# 基于扩展短文本词特征向量的分类研究

孟 涛,王 诚
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

**摘 要**:由于短文本的文档长度较短,短文本中词语的共现信息非常匮乏,造成短文本信息稀疏性问题。信息稀疏性也成为了传统主题模型在短文本上难以取得突破性进展的瓶颈之一。针对短文本分类,充分利用短文本中的每一个词语并解决其稀疏性成为关键。为了解决这一问题,基于 Word2vec 模型对短文本进行词嵌入扩展以解决其稀疏性,并将词向量转换成概率语义分布来测量语义关联性;针对短文本扩展后的特征向量,利用改进后的特征权重算法并引入语义相关度去处理扩展后的词特征向量。该方法可以区分出扩展后的短文本中词的重要程度,以便获得更准确的语义相关性。短文本分类研究采用 KNN 算法分类,实验结果表明,通过在外部语料集上学习得到的语义相关性扩展来处理短文本特征,可以有效提高短文本的分类效果。

**关键词**:短文本;Word2vec 模型;词嵌入;改进后的特征权重算法;语义相关度
**中图分类号**:TP301 **文献标识码**:A **文章编号**:1673-629X(2019)04-0057-06
doi:10.3969/j.issn.1673-629X.2019.04.12

# Research on Short Text Classification Based on Extended Word Feature Vectors

MENG Tao, WANG Cheng
(School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract**: Due to the short length of the document, the co-occurrence information of the words in the short text is very scarce, which causes the sparseness of the short text. The sparseness of information has also become one of the reasons why the traditional topic model is difficult to make breakthrough progress on short texts. For short text classification, it is very important to make full use of every word in essay and solve its sparseness. For this, word embedding is extended based on Word2vec model to solve its sparsity, and word vectors are converted into probabilistic semantic distribution to measure semantic relevance. For the extended feature vector of short text, the improved feature weight algorithm is used and the semantic relevance is introduced to handle the extended word feature vector. This method can distinguish the importance degree of words in the extended short text so that we can get more accurate semantic relevance. In this paper, we adopt KNN algorithm to study the short text classification. The experiment shows that we can extend short text features by learning semantic correlation obtained from external corpus, which can effectively improve the classification effect of short texts.

**Key words**: short texts; Word2vec model; word embedding; improved feature weight algorithm; semantic relevance

## 1 概 述

随着社交网络和电子商务的飞速发展,微博、Twitter、商品评价、实时新闻推送等短文本形式已成为互联网的主流内容。短文本通常长度较短,范围在 10 到 140 个字。研究短文本中热点话题的分类挖掘以及监测网络舆情信息对各种领域决策方面有着重要的应用前景,因此如何高效正确地挖掘短文本成为了一个研究的热门方向[1]。

针对常规文本分类,大多是利用传统的向量空间模型(vector space model, VSM)将文本向量化并按向量之间的欧氏或余弦距离计算文本间的关系,在处理长文本时取得了很好的分类效果[2]。但是由于短文本文档长度较短,词项共现信息相对于常规文本非常匮乏,会存在向量空间信息稀疏问题。而 VSM 忽略了词语之间的语义相似度,词本身无法存储语义信息,会严重限制短文本主题分类的质量[3]。

**收稿日期**:2018-05-28 **修回日期**:2018-09-19 **网络出版时间**:2018-12-20
**基金项目**:国家自然科学基金青年项目(71301081)
**作者简介**:孟 涛(1993-),女,硕士研究生,研究方向为网络信息安全与数据挖掘;王 诚,副教授,硕导,研究方向为互联网大数据挖掘。

对于缺乏语境信息而导致向量空间信息稀疏性的短文本问题,现有方法主要遵循两个方向来丰富短文本。第一种是仅使用隐藏在当前短文本上下文中的规则或统计信息来扩展特征空间[4],称为基于自我资源的方法。另一种是通过外部资源扩展特征空间,称为基于外部资源的方法[5],该方法可以分为四类,包括基于链接的方法[6]、基于 Web 搜索的方法[7]、基于分类的方法[8-10]以及基于主题的方法[11]。其次还有一种方式是将词矢量作为输入特征,利用卷积神经网络进行训练[12]。

现有的短文本研究存在的问题:引入背景知识和对外部相关数据的过度依赖,未从句子语义层面出发,无法深度挖掘短文本所表达的语义;改进短文本词向量的权重计算方法,但忽略了上下文因素。

针对短文本特征向量较少的问题,提出使用神经概率语言模型中的 Word2vec 技术进行词嵌入来训练扩展短文本中的词向量[13]。词嵌入也被称为词向量和词的分布式表示,已被证明在捕获语言中的语义规则方面是有效的,具有相似语义和语法属性的词被投影到向量空间中的相同区域,由此产生的语义特征被用作补充信息来克服短文中语境信息的局限性。词嵌入具有两个优点:维度缩减,上下文相似性。为了更好地利用词嵌入后向量空间中的词矢量,进一步将背景语料与词的语义相关度相结合,并用改进后的特征权重的计算方式去区分词汇的重要程度,去除大多数背景词在语义上没有关联的词汇。最后通过 KNN 分类方法对所提出的改进算法进行实验分析比较,以验证该方法可提高短文本分类精度的有效性。

## 2 相关工作

### 2.1 Word2vec

Word2vec 是 Google 发布的一个 NLP(neuro-linguistic programming,神经语言程序学)开源工具。Word2vec 采用 distributed representation 表示词向量,这种表达方式不仅可以避免 one-hot representation 的维数灾难问题,使词向量的维度变小,而且可以比较容易地用普通统计学方法来分析词与词之间的关系。

Word2vec 模型使用 CBOW(continuous bag-of-words)和 skip-gram[14] 两种结构作为学习模型。CBOW 模型是使用上下文来预测目标词,而 skip-gram 模型的思路是使用特定的词去预测对应的上下文。这两种方法都需要利用神经网络对大规模语料进行语言模型训练,同时能够得到描述语义和句法关系的词矢量。CBOW 模型使用上下文来预测目标词,会因为窗口大小的限制丢失短文本训练语料集中的相关语义信息,而 Skip-Gram 模型使用当前词来预测目标上下文可避免该问题,能够高效进行词矢量的训练,所以文中选用 Word2vec 的 skip-gram 模型。

### 2.2 针对短文本的权重改进

TF-IDF[15](term frequency - inverse document frequency,词频-逆文档频率)是一种用于信息检索与数据挖掘的传统加权方式,如式 1 所示:

$$W_{ij} = \mathrm{tf}_{ij} \times \mathrm{idf}_j = \mathrm{tf}_{ij} \times \log\left(\frac{N}{n_j}\right) \tag{1}$$

其中,$W_{ij}$ 是词 $t_j$ 在短文本 $d_i$ 中的权重;$\mathrm{tf}_{ij}$ 是词 $t_j$ 在短文本 $d_i$ 中的词频表示;$\mathrm{idf}_j$ 是词 $t_j$ 的逆文档频率;$N$ 是语料库中短文本文档的总数;$n_j$ 是训练背景语料库中出现 $t_j$ 的短文本数量。

由于短文本的数据稀疏性,导致 TF-IDF 权重算法的特征权重区分度不足。因此文中提出改进词频 TF 以及 IDF 权重以解决短文本分类中的数据稀疏性与文本区分度问题。

针对 IDF 权重算法,Basili 提出了 IWF 权重算法[16],表达式如下:

$$W'_{ij} = \mathrm{tf}_{ij} \times \left(\log\frac{N}{n_j}\right)^2 \tag{2}$$

相对于传统的 IDF 部分,显著区别在于 IWF 权重算法中对词逆文档频率做平方处理,其共同目的都在于降低高频率出现且相对来说无意义的词。Basili 认为 IDF 的逆频率方面太过绝对,因此为了平衡各词的贡献,采用了平方去降低并缓和语料库中包含出现该词的文档对该词权重的影响,该方法也被证明相对 IDF 有更好的分类效果。同时,IWF 应用到短文本领域对关键词快速提取也起到了更精确的效果,因为短文本本身的特性,该方法可以有效缓解短文本中计算词权重的偏向程度。

相比 IDF 与 IWF 的改进方式,对于 TF 采取类似的改进方式。$\mathrm{tf}_{ij}$ 的定义是短文本 $d_i$ 中词 $t_j$ 的词频,由于短文本相对于长文本字数较少的特性,其词频也相对较低,所以单纯使用原公式的 TF 相对于短文本并不适用。针对短文本的稀疏性使得难以从词频信息来判断该词的重要程度,需要降低并削弱词频对权重的影响,所以做出以下改进:

(1)考虑采用对数函数 log(底数为 10)对 $\mathrm{tf}_{ij}$ 作处理,由于 $\mathrm{tf}_{ij}$ 是小数,直接进行 log 处理会是负数,所以做加 1 处理,即 $\log(1+\mathrm{tf}_{ij})$。

(2)考虑到扩展后的短文本中关键词相对扩展的集中性,用对数函数削弱会导致词频降低的太过绝对,词频之间相差会比较大,从而影响短文本分类的性能。所以引入方根来缓和差异性较大的问题,但是对于缓和程度,即方根的次数 $\theta$,需要结合实际短文本做实验来确定,即 $(\log(1+\mathrm{tf}_{ij}))^{1/\theta}$。

该方法改进的意义在于,选用对数函数来降低词频的影响,但是对于短文本词频来说对数函数削弱的太过绝对,所以引入方根来做缓和,该方法的有效性将通过具体实验进行验证。综合以上改进,得到改进后的权重计算公式,记为 TF'-IWF。

$$W_{ij} = \sqrt[\theta]{\lg(\mathrm{tf}_{ij}) + 1.0} \times \left(\log \frac{N}{n_j}\right)^2 \qquad (3)$$

## 3　改进后的特征权重算法处理扩展后的短文本词特征向量

对于短文本稀疏的特性,使用外部语料集来扩展短文本。首先,通过收集相关训练语料库作为训练短文本的扩展词典,并使用 Word2vec 模型处理该语料库,并测试扩展后的语义相关性;然后通过 Word2vec 训练词的语境相关概念并用来处理扩展短文本的词向量。该方法的一般框架如图 1 所示。

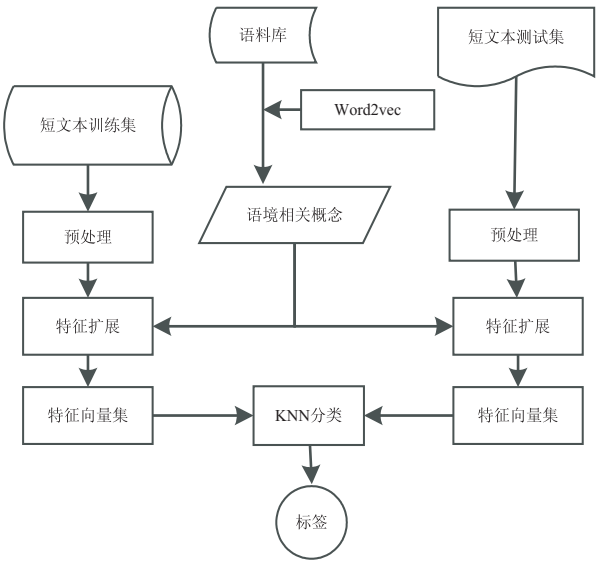


图 1　短文本分类过程

### 3.1　预处理

在对短文本做功能扩展之前对短文本进行预处理,包括三个主要步骤:中文版分词、停用词过滤和特征选择。文中采用比较成熟的中文分词工具-结巴分词将短文本收集成分词;之后对停用词进行过滤,通过功能选择保留了有代表性的词;最后,使用式 3 对特征进行加权,得到短文本的向量表示。

### 3.2　构建扩展词向量的相关工作

文中使用整理好分类的新闻语料库,总共包含 39 247篇新闻,分为历史、军事、文化、经济、教育、IT、娱乐、法制总共八个类别。对于词嵌入扩展的新闻数据集,使用中文的百度百科网站上抓取的 10 万篇文档信息和已整理好的新闻语料库的新闻内容当作训练语料词典,即语料库。

首先,使用 Word2vec 模型去训练整理好的新闻语料库,得到词的相关语境概念。对于使用 Word2vec 训练得到的词相关语境概念只是为了预测之后测试短文本过程中的副产物,而不是结果。

(1)训练语料库中词的相关语境概念。

定义训练的新闻语料库为词库 vocab,其中文本数定义为 $d_i$ ,文本中出现的词定义为 $t_k$ , $M$ 是指词向量维数。

$$\mathrm{vocab} = \{\, t_k \mid k \in [1, M] \,\}$$

使用 Word2vec 模型训练该语料库,将得到词相关语境的向量空间 $C_k$ ( $\mathrm{dis}_{k,1}, \mathrm{dis}_{k,2}, \cdots, \mathrm{dis}_{k,n}$ ),一个输入单词,有 $n$ 个输出概率结果, $n$ 是指该词在相关词库出现的次数。该训练过程只包括该词的语义环境,并不针对整个词库。

该相关语境的词向量空间 $C_k$ 不是预测的目标(语义相关度才是需要的),该词向量是为之后测试输入词用来学习的表达,是为了测试短文本任务中预测所给单词的语义环境。该相关语境的词向量 $C_k$ 是 Word2vec 模型在试图通过不断学习该词的一个好的数学表达来提高预测词分类的精确结果。在模型重复循环的过程中,不断调整完善词的数学表达。

(2)测试短文本与相关语境之间的语义相关度。

在得到训练语料库中词的相关语境概念后,需要测量测试短文本中词的语境向量 $C_i$ 与其相关词语境向量空间 $C_k$ 之间的语义相关度。其中测试短文本词的语境向量 $C_i$ 是用 Word2vec 得到词 $t_i$ 在该文本语境中的矢量表示,表示形式为 $C_i(\mathrm{dis}_{i,1}, \mathrm{dis}_{i,2}, \cdots, \mathrm{dis}_{i,n})$。

文中用 Word2vec 来处理,得到测试文本中词的语境向量 $C_i$ 与训练语料库中词的相关语境向量概念 $C_k$ ,并测量两者的语义相关度,具体过程如图 2 所示。

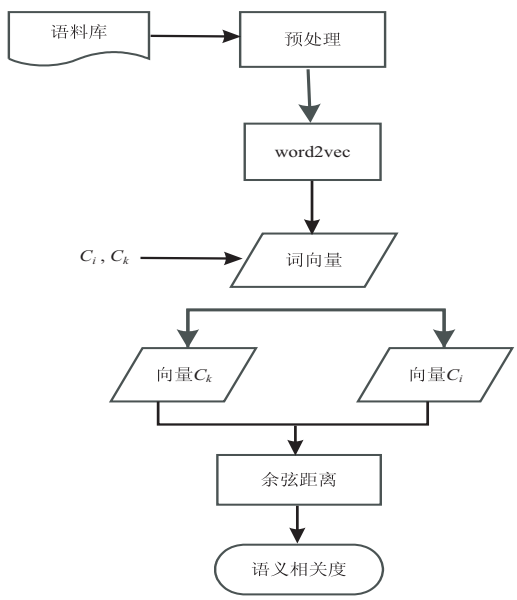


图 2　语义相关性的具体过程

语义相关度的定义如下：

$$R_i = \text{Cosine}(C_k, C_i) = \frac{C_k, C_i}{\| C_k \| \times \| C_i \|} = \frac{\sum_{j=1}^{n} (\text{dis}_{k,j} \times \text{dis}_{i,j})}{\sqrt{\sum_{j=1}^{n} (\text{dis}_{k,j})^2} \times \sqrt{\sum_{j=1}^{n} (\text{dis}_{i,j})^2}} \quad (4)$$

经过以上处理，最终得到测试短文本的文档向量表示 $C(d_i)$，$C(d_i) = ((C_1, R_1), (C_2, R_2), \cdots, (C_n, R_n))$，其中 $R_i (1 \leqslant i \leqslant n)$ 是词 $t_i$ 的语境向量 $C_i$ 和涉及到的 $C_k$ 的语义相关度。

### 3.3 扩展短文本的特征权重处理

扩展特征项是短文本分类的核心，基于语料库扩展短文本特征的方式有助于解决数据稀疏性问题和传统缺乏描述短文本类特征的能力。在对短文本进行预处理后，得到特征项列表及其加权值为 $((t_1, \text{tf}'\text{iwf}_1), (t_2, \text{tf}'\text{iwf}_2), \cdots, (t_m, \text{tf}'\text{iwf}_m))$，$\text{tf}'\text{iwf}_i$ 是特征项 $t_i$ 的加权值，由式 3 计算，$m$ 是短文本中特征项的数量。基于语料库的 $C_k$ 扩展特征空间，该方法由以下步骤完成。

(1)确定特征词 $t_i$ 有无语料中对应的词嵌入。如果有该词 $t_i$ 的相关语义知识，则继续下一步；如果不是，则更改为下一个特征词。

(2)将相关语义加到特征向量中，得到语义词 $C_i$ 和相关语境集合 $C_i$ $((C_1, R_1), (C_2, R_2), \cdots, (C_n, R_n))$ 的特征项 $t_i$。对于扩展性能是否合理以及最佳性，需要设定阈值 $\varphi$ 来判断词嵌入的质量。

(3)使用特征权重定义扩展后短文本集。为了准确衡量扩展后的词对短文本原始语义的影响，结合短文本特征的重要性和扩展语境之间的相关性，通过式 5 计算扩展项的权重值。

$$\text{weight}_{i,j} = \frac{\text{tf}'\text{iwf}_i}{m} \times R_j \quad (5)$$

其中，$\text{weight}_{i,j}$ 是扩展项 $j$ 的权重值；$\text{tf}'\text{iwf}_i$ 是短文本中特征词 $t_i$ 的加权值；$R_j$ 是语义相关度。如式 5 所示，记为 TF'IWF-$R_j$。

从以上的处理分析可得到短文本的向量空间包含原始特征项和上述处理之后扩展的词向量。

## 4 实 验

### 4.1 实验数据集

文中使用整理好分类的新闻语料库，总共包含 39 247篇新闻，分为历史、军事、文化、经济、教育、IT、娱乐、法制共八个类别。数据集包括新闻标题及新闻内容，文本采用原新闻标题数据集作为短文本数据集，内容数据集作为背景语料库数据集。按照 $u:v(u+v=1)$ 的比例将短文本数据集进行分类，其中 $u$ 是用于训练的数据集，$v$ 是用于测试的数据集。

实验平台：两台 Linux 操作系统的计算机搭建 Spark 集群，这两台计算机一个 Master 节点，一个 Slave 节点，两台计算机都同时部署 Hadoop 2.6.4 和 Spark 2.1.0，Hadoop 提供 HDFS 等底层文件支持。

### 4.2 分类性能评价指标

分类评价指标主要包括预测准确率 Precision、召回率 Recall 和二者的综合评价指标调和平均数 $F_1$。

$$\text{Precision} = \frac{\text{分类器正确判断为该类的样本数}}{\text{分类器判断属于该类的样本数}}$$

$$\text{Recall} = \frac{\text{分类器正确判断为该类的样本数}}{\text{属于该类的样本数}}$$

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

### 4.3 实验设计和结果

通过 Word2vec 获得百度百科网站等新闻语料库中概念的特征矢量。扩展短文本时，根据筛选出的语义相关词的质量来决定阈值 $\varphi$。设置扩展阈值 $\varphi = 0.5$，该扩展的参数调整使分类准确率达到最高，并设置 $\theta$ 来权衡扩展短文本权重比例对分类的影响。最后使用 Spark 中可用的 KNN 分类器实现实验[17]。

#### 4.3.1 改进词频对分类结果的影响

为了验证所提降低词频的算法对短文本的实用性，从语料库中选取已标注关键词的 8 个类别中不相关的短文本作为测试文本集。

对于式 3 中方根次数 $\theta$ 的选定，依次选择 $\theta$ 值等于 1、2、3、4，并进行实验分析不同方根值对短文本分类结果的影响。比较数据如表 1 所示。

表 1 不同 $\theta$ 值对分类结果的影响

| 类别 | $\varphi=1$ | | | $\varphi=2$ | | | $\varphi=3$ | | | $\varphi=4$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| 历史 | 0.692 | 0.732 | 0.711 | 0.786 | 0.832 | 0.808 | 0.735 | 0.852 | 0.789 | 0.694 | 0.861 | 0.769 |
| 军事 | 0.835 | 0.798 | 0.816 | 0.928 | 0.895 | 0.911 | 0.928 | 0.872 | 0.899 | 0.926 | 0.869 | 0.897 |
| 文化 | 0.603 | 0.809 | 0.691 | 0.847 | 0.802 | 0.824 | 0.821 | 0.792 | 0.806 | 0.845 | 0.781 | 0.812 |
| 经济 | 0.762 | 0.776 | 0.769 | 0.769 | 0.913 | 0.835 | 0.751 | 0.911 | 0.823 | 0.751 | 0.912 | 0.824 |
| 教育 | 0.721 | 0.802 | 0.759 | 0.899 | 0.836 | 0.866 | 0.887 | 0.814 | 0.849 | 0.886 | 0.815 | 0.849 |

续表 1

| 类别 | $\varphi=1$ | | | $\varphi=2$ | | | $\varphi=3$ | | | $\varphi=4$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| IT | 0.817 | 0.671 | 0.737 | 0.927 | 0.891 | 0.909 | 0.941 | 0.862 | 0.900 | 0.941 | 0.845 | 0.89 |
| 娱乐 | 0.832 | 0.773 | 0.801 | 0.887 | 0.834 | 0.860 | 0.902 | 0.815 | 0.856 | 0.901 | 0.793 | 0.844 |
| 法制 | 0.783 | 0.765 | 0.774 | 0.858 | 0.871 | 0.864 | 0.791 | 0.786 | 0.788 | 0.807 | 0.821 | 0.814 |
| 平均值 | 0.756 | 0.766 | 0.757 | 0.863 | 0.859 | 0.860 | 0.845 | 0.838 | 0.839 | 0.844 | 0.837 | 0.837 |

从图 3 中可以发现,当 $\theta=2$ 时,分类结果相较于其他值较好,同时也验证了考虑方根值缓和对数函数削弱词频太过绝对性方面是有意义的。

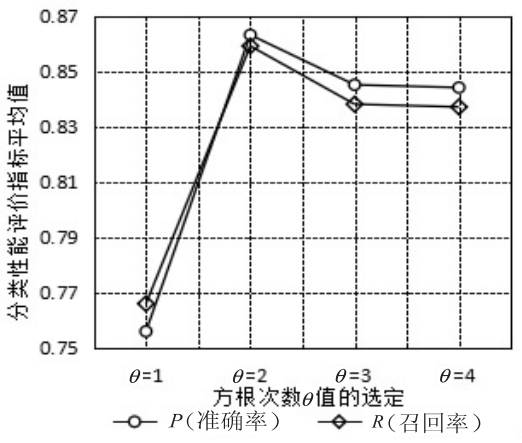


图 3　词频中不同 $\theta$ 值对分类性能的影响

4.3.2　改进的关键词提取算法比较

对于关键词加权算法 TFIDF 与 TFIWF 和 TF'IWF 的比较,采用确定值 $\theta=2$,通过实验来具体比较它们的性能。一般关键词提取算法的评估是通过精确率和召回率来评测算法性能,比较如图 4 所示。

从图 4 中可看出,文中算法提取的关键词在分类的准确率和召回率上相对于传统 TFIDF 算法提高了 10% 左右,验证了该算法在短文本上的适用性更可取。

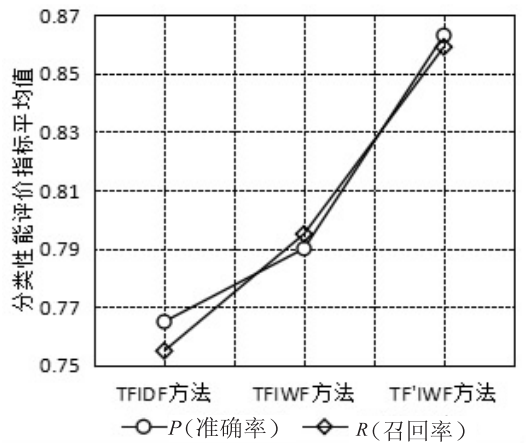


图 4　不同关键词提取算法对分类结果的影响

4.3.3　引入 Word2vec 方法分类对比

通过三个实验来验证该方法的效果。其中使用原始的 TFIWF+Word2vec 的词嵌入特征测试短文本的分类性能;用 Word2vec 与改进后的 TF'IWF 加权方法作为第二组对比实验;最后用文中改进的 TF'IWF-$R_j$ 方法与 Word2vec 进行对比。实验数据如表 2 所示。

从表中可以看出,文中方法比 Word2vec 和 TF'IWF 相融合的方法在准确率、召回率都有提高,其中 $F_1$ 值从 89.2% 上升到 91.7%,整体提高了 2.5%;与 Word2vec 融和 TFIWF 方法相比较,在 $F_1$ 值指标中有 8.7% 左右的提高。

表 2　引入 Word2vec 的分类结果对比

| 类别 | Word2vec 与 TFIWF 方法 | | | Word2vec 与 TF'IWF 方法 | | | Word2vec+TF'IWF-$R_j$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| 历史 | 0.814 | 0.824 | 0.819 | 0.892 | 0.853 | 0.872 | 0.913 | 0.897 | 0.905 |
| 军事 | 0.927 | 0.831 | 0.876 | 0.939 | 0.937 | 0.938 | 0.951 | 0.945 | 0.948 |
| 文化 | 0.802 | 0.825 | 0.813 | 0.877 | 0.837 | 0.857 | 0.898 | 0.891 | 0.894 |
| 经济 | 0.831 | 0.822 | 0.826 | 0.831 | 0.879 | 0.854 | 0.872 | 0.901 | 0.886 |
| 教育 | 0.822 | 0.807 | 0.814 | 0.906 | 0.891 | 0.898 | 0.927 | 0.928 | 0.927 |
| IT | 0.842 | 0.812 | 0.827 | 0.931 | 0.931 | 0.931 | 0.943 | 0.957 | 0.950 |
| 娱乐 | 0.839 | 0.824 | 0.831 | 0.889 | 0.896 | 0.892 | 0.894 | 0.923 | 0.908 |
| 法制 | 0.842 | 0.821 | 0.831 | 0.892 | 0.889 | 0.890 | 0.917 | 0.924 | 0.920 |
| 平均值 | 0.840 | 0.821 | 0.830 | 0.895 | 0.889 | 0.892 | 0.914 | 0.921 | 0.917 |

4.3.4 综合比较

由于以上对比改进的关键词提取算法和引入 Word2vec 都是基线比较,为了直观表现该方法的有效性,将比较只引入词嵌入的方法 1:Word2vec,只引入关键词最优的提取算法作为方法 2:TF'IWF,以及与文中所提出的方法 3:TF'IWF- $R_j$ +Word2vec 作比较。比较结果如图 5 所示,可以直观地看出文中算法的有效性。

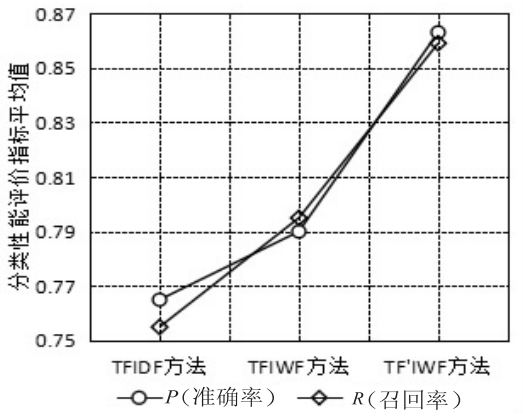


图 5 方法综合比较

## 5 结束语

与传统文本向量空间模型相比,通过 Word2vec 扩展短文本特征向量,使用语义相关度并融合改进后的权重计算可以有效地改进短文本的稀疏性和缺少上下文语义分析的问题,进而提高了短文本分类的性能。下一步可以考虑分类主题模型的创新以及对扩展训练语料进行研究,从而继续提高算法的精度和适用性。

**参考文献:**

[1] CHEN Mengen,JIN Xiaoming,SHEN Dou. Short text classification improved by learning multi-granularity topics[C]// International joint conference on artificial intelligence. Barcelona:AAAI Press,2011:1776-1781.

[2] NIGAM K,MCCALLUM A,THRUN S,et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning,2000,39(2-3):103-134.

[3] WANG Bingkun,HUANG Yongfeng,YANG Wanxia,et al. Short text classification based on strong feature thesaurus [J]. Journal of Zhejiang University Science C,2012,13(9):649-659.

[4] CHENG Xueqi, YAN Xiaohui, LAN Yanyan,et al. BTM: topic modeling over short texts[J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(12):2928-2941.

[5] GAO Longwen,ZHOU Shuigeng,GUAN Jihong. Effectively classifying short texts by structured sparse representation with dictionary filtering[J]. Information Sciences,2015,323:130-142.

[6] WANG Ying, WANG Xin, ZUO Wanli. Exploring social context for topic identification in short and noisy texts[C]// Twenty-ninth AAAI conference on artificial intelligence. Austin Texas,USA:AAAI,2015.

[7] BOLLEGALA D, MATSUO Y, ISHIZUKA M. A web search engine based approach to measure semantic similarity between words[J]. IEEE Transactions on Knowledge and Data Engineering,2011,23(7):977-990.

[8] SHIRAKAWA M A,NAKAYAMA K,HARA T,et al. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes[J]. IEEE Transactions on Emerging Topics in Computing,2017,3(2):205-219.

[9] 刘文军,郑国义,张小琼. 基于粗糙集与统计学习理论的样本分类算法[J]. 模糊系统与数学,2015,29(1):183-190.

[10] 王义真,郑啸,后盾, 等. 基于 SVM 的高维混合特征短文本情感分类[J]. 计算机技术与发展,2018,28(2):88-93.

[11] PHAN X,NGUYEN C,LE D,et al. A hidden topic-based framework toward building applications with short web documents[J]. IEEE Transactions on Knowledge and Data Engineering,2011,23(7):961-976.

[12] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha,Qatar:Association for Computational Linguistics,2014:1746-1751.

[13] WANG Peng,XU Bo,XU Jiaming,et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing,2016,174:806-814.

[14] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th international conference on neural information processing systems. Lake Tahoe,Nevada,USA:Curran Associates Inc. ,2013:3111-3119.

[15] 李学明,李海瑞,薛 亮,等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程,2012,38(8):37-40.

[16] 耿丽娟,李星毅. 用于大数据分类的算法研究[J]. 计算机应用研究,2014,31(5):1343-1344.