基于改进 TF-IDF 特征的中文文本分类系统*

但唐朋　许天成　张姝涵
（华中师范大学计算机学院　武汉　430079）

摘　要　随着 Internet 技术的发展，人们不仅可以从网络获取信息，也能够在网络上表达个人观点、分享自身体验。自 Web2.0 以来网络已经由原来的阅读式网络转换成为了当今的交互式网络。而伴随网络发展的是成几何速率增长的网络信息。文本信息是网络信息的重要组成部分，不同文本信息可以分成新闻、娱乐、时评、财经等不同类别。进行中文文本分类不仅能为建立文本语料库提供便利还能够应用到其它数据挖掘领域。论文基于改进 TF-IDF 特征并结合 SVM 模型设计了一种自动化的中文文本分类系统。实验证明，对比传统特征提取方式，采用改进 TF-IDF 特征策略进行文本分类能够获得更高的准确度。

关键词　文本分类；自然语言处理；BOW 模型；机器学习；改进 TF-IDF 特征
中图分类号　P315.69　　**DOI**：10. 3969/j. issn. 1672-9722. 2020. 03. 011

A Chinese Text Classification System Based on Improved TF-IDF Feature

DAN Tangpeng　XU Tiancheng　ZHANG Shuhan
（School of Computer, Central China Normal University, Wuhan　430079）

Abstract　With the development of Internet technology, people can not only obtain information from the Internet, but also express personal opinions and analyze their own experiences on the Internet. Since Web2.0, the network has been transformed from the original reading network to today's interactive network. What's more, with the development of network, the network information of geometric growth rate is growing. Text information is an important part of network information. Different text information can be divided into different categories such as news, entertainment, commentary, finance and so on. Chinese text classification can not only facilitate the establishment of a text corpus, but also can be applied to other data mining areas. In this paper, an automatic Chinese text classification system is designed based on improved TF-IDF features and SVM model. Experiments show that the classification system constructed by machine learning algorithms can achieve high degree of accuracy and meets practical needs.

Key Words　text classification, natural language processing, BOW model, machine learning, improved TF-IDF feature
Class Number　P315.69

1　引言

随着当下移动互联网技术的高速发展，网络用户的身份由单一的信息获取者转变成为双向的信息产生者和传递者。这一身份的改变加快了信息的传播速度，扩大了信息传播广度。与此同时，大量的文本信息也出现在网络当中。作为数据收集的重要步骤，在录入文本信息时往往需要对应其类别。然而不是每一个信息来源都会标注原始数据的类别。所以就需要设计一种自动文本分类系统来进行分类。更进一步说，如果对某一个用户所发布的信息进行分类处理，研究者们就可以探寻该用户对网络信息的关注程度，同时将其感兴趣的部分应用到信息推送等应用当中。总体而言，文本的分类研究拥有广泛的应用前景。

目前国内外已有不少学者在进行相关领域的

* **收稿日期**：2019 年 9 月 13 日，**修回日期**：2019 年 10 月 26 日
基金项目：华中师范大学国家级大学生创新创业训练计划（编号：201810511002）；华中师范大学院级大学生创新创业训练计划（编号：CA20180418221834349C）资助。
作者简介：但唐朋，男，研究方向：空间数据库、机器学习。许天成，男，研究方向：自然语言处理、人工智能。张姝涵，女，研究方向：大数据、机器学习。

研究，如魏芳芳等[1]基于支持向量机对农业文本进行分类，其缺点是分类的范围被限定在了农业领域不具备常规性和通用性。刘怀亮等[2]利用知网语义相似度进行中文文本的分类，提升了中文文本分类的准确度，但在进行向量模型构建时仅通过评论筛选维度没有考虑不同维度间可能存在的内在联系。李锋刚等[3]将 LDA 主题模型与传统 SVM 分类模型相结合解决了传统分类问题中相似性度量和主题单一的问题，但 LDA 主题模型[4]的效能会跟随所设置参数而变化，通过模型不一定能够建立起完全匹配文本的向量集。为了解决上述问题，不仅需要在通用性质的类别下进行分类，还需要充分考虑文本中各个维度特征之间的关系。徐明等[5]针对微博短本文特征选择提出了一种改进的卡方统计算法，其能够增加分类的准确度但还是不能挖掘文本间的潜在联系。为了提高中文文本分类的准确度，本文采用 BOW 模型[6~7]进行文本向量化操作，并利用 TF-IDF 策略[8~9]进行向量维度的权重计算以保证最终向量化的结果能够对应原始文字数据。其后对传统 TF-IDF 策略进行改进并联合基于 SVM 模型机器学习算法完成自动化文本分类系统的设计。

2　相关理论

2.1　BOW 模型简介

BOW 模型是一种文本向量化模型，将文档表示成特征矢量。它的基本思想是对于任何一个文本，不考虑其词序、语法以及句法，仅仅将其看作是各个不同词汇的集合，而文本的每个词汇都是独立的。通过对不同语句中不同词汇的提取构建空间向量集。BOW 的构造原理如下所示：

文本内容：不错不错

他去了北京读书

文本 1：不错

文本 2：去了、北京、读书

对文本 1、文本 2 构建向量：

表 1　BOW 向量集的构建方法

文本/词	不错	去了	北京	读书
文本 1	1	0	0	0
文本 2	0	1	1	1

2.2　传统 TF-IDF 特征

传统 TF-IDF（Term Frequency-Inverse Document Frequency）即词频-反文档频率是一种用于文本挖掘的常用技术。其中 DF 指的是文本集中所含有该特征的文档数目。IDF 反文档频率则反应了特征词在整个文档集合中的分步，可以在一定程度上体现这个特征的区分能力。某一个特定词语的 IDF 可以由总文档数除以包含该词语的文档数再取商对数获得。即

$$idf_t = \log(\frac{N}{df_t}) \tag{1}$$

其中 N 指所有文档总数，df_t 表示含有特征词 t 的文档数目。

TF 为词频，指的是某一个给定词语在该文件中出现的频率。所以可以通过推导得到计算公式。

$$w_{ij} = tf_{ij} \times idf_t \tag{2}$$

归一化计算公式为

$$W_{ij} = \frac{tf_{ij} \times \log(N/df_t)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ij} \times \log(N/df_t)\right]^2}} \tag{3}$$

其中 tf_{ij} 表示特征词 t_j 在文本中出现的次数，即为词频。在完成权值计算之后，将得到拥有如下性质的矩阵。

1）列是所有文档总共词的集合。

2）每一行代表每一个评论文本。

3）每行是一个向量，这个向量代表了词的权重。

若仅使用 BOW 模型对文本进行向量化处理会导致所得文本向量过于庞大，所得的维度也越高，高纬度的向量组不利于文本分类，所以本文引入 TF-IDF 策略进行处理，在降低维度的同时也能得到尽可能符合原文的向量组。

2.3　SVM 分类器

SVM 分类器[10]是一种经典的机器学习分类模型，它能够准确地对高纬度信息进行分类，将文本的待处理数据表示为空间中的向量 x_i。通过在这个空间中创建一个超平面来达到将不同向量分类的目的，超平面的法向量表示为 w。y_i 表示对应数据 x_i 的类别且 $y_i \in \{-1, 1\}$。下面分别给出其目标函数和对应最优解。

目标函数：

$$\begin{cases} \min \dfrac{\|w\|^2}{2} \\ s.t.y_i(wx_i + b) \geq 1, i = 1, 2, 3, \cdots, m_i \end{cases} \tag{4}$$

最优解：

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{5}$$

其中 α_i 表示拉格朗日算子，大于 0 的拉格朗日算子被称为支持向量，其余的拉格朗日算子等于 0，SVM

分类器根据计算待测数据并以超平面为界划分类别。本文将利用SVM分类器对已经处理好的文本信息就行类别分类。

3 改进TF-IDF特征

运用传统的TF-IDF特征对文本进行向量构建时只考虑了特征项在各个文本中的分布情况,而忽略了特征项词语间的近义、同义情况。不同于英文文本,中文文本中往往含有大量近同义词,这些词语的存在势必会影响到文本分类器的分类效能。如果仅使用传统特征计算方法反而会丢失文本关键特征,文献[11~12]通过知识语言分析提出了词语级的相似度分析方法,并采用"知网"相似度算法来计算词语间的相似程度进一步提高词语相似度计算的准确程度。所以为了解决传统TF-IDF特征不能解决文本中含有近义词、同义词的情况,本文将相似度计算应用到TF-IDF特征计算方法中。并以此来增加特征项的权重。使其能够反应整个文本的特征。为了规范对于同义词、近义词的判断。我们规定相似度计算大于0.8的两个词可以被认为是同义词或近义词,并定义相似度因子 α。α 代表文本中一个特征项的数量与其相似特征之和在所有特征项总数中的占比情况。我们将使用相似度因子 α 来调节TF-IDF特征计算公式。相关公式如下所示:

1)相似度计算

$$sim(x,y)=\frac{x\cdot y}{\|x\|\cdot\|y\|} \tag{6}$$

其中 x 和 y 分别表示文本向量中的两组特征。

2)相似度因子

$$\alpha=\frac{P+Q}{U} \tag{7}$$

其中 P 表示某一文本中特征项 t 的个数,Q 表示与特征项 t 相似度大于0.8的特征项的个数,我们认为这部分特征词与原提特征词能够表示相似的文本特征。U 是所有特征的数量。

3)融合语义的TF-IDF策略

$$W_{ij}=\frac{\sqrt{tf_{ij}\times\alpha}\times\log(N/df_t)}{\sqrt{\sum_{j=1}^{M}\left[\sqrt{tf_{ij}\times\alpha}\times\log(N/df_t)\right]^2}} \tag{8}$$

其中 W_{ij} 表示某一特征词的权重,tf_{ij} 表示特征词 t_j 在文本中出现的次数,即词频。N 是整个文本的总数量,df_t 表示含有特征词 t 的文档数目。

4 中文文本分类系统的构建

文本的分类系统[13~15]主要由以下几个部分构成:1)文本获取器,主要通过网络爬虫技术定向从互联网自动获得文本信息;2)分类训练器,使用一定量的训练数据来训练机器学习模型以达到进行文本分类的目的;3)分类器,与训练器相似,不同的是分类器将直接对所收集到的未标记数据进行分类。本文将重点介绍分类训练器的构建。

4.1 分类训练器

分类训练器由5个步骤完成:1)原始文本信息;2)预处理,由于文本中含有大量无用信息,如人称"我"、助词"的"等对文本类别意义没有帮助的字词所以需要对原始信息进行分词与去停用词处理;3)降维和向量化,分别利用第二部分所介绍的BOW模型、TF-IDF策略对文本信息进行向量化操作和降维处理;4)构建Bunch库使组成的向量能够被SVM分类器所训练。5)利用机器学习算法中的SVM分类器对数据元素进行学习,使其能够对中文文本进行分类。分类器的步骤由图1所示。

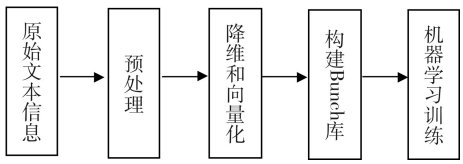

图1 分类训练器工作流程

4.2 Bunch类库

Bunch类库用来存储进行降维和向量化后的文本数据同时对应其标签。在Bunch类库中有4个成员:对象名称、标签、文件名与文本内容。这4个成员之间含有两层映射:一是对象名称与文件名的映射;二是标签类与文本内容的映射。构建Bunch类主要是利用其机理来规范化机器学习训练的操作。Bunch类如表2所示。

表2 Bunch类库说明

对象名称	文件名	标签	文本内容
Art	Art/1.txt	art	优美、油画
Computer	Computer/2.txt	computer	决定、速度 改进

5 实验验证

为了更加便利地构建文本分类系统,本文在如下平台进行实验,CPU:Intel Core i7 6700,内存:DDR4 8G,硬盘:固态硬盘120G,操作系统:Windows 7,开发环境:python 3.4.4,sklearn机器学习库。且训练数据选择了复旦大学中文文本分类语

料库进行训练。文本分类普遍使用的评价标准有准确度、召回率、F1 值，具体定义如下：

$$P=\frac{TP}{TP+FP} \tag{9}$$

$$R=\frac{TP}{TP+FN} \tag{10}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{11}$$

上述公式中，*TP* 表示某特征被正确分类的正样本，*FP* 表示某个特征被错误分类的负样本，*FN* 表示某一特征在分类时被错误分类的正样本。为了说明改进特征的有效性，本文将分别利用传统 TF-IDF 策略和改进 TF-IDF 策略进行文本分类实验实验结果由表 3、表 4、图 2、图 3、图 4 所示。

表 3　传统 TF-IDF 策略的实验结果

维度数	正确率 *P*	召回率 *R*	*F*1 值
500	0.841	0.725	0.742
1000	0.854	0.742	0.775
1500	0.866	0.728	0.783
2000	0.892	0.725	0.793

表 4　基于改进 TF-IDF 的实验结果

维度数	正确率 *P*	召回率 *R*	*F*1 值
500	0.894	0.731	0.758
1000	0.912	0.740	0.793
1500	0.917	0.747	0.836
2000	0.904	0.739	0.843

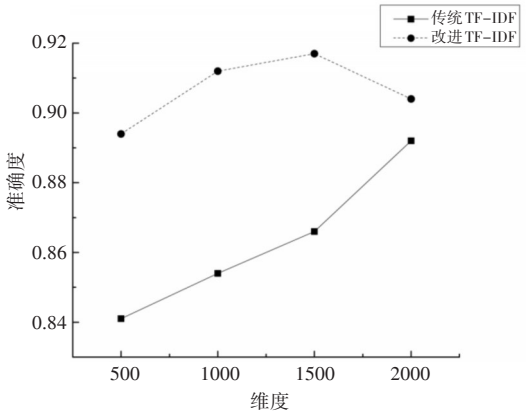

图 2　不同维度准确度的对比

由实验结果可知，应对不同纬度下的中文文本分类任务，该文本分类系统的准确度都处于较高水平，说明系统的建立是成功的，且应用改进策略的结果明显优于应用传统方法。值得注意的是，当选择不同维度来表达中文文本时得到了不同的分类准确度。这是因为在较低维度时用于衡量整篇文章的特征向量较少，不利于机器学习进行分类。而较高的维度虽然能够从各个方面完善的表示文本信息，但过高的纬度会伴随噪声，即向量中的无关信息会影响机器学习模型的判断。从实验结果来看在维度数为 1500~2000 时分类的效果能够达到最好的情况，但这可能并不是绝对。需要重复试验找到最优情况。

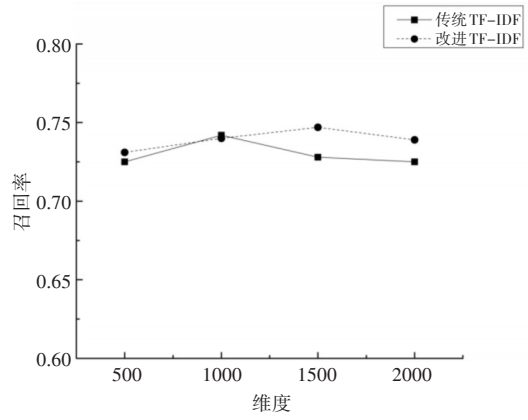

图 3　不同维度召回率的对比

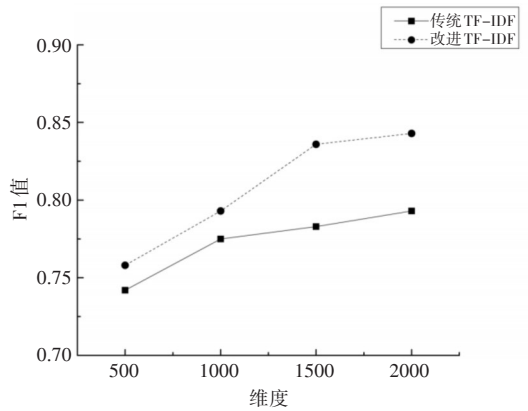

图 4　不同维度 *F*1 值的对比

6　结语

本文以文本分类为切入点，基于改进 TF-IDF 特征和机器学习原理设计了一种自动化中文文本分类系统。实验证明，改进后基于 TF-IDF 特征的文本分类性能要优于传统文本分类方法。在本方法下设计的文本分类系统具有较高的准确度，能够用于实际应用满足了研究的需要。值得关注的是当维度提升至一定程度后各项指标不再增长，我们将在后续的研究中继续相关工作直至解决问题。

参 考 文 献

[1] 魏芳芳，段青玲，肖晓琰，等. 基于支持向量机的中文农业文本分类技术研究[J]. 农业机械学报，2015（S1）：174-179.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification Technique of Chinese Agricultural Text Information Based on SVM[J].Transactions of the Chinese Society for Agricultural Machinery, 2015, 46（S1）：174-179.

[2] 刘怀亮, 杜坤, 秦春秀. 基于知网语义相似度的中文文本分类研究[J].现代图书情报技术, 2015, 31(2): 39-45.
LIU Huailiang, DU Kun, QIN Chunxiu. Research on Chinese Text Categorization Based on Semantic Similarity of HowNet. New Technology of Library and Information Service, 2015, 31(2): 39-45.

[3] 李锋刚, 梁钰. 基于LDA—WSVM模型的文本分类研究[J]. 计算机应用研究, 2015, 32(1): 21-25.
LI Fenggang, LIANG Yu, GAO Xiaozhi, et al. Research on text categorization based on LDA-Wsvm model[J]. Application Research of Computers, 2015, 32(1): 21-25.

[4] 胡吉明, 陈果. 基于动态LDA主题模型的内容主题挖掘与演化[J]. 图书情报工作, 2014, 58(02): 138-142.
HU Jiming, CHEN Guo. Mining and Evolution of Content Topics Based on Dynamic LDA.LIS, 2014, 58(02): 138-142.

[5] 徐明, 高翔, 许志刚, 等. 基于改进卡方统计的微博特征提取方法[J]. 计算机工程与应用, 2014(19): 113-117.
XU Ming, GAO Xiang, XU Zhigang, et al. Feature selection methods of microblogging based on improved CHI-square statistics[J]. CEA, 2014, 50(19): 113-117.

[6] Ngo-Ye T L, Sinha A P. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model[J]. Decision Support Systems, 2014, 61: 47-58.

[7] González L C, Moreno R, Escalante H J, et al. Learning roadway surface disruption patterns using the bag of words representation[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(11): 2916-2928.

[8] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1-4.
LIU Zhiming, LIU Lu. Empirical study of sentiment classification for Chinese microblog based on machine learning[J]. CEA, 2012, 48(1): 1-4.

[9] Qu Z, Song X, Zheng S, et al. Improved Bayes Method Based on TF-IDF Feature and Grade Factor Feature for Chinese Information Classification[C]//Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. IEEE, 2018: 677-680.

[10] Vapnik V, Izmailov R. Knowledge transfer in SVM and neural networks[J]. Annals of Mathematics and Artificial Intelligence, 2017, 81(1-2): 3-19.

[11] 任姚鹏, 陈立潮, 张英俊, 等. 结合语义的特征权重计算方法研究[J]. 计算机工程与设计, 2010(10): 2381-2383.
REN Yaopeng, CHEN Lichao, ZHANG Jianjun, et al. Research on the calculation methods of feature weights combined with semantics[J]. Computer Engineering and Design, 2010, 31(10): 2381-2383, 2387.

[12] 马莹, 赵辉, 李万龙, 等. 结合改进的CHI统计方法的TF-IDF算法优化[J]. 计算机应用研究, 2019, 36(9): 2596-2598.
MA Ying, ZHAO Hui, LI Wanlong, et al. Optimization of TF-IDF algorithm combined with improved CHI statistical method[J]. Application Research of Computers, 2019, 36(9): 2596-2598.

[13] Al-Anzi F S, AbuZeina D. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing[J]. Journal of King Saud University-Computer and Information Sciences, 2017, 29(2): 189-195.

[14] Cao Z, Li W, Li S, et al. Improving Multi-Document Summarization via Text Classification[C]//AAAI. 2017: 3053-3059.

[15] El-Halees A M. Arabic text classification using maximum entropy[J]. IUG Journal of Natural Studies, 2015, 15(1).

(上接第555页)

[11] Xi N, Yang Y. The rise of the robot industry in China[J]. HKIE Transactions, 2015: 22(2): 98-102.

[12] 梁文莉. 全球机器人市场统计数据分析[J]. 机器人技术与应用, 2018(2): 43-48.
LIANG Wenli. Statistical data analysis of the global robot market[J]. Robot Technique and Application, 2018(2): 43-48.

[13] 张红霞. 国内外工业机器人发展现状与趋势研究[J]. 电子世界, 2013, 12: 5-7.
ZHANG Hongxia. Research on the Present Situation and Trend of Industrial Robots at Home and abroad[J]. Electronics World, 2013, 12: 5-7.

[14] 郑耀锋, 肖宏, 李旭. 基于DSP的视频快速特征匹配跟踪算法研究[J]. 电子技术与软件工程, 2014(15): 114-115.
ZHENG Yaofeng, XIAO Hong, LI Xu. Research on Video Fast Feature Matching Tracking Algorithm Based on DSP. Electronic Technology & Software Engineering, 2014(15): 114-115.

[15] 陈宁, 杨永全. 基于纹理特征匹配的快速目标分割方法[J]. 电子设计工程, 2017, 25(23): 39-40.
CHEN Ning, YANG Yangquan. Fast object segmentation based on texture matching[J]. Electronic Design Engineering, 2017, 25(23): 39-40.