基于 word embedding 和 CNN 的情感分类模型*

蔡慧苹，王丽丹†，段书凯
(西南大学 电子信息工程学院，重庆 400715)

摘 要： 尝试将 word embedding 和卷积神经网络(CNN)相结合来解决情感分类问题。首先，利用 skip-gram 模型训练出数据集中每个词的 word embedding，然后将每条样本中出现的 word embedding 组合为二维特征矩阵作为卷积神经网络的输入，此外每次迭代训练过程中，输入特征也作为参数进行更新；其次，设计了一种具有三种不同大小卷积核的神经网络结构，从而完成多种局部抽象特征的自动提取过程。与传统机器学习方法相比，所提出的基于 word embedding 和 CNN 的情感分类模型成功地将分类正确率提升了 5.04%。

关键词： 卷积神经网络；自然语言处理；深度学习；词嵌入；情感分类
中图分类号： TP183 **文献标志码：** A **文章编号：** 1001-3695(2016)10-2902-04
doi:10.3969/j.issn.1001-3695.2016.10.005

Sentiment classification model based on word embedding and CNN

Cai Huiping, Wang Lidan†, Duan Shukai
(*College of Electronic & Information Engineering, Southwest University, Chongqing* 400715, *China*)

Abstract: This paper tried to propose a method to solve the problem of sentiment classification by integrating word embedding and convolutional neural network (CNN). First of all, the method accomplished a training process with skip-gram model to generate word embedding of each word in the dataset. Then, it created a two-dimensional feature matrix which was the combination of word embedding of each word in a training sample as the input of CNN model. Each iteration process of training, entries of feature matrix would also update as part of model parameters. Secondly, this paper proposed a CNN structure which was mainly composed of three different sizes of convolution kernels so as to complete the automatic extraction process of a variety of local abstract features. Compared with traditional machine learning algorithms, the proposed word embedding and CNN based sentiment classification model has successfully improved classification accuracy by 5.04%.

Key words: convolutional neural network; natural language processing(NLP); deep learning; word embedding; sentiment classification

0 引言

随着近年来互联网的高速发展，正在逐步走向成熟的社交平台和电商网站使得人们与外界的沟通变得更加方便，可以自由地对电商产品和时事政治发表自己的观点。情感分析能够提取出这些短文本数据中所蕴涵的语义信息，从而挖掘出用户的情感倾向。随着互联网的不断发展，文本数据也越来越多，情感分析成为了自然语言处理中一个非常重要的研究方向。目前，国内外的研究学者已经针对情感分析问题作出了一些探索，使用的方法主要包括传统的机器学习和现在比较流行的深度学习方法两大类。

李婷婷等人[1]尝试从文本数据中人工构建若干特征，再利用传统的机器学习方法进行情感分类。这种方法本质上属于传统的机器学习范畴，其分类效果严重依赖于所构建特征的质量和模型参数的调优，整个过程非常耗时耗力。陈翠平[2]引入了深度学习的思想来完成文本分类任务，利用深度信念网络自动提取文本特征。相比于人工构建特征的方式，深度学习能够更加高效地完成特征提取任务，但只有足够深的模型才能够提取出能较好地反映出文本语义信息的特征，这就造成了模型参数数量和训练时间的显著增加。Kim[3]将 word embedding 与卷积神经网络结合应用在情感分析和问题分类等若干自然语言处理任务中，获得了非常好的效果。其思路与本文非常相似，但并未针对数据集重新训练 word embedding，这就造成许多词没有 word embedding，导致对原始数据特征的描述不够充分。

为了解决传统机器学习方法特征提取困难的问题，本文尝试利用 word embedding 和卷积神经网络模型完成自然语言处理中的情感分类任务。卷积神经网络最早用于图像特征提取，因此需要将文本表示为类似图像数据的二维特征形式。首先，针对数据集训练出每个词的 word embedding；然后将训练好的 word embedding 进行组合得到每条评论的 embedding；最后，将其作为卷积神经网络的输入特征，并作为网络参数一起参与迭代训练过程。为了验证 word embedding 对于卷积神经网络模型的性能提升情况，添加了一组对比实验：将卷积神经网络的输入特征按高斯分布进行随机初始化。此外，在相同数据集上，本文引入了几种传统的机器学习模型，同样利用 word em-

收稿日期：2015-06-03；修回日期：2015-07-31 基金项目：国家自然科学基金资助项目(61372139)；国家教育部“春晖计划”科研资助项目(z2011148)

作者简介：蔡慧苹(1991-)，女，江西上饶人，硕士，主要研究方向为机器学习、深度学习、数据挖掘；王丽丹(1976-)，女(通信作者)，河南长垣人，教授，博导，博士，主要研究方向为人工神经网络、非线性系统与电路设计、生物电子电路(ldwang@swu.edu.cn)；段书凯(1973-)，男，重庆奉节人，教授，博导，博士，主要研究方向为人工神经网络、非线性系统与电路设计、生物电子电路.

bedding 来构建输入特征。实验证明,与传统的机器学习方法相比,本文提出的基于 word embedding 和卷积神经网络的情感分类模型能够极大地提升分类的准确率。

1　相关理论

1.1　深度学习与卷积神经网络

深度学习概念最初由 Hinton 等人[4,5]在 2006 年提出,其模拟人脑对于视觉信号的分层处理机制,能够从复杂的原始数据中提取出层级特征。利用深度学习提取出来的特征可以看做原始数据在更高层面上的抽象表示,非常适合解决一些比较抽象的识别类任务。自其诞生以来,深度学习已经在计算机视觉[6]、语音识别[7]和自然语言处理[8,9]等诸多领域中取得了许多优秀的研究成果。本文涉及到的卷积神经网络是目前应用最为广泛的一种深度学习结构,其底层由卷积层、pooling 层交替组成,顶端一般会利用全连接层来完成具体的任务。这种特殊的网络结构使得卷积神经网络有如下显著的优点:

a) 卷积层及 pooling 层的交替叠加使得卷积神经网络对于局部微小特征非常敏感;

b) 特征提取和模式分类同时进行,并同时在训练中产生;

c) 利用局部感受野和权值共享减少网络的训练参数、降低网络结构的复杂度,使得其适应性更强。

卷积神经网络的这些优点使其在图像识别领域大获成功。Krichevsky 等人[6]设计的卷积神经网络结构在 2012 年的 ImageNet 挑战赛中一举夺冠,将 Top5 错误率由 26% 大幅降低至 15%。LeCun 等人[10]利用卷积神经网络成功解决了手写体数字识别问题。鉴于卷积神经网络模型的诸多优秀表现,研究人员开始尝试将其引入到更多研究领域中。本文即尝试利用卷积神经网络解决自然语言处理中的情感分类问题。

1.2　卷积神经网络与自然语言处理

互联网的发展使得社交网络和电商平台产生了大量文本数据,其中蕴涵的宝贵信息亟待有效的处理方法进行挖掘。因此,自然语言处理(NLP)作为人工智能领域的一个重要方向,越来越受到学术界和工业界的重视。不同于图像和语音,文本在很多方面有其自身特殊的特点。例如,文本数据中包含了人类更高层的情感特征,由此造成的歧义与多义使得自然语言处理在研究过程中遇到了更多的困难。

随着深度学习领域的高速发展,已经开始有学者尝试利用卷积神经网络解决自然语言处理中的一些难题。Collobert 等人[11]将卷积神经网络引入到了自然语言处理中的许多任务中,并且证明其提出的模型在各项任务中都获得了很好的表现; Shen 等人[12]利用卷积神经网络解决信息检索中的语义分析问题; Kalchbrenner 等人[13]利用卷积神经网络对句子进行建模,并且提出一种新的 pooling 方式。这些工作证明了卷积神经网络在自然语言处理领域中同样具有广阔的应用前景。

1.3　Word embedding

词向量概念的引入是为了将自然语言中的词转换成计算机能够"认识和理解"的形式,从而可以利用各种算法进行下一步处理,完成各种自然语言处理任务。One-hot representation 是一种最经典的词向量表示方法,针对特定的任务能够很快生成词向量。但其维度过高,对短文本数据采用这种方法构建词向量往往会造成词向量非常稀疏,且存在"词汇鸿沟"现象,很难反映词与词之间的语义关系。

因此,Hinton[14]提出了一种叫做 word embedding 的词向量表示方法,这种方法的主要思想是将词分布式地映射到低维空间中,从而解决了向量稀疏问题。此外,该低维空间中词向量之间的位置关系可以很好地反映它们在语义层面上的联系,非常适合作为文本的高层抽象特征。

Mikolov 等人[15]提出的 Skip-gram 模型能够针对某个文本数据集快速高效地训练出 word embedding,其结构如图 1 所示。

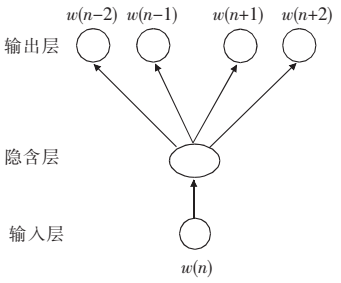

图 1　Skip-gram 模型

Skip-gram 模型的主要思想是根据当前单词来预测上下文。假设存在某一词组序列为 $w_1, w_2, w_3, \ldots, w_N$,Skip-gram 的目标是使式(1)最大。

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leqslant i\leqslant c, i\neq 0}\log p(w_{n+i}\mid w_n) \tag{1}$$

其中: c 是以当前词语为中心的前后文词数,c 的值越大,模型的训练效果越好,但同时也会造成训练时间的增加。

在实际应用中,只要训练所用的语料库足够大,窗长度 c 选择恰当,就能够在较短训练时间内得到高质量的 word embedding。

2　基于 word embedding 和 CNN 的情感分类模型

卷积神经网络中的卷积层能够很好地描述数据的局部特征,通过 pooling 层可以进一步提取出局部特征中最具有代表性的部分。这种重点关注局部信息的处理思想与自然语言处理中的 n-gram 语言模型的处理思想非常类似。此外,情感属于人类特有的高层意识形态,解决情感分析问题首先需要提取出高层抽象特征。传统的人工提取方法在描述抽象特征方面能力十分有限,而卷积神经网络结构已经在图像和语音识别中被证明可以出色地完成此类特征提取任务。

基于以上分析,本文尝试引入卷积神经网络模型来解决情感分类问题。

2.1　模型的输入处理

卷积神经网络最早用来对图像进行识别,而图像数据是由二维数据组成。因此,首先需要将文本数据组合成二维数据矩阵的形式以作为模型的输入进行处理。

假设数据集中长度最长的评论包含 m 个词,$a_i \in \mathrm{R}^k$ 是该条评论中的第 i 个词所对应的 word embedding,卷积神经网络的输入为由 m 个 k 维向量组合而成的 $m \times k$ 的二维数据矩阵。对于其余长度小于 m 的样本,需要进行补零处理。本文所采用的数据集中最长评论包含 117 个词,word embedding 设为 300 维,即 $m = 117, k = 300$。

图 2 为 word embedding 组合方式。输入的文本中的每个词从训练好的词向量矩阵中找到自己的 word embedding,再通过"纵向累加"的方式组合成适合 CNN 处理的二维特征矩阵。

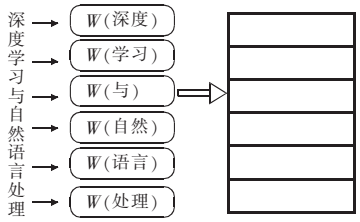

图 2 word embedding 组合方式

2.2 模型具体结构

本文所采用的卷积神经网络的具体结构如图 3 所示。

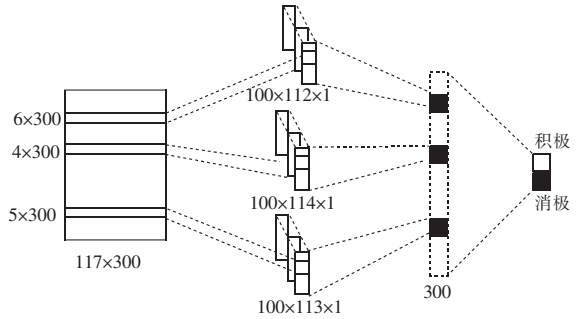

图 3 本文 CNN 模型结构

图 3 为本文 CNN 模型结构。卷积层由三种不同大小的卷积核来提取出多组局部特征图; 随后的 max-pooling 层提取出每张特征图中最有代表性的特征; 最后通过全连接层来完成正负情感的映射。

卷积核 $w \in \mathrm{R}^{h \times k}$ 在长度为 h 的窗内的 h 个词进行卷积操作,则输出特征为

$$s_i = f(w \times a_{i:i+h-1} + b) \tag{2}$$

其中: b 为偏置项, f 为激活函数。神经网络中常用的激活函数有多种,如 sigmod 函数、tanh 函数等。为加快训练收敛速度,本实验中采用 ReLu 函数作为激活函数:

$$f(x) = \max(0, x) \tag{3}$$

卷积核 w 对输入数据进行卷积,就可以得到一张特征图 ($S \in \mathrm{R}^{m-h+1}$)

$$S = [s_1, s_2, \cdots, s_{m-h+1}] \tag{4}$$

为了尽可能充分地考虑到每个词的前后文信息,从而提取出不同粒度大小的局部特征,本文通过改变窗长度 h 分别设计了 4×300、5×300 和 6×300 三种不同大小的卷积核结构,经过卷积操作得到的特征图尺寸分别为 114×1、113×1 和 112×1,每种卷积核各提取出 100 张特征图。

通常,自然语言中的每句话中具有代表性的局部少数几个词就能够大致反映出该句话的情感类别。同样地,为了能够从每张特征图中找到最具代表性的局部最优特征,需要对卷积层提取出来的特征图进行 max-pooling 操作,完成由输入数据中提取出某个局部特征的全部过程。所有的 300 张卷积图都经过 max-pooling 操作后得到 300 个局部最优特征,再通过全连接层连接到最后一层的两个输出节点,从而能够综合考虑提取出来的所有特征,进而完成情感分类任务。

2.3 模型的训练

目前,卷积神经网络的训练方法主要采用的依然是传统的梯度下降法。其中,批量梯度下降法能够寻找到最优解,但由于每次更新权值都需要全部样本参与运算,收敛速度非常慢; 随机梯度下降法每次更新权值只需要一个样本参与运算,所以收敛速度能够显著加快,但很容易造成收敛到局部最优解。为了能够兼顾两种方法的优点,本文采用 mini-batch 梯度下降法进行训练,即每次更新权值需要一小批样本参与计算,这样能够在保证尽可能寻找到最优解的同时加快训练速度。本数据集中共有 4 000 条样本,经过反复实验,batch 的大小设置为 50 时能够获得比较好的折中。

为了防止过拟合,利用 L_2 正则化对网络参数进行约束; 全连接层的训练过程中引入了 Dropout 策略,即每次迭代中随机放弃一部分训练好的参数[16]。本文在训练过程中设置 Dropout = 0.5,即随机放弃一半参数。

最后,为了得到稳定可靠的模型,实验过程中采用了 10 折交叉验证,最后的结果取 10 次结果的平均值。

3 实验过程、结果及分析

3.1 实验环境

本文中的所有实验均在如表 1 所示的实验环境中完成。

表 1 实验环境及配置

实验环境	环境配置
操作系统	Ubuntu 14.04
CPU	Intel Core i5-3470 3.20 GHz
内存	4 GB
编程语言	Python 2.7
分词工具	jieba 0.36
深度学习框架	Theano 0.7
word embedding 训练工具	word2vec

3.2 数据集的选择与处理

实验数据来自于京东商品的用户评论数据集。该数据集中包括 4 000 条评论样本,其中正负样本各 2 000 条,分别为用户对某商品的积极评价和消极评价。通过 10 折交叉验证来对模型预测的准确率进行评估,即将 4 000 条样本平均分为 10 份,每份中包含正负样本各 200 条; 总共进行 10 次实验,每次取出 9 份作为训练集,1 份作为测试集; 最后模型的分类准确率是 10 次测试结果的平均值。

数据集划分完成后,还需要对其进行一系列预处理:

a) 过滤所有标点符号和特殊字符,只保留含有较多语义信息的中英文文本。

b) 格式转换。原始数据中存在很多不规范的文本格式,如全角状态下输入的英文字符。为了方便下一步的分词处理,需要对数据中格式不正确的文本进行转换。

c) 分词处理。英语中词与词之间是用空格连接的,因此对每个词进行各种处理过程是非常简单的。但中文文本中词与词之间没有办法区分,想要建模需要首先进行分词处理,最终的实验效果与分词的质量有关。

d) 训练 word embedding。利用 Mikolov 等人提出的 Skip-gram 模型训练 word embedding,该模型已经被 word2vec 工具实现,可以直接使用。值得一提的是,为了能够得到较高质量的 word embedding,训练所用的语料库规模不能太小。因此,除了用户评论数据外,本文将中文维基百科数据也加入到数据集中作为语料库进行训练[17]。

3.3 实验设计

在 CNN 模型的基础上,本文尝试将 word embedding 与之相结合,从而针对性地优化 CNN 在情感分类任务上的性能。具体实验设计如下:

a) CNN + Skip-gram。在 Skip-gram 模型训练好的词向量表中查找每条样本中出现的每个词的 word embedding,并组合成 $m \times k$ 的二维数据矩阵作为 CNN 的输入。其中 m 为数据集中最长评论所包含的词数,对于长度小于 m 的样本需要补零; k 为 word embedding 长度。

b) CNN + rand。CNN 模型部分保持不变,按高斯分布随机初始化 word embedding。实验目的是通过与 CNN + Skip-gram 模型的结果相比较,从而验证 word embedding 在描述原始数据特征分布方面的性能。

c) 传统机器学习模型。在相同数据集上,利用几种常用的机器学习模型作为对比来证明 CNN 在情感分类任务上的性能优势。为了排除由于特征构建方式的不同而导致实验结果无法比较,传统模型的特征构建方式同样基于 word embedding,每条样本的特征为该样本中所有 word embedding 的均值。

3.4　实验结果与分析

3.4.1　实验结果

三组实验结果如表 2 所示。实验结果全部经过交叉验证得出,具有较高的可信度。

表 2　实验结果

模型	准确率/%	模型	准确率/%
Linear regression	48.37	Linear SVM	86.30
SVM(RBF)	60.98	CNN + rand	89.41
Random forest	83.25	CNN + Skip-gram	91.34
Logistic regression	86.15		

3.4.2　实验结果分析

1) CNN 与传统模型

相比于传统的机器学习方法,本文所提出的 CNN 模型在情感分类任务上获得了出色的性能提升。随机初始化 CNN 的输入数据,仅靠模型自身的特征提取能力就能够得到 89.41% 左右的正确率,远远超过表现最好的 Linear SVM。这也同时证明了卷积神经网络结构在噪声数据环境下的健壮性。

2) CNN + rand 与 CNN + Skip-gram

与 CNN + rand 相比,CNN + Skip-gram 通过引入 word embedding 使分类正确率由 89.41% 上升至 91.34%。一方面,word embedding 能够在更抽象的层面上描述原始输入数据的特征分布情况,这是人工特征提取方式很难做到的; 另一方面,将其作为 CNN 的输入特征并在迭代训练过程中不断更新,这相当于引入了一定先验知识,能够在训练过程中引导模型按照更好的方向收敛到最优解。

3) Linear regression 与 SVM(RBF)

实验中 Linear regression 与 SVM(RBF) 虽然经过了参数调优,但两者依然表现出了较差的性能。传统的机器学习方法中,模型在具体任务中的表现与特征的构建方式有直接关系,不同的模型需要构建与其特点相匹配的特征来使其能力最大化。而在情感分类任务中,word embedding 这种分布式的特征表示方法或许无法充分发挥这两种模型的优势。

3.5　训练集规模及分布的影响分析

3.5.1　Word embedding 的训练

在确定 word embedding 的维度时,需要根据具体的应用需求和训练集规模来确定。为了更精确地反映每个词在低维空间中的语义分布情况,要求 word embedding 的维度应该尽可能地高,这需要规模更大且分布更均匀的训练语料库来支持; 但这同时也对硬件计算能力和模型的表达能力提出了更高的要求。本文训练 word embedding 时所采用的维基百科中文数据包括 232 894 篇中文文献,共包含超过一亿个中文词汇及少量英文词汇。其语料库规模已经足够训练出高维度的 word embedding,这种情况下每个词所携带的语义信息也能够被充分表达。虽然使用维度更高的 word embedding 可以更好地对语义特征进行描述,但这也会造成卷积神经网络模型参数显著增多,从而增大过拟合的风险。因此经过综合考虑,本文最终采用的是 300 维的 word embedding。

3.5.2　卷积神经网络的训练

同样地,在卷积神经网络结构的设计方面,依然要考虑训练样本的规模所带来的影响。卷积神经网络的层数越多,每一层设计得越复杂,其表达能力就越强。但同样需要数量更多、覆盖范围更大的训练样本使得网络通过训练最终能够顺利收敛。本文所采用的训练集包含了 4 000 条正负样本。为了防止模型太过复杂从而导致过拟合,只设计了一层卷积层和一层 max-pooling 层,尽可能使模型简单化从而提高泛化能力; 训练过程中,采用 L_2 正则和 Dropout 策略控制模型的训练过程,防止陷入局部最优; 通过 10 折交叉验证能够保证实验结果具有较强的说服力。理论上,若能够获得规模更大、覆盖面更广的训练集,则可以在避免过拟合的情况下,通过增加卷积层数等方式获得表达能力更强的卷积神经网络模型。

4　结束语

情感分类一直是自然语言处理中的重要任务之一,本文尝试利用深度学习的思想解决情感分类问题。首先,本文采用 Skip-gram 模型自动训练出数据集中所有中文词语的 word embedding,并将其组合为二维矩阵作为每条样本的情感特征,省去了手工提取特征的繁琐步骤; 其次,提出了一种卷积神经网络结构来完成情感分类任务,进一步提取层级局部特征,并设计了三种不同的卷积核,从而获得多种粒度大小的特征表示。与传统的机器学习模型的对比实验证明,本文提出的基于 word embedding 的卷积神经网络模型在情感分类任务上表现出了更加出色的性能。此外,通过随机初始化 word embedding 的对比实验,证明了单纯依靠卷积神经网络提取的层级局部特征已经能够获得良好的分类性能,而针对性地训练 word embedding 能够更好地描述原始文本数据的情感特征分布,因而能够获得更好的分类效果。

参考文献:

[1] 李婷婷,姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究,2015,32(4):978-981.

[2] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用,2015,24(2):121-126.

[3] Kim Y. Convolutional neural networks for sentence classification[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2014:1746-1751.

[4] Hinton G E,Salakhyurdinov R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504-507.

[5] Hinton G E,Osindero S,Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation,2006,18(7):1527-1554.

[6] Krichevsky A,Sutskever I,Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012:1097-1105.

[7] Hinton G,Deng Li,Yu Dong,*et al*. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine,2012,29(6):82-97.

(下转第 2909 页)

c) 提及数。根据用户提及其他用户的数量进行用户影响力排名。

根据以上三个用户行为数据进行微博用户影响力实验，对比各个用户行为的影响大小，结果如图 4 所示。

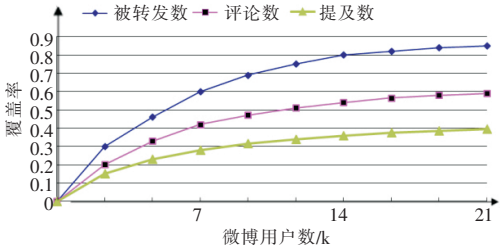

图 4　微博用户行为因子影响力对比

图 4 的纵坐标表示影响覆盖率，横坐标表示微博用户数的变化。通过图 4 结果对比，微博被转发数所占的覆盖率明显比评论数和提及数高。微博的转发是信息传播的主要途径，显示用户较高的影响力。用户转发微博，一半以上会进行评论，但是微博用户通过@ 其他人的概率就很小，所以提及数的覆盖率较低。

3.3　算法结果分析

通过已进行的数据预处理、参数设置和计算，本文通过对比基于用户行为综合分析的 PageRank 算法[8]、基于用户内容的 CELF 算法[9]，用来测试本文融合用户行为和内容的算法。

通过上述三种方法进行用户影响力的排名，然后根据三种指标所取得的用户影响力的覆盖率对比，分析各个算法之间的优劣，具体结果如图 5 所示。

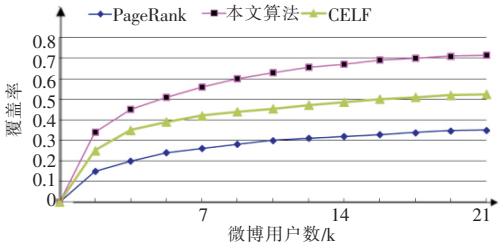

图 5　各个算法覆盖率对比

从图 5 中可以看出，随着微博用户数的增加，通过用户行为和内容所形成的用户关系结构变得更加紧密，信息传播范围扩大，覆盖率增大。当微博用户数量过多时，会出现一些孤立的节点，覆盖率增大变得缓慢，进而趋近某一值。对于覆盖率，本文算法获得的值明显要高于 PageRank 算法和 CELF 算法。

在图5的对比结果中，PageRank算法评价指标性能较低，这是因为单从用户行为中的粉丝数和被转发数判定用户影响力的大小，忽略了用户之间的兴趣也是判断影响力大小的标准；CELF 算法则没有考虑网络结构和行为。在图 5 对比结果中，本文算法的性能都高于其他算法。由于本文算法不仅考虑到了用户行为和微博内容的联系，而且对用户进行兴趣建模，从而达到了降维的效果，提高了算法的精度。

4　结束语

在社交网络中，信息传播引起了广泛的关注，为避免恶意信息造成较大的影响，快速检测对信息传播有较大影响力的用户节点显得尤为重要。本文基于用户的行为，结合用户的内容关系，提出了一种融合用户行为和内容的微博用户影响力算法。通过 LDA 模型对用户建立兴趣模型，利用 KL 散度计算用户的相似性，进而结合用户行为因子的加权得到用户节点的影响力。实验结果显示，本文算法能够很好地检测到微博信息的传播情况。

参考文献：

[1] Pinto P，Thiran P，Vetterli M. Locating the source of diffusion in large scale network [J]. Physical Review Letters，2012，109(6)：068702.

[2] 曹玖新，吴江林，石伟，等. 新浪微博网信息传播分析与预测 [J]. 计算机学报，2014，37(4)：779-790.

[3] 王悦，黄威靖. ELPS：一种高效的微博信息传播轨迹提取算法 [J]. 计算机科学，2014，41(4)：233-238，255.

[4] 齐超，陈鸿昶，于洪涛. 基于用户行为综合分析的微博用户影响力评价方法 [J]. 计算机应用研究，2014，31(7)：2004-2007.

[5] Salton G. The SMART retrieval system experiments in automatic document processing [M]. Englewood Cliffs：Prentice Hall Inc，971：337-354.

[6] Blei D M，Ng A Y，Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research，2003，3(3)：993-1022.

[7] 周东浩，韩文报. DiffRank：一种新型社会网络信息传播检测算法 [J]. 计算机学报，2014，37(4)：884-893.

[8] Sun Jimeng，Tang Jie. A survey of models and algorithms for social influence analysis [M] //Social Network Data Analytics. New York：Springer，2011：177-214.

[9] Leskover J，Krause A，Guestrin C，*et al*. Cost effective out-break detection in networks [C] //Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM Press，2007：420-429.

(上接第 2905 页)

[8] Graves A，Mohamed A R，Hinton G. Speech recognition with deep recurrent neural networks [C] //Proc of IEEE International Conference on Acoustics，Speech and Signal Processing. [S. l.]：IEEE Press，2013：6645-6649.

[9] Socher R，Perelygin A，Wu J Y，*et al*. Recursive deep models for semantic compositionality over a sentiment Treebank [C] //Proc of Conference on Empirical Methods in Natural Language Processing. 2013：1631-1642.

[10] LeCun Y，Bottou L，Bengio Y，*et al*. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[11] Collobert R，Weston J，Bottou L，*et al*. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research，2011，12(3)：2493-2537.

[12] Shen Yelong，He Xiaodong，Gao Jianfeng，*et al*. Learning semantic representations using convolutional neural networks for Web search [C] //Proc of the 23rd International Conference on World Wide Web. New York：ACM Press，2014：373-374.

[13] Kalchbrenner N，Grefenstette E，Blunsom P. A convolutional neural network for modelling sentences [C] //Proc of the 52nd Annual Meeting of the Association for Computational. 2014：655-665.

[14] Hinton G E. Learning distributed representations of concepts [C] //Proc of the 8th Annual Conference of the Cognitive Science Society. 1986.

[15] Mikolov T，Sutskever I，Chen Kai，*et al*. Distributed representations of words and phrases and their compositionality [EB/OL]. (2013-10-16). http：//arxiv. org/pdf/1310. 4546. pdf.

[16] Srivastava N，Hinton G，Krizhevsky A，*et al*. Dropout：a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research，2014，15(1)：1929-1958.

[17] https：//dumps. wikimedia. org/zhwiki/latest/zhwiki-latest-pages-articles. xml. bz2 [EB/OL].