



融合单词贡献度与 Word2Vec 词向量的文档表示

彭俊利, 谷雨, 张震, 耿小航

(杭州电子科技大学 通信信息传输与融合技术国防重点学科实验室, 杭州 310000)

摘要: 针对现有文档向量表示方法受噪声词语影响和重要词语语义不完整的问题, 通过融合单词贡献度与 Word2Vec 词向量提出一种新的文档表示方法。应用数据集训练 Word2Vec 模型, 计算数据集中词语的贡献度, 同时设置贡献度阈值, 提取贡献度大于该阈值的单词构建单词集合。在此基础上, 寻找文档与集合中共同存在的单词, 获取其词向量并融合单词贡献度生成文档向量。实验结果表明, 该方法在搜狗中文文本语料库和复旦大学中文文本分类语料库上分类的平均准确率、召回率和 F1 值均优于 TF-IDF、均值 Word2Vec、PTF-IDF 加权 Word2Vec 模型等传统方法, 同时其对英文文本也能进行有效分类。

关键词: 单词贡献度; Word2Vec 词向量; 词嵌入; 文档表示; 文本分类

开放科学(资源服务)标志码(OSID):



中文引用格式: 彭俊利, 谷雨, 张震, 等. 融合单词贡献度与 Word2Vec 词向量的文档表示[J]. 计算机工程, 2021, 47(4): 62-67.

英文引用格式: PENG Junli, GU Yu, ZHANG Zhen, et al. Document representation fused with term contribution and Word2Vec word vector[J]. Computer Engineering, 2021, 47(4): 62-67.

Document Representation Fused with Term Contribution and Word2Vec Word Vector

PENG Junli, GU Yu, ZHANG Zhen, GENG Xiaohang

(National Defense Key Discipline Laboratory of Communication Information Transmission and Fusion Technology, Hangzhou Dianzi University, Hangzhou 310000, China)

[Abstract] The existing document vector representation methods are affected by noise words and the semantics of important words is incomplete. To address the problems, this paper proposes a new document representation method by fusing Term Contribution (TC) and Word2Vec word vector. Trained with a dataset, the Word2Vec model calculates the TC of words in the data set. Then the contribution threshold is set and the words whose TC is greater than the threshold are extracted to construct a word set. On this basic, the word that exists both in the document and the set is extracted, and its word vector is fused with the TC to generate the document vector. Experimental results show that the average accuracy, recall rate and F1 value of the proposed method on Sogou Chinese text corpus and Fudan University Chinese text classification corpus are better than those of traditional methods such as TF-IDF, mean Word2Vec and PTF-IDF weighted Word2Vec models. Meanwhile, it can also effectively classify English texts.

[Key words] Term Contribution (TC); Word2Vec word vector; word embedding; document representation; text classification

DOI: 10.19678/j.issn.1000-3428.0056370

0 概述

随着深度学习技术的快速发展, 文档表示方法已由基于词频信息的词袋模型(Bag-of-Words, BOW)^[1-2]逐渐转向基于词嵌入(Word Embedding)的表示法。词袋模型将文档视为多个词的集合, 其不考虑词的顺序和语义等信息, 使用与词集合相同维度的向量来表示

文档, 向量中每一维所包含的数值即为该位置所表示词的权重^[3-4]。虽然词袋模型在支持向量机(Support Vector Machine, SVM)、贝叶斯分类器和逻辑回归分类器中可得到较好效果, 但仍存在一些问题^[5-6]。当数据集较大时, 采用词袋模型获得的文档向量维度会很高, 从而导致维度灾难, 而且文档中出现的词语数量较多, 但在表示为高维向量时只有极少数的维度存在有效权

基金项目: 国家自然科学基金(61673146)。

作者简介: 彭俊利(1993—), 男, 硕士研究生, 主研方向为机器学习、自然语言处理; 谷雨(通信作者), 副教授、博士; 张震、耿小航, 硕士研究生。

收稿日期: 2019-10-22 修回日期: 2020-01-02 E-mail: guyu@hdu.edu.cn

重^[7-8]。例如,文档中出现 1 000 个词,词向量维度为 10 万,但其中仅有 1 000 个维度存在有效权重。此外,词袋模型仅考虑了词语的频次信息,没有在词语与上下文之间建立联系,导致词语的语义信息不足,从而无法区分一词多义或多词一词的情况。

以词嵌入为代表的基于深度学习的词向量表示法在词语与上下文之间建立联系,把维数为所有词语数量的高维空间嵌入到一个维数较低的连续向量空间中,每个词语都被映射为实数域上的向量,弥补了词频统计法存在的不足^[9-10]。MIKOLOV 等人于 2013 年构建 Word2Vec 词嵌入模型^[11],其利用词语与上下文的关系将词语转化为一个低维实数向量,从而有效地区分了一词多义或多词一词的情况。此后,EMLO、BERT 等词嵌入模型相继出现。虽然这些模型在多项自然语言处理任务中均获得了性能提升,但由于 Word2Vec 能够简单、高效地获取词语的语义向量,因此其依然被广泛应用于分类任务、推荐系统和中文分词等方面。

文档由大量词语构成,但其中只有少数词语能代表文档,而大部分为噪声词语,如何去噪声词语,充分利用具有表征性词语的语义信息构建文档向量是一个难题。现有研究多将词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)算法^[12]和 Word2Vec 相结合构建文档向量。文献[3]通过考虑词语的重要性,提出一种融合 Word2Vec 词向量与 TF-IDF 权重的文档表示法,并在搜狗中文文本语料库上验证了方法的有效性。文献[13]采用 Word2Vec 对微博文本进行扩展后以 TF-IDF 方法表示句向量,将句子中每个词的词向量相加形成句向量。文献[14]将词性引入 TF-IDF 算法,结合 Word2Vec 生成文本向量,并在复旦大学中文文本分类数据集上验证方法的有效性。文献[15]通过融合 Word2Vec 模型与改进 TF-IDF 算法获取文本向量,利用卷积神经网络进行分类,并在 THUCNews 数据集上验证方法的有效性。

上述方法结合了词语的 TF-IDF 权重与 Word2Vec 词向量,但均未考虑文档中只有少数词语具有表征性的事实,影响了分类性能。针对该问题,本文提出一种新的文档表示方法。设计改进的单词贡献度(Term Contribution, TC)算法筛选具有表征性的词语集合,将其中所有词语的单词贡献度与 Word2Vec 词向量相结合构建文档向量。为验证该方法的有效性,在搜狗中文文本语料库、复旦大学中文文本分类语料库和 IMDB 英文语料库上进行实验,并与相关方法进行对比。

1 相关工作

1.1 词的量化表示

1.1.1 TF-IDF 算法

TF-IDF 算法以概率统计为基础,计算一个词语在数据集中的重要程度。当一个词语在某篇文档中出现的频率越高而在数据集其他文档中出现的频率越低,则该词语的表征性越强,其 TF-IDF 值越大。在 TF-IDF 算法中,TF 指词频,即词语在文档中出现

的频率,IDF 指逆文档频率。TF 和 IDF 的计算公式分别如式(1)和式(2)所示:

$$tf_{ij} = \frac{|w_i|}{|d_j|} \quad (1)$$

$$idf_i = \log_a \frac{|D|}{n_w + 1} \quad (2)$$

其中, $|w_i|$ 表示词语 w_i 在文档 d_j 中出现的次数, $|d_j|$ 表示文档 d_j 中所有词语的总数, $|D|$ 表示数据集中的文档总数, n_w 表示包含词语 w_i 的文档数目。由此可以得出词语 w_i 在文档 d_j 中归一化后的 TF-IDF 权重计算公式,如式(3)所示:

$$f(w_i, d_j) = \frac{tf_{ij} \times idf_i}{\sqrt{\sum_{w_i \in d_j} (tf_{ij} \times idf_i)^2}} \quad (3)$$

然而,TF-IDF 权重只考虑了词语在数据集中的频次信息,无法表达词语的语义信息。例如,“番茄”和“西红柿”表示的意义相同,若在数据集中出现的频次不同,则其 TF-IDF 权重可能会相差很大。

1.1.2 Word2Vec 模型

受文献[16]提出的 NNLM 模型启发,MIKOLOV 等人^[11]提出了 Word2Vec 模型。Word2Vec 与 NNLM 的区别在于:NNLM 是一个语言模型,词向量只是“副产品”,而 Word2Vec 是一种用于获取词向量的词嵌入模型。

Word2Vec 主要包括 CBOW 和 Skip-gram 两种模型^[17]:CBOW 模型利用词 w_i 的前后各 c 个词来预测当前词,如图 1(a)所示;Skip-gram 模型利用 w_i 预测其前后各 c 个词,如图 1(b)所示。

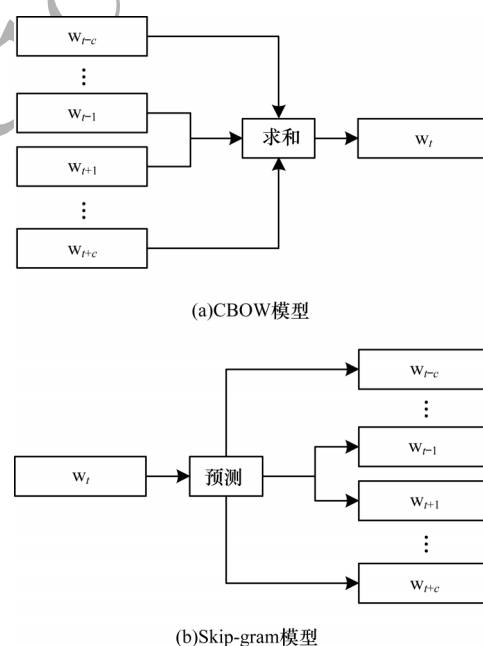


图1 Word2Vec 模型示意图

Fig.1 Schematic diagram of Word2Vec model

CBOW 模型的输入层是词 w_i 的前后 $2c$ 个 one-hot 词向量,投影层将这 $2c$ 个词向量累加求和,输出层是一棵

以训练数据中所有的词作为叶子节点,以各词在数据中出现的次数作为权重的 Huffman 树^[3],此模型应用随机梯度上升法预测投影层的结果作为输出。Skip-gram 模型与之类似。当获得所有词的词向量后,可发现类似如下规律:“king”-“man”+“woman”=“queen”^[18],可见词向量能够有效表达词语的语义信息。

1.2 单词贡献度

单词贡献度^[19]用于计算数据集中特征词 t 对于文本相似性的贡献程度,计算公式如式(4)所示:

$$T_{TC}(w) = \sum_{i,j \in T \neq j} f(w, d_i) \times f(w, d_j) \quad (4)$$

其中, $f(w, d)$ 表示单词 w 在文档 d 中的 TF-IDF 权重。

词语的文档频数越高其 IDF 值越低,当所有文档都包含该特征词时,其 IDF 值为 0。而那些在大部分文本中出现的特征词因为 IDF 值非常小,所以 TC 值也会较小。当特征词只出现在一个文本或所有文本中,其 TC 值为 0。在文档频数适中的情况下,TF-IDF 权重较大的特征词具有较大的 TC 值。

1.3 文档的量化表示

文档的量化表示就是将非结构化的文本信息转化为计算机可处理的数字信息^[20],本文主要介绍以下两种表示方法。

1.3.1 基于词频信息的表示方法

经典的文档量化方法是 BOW 模型,主要原理是词 one-hot 编码的叠加^[3],如数据集中共 9 个词语,其中,“番茄”的 one-hot 编码是 $[0, 0, 0, 1, 0, 0, 0, 0, 0]$,“西红柿”的 one-hot 编码是 $[0, 1, 0, 0, 0, 0, 0, 0, 0]$,若一篇文档仅包含“番茄”和“西红柿”,则文档向量表示为 $[0, 1, 0, 1, 0, 0, 0, 0, 0]$ 。BOW 模型中仅体现了词语是否出现在文档中,可以看出很难利用该向量计算出与实际相符的文档相似度。因此,研究者提出了许多改进方法。应用词的 TF-IDF 权重代替 BOW 模型中的非“0”值是最常用的一种方法,它能够有效计算文本间的相似度,但这种基于词频统计的方法容易导致向量的高维性和稀疏性。

1.3.2 基于 Word2Vec 的表示方法

基于 Word2Vec 的文档表示方法解决了传统词向量高维性和稀疏性的问题,并引入了语义信息,因此其被广泛应用。此类方法的基本流程是先获取文档中所有词的词向量,再通过聚类或取平均值的方法进行文档向量表示^[3]。很多研究者考虑到单词的重要性,将 TF-IDF 权重与 Word2Vec 相结合,这样虽然可以使性能得到提升,但很多无区分度的词语依然会影响文档向量的表征性。

2 融合改进 TC 算法与 Word2Vec 的文档表示

考虑到文档中无区分度词语对文档向量表征性的影响以及词语在文档中的重要性,本文对传统单词贡献度算法进行改进,在 Word2Vec 词向量的基础上,结合改进算法提出一种新的文档向量表示方法。

2.1 改进的单词贡献度算法

由式(4)可知,传统的单词贡献度算法是将不同文本中相同单词的 TF-IDF 值两两相乘再相加,这样会严重弱化 IDF 值所包含的语义信息,即弱化单词在整个数据集中的重要程度。为解决这一问题,本文提出一种新的计算方法:先将每篇文档中相同单词的 TF 值进行两两相乘再相加的操作,得到根据 TF 值计算出的单词权值,再将该值与 IDF 相乘。这样得到的单词贡献度不仅保留了由 TF 值计算得到的权值,同时也保留了 IDF 值包含的完整语义信息,提高了特征词与噪声词的区分度。单词贡献度的计算公式如式(5)所示:

$$T_{TC}(w_i) = \sum_{i,j \in T \neq j} (tf_{ii} \times tf_{ij}) \times idf_i \quad (5)$$

其中, tf_{ii} 和 tf_{ij} 表示单词 w_i 在第 i 篇文档和第 j 篇文档中的文档频率 TF 值, idf_i 表示 w_i 的逆文档频率 IDF 值。

2.2 Word2Vec 词向量与改进单词贡献度的融合

设数据集 D 中包含 M 个文档,首先将 M 个文档中的内容采用分词工具进行分词,利用 Word2Vec 模型进行训练,获取每个词语对应的 n 维词向量 $V = (v_1, v_2, \dots, v_n)$,同时应用改进的单词贡献度算法计算每个单词的贡献度。对所有单词按贡献度大小进行降序排列,设置贡献度阈值 x ,选出贡献度大于 x 的前 W 个单词构建单词集合 T 。

对于文档 $d_j (j=1, 2, \dots, M)$,找出集合 T 中存在的单词,文档向量可表示为:

$$V_{d_j} = \sum_{w_i \in d_j} V_{w_i} T_{TC} w_i \quad (6)$$

其中, V_{w_i} 表示词语 w_i 的词向量, TC_{w_i} 表示词语 w_i 的单词贡献度。通过改进 TC 与 Word2Vec 的融合,能够将单词对文档的重要性权值融入包含语义信息的词向量,使词向量更具表征性。本文方法流程如图 2 所示。

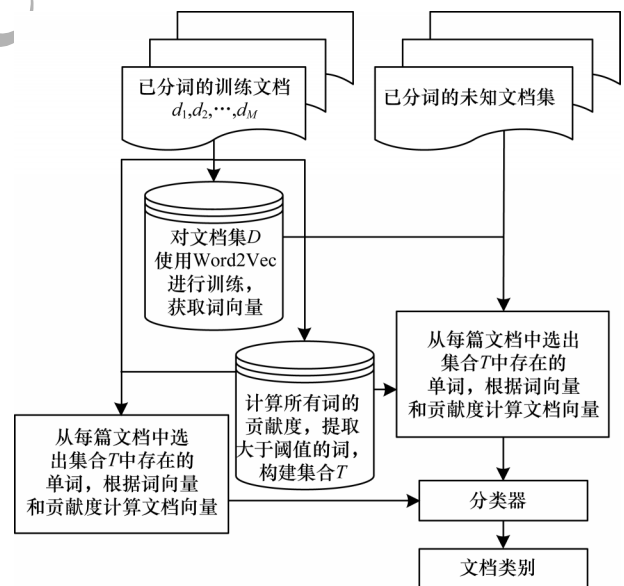


图2 本文方法流程

Fig.2 Procedure of the proposed method

3 实验

3.1 分类性能评价指标

本文采用的是分类任务最常用的评价指标准确率、召回率和 F_1 值。

准确率是指分类结果中某类被正确分类的文档数目与所有被分入该类文档总数的比值,计算公式如式(7)所示:

$$P = \frac{TP}{TP + FP} \quad (7)$$

召回率是指某类被正确分类的文档数与该类实际文档数的比值,计算公式如式(8)所示:

$$R = \frac{TP}{TP + FN} \quad (8)$$

F_1 值是综合准确率与召回率的评价指标,计算公式如式(9)所示:

$$F_1 = \frac{2PR}{P + R} \quad (9)$$

在上述评价指标中,TP 表示分入 A 类实际也为 A 类的文档数,FN 表示未分入 A 类而实际为 A 类的文档数,FP 表示分入 A 类而实际不为 A 类的文档数。

3.2 实验结果与分析

选用搜狗实验室整理的中文文本分类语料库作为实验数据集,语料库内文本被分为 9 类,分别为财经(C1)、互联网(C2)、健康(C3)、教育(C4)、军事(C5)、旅游(C6)、体育(C7)、文化(C8)和招聘(C9),每类包含 1 990 篇文档。

分别采用 TF-IDF、均值 Word2Vec、TF-IDF 加权 Word2Vec、TC 加权 Word2Vec 模型构建的文档向量与本文模型构建的文档向量进行分类结果对比。经过多次试验,在计算单词贡献度时将贡献度阈值设置为效果最佳的 0.009,本文模型提取 19 082 个具有表征性的词语。选用 lib-svm 作为分类器,所有实验采用五折交叉验证。各模型的分类性能对比如表 1~表 5 所示。

由表 1、表 2 和表 5 可以看出,本文模型的分类性能优于 TF-IDF 模型和均值 Word2Vec 模型,表明 TF-IDF 与 Word2Vec 融合后生成的词向量包含更丰富的语义信息,能够更准确地进行分类。由表 3、表 4 和表 5 可以看出,本文模型在 SVM 分类器上平均准确率、召回率和 F_1 值较 TF-IDF 加权 Word2Vec 模型分别提升 2.27%、2.24% 和 2.26%,较 TC 加权 Word2Vec 模型分别提升 1.32%、1.29% 和 1.25%。通过比较 5 种模型分类性能评价的平均值可以看出,本文模型在准确率、召回率、 F_1 值指标上均获得

了最佳的效果,验证了本文方法在中文文档表示方面的有效性。

表 1 TF-IDF 模型分类性能

Table 1 Classification performance of TF-IDF model %

类别	准确率	召回率	F_1 值
C1	74.75	76.63	75.68
C2	84.97	82.41	83.67
C3	84.70	77.89	81.15
C4	86.27	89.95	88.07
C5	78.32	77.14	77.72
C6	70.88	69.10	69.97
C7	84.66	80.40	82.47
C8	75.37	76.13	75.75
C9	72.48	81.41	76.69
平均值	79.16	79.01	79.02

表 2 均值 Word2Vec 模型分类性能

Table 2 Classification performance of average Word2Vec model %

类别	准确率	召回率	F_1 值
C1	77.86	80.40	79.11
C2	85.50	85.93	85.71
C3	91.29	81.66	86.21
C4	88.52	92.96	90.67
C5	83.65	78.39	80.93
C6	74.62	73.12	73.86
C7	83.68	79.90	81.75
C8	81.07	83.92	82.47
C9	76.24	84.67	80.24
平均值	82.49	82.33	82.33

表 3 TF-IDF 加权 Word2Vec 模型分类性能

Table 3 Classification performance of TF-IDF weighted Word2Vec model %

类别	准确率	召回率	F_1 值
C1	85.28	84.42	84.85
C2	85.56	89.20	87.55
C3	91.78	84.17	87.81
C4	91.57	95.48	93.48
C5	89.92	82.91	86.27
C6	79.60	79.40	79.50
C7	87.86	83.67	85.71
C8	84.60	86.93	85.75
C9	81.26	90.45	85.61
平均值	86.42	86.29	86.28

表4 TC加权Word2Vec模型分类性能

Table 4 Classification performance of
TC weighted Word2Vec model %

类别	准确率	召回率	F_1 值
C1	87.21	85.68	86.44
C2	90.48	90.70	90.59
C3	91.85	84.92	88.25
C4	88.42	93.97	91.11
C5	90.21	85.68	87.89
C6	83.29	82.66	82.98
C7	88.05	85.18	86.59
C8	86.28	86.39	86.61
C9	80.54	89.45	84.76
平均值	87.37	87.24	87.25

表5 本文模型分类性能

Table 5 Classification performance of
the proposed model %

类别	准确率	召回率	F_1 值
C1	88.97	87.19	88.07
C2	91.67	91.21	91.44
C3	93.65	85.18	89.21
C4	88.00	93.97	90.89
C5	92.06	87.44	89.69
C6	90.91	87.94	89.40
C7	84.92	84.92	84.92
C8	85.57	87.94	86.74
C9	82.46	90.95	86.50
平均值	88.69	88.53	88.54

此外,采用复旦大学中文文本分类语料库中文档数目较多的Art类、Agriculture类、Economy类和Politics类数据集进行实验,将本文模型与文献[13]提出的PTF-IDF加权Word2Vec模型进行对比,如图3所示。可以看出,本文模型在准确率、召回率和 F_1 值上均具有优势,进一步验证了本文方法的有效性。

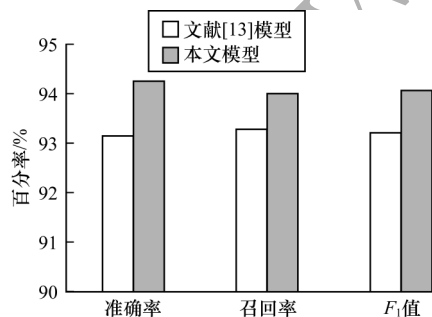


图3 本文模型与文献[13]模型的性能对比

Fig.3 Classification performance comparison of
the proposed model and the model in literature[13]

上述实验均采用了中文语料,仅验证了本文方法在中文文本表示方面的有效性。为验证其对英文文本表示的有效性,选取IMDB文本情感分类语料库作为英文实验数据集,其中包括积极和消极两类数据,从每类中选取2 500篇英文数据,采用五折交叉验证。结果表明,本文模型分类准确率为76.26%,验证了其在英文文本表示方面的有效性。但由结果可知,英文文本分类准确率低于中文文本分类,这可能有两点原因:1)由于英文数据集较少,本文仅用3万多篇英文文本训练Word2Vec模型,训练语料的不足导致模型无法学习到单词较为准确的语义信息;2)由于英文存在各类语态的表达方法,单词在不同语态下需要添加不同后缀,使得多词一义的情况进一步加重,Word2Vec模型可能会将加了不同后缀的单词理解为不同的单词,单词贡献度的计算也因此受到影响。

4 结束语

针对当前文档向量表示方法存在的不足,本文综合考虑单词的重要程度和语义信息,将计算出的贡献度权值与Word2Vec词向量进行融合,提出一种新的文档表示方法。实验结果表明,应用于中文文本分类任务时,本文模型较基于TF-IDF模型、均值Word2Vec模型、TF-IDF加权Word2Vec模型和传统TC加权Word2Vec模型效果更好,并且其对英文文本也可实现有效分类。后续将从降低单词贡献度算法的时间复杂度出发,进一步优化本文方法。

参考文献

- [1] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval [M]. New York, USA: ACM Press, 1999.
- [2] MANNING C D, SCHUTZE H. Foundations of statistical natural language processing [M]. Cambridge, USA: MIT Press, 1999.
- [3] MA Linjin, WAN Liang, MA Shaoju, YANG Ting. Abnormal traffic identification method based on bag of words model clustering [J]. Computer Engineering, 2017, 43(5): 204-209. (in Chinese)
马林进, 万良, 马绍菊, 等. 基于词袋模型聚类的异常流量识别方法 [J]. 计算机工程, 2017, 43(5): 204-209.
- [4] LEI Shuo, LIU Xumin, XU Weixiang. Chinese short text classification based on word vector extension [J]. Computer Applications and Software, 2018, 35(8): 269-274. (in Chinese)
雷朔, 刘旭敏, 徐维祥. 基于词向量特征扩展的中文短文本分类研究 [J]. 计算机应用与软件, 2018, 35(8): 269-274.
- [5] HWANG M, CHOI C, YOUN B, et al. Word sense disambiguation based on relation structure [C] // Proceedings of 2008 International Conference on Advanced Language Processing and Web Information Technology. New York, USA: ACM Press, 2008: 15-20.
- [6] WANG X, McCALLUM A, WEI X. Topical N-grams: phrase and topic discovery, with an application to information retrieval [C] // Proceedings of IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2007: 697-702.

- [7] CHEN Xingjian, HU Xuejiao, XUE Wei. Improved bag of words model based on relational expansion[J]. Journal of Chinese Computer Systems, 2019, 40(5): 1040-1044. (in Chinese)
陈行健, 胡雪娇, 薛卫. 基于关系拓展的改进词袋模型研究[J]. 小型微型计算机系统, 2019, 40(5): 1040-1044.
- [8] CHEN Wenshi, LIU Xinhui, LU Mingyu. Feature extraction of deep topic model for multi-label text classification[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(9): 785-792. (in Chinese)
陈文实, 刘心惠, 鲁明羽. 面向多标签文本分类的深度主题特征提取[J]. 模式识别与人工智能, 2019, 32(9): 785-792.
- [9] HAN Xuli, ZENG Biqin, ZENG Feng, et al. Sentiment analysis based on word embedding auxiliary mechanism[J]. Computer Science, 2019, 46(10): 258-264. (in Chinese)
韩旭丽, 曾碧卿, 曾锋, 等. 基于词嵌入辅助机制的情感分析[J]. 计算机科学, 2019, 46(10): 258-264.
- [10] ZHENG Cheng, HONG Tongtong, XUE Manyi. BLSTM_MLPCNN model for short text classification[J]. Computer Science, 2019, 46(6): 206-211. (in Chinese)
郑诚, 洪彤彤, 薛满意. 用于短文本分类的 BLSTM_MLPCNN 模型[J]. 计算机科学, 2019, 46(6): 206-211.
- [11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of International Conference on Learning Representations. Scottsdale, USA: [s. n.], 2013: 1-12.
- [12] TU Shouzhong, HUANG Minlie. Mining microblog user interests based on TextRank with TF-IDF factor[J]. The Journal of China Universities of Posts and Telecommunications, 2016, 23(5): 40-46.
- [13] DUAN Xulei, ZHANG Yangsen, SUN Yizhuo. Research on sentence vector representation and similarity calculation method about Microblog texts[J]. Computer Engineering, 2017, 43(5): 143-148. (in Chinese)
段旭磊, 张仰森, 孙卓. 微博文本的句向量表示及相似度计算方法研究[J]. 计算机工程, 2017, 43(5): 143-148.
- [14] WANG Jing, LUO Lang, WANG Deqiang. Research on Chinese short text classification based on Word2Vec[J]. Computer Systems and Applications, 2018, 27(5): 209-215. (in Chinese)
汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究[J]. 计算机系统应用, 2018, 27(5): 209-215.
- [15] WANG Gensheng, HUANG Xuejian. Convolution neural network text classification model based on Word2vec and improved TF-IDF[J]. Journal of Chinese Computer Systems, 2019, 40(5): 1120-1126. (in Chinese)
王根生, 黄学坚. 基于 Word2vec 和改进型 TF-IDF 的卷积神经网络文本分类模型[J]. 小型微型计算机系统, 2019, 40(5): 1120-1126.
- [16] BENGIO Y, SCHWENK H, SENEAL J S, et al. Neural probabilistic language models[M]. Berlin, Germany: Springer, 2006.
- [17] GAO Mingxia, LI Jingwei. Chinese short text classification method based on word2vec embedding[J]. Journal of Shandong University(Engineering Science), 2019, 49(2): 34-41. (in Chinese)
高明霞, 李经纬. 基于 word2vec 词模型的中文短文本分类方法[J]. 山东大学学报(工学版), 2019, 49(2): 34-41.
- [18] MIKOLOV T, YIH W, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Atlanta, USA: NAACL Press, 2013: 746-751.
- [19] LIU Tao, LIU Shengping, CHEN Zheng, et al. An evaluation on feature selection for text clustering[C]//Proceedings of the 20th International Conference on International Conference on Machine Learning. Washington D. C., USA: AAAI Press, 2003: 488-495.
- [20] NIE Weimin, CHEN Yongzhou, MA Jing. A text vector representation model merging multi-granularity information[J]. Data Analysis and Knowledge Discovery, 2019, 3(9): 45-52. (in Chinese)
聂维民, 陈永洲, 马静. 融合多粒度信息的文本向量表示模型[J]. 数据分析与知识发现, 2019, 3(9): 45-52.

编辑 金胡考