计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):256−261 [doi: 10.15888/j.cnki.csa.007030]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Word2Vec 的微博文本分类研究①

牛雪莹, 赵恩莹

(太原科技大学 计算机科学与技术学院, 太原 030024)
通讯作者: 牛雪莹, E-mail: 943863079@qq.com

摘　要: 以微博为代表的社交平台是信息时代人们必不可少的交流工具. 挖掘微博文本数据中的信息对自动问答、舆情分析等应用研究都具有重要意义. 短文本数据的分类研究是短文本数据挖掘的基础. 基于神经网络的 Word2vec 模型能很好的解决传统的文本分类方法无法解决的高维稀疏和语义鸿沟的问题. 本文首先基于 Word2vec 模型得到词向量, 然后将类别因素引入传统权重计算方法 TF-IDF (Term Frequency-Inverse Document Frequency) 设计词向量权重, 进而用加权求和的方法得到短文本向量, 最后用 SVM 分类器对短文本做分类训练并且通过微博数据实验验证了该方法的有效性.

关键词: Word2Vec; 短文本分类; TF-IDF

引用格式: 牛雪莹,赵恩莹.基于 Word2Vec 的微博文本分类研究.计算机系统应用,2019,28(8):256–261. http://www.c-s-a.org.cn/1003-3254/7030.html

Research on Chinese Weibo Text Classification Based on Word2Vec

NIU Xue-Ying, ZHAO En-Ying

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: The Chinese Weibo is an indispensable communication tool for people today. Mining information in Weibo text is of great significance to automatic question and answer, public opinion analysis and other applied research. The short text classification study is the basis of short text mining. The neural network-based Word2Vec can solve problems of high-dimensional sparseness and semantic gap that traditional text categorization methods cannot solve. This study obtains the word vector based on Word2Vec, then the class factor is introduced into the traditional weight calculation method TF-IDF (Term Frequency-Inverse Document Frequency) to design the word vector weight. Finally, the SVM classifier is used for classification. The effectiveness of the method is verified by experiments on Weibo data.

Key words: Word2Vec; short text classification; TF-IDF

移动互联网的高速发展让人们随时发表言论成为了可能. 以微博、微信等为代表的社交平台成为人们沟通交流的主要方式, 同时积累了越来越多的文本数据, 特别是短文本数据. 这些数据中蕴含着很多重要的信息, 对这些信息的分类和挖掘吸引了很多学者关注. 对短文本的分类研究是自然语言处理的一个重要分支, 在搜索引擎、自动问答、情感分析和舆情分析等方面有重要意义[1].

传统的向量空间模型 (vector space model)[2]对长文本的分类表现出很好的效果, 但用于短文本分类却存在特征稀疏和维度灾难的问题, 所以直接应用向量空间模型解决短文本分类问题效果并不理想. 面对这个问题, 国内外研究者主要从特征扩展和抽象语义特征两方面给出了解决方案.

特征扩展包括利用主题模型扩展和借助外部知识库扩展特征. 由于主题是词语的高层次语义抽象, 主题

① 收稿时间: 2019-02-17; 修改时间: 2019-03-08; 采用时间: 2019-03-14; csa 在线出版时间: 2019-08-08

相对词语来说会少很多, 这样就能很好的解决维度灾难问题. Phan XH 等[3]通过分析文本的主题, 并且结合 TF-IDF 来确定特征. 张志飞等[4]提出基于文档主题生成模型 LDA 的文本分类模型. 这些都是直接用主题分布来做文本特征.

很多学者[5–9]希望通过外部知识库 (例如知网、维基百科、WordNet 等) 对词语进行扩展, 以期解决特征稀疏的问题, 但是这个方法受到所用知识库质量的影响. Bouaziz[10]提出先利用 LDA 模型学习维基百科数据上的主题以及主题在词语上的分布, 然后用这些来扩展短文本, 再使用语义随机森林对扩展特征进行选抽象高层语义的方法. 这是结合了主题模型和外部知识库来进行特征扩展的方法.

也有很多学者[11–15]希望抽象文本语义特征来进行文本分类研究. 近几年深度学习通过深层次的神经网络实现对特征的高层语义抽象在自然语言处理方面表现突出. 韩栋[16]、冯国明[17]分别采用深度学习的 CNN 和 CapsNet 网络进行中文文本分类研究都取得了较好的结果. 以 Word2Vec 为代表的词向量模型是通过神经概率语言模型学习到词语的向量表达, 很多学者[16,17]在此基础上利用一定的权重组合方式得到文本的向量表达, 进而进行分类研究.

本文采用的是基于 Word2Vec 的词向量模型, 首先用 Word2Vec 在维基百科中进行学习得到词向量, 然后用改进的 TF-IDF 设计权重进而得到文本向量, 最后用 SVM 分类器进行文本分类训练, 并且通过实验表明该方法与传统方法相比较分类效果有明显提高.

1 基于 Word2Vec 的短文本分类模型

短文本自动分类是一个有监督的机器学习模型. 让机器根据词语的特征学习模型然后预测文本所属的类别. 在自动文本分类领域常用的技术有朴素贝叶斯分类器、决策树、支持向量机、KNN 等. 本文结合 Word2Vec 和 TF-IDF 提出短文本分类算法, 并验证其有效性.

1.1 Word2Vec 词向量模型

Word2Vec 是 2013 年 Google 的研究员发布的一种基于神经网络的词向量生成模型. 模型是用深度学习网络对语料数据的词语及其上下文的语义关系进行建模, 以求得到低维度的词向量. 该词向量一般在 100–300 维左右, 能很好的解决传统向量空间模型高维稀疏的问题. 因为深度的神经网络模型能对特征的高层语义进行很好的抽象所以模型能很好的避免语义鸿沟. 所以 Word2Vec 是目前应用在自然语言处理方面表现较优秀的方法.

Word2Vec[18,19]主要有 Continuous Bag-of-Words Model (CBOW) 和 Continuous Skip-gram Model (Skip-gram) 两种模型, CBOW 模型是在已知上下文 Context(t) 的情况下预测当前词 t, 而 Skip-gram 模型则是在已知当前词 t 的情况下预测其上下文词 Context(t). 这两个模型都包括输入层、隐藏层和输出层, 如图 1 所示. CBOW 模型的输入层是选定窗口个数 w 的上下文词 one-hot 编码的词向量, 隐藏层向量是这些词向量、连接输入和隐含单元之间的权重矩阵计算得到的, 输出层向量可以通过隐含层向量、连接隐含层与输出层之间的权重矩阵计算得到. 最后输出层向量应用 SoftMax 激活函数, 可以计算出每个单词的生成概率. 但是由于 SoftMax 激活函数中存在归一化项的缘故, 推导出来的迭代公式需要对词汇表中的所有单词进行遍历, 使得每次迭代过程非常缓慢, 可使用 Hierarchical Softmax 来提升速度.

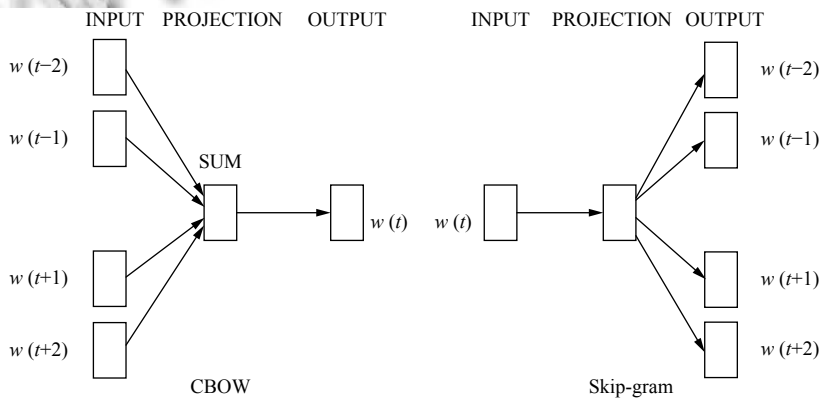

图 1 Word2Vec 模型

1.2 TF-IDF

TF-IDF[20](Term Frequency-Inverse Document Frequency) 是组合了词频和逆文档频率是一种统计方法.

词频 (Term Frequency, TF) 是指某个给定的词t_i在文档d_j中出现的频率, 频率越高对文档越重要, 数学表达公式如式 (1) 所示:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中, $n_{i,j}$表示词t_i在文档d_j中出现的次数, $\sum_k n_{k,j}$表示文档d_j中所有 k 个词出现次数的总和.

逆文档频率 (Inverse Document Frequency, IDF) 是指包含该词t_i的文档占总文档 D 的比重的倒数. 逆文档频率的出现是为了避免一些类似“我”、“的”、“他”等出现频率很高但是对文档分类作用较小的词获得高权重. 数学表达公式如式 (2) 所示:

$$idf_i=\log\frac{|D|}{\left|\left\{j:t_i\in d_j\right\}\right|} \tag{2}$$

其中, $\left|\left\{j:t_i\in d_j\right\}\right|$表示出现词$t_i$的文档数.

$$TF-IDF=tf_{i,j}\cdot idf_i \tag{3}$$

表示词语对于文本的重要性, 随着词频的增加而增大, 随着文档频率的增大而减小. 也就是说在当前文本中出现频率高且在其他文本中出现的少的词对文本的意义大, 均匀出现在各个文本中的词对文本的意义小.

1.3 短文本向量模型

很多学者[21–25]提出基于词向量生成短文本向量的方法. Le[25]等人根据 Word2Vec 生成词向量的方法扩展到语句、段落、文档的层面上提出 PV-DM 和 PV-DBOW 模型; 词向量组合法是将文本中所有词语的词向量加权求和的方法. 其中权重确定的方法包括: 直接采用词语的 TF-IDF 值为权重[21]; 采用语法、词性标注结果设置权重[22]等.

对文本分类来说词语对类别的影响更重要, 而 TF-IDF 衡量词语对某个文本的重要性并没有考虑词语在类内和类间分布情况, 所以本文考虑在 TF-IDF 的基础上加入类别因素 c, 提出新的权重确定方法 CTF-IDF, 数学表达式为式 (4):

$$CTF-IDF=c\cdot tf_{i,j}\cdot idf_i \tag{4}$$

其中,

$$c=\frac{p}{p+q},\ p=\frac{n}{n+m},\ q=\frac{k}{k+l} \tag{5}$$

类别因素 c, 随着词语 t 在类 r 中出现频率 p 的增加而增加; 随着词语 t 在非 r 类别中出现频率 q 的增加而减小, 理想情况下词语 t 都出现在某一个类别中, 类别因素 c=1. n 表示出现词语 t 且属于类别 r 的短文本数量; m 表示属于类别 r, 但没出现词语 t 的短文本数量; k 表示出现词语 t 但不属于类别 r 的短文本数量; l 表示没出现词语 t 也不属于类别 r 的短文本数量.

确定词向量权重算法 CTF-IDF 之后, 采用加权求和的方法得到短文本的向量表示, 数学表达式为 (6).

$$v_{d_j}=\sum_{i\in d_j} v_i\cdot CTF-IDF \tag{6}$$

其中, v_{d_j}表示文本d_j的向量, v_i表示词语t_i的词向量.

1.4 短文本分类流程

微博短文本的分类流程如图 2 所示. 首先对文本进行预处理, 包括去停用词、分词. 然后用 Word2Vec 模型对维基百科进行训练, 得到大量词语结合上下文和语义的词向量. 下一步是计算短文本的向量, 需要把 Word2Vec 生成的与文本对应的词向量加权求和, 权重通过词的词频和分类计算 CTF-IDF 得到. 最后进入分类器分类, 很多研究表明, 与其他分类系统相比, SVM 在分类性能上和系统健壮性上表现出很大优势[26–28], 因此选用 SVM 分类器作为分类工具, 根据短文本向量及其对应的标签训练出分类器. 测试过程与训练过程相似, 只是最后通过已训练好的分类器预测测试短文本的标签.

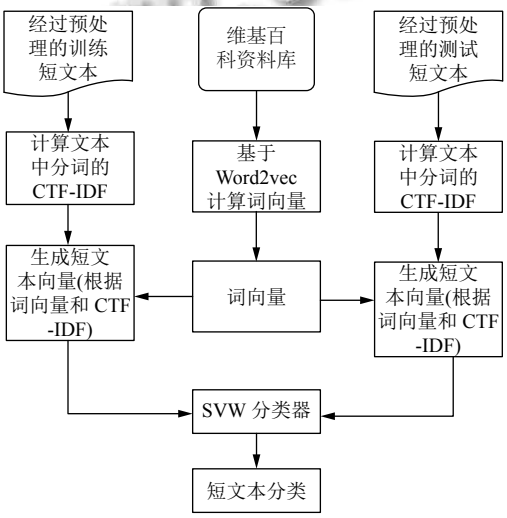

图 2 短文本分类流程图

2 微博短文本分类实验

本文前面介绍了短文本分类的流程以及通过词向

量生成短文本向量的方法, 现通过实验验证本文提出的方法的有效性.

2.1 数据来源和预处理

本文收集了从新浪微博上用八爪鱼爬取到的微博数据 29 000 条分为以下 10 个类别, 其中 IT、财经、时尚、健康、母婴、体育各 3500 条, 医疗、动漫、文学、教育各 2000 条. 所有类别 80% 的数据用于训练, 20% 的数据用于测试. 所有的数据都经过去停用词、去表情符号预处理, 并用结巴分词对数据进行了分词处理.

2.2 实验评价指标

分类任务的常用评价标准有准确率 (precision)、召回率 (recall) 和 Fl 评分[21]. 表 1 是两分类器混淆矩阵 (confusion Matrix), 其中 TP 表示实际是正类、预测也为正类的样本数量; FN 表示实际为正类、预测为反类的样本数量; FP 表示实际为反类、预测为正类的样本数量; TN 表示实际为反类、预测也为反类的样本数量. 准确率是指分类结果中被正确分类的样本个数与所有分类样本数的比例, 如式 (7) 所示.

$$P = \frac{TP}{TP + FP} \tag{7}$$

召回率是指分类结果中被正确分类的样本个数与该类的实际文本数的比例, 如式 (8) 所示.

$$R = \frac{TP}{TP + FN} \tag{8}$$

Fl 评分是综合考虑准确率与召回率的一种评价标准, 如式 (9) 所示.

$$F1 = \frac{2PR}{P + R} \tag{9}$$

表 1 两分类混淆矩阵

	预测正例	预测反例
实际正例	TP	FN
实际反例	FP	TN

2.3 分类实验和分析结果

实验分别用 TF-IDF 模型、均值加权 Word2vec 模型、TF-IDF 加权 Word2vec 模型、CTF-IDF 加权 Word2vec 模型对微博数据进行分类实验, 试图验证文章提出的 CTF-IDF 加权的有效性, 并分析分类数量对模型的影响.

对于 TF-IDF 分类模型, 使用 Scikit-learn 提供的 TfidfVectorizer 模块提取文本特征并将短文本向量化. 剩余三种都是在 Word2Vec 模型的基础上, 加权求和得到微博文本向量, 只是各自的权重确定方式不同.

均值加权 Word2Vec 模型是取全部词语向量的平均值; TF-IDF 加权 Word2Vec 模型是用对应词的 TF-IDF 为权重; CTF-IDF 加权 Word2Vec 模型是用本文提出的结合了类别因素的 CTF-IDF 为权重.

表 2 SVM 微博文本分类实验结果

类别	TF-IDF F1	Averge+Word2Vec F1	TF-IDF+Word2Vec F1	CTF-IDF+Word2Vec F1
IT	0.849	0.856	0.867	0.881
财经	0.873	0.876	0.891	0.911
时尚	0.839	0.845	0.851	0.866
法律	0.905	0.913	0.932	0.940
母婴	0.893	0.893	0.904	0.913
体育	0.849	0.851	0.860	0.874
医疗	0.913	0.922	0.933	0.947
动漫	0.854	0.860	0.871	0.890
文学	0.802	0.834	0.840	0.857
教育	0.837	0.846	0.857	0.870

从上表可以看出: 均值加权的 Word2Vec 模型比 TF-IDF 模型在 SVM 分类器的表现稍好, F1 值稍有提升, 说明 Word2Vec 模型比传统的模型生成的词向量能更好的表示文本特征, 更适应文本分类.

TF-IDF 加权的 Word2Vec 模型的表现相比均值加权的 Word2Vec 又有所提高, 这是因为相较于平均词向量, TF-IDF 加权的方法更能准确的表现词语对于文档的重要性, 所以其形成的文档向量在 SVM 分类器上表现更好. 本文提出的基于 CTF-IDF 加权的 Word2Vec 模型表现最好, 这是因为虽然 TF-IDF 考虑了不同词语对文档重要性不一样, 但是忽略了对类别的影响, 当使用加入类别因素的 CTF-IDF 权重之后文本在 SVM 分类器上表现不错. 这说明本文所提出的 CTF-IDF 加权的 Word2Vec 模型在短文本分类上的有效性.

从图 3 可以看出, Word2Vec 分类模型准确度与分类类别、类别数量等因素有关, 类别数越少模型分类准确度越高.

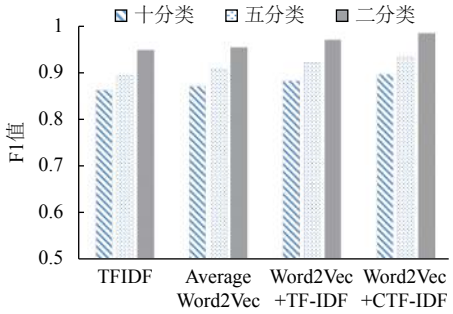

图 3 多类别微博文本分类效果图

3 结论与展望

基于 Word2Vec 的微博文本分类模型与传统的向量空间模型相比在微博短文本分类上表现良好. Word2Vec 在短文本分类问题上既可以解决高维稀疏问题又可以结合上下文语义, 但是对于词语的权重问题无法解决. 本文在 TD-IDF 的基础上提出 CTF-IDF 加权的 Word2Vec 模型, 既考虑了词频又考虑类别因素. 从实验结果可见相较于均值加权的 Word2Vec 模型、TF-IDF 加权的 Word2Vec 模型, 本文提出的 CTF-IDF 加权的 Word2Vec 模型在微博短文本分类问题上表现相对最好. 但文章也存在一些不足之处, 算法中权重确定方法忽略了词语的位置信息, 而词语的位置信息可能对于文档的语义有一定作用, 有待后续研究和实验.

参考文献

1 盛成成, 朱勇, 刘涛. 基于微博社交平台的舆情分析. 智能计算机与应用, 2019, 9(1): 57–59, 64. [doi: 10.3969/j.issn.2095-2163.2019.01.013]

2 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613–620. [doi: 10.1145/361219.361220]

3 Phan XH, Nguyen LM, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceedings of the 17th International Conference on World Wide Web. New York, NY, USA. 2008. 91–100.

4 张志飞, 苗夺谦, 高灿. 基于 LDA 主题模型的短文本分类方法. 计算机应用, 2013, 33(6): 1587–1590.

5 王细薇, 樊兴华, 赵军. 一种基于特征扩展的中文短文本分类方法. 计算机应用, 2009, 29(3): 843–845. [doi: 10.3969/j.issn.1001-3695.2009.03.012]

6 王盛, 樊兴华, 陈现麟. 利用上下位关系的中文短文本分类. 计算机应用, 2010, 30(3): 603–606, 611.

7 张振豪, 过弋, 韩美琪, 等. 基于关键词相似度的短文本分类方法研究. 计算机应用研究, 1–6. https://doi.org/10.19734/j.issn.1001-3695.2018.04.0440,2019-01-25.

8 范云杰, 刘怀亮. 基于维基百科的中文短文本分类研究. 现代图书情报技术, 2012, (3): 47–52. [doi: 10.11925/infotech.1003-3513.2012.03.08]

9 孟涛, 王诚. 基于扩展短文本词特征向量的分类研究. 计算机技术与发展, 2019, 29(4): 57–62. [doi: 10.3969/j.issn.1673-629X.2019.04.12]

10 Bouaziz A, Dartigues-Pallez C, Da Costa Pereira C, *et al*. Short text classification using semantic random forest. Bellatreche L, Mohania M K. Data Warehousing and Knowledge Discovery. Cham: Springer, 2014. 288–299.

11 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301. 3781, 2013.

12 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

13 Collobert R, Weston J, Bottou L, *et al*. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 2011, 12: 2493–2537.

14 Yan DF, Ke N, Gu C, *et al*. Multi-label text classification model based on semantic embedding. The Journal of China Universities of Posts and Telecommunications, 2019, 2(1): 95–104.

15 孙建旺, 吕学强, 张雷瀚. 基于语义与最大匹配度的短文本分类研究. 计算机工程与设计, 2013, 34(10): 3613–3618. [doi: 10.3969/j.issn.1000-7024.2013.10.048]

16 韩栋, 王春华, 肖敏. 基于句子级学习改进 CNN 的短文本分类方法. 计算机工程与设计, 2019, 40(1): 256–260, 284.

17 冯国明, 张晓冬, 刘素辉. 基于 CapsNet 的中文文本分类研究. 数据分析与知识发现, 2018, 2(12): 68–76. [doi: 10.11925/infotech.2096-3467.2018.0391]

18 Nyberg K, Raiko T, Tiinanen T, *et al*. Document classification utilising ontologies and relations between documents. Proceedings of the Eighth Workshop on Mining and Learning with Graphs. New York, NY, USA. 2010. 86–93.

19 江大鹏. 基于词向量的短文本分类方法研究[硕士学位论文]. 杭州: 浙江大学, 2015.
20 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述. 计算机应用, 2009, 29(S1): 167–170, 180.
21 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究. 计算机系统应用, 2018, 27(5): 209–215.
22 张谦, 高章敏, 刘嘉勇. 基于 Word2Vec 的微博短文本分类研究. 信息网络安全, 2017, (1): 57–62. [doi: 10.3969/j.issn.1671-1122.2017.01.009]
23 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究. 中文信息学报, 2004, 18(3): 17–23. [doi: 10.3969/j.issn.1003-0077.2004.03.003]
24 刘小敏, 王昊, 李心蕾, 等. 不同特征粒度在微博短文本分类中作用的比较研究. 情报科学, 2018, 36(12): 126–133.
25 Le QV, Mikolov T. Distributed representations of sentences and documents. arXiv prepfint arXiv: 1405. 4053, 2014.
26 李玲俐. 数据挖掘中分类算法综述. 重庆师范大学学报 (自然科学版), 2011, 28(4): 44–47.
27 Kotsiantis S B. Supervised machine learning: a review of classification techniques. Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Amsterdam, the Netherland. 2007. 3–24.
28 王杨, 许闪闪, 李昌, 等. 基于支持向量机的中文极短文本分类模型. 计算机应用研究 (优先出版), 1–5. https://doi.org/10.19734/j.issn.1001-3695.2018.06.0514,2018-12-13/2019-02-17.