关联词分析每个子节点是否与根节点表达的情感倾向一致，将具有相同情感倾向的语料扩展到原始文本中以丰富文本的特征密度，具体过程如图 2 所示。图 2 中“合理涨价，还是能够接受，北京地铁之前确实便宜，物价上涨，票价稍涨无可厚非”为扩展的关联文本。

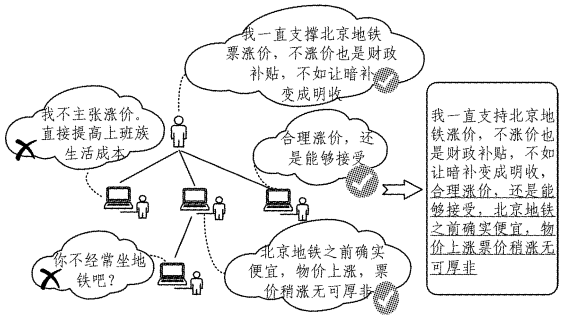

图 2 上下文语义关联词的文本扩展

4 特征扩展及基于 DBN 的深度特征学习

4.1 情感词语义相似度模型的构建

本文运用 Google 开源的基于神经网络的文本处理工具 Word2vec 构建同义词语义相似度模型。Word2vec 能够将单

词转化为向量的表示形式，将复杂的文本内容简化为向量的计算，通过深度学习训练及空间向量相近度计算来获得词语之间的语义相似度，与潜在狄立克雷分布（Latent Dirichlet Allocation，LDA）、潜在语义分析（Latent Semantic Index，LSI）等相比，Word2vec 能够更好地挖掘词的上下文关联信息。

本文主要采用 Word2vec 工具中的 Skip-gram 模型（Continuous Skip-gram Model），该模型能够根据给定的中心词获取语义相似的上下文关联词，以扩充情感特征向量。

Skip-gram 模型分为 3 层：输入层、投影层、输出层。输入层需要输入中心词 w 的词向量 $v(w)=R^m$；投影层为输入层的恒等投影，即 $Input\_v(w)$ 到 $Project\_v(w)$；输出层为一棵 Huffman 二叉树，此二叉树的叶子节点为输入数据中所有的词语，Skip-gram 模型网络结构如图 3 所示。

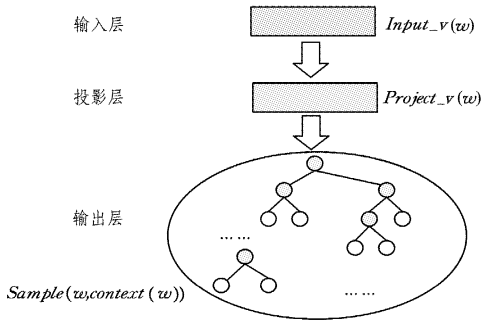

图 3　Skip-gram 模型网络结构示意图

经过学习训练，Word2vec 能够较好地分离不相关的单词，并标注同义词间的相似度权值，能够根据用户输入的中心词预测其语义近似词。

为了更好地丰富短文本的特征密度，本文利用上下文挖掘语义递进语料，且实现了候选特征集合的同义词扩展，利用搜狗全网新闻数据、原始微博训练集、情感词典及哈工大同义词林混合训练词语余弦相似度计算模型。该模型能够通过中心词获得与其相似程度倒序排列的词语集合。语义相似度模型样例如图 4 所示，其中距离中心点越近表示相似度值越高。

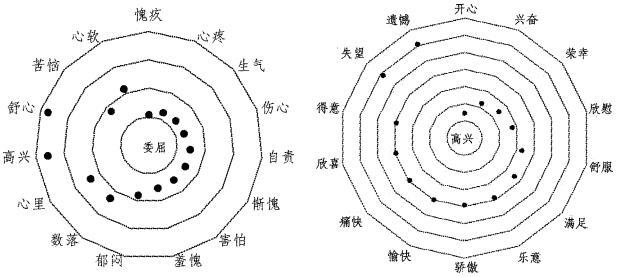

图 4　语义相似度模型样例分布示意图

4.2　粗粒度候选特征词筛选

本文利用深度信念网络 DBN 模型进行特征选取学习，鉴于特征选取的准确度及深度学习训练网络的效率，本文在进行特征深度学习训练之前，首先抽取具有潜在情感色彩词性的候选特征词并结合卡方统计及词频过滤部分的候选特征词，最终获得粗粒度候选特征词，特征词权值的计算公式如式(2)、式(3)所示：

$$feature\_weight_{(i)}=CHI*freq(i) \tag{2}$$

$$CHI=\chi^2(t,c)=\frac{N*(AD-CB)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \tag{3}$$

其中，$feature\_weight_{(i)}$ 为候选词 i 的情感特征权值；CHI 为该词的卡方统计量权值，N 为文档总数，A 为一个 $Label$ 中含有某个词的文本数目，B 为其余类别中含有该词的文本数目，C 为一个 $Label$ 中不包含某个词的文本数目，D 为其余类别中不包含某个词的文本数目；$freq(i)$ 为某个词出现的频率依照词语权值大小过滤掉低于设定阈值的候选特征词。

4.3　候选特征词扩展

深度信念网络可见层（输入层）的输的入要求，将本文提取的候选特征词转换为向量的表示形式，由于深度信念网络要求输入的数据格式为整数，因此按照词频（Term Frequency，TF）的形式表示向量，且向量中候选特征词的表示顺序为正向、中性、负向。为了进一步丰富候选特征词的特征密度，提高深度学习的准确度，本文采用 4.1 节中的同义词语义相似度模型，依次扩充每个候选特征词余弦相似度大于 0.8 的相似词，具体过程如图 5 所示。首先通过词聚类模型得到特征词原始数据中的情感词“开心”“梦寐以求”的语义相近词“高兴”“渴望”，由此更新其在特征向量中的权值。实验证明，该方法通过预测并扩充相似情感词提高了特征表示密度，同时也提高了 DBN 特征学习的质量。

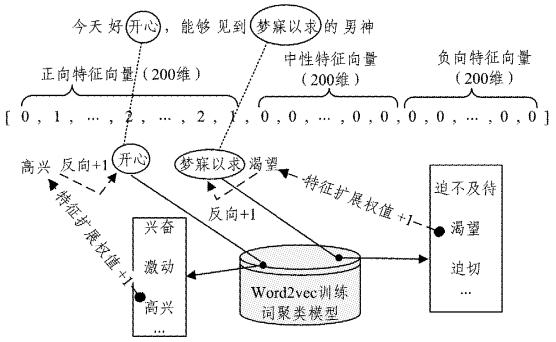

图 5　粗粒度候选特征词扩展模型

4.4　深度信念网络（DBN）特征自适应学习

深度信念网络（Deep Belief Network，DBN）算法是 Hinton 等[19]于 2006 年提出的一种深度学习算法，该算法训练神经元节点之间的权值，并将其用于特征识别、数据分类等领域。本文将其用于特征学习，输入候选特征词向量，经过深度学习结构得到降维后的抽象特征向量，DBN 由多层受限玻尔兹曼机（Restricted Boltmann Machine，RBM）组成，其主要结构如图 6 所示。

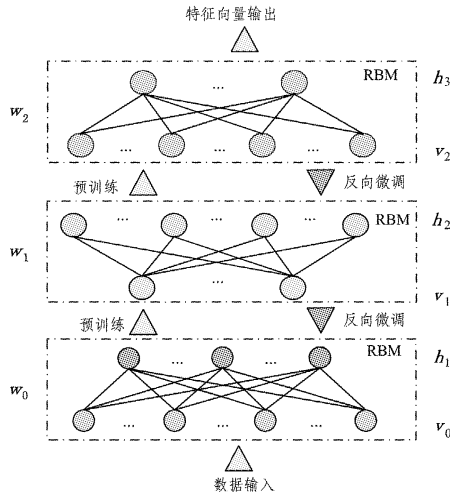

图 6　DBN 深度信念网络结构图

DBN 模型的训练过程主要分为两步。

(1)预训练(Pre-Training):每一层 RBM 网络独立执行无监督训练,确保特征能够对应到不同特征空间中,最大限度地保留特征的完整性。

(2)微调(Fine-Tuning):每一层的 RBM 网络只能保证该层权值映射达到局部最优,并不能保证整个 DBN 网络全局最优,因此微调过程会将每一层的误差由上而下传播至每一层 RBM 网络,以便达到微调整个全局训练网络的目的,弥补了某些分类器容易陷入局部最优及训练时间长的缺点。

经过预训练与微调后的特征向量可以应用于任何分类器模型,例如 BP 神经网络、逻辑回归(Logistic Regression,LR)分类器、支持向量机(Support Vector Machine,SVM)分类器等。

对候选特征词进行语义相似词特征扩展后,便可以利用 DBN 深度学习算法进行特征选取。首先设置文档特征向量对应的 Label、预训练与微调迭代次数及各个隐藏层对应节点的个数,然后对候选特征向量进行训练。

本文最终利用 SVM 及逻辑回归(Logistic Regression,LR)对输出的特征向量验证分类的准确度。设置多组迭代次数及最终输出特征维度,以验证不同参数对最终分类结果准确率的影响,主要过程如图 7 所示。

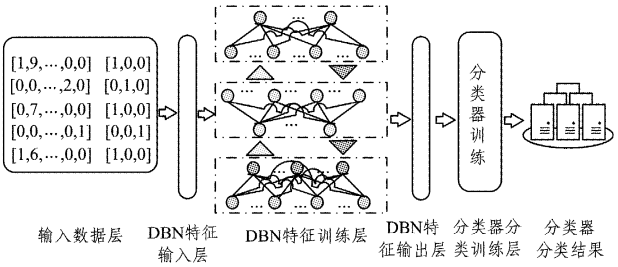

图 7 DBN 特征学习及情感分类过程

5 实验分析

本文以第七届中文文本倾向性分析评测(COAE2015)任务一提供的中文微博评测语料为实验数据集,经过分析整理本文实验所采用的标注训练集约为 8000 条,测试集约为 4000 条,计算准确率、召回率、F1 值并与 COAE2015 评测委员会提供的评测结果做对比,同时本实验还分析了不同特征数量、不同分类器对分类准确度的影响。

5.1 评价标准

本文采用较为通用的评测标准,通过计算准确率(Precision)、召回率(Recall)以及 F1 值对分类结果的准确性进行评测,如式(4)—式(6)所示:

$$Precision=\frac{Sytem\_Correct}{System\_Output} \tag{4}$$

$$Recall=\frac{System\_Output}{Labeled\_Human} \tag{5}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{6}$$

5.2 实验结果分析

本实验计算情感倾向为正向(Pos)、中性(Neu)、负向(Neg)的识别准确率、召回率及 F1 值,并将实验结果与 COAE2015 任务一的评测结果进行对比分析,结果表明基于特征扩展与深度学习的方法(BJUT_Feature Extension+DBN)的准确度较基于词典(BJUT_Dictionary_2015)的方法有较大的提高,并且明显优于 COAE2015 任务一评测的中位水平。同该任务的最优结果相比,一方面在语义递进语料与候选特征词选取方面未达到最优,并在扩展特征的同时引入了部分噪音数据。另一方面,在短文本领域运用深度学习的探索能够较大幅度地弥补基于规则方法需大量人工干预的缺点,同时较传统特征选择算法能够更好地提高特征挖掘的质量。性能对比结果如表 2 所列。

表 2 不同方法与 COAE2015_Task1 的性能对比

项目	准确率			召回率			F1 值		
情感倾向	Pos	Neu	Neg	Pos	Neu	Neg	Pos	Neu	Neg
BJUT_Feature Extension+DBN	0.606	0.591	0.621	0.601	0.634	0.575	0.603	0.612	0.597
BJUT_Dictionary_2015	0.554	0.569	0.471	0.681	0.510	0.392	0.515	0.510	0.428
COAE2015_Task1_Best	0.688	0.837	0.645	0.970	0.686	0.609	0.699	0.643	0.627
COAE2015_Task1_Medium	0.514	0.535	0.521	0.518	0.512	0.515	0.542	0.538	0.514

本文利用挖掘递进关系上下文以及候选特征词扩展方式丰富特征表示密度,为了证明特征扩展方法的有效性,本文分别对比了特征扩展前后的情感分类准确度,实验结果如图 8 所示。

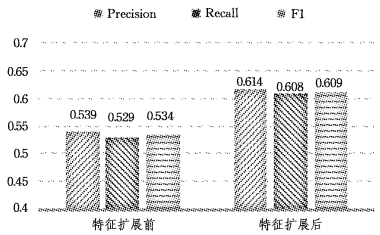

图 8 特征扩展前后的性能对比

本实验中特征扩展前后的分类准确度相差约 10%,从而证明了特征扩展的有效性。

为了验证 DBN 深度学习有助于提高情感特征选择的质量,本文对比分析了传统特征选择算法 TF*IDF 及深度学习算法 DBN 对情感分类准确度的影响。

实验结果如图 9 所示,使用 DBN 深度学习后情感分类准确度的各项指标均有所提升,能够有效地提升分类的准确度。

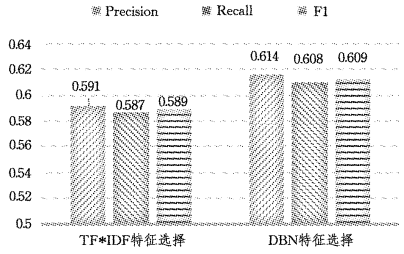

图 9 DBN 与 TF*IDF 特征选择的性能对比

本文同时验证了不同特征维度对分类准确度的影响。本实验在保证其他实验条件相同的条件下,通过调整深度学习输出特征维度,以 50 为梯度获得 100～600 维的结果,经过对比分析,该数据集在 100～200 维时的平均准确率较高,因此我们以 10 为梯度细化分析 100～200 维的准确度,以获取最佳的特征维度。实验结果如图 10 所示。

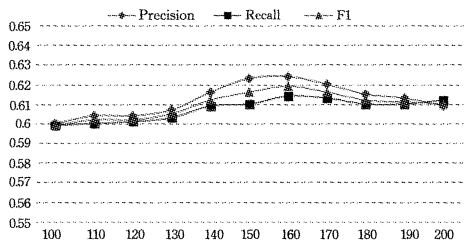

图 10　不同特征维度的分类结果比较

由图 10 可知，特征维度在 150～170 之间的分类效果最佳，在 100～160 维之间，分类的准确度逐渐提高，特征维度超过 160 之后，分类器的准确度略微下降并逐渐呈稳定趋势。

在获得深度信念网络(DBN)经过学习过程输出的特征表示后，尝试采用不同的分类器对其特征训练的质量进行评测，本实验主要采用了逻辑回归(Logistic Regression，LR)与支持向量机(Support Vector Machine，SVM)对特征输出进行分类评测，实验结果如表 3 所列。

表 3　不同分类器的分类性能比较

分类器	特征维度	Precision	Recall	F1	Avg_Precision	Avg_Recall	Avg_F1
LR	100	0.590	0.596	0.600	0.606	0.605	0.608
	130	0.606	0.604	0.605			
	160	0.622	0.613	0.617			
	190	0.606	0.608	0.610			
SVM	100	0.61	0.603	0.598	0.614	0.608	0.609
	130	0.608	0.602	0.605			
	160	0.626	0.615	0.620			
	190	0.613	0.612	0.612			

由表 3 可知，虽然两种分类器的各项指标的平均值都能达到 60%以上，但支持向量机分类器的性能较逻辑回归略胜一筹，同样地，两种分类器在特征维度为 160 时取得了较好的性能。

结束语　本文面向微博短文本展开研究，基于评论转发等上下文关联信息进行语义递进关系语料挖掘以缓解数据稀疏问题，采用基于 Word2vec 的深度学习词相似度训练工具进行特征词扩展，并进一步基于深度信念网络 DBN 进行特征自适应学习。实验结果表明，本文工作能够有效地扩展短文本特征密度，并能够挖掘文本深层的隐含特征，最终使得微博短文本的分类性能得到较大的提升。

今后工作将充分利用短文本的上下文信息对微博数据的关联重构进行优化，并深入研究深度学习的相关算法，实现有效特征的学习与扩展，提高情感分类性能。

参 考 文 献

[1] ZHANG C G, LIU P Y, ZHU Z F, et al. A sentiment analysis method based on a polarity lexicon[J]. Journal of Shandong University(Natural Science), 2012, 47(3): 47-50. (in Chinese)
张成功，刘培玉，朱振方，等. 一种基于极性词典的情感分析方法[J]. 山东大学学报(理学版)，2012，47(3)：47-50.

[2] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniues[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP). Philadelphia, PA, USA, 2002: 79-86.

[3] SUN Y, ZHOU X G, FU W. Unsupervised Topic and Sentiment Unification Model for Sentiment Analysis[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 102-108. (in Chinese)
孙艳，周学广，付伟. 基于主题情感混合模型的无监督文本情感分析[J]. 北京大学学报(自然科学版)，2013，49(1)：102-108.

[4] MEI Q Z, LING X, WONDRA M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs[C]// International Conference on World Wide Web. 2007: 171-180.

[5] YANG Z, LAI Y X, DUAN L J, et al. Short Text Sentiment Classification Based on Context Reconstruction[J]. Acta Automatica Sinca, 2012, 38(1): 55-67. (in Chinese)
杨震，赖英旭，段立娟，等. 基于上下文重构的短文本情感极性判别研究[J]. 自动化学报，2012，38(1)：55-67.

[6] WANG M, LIN L F, WANG F. Short text expansion and classification based on pseudo-relevance feedback[J]. Journal of Zhejiang University (Engineering Science), 2014, 48(10): 1835-1842. (in Chinese)
王蒙，林兰芬，王锋. 基于伪相关反馈的短文本扩展与分类[J]. 浙江大学学报(工学版)，2014，48(10)：1835-1842.

[7] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA, 2002: 417-424.

[8] HE T X, ZHANG H, LI B, et al. Sentiment classification combined with sentiment lexicon network for Chinese short texts [J]. Application Research of Computers, 2015, 32(10): 2905-2909. (in Chinese)
何天翔，张晖，李波，等. 结合情感词网的中文短文本情感分类[J]. 计算机应用研究，2015，32(10)：2905-2909.

[9] HE F Y, HE Y X, LIU N, et al. A Microblog Short Text Oriented Multi-class Feature Extraction Method of Fine-Grained Sentiment Analysis[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 48-54. (in Chinese)
贺飞艳，何炎祥，刘楠，等. 面向微博短文本的细粒度情感特征抽取方法[J]. 北京大学学报(自然科学版)，2014，50(1)：48-54.

[10] XIA M N, DU Y P, ZUO B X. Micro-blog opinion analysis based on syntactic dependency and feature combination[J]. Journal of Shandong University (Natural Science), 2014, 49(11): 22-30. (in Chinese)
夏梦南，杜永萍，左本欣. 基于依存分析与特征组合的微博情感分析[J]. 山东大学学报(理学版)，2014，49(11)：22-30.

[11] HE Y X, LIU J B, SUN S T, et al. Product reviews sentiment classification in Micro-blog based on cascaded conditional ran-

dom field[J]. Journal of Shandong University (Natural Science), 2015,50(11):67-73. (in Chinese)
何炎祥,刘健博,孙松涛,等. 基于层叠条件随机场的微博商品评论情感分类[J]. 山东大学学报(理学版),2015,50(11):67-73.

[12] RAO Y H,XIE H R,LI J,et al. Social emotion classification of short text via topic-level maximum entropy model[J]. Information & Management,2016,53(8):978-986.

[13] ODBAL,WANG Z F. Emotion Analysis Model Using Compositional Semantics[J]. Acta Automatica Sinica, 2015, 41(12): 2125-2137.

[14] WANG X R,ZHANG Q H. Text Emotion Cassification Research Based on Improved Latent Semantic Analysis Algorithm[C]// Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Hangzhou,China,2013:210-213.

[15] ZHANG J M,WANG B,TANG H H,et al. Unsupervised Sentiment Orientation Analysis on Micro-blog Based on Biterm Topic Model[J]. Computer Engineering,2015,41(7):219-223. (in Chinese)
张佳明,王波,唐浩浩,等. 基于 Biterm 主题模型的无监督微博情感倾向性分析[J]. 计算机工程,2015,41(7):219-223.

[16] SU Y,JU S F,WANG Z Q,et al. Semi-supervised Sentiment Classification with Random Feature Subspace Method[J]. Journal of Chinese Information Processing,2012,26(4):85-90. (in Chinese)
苏艳,居胜峰,王中卿,等. 基于随机特征子空间的半监督情感分类方法研究[J]. 中文信息学报,2012,26(4):85-90.

[17] Google 开源深度学习工具 Wordvec[OL]. https://code.google.com/p/word2vec.

[18] 搜狗实验室全网新闻数据(SogouCA)[OL]. http://download.labs.sogou.com/dl/ca.html.

[19] HINTON G E,OSINDERO S,THE Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7):1527-1554.

(上接第 282 页)

结束语 数据预处理方法是影响文本挖掘性能的关键因素。为解决文本挖掘中由文本特征“高维-稀疏”矛盾导致的高时空复杂度与低效率问题,提出基于词频统计规律的数据预处理方法,首先推导出文本同频词表达式,然后探究各频次词语在文中的分布规律,最后以此为基础进行数据预处理。该方法与传统文本挖掘算法在预处理阶段只进行简单的分词和去停用词操作相比,从词频规律入手,进一步探究可改善数据预处理性能的方法。研究发现,词频为 1 或 2 的低频词与文档的关联度较低,但在文中所占比重约为 2/3,在预处理过程中对低频词进行去噪,可在保证文本挖掘精度的前提下,大大减少特征维度,使时空复杂度明显下降,平均运行时间降低了 70%以上,有效提升了文本挖掘性能。该方法所提出的基于词频统计规律进行数据预处理的思想对文本挖掘算法的改进具有重要意义。

参 考 文 献

[1] HAN J,FAN J,et al. Semanti-Enhanced Spatial Keyword Search [J]. Journal of Computer Research and Development,2015,52(9): 1954-1964. (in Chinese)
韩军,范举,等. 一种语义增强的空间关键词搜索方法[J]. 计算机研究与发展,2015,52(9):1954-1964.

[2] HU J,FAN J,LI G L,et al. Top-k Fuzzy Spatial Keyword Search[J]. Chinese Journal of Computers,2012,35(11):2237-2246. (in Chinese)
胡骏,范举,李国良,等. 空间数据上 Top-k 关键词模糊查询算法[J]. 计算机学报,2012,35(11):2237-2246.

[3] REN P J,CHEN Z M,et al. Search Result Diversification Combing Semantic and Temporal Intent[J]. Chinese Journal of Computers, 2015,38(10):2076-2091. (in Chinese)
任鹏杰,陈竹敏,等. 一种综合语义和实效性意图的检索多样化方法[J]. 计算机学报,2015,38(10):2076-2091.

[4] DING Z Y,JIA Y,et al. Survey of Data Mining for microblogs [J]. Journal of Computer Reaserch and Development,2014,51(4):691-706. (in Chinese)
丁兆云,贾焰,等. 微博数据挖掘综述[J]. 计算机研究与发展,2014,51(4):691-706.

[5] SONG Q,NI J,WANG G. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering,2013,25(1):1-14.

[6] ZHAO Z,HE X F,CAI D,et al. Graph Regularized Feature Selection with Data Reconstruction[J]. IEEE Transactions on Knowledge and Data Engineering,2016,28(3):689-700.

[7] ZIPF G K. Human behavior and the principle of least effort:an introduction to human ecology[M]. Addison-Wesley Press,1949:23.

[8] BOOTH A D. A law of occurrences for words of low frequency [J]. Information and Control,1967,10(4):386-393.

[9] EGGHE L. A new short proof of Naranan's theorem, explaining Lotka's law and Zipf's law[J]. Journal of the American Society for Information Science & Technology,2010,61(12):2581-2583.

[10] CHAN P,HIJIKATA Y,NISHIDA S. Computing semantic relatedness using word frequency and layout information of wikipedia[C]// Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM,2013:282-287.

[11] SURYASEN R,RANA M S. Content analysis and application of Zipf's law in computer science literature [C]// 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS). IEEE,2015:223-227.

[12] GEORGE K Z. Human Behavior and the Principle of Least Effort:An Introduction to Human Ecology[M]. New York:Addison-Wesley Press,1949:573-584.

[13] 邱均平. 文献计量学[M]. 科学技术文献出版社,1988:157.

[14] BOOTH A D. A law of occurrences for words of low frequency [J]. Information and Control,1967,10(4):386-393.

[15] AGRAWAL R,GOLLAPUDI S,KENTHAPADI K. Enriching Textbooks Through Data Mining [C]//Proceedings of the First ACM Symposium on Computing for Development. 2010:1-9.

[16] AGRAWAL R,GOLLAPUDI S,KANNAN A,et al. Data mining for improving textbooks[J]. ACM SIGKDD Explorations Newsletter,2012,13(2):7-19.