文章编号:1674-2974(2016)04-0141-06

不同情境下中文文本分类模型的表现及选择*

兰秋军†,李卫康,刘文星
(湖南大学 工商管理学院,湖南 长沙 410082)

摘 要:针对中文文本分类任务中N-Gram,素贝叶斯、K最近邻和TF-IDF等经典而广泛使用的文本分类模型的选择困惑问题,基于万余篇中文新闻文本语料数据,设计了一系列的对比实验,考察了各模型在不同参数、不同训练数据规模、不同训练文本长度、类别是否偏斜等多种情境下分类性能的表现,总结了各模型的特性,为中文文本分类模型的选择和应用提供了实践依据和参考.

关键词:中文文本;文本分类;数据挖掘;情报分析
中图分类号:TP274;TP302 **文献标识码**:A

Performance and Choice of Chinese Text Classification Models in Different Situations

LAN Qiu-jun†, LI Wei-kang, LIU Wen-xing
(Business School, Hunan Univ, Changsha, Hunan 410082, China)

Abstract: N-Gram, Naïve Bayes, K nearest neighbors and TF-IDF are classical text classification models with a wide range of applications. People are often puzzled about which classification model should be used in a certain Chinese text classification task. This paper collected more than ten thousand Chinese news texts, and designed a series of experiments to analyze the performance of these models in varied situations from classification parameters, training data scale, text length and skewed data sets. The characteristics of these models were summarized, which provides a practical guide for the model selection in Chinese text classification tasks.

Key words: Chinese text; text classification; data mining; information analysis

文本挖掘是语言学、统计学以及计算机技术相结合的产物,是对海量文本信息进行自动处理,获取人们感兴趣的、隐含的、有用信息的过程[1-2],在信息检索、生物医学、情报获取、舆情分析和市场营销等众多领域备受关注[3-5].文本分类作为文本挖掘领域中的核心技术,是各种自然语言处理、应用的基础.其中分类模型的选择对最终结果具有至关重要的影响.然而,因所基于的原理、参数、应用场合各不相同,即使相同的模型其性能表现也往往大相径庭.

新闻文本是一类常见的文本形式,其蕴含的信息量大,是各种情报分析的重要数据源[6-7].尽管现有的各个新闻网站以栏目形式对新闻进行了人工划分,然而各网站的分类体系和栏目形式各不相同,因此在具体的新闻挖掘应用项目中,常需将采集的新闻数据重新进行组织和划分.中文文本分类领域中具有代表性的模型是朴素贝叶斯、N-Gram,K最

* **收稿日期**:2014-11-08
基金项目:国家自然科学基金资助项目(71171076),National Natural Science Foundation of China(71171076)
作者简介:兰秋军(1972—),男,湖南娄底人,湖南大学副教授,博士
† 通讯联系人,E-mail:lanqiujun@hnu.edu.cn

近邻和 TF-IDF. 这些不同的模型各具有怎样的特性? 分别适合哪些场合? 在使用时应如何选取合适的参数? 人们往往面临困惑. 由于各方法在处理细节上有不少差异,很难从理论分析的角度来比较各方法的优劣. 因此,基于典型数据,采用实验的方式进行比较是比较通行的做法[8-9]. 本文精心构造了多组实验,从模型参数选取、训练数据规模、训练文本长度、数据是否偏斜等几个情境来考察各模型在不同情境下的性能表现,其结论对中文文本分类模型的选择与参数设置等具有实践指导意义.

1 模型概述

文本分类就是通过计算机程序自动将某个文档归属到事先给定的类别体系中一个或多个类别. 现有的文本分类方法大致可归为两类:基于规则的方法和基于机器学习的方法. 其中基于规则的方法早在 20 世纪 70 年代就已出现,但因规则制定的困难,目前普遍采用的是基于机器学习的方法. 而机器学习方法中,基于统计的方法是最具有代表性和使用最为广泛的[10]. 其中,N 元模型(N-Gram)[11]、朴素贝叶斯(NB),K 最近邻(KNN)和 TF-IDF 又是其中最经典的几个模型.

N-Gram 模型基于马尔科夫假设,即下一词的出现概率仅依赖于它前面的 N 个词,统计 N 元词串在各类别中出现概率,以此确定文档归属于哪个类别[11-12]. 朴素贝叶斯模型基于贝叶斯定理,假设单词两两独立,获得各文档类别的后验概率,哪个类别概率值大,文档即归属于该类别[13]. KNN 的主要思想是先将文档内容转化为特征空间中的特征向量,计算待分类文档与训练文档中每个样本的相似度,找出其中的 k 个最近邻居,据此判别文档所属类别[14]. TF-IDF 则先将文本内容转化为特征向量,然后计算其与类别特征向量间的余弦相似度,以此作为其所属类别的判据[15]. 分析上述模型,不难发现,N-Gram 模型主要是提取了不同类别文档中字与字之间的顺序依赖关系来构造分类特征,朴素贝叶斯则提取了不同类别文档中词与词之间的概率依赖关系构成分类依据,K 近邻直接利用了空间向量模型,以文档相似性特征作为分类依据,而 TF-IDF 则同时考虑了词在文档本身中的出现频度以及其在不同文档中的出现频度信息. 几个模型所抓取的文档类别信息特征明显不同,很难在理论上判别哪个模型更好,更适合哪些情境. 因此,从实验的角度来考察是更为切合实际的方案.

2 基础与准备

2.1 算法实现工具

LingPipe 是基于 Java 语言的自然语言处理的开源软件包,提供了文本挖掘各阶段的基本功能. 由于该软件包的数据处理都基于一个共同框架,采用了相同的基础源代码模块,故本文以其作为算法实现工具,可尽量减少模型本身之外的因素(如文本预处理阶段的分词、特征提取、文本表示等)给模型性能带来的影响.

2.2 实验数据

本文实验数据采集来自新浪、腾讯、凤凰等主流网站. 特地挑选了历史、军事、文化、读书、社会几个比较近似,甚至人工也容易分错的文本类别. 其中,历史类和军事类的文章比较相近,而文化类和读书类的也常相似. 数据采集跨时 2 个月,去除了所有 Html 标记和网页中的噪声文本,只包含标题、正文内容以及标点符号. 共采集 16 000 篇,去除了部分重复和校验过程中有问题的文档,最终保留 14 000 篇作为本文研究的语料数据. 其中含历史 1 900 篇,军事 1 600 篇,文化 2 500 篇,读书 4 000 篇,社会 4 000篇. 每篇文章按类别以 txt 文件的形式保存. 各类别、不同文本长度的文档篇数分布情况如表 1 所示,所有文档的长度介于 10 000 字节以内,涵盖了网页中的绝大多数新闻文本长度.

2.3 分词与特征项

尽管特征选择和预处理措施都是影响文本分类性能的关键因素,但因各模型算法原理相差太大,无法基于统一的特征项和预处理进行比较[16]. 因此,各模型的特征项均以词频为基础,采取各模型常用的特征形式和预处理方式. N-Gram 模型本身不需分词,因此未做分词处理,而其它模型则应用中科院分词系统 NLPIR_2014 进行分词处理.

2.4 分类评价指标

LingPipe 提供了一系列指标对模型性能进行评估. 本文实验主要采用宏平均、微平均下的 F 值进行评价,它综合考虑了准确率、召回率两个被广泛认可的分类器评价指标. 其详细定义和含义可参见相关文献[17]. 此外,实验过程中,还记录了各模型的训练和分类运行时间. 这也是反映分类模型性能的一个方面.

表 1　数据集文本长度分布
Tab. 1　Text length distribution of dataset

	1000 以下	1000—2000	2000—3000	3000—4000	4000—5000	5000—6000	6000—7000	7000—8000	8000—9000	9000—10000	10000 以上
历史	220	206	218	226	177	164	121	99	85	71	313
军事	524	384	187	108	95	90	51	43	29	25	64
文化	244	356	516	393	289	203	150	98	69	72	210
读书	223	599	646	490	334	236	169	163	140	147	853
社会	828	1328	857	458	217	138	82	35	21	12	24

注:1000 以下表示 1000 字节以下,1000—2000 表示 1000 字节—2000 字节,后同;单元格内数字为文档数.

3　实验方案与结果

从应用角度来看,分类准确度和处理效率是用户最为关注的两个方面. 而影响这两方面的因素无外乎模型本身和待处理的数据,如图 1 所示. 模型本身因素具体包括模型的构造机制和模型参数. 其中,模型机制对用户而言是封装的,要提升分类性能,用户只能调整模型参数. 而数据方面,文本的词语和语义特点太过复杂和精细,用户难以据此选择模型. 然而待处理文本的长度、规模和偏斜程度等是影响分类性能的重要因素,用户可以据此选择最合适的模型. 因此,本文主要设计了 4 组实验考察不同情境下中文文本分类模型的表现. 下面具体阐述各组实验的具体方案及结果.

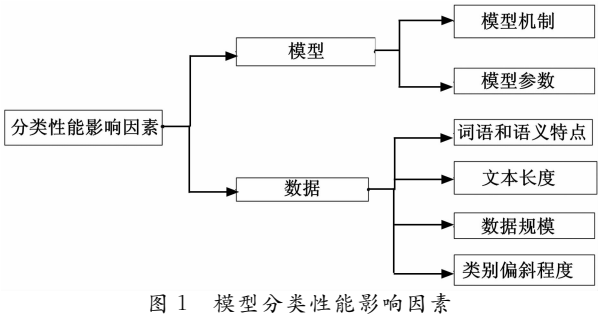

图 1　模型分类性能影响因素
Fig. 1　Influential factors of model classification performance

3.1　模型参数与模型性能

所述的 4 个分类模型中,NB 模型和 TF-IDF 模型没有参数,而 N-Gram 和 KNN 模型则分别有一个关键参数 N 和 K. 先对这两个模型进行不同参数取值下的实验. 所采用的数据集文档篇数情况如表 2 所示.

表 2　不同模型参数下的实验数据分布情况
Tab. 2　Document distribution for model parameters selection

	历史	军事	文化	读书	社会	总计
训练集	1 500	1 500	1 500	1 500	1 500	7 500
测试集	200	200	200	200	200	1 000

3.1.1　关于 N-Gram 模型参数 N 的实验

根据 N-Gram 模型中参数 N 的含义,字与字之间的概率依赖关系主要由词组造成,汉语超过 6 个字的词组已经相当稀少. 本组实验考察了 N 取值为 1,2,4,6,7 的情形,获得结果如表 3 所示. 可见,随着 N 取值的加大,该分类器的分类性能也越高,但在 N 超过 4 以后,分类性能改善幅度已相当小,而训练时间和测试时间却成倍增长,为此,后述实验均取 $N=4$,不再赘述.

表 3　N-Gram 模型 N 值大小与性能关系
Tab. 3　N-Gram performance under different values of N　　ms

N 值	1	2	4	6	7
Micro-F	0.661	0.816	0.837	0.852	0.862
训练时间	23 594	40 031	138 734	351 078	553 469
测试时间	7 219	7 844	15 063	35 718	55 953

3.1.2　关于 KNN 模型参数 K 的实验

参数 K 为经验参数,表示选取的近邻个数,其值的大小对于模型的分类性能有显著的影响. 实验中,为确定 K 最佳值,将 K 分别取值 1,3,5,10,20,获得如表 4 所示结果. 显见,随着 K 值的增加,分类性能有缓慢下降趋势,表明并非选取的近邻数越多越好. 原因在于 KNN 基于向量空间模型,维数较高,数据比较稀疏,K 值越大,反而可能带来更多的误判信息. 本实验中,K 取值为 1 时分类性能最优,因此在后述实验中均取该最优值.

表 4　KNN 模型 K 值大小与性能关系
Tab. 4　KNN performance under different values of K　　ms

K 值	1	3	5	10	20
Micro-F	0.666	0.636	0.628	0.618	0.616
训练时间	210 062	210 891	211 750	211 766	230 250
测试时间	50 719	50 516	51 250	50 187	51 578

3.2　训练集规模与模型性能

为了考察各模型在不同训练集规模情境下的性能表现,分 8 次小实验,每次从实验语料中抽取 1 000,2 000,3 000,4 000,5 000,6 000,7 000,8 000

篇文档构成训练集,抽取 1 000 篇文档构成测试集.抽样过程中,为了消除各类别新闻分布不一致、文本长度不一致所带来的影响,进行了适度控制,即确保 8 次实验中,各类别新闻的比例保持一致,各文本长度所占比例也保持一致.在此控制下,随机抽取样本,每次实验进行 3 次,以其平均值作为最后结果.实验结果如表 5 所示,图 2 展示了不同训练集规模情境下,模型性能(微平均下 F 值)的情况;图 3 和图 4 则分别展示了分类训练时间和测试时间与训练集规模的关系.

由表 5 以及图 2～图 4 可知,在各类别数据相近的情况下,随着训练集规模的增大,各模型的分类性能也均得到提升并渐趋于平稳,与文献[9]结论相同.KNN 改善效果最为明显.就运行时间而言,各分类模型随着训练集规模的增大,训练时间明显增加,而测试时间仅 KNN 分类模型显著增加,其它则变化微小.通过逐渐加大训练集规模,实验还发现,在测试集不变的情况下,训练集达到一定规模后(例如 7 000 篇),即使再显著增大训练集规模,分类性能的改善也非常微弱.

表 5 模型性能与训练集规模间关系

Tab. 5 Model performance under different training dataset scales ms

训练集规模(文档篇数)		1000	2000	3000	4000	5000	6000	7000	8000
N-Gram	Macro-F	0.655	0.708	0.742	0.742	0.759	0.763	0.780	0.791
	Micro-F	0.655	0.708	0.741	0.744	0.761	0.765	0.783	0.793
	训练时间	20 数 485	39 781	59 156	75 187	108 875	125 750	136 297	138 734
	测试时间	11 235	11 765	12 516	13 156	14 656	15 235	14 688	15 063
NB	Macro-F	0.704	0.754	0.751	0.781	0.777	0.787	0.796	0.807
	Micro-F	0.711	0.745	0.752	0.783	0.781	0.792	0.801	0.809
	训练时间	34 703	58 938	90 938	119 219	147 688	177 485	201 437	221 437
	测试时间	14 125	13 938	14 141	13 984	14 078	13 985	13 984	14 046
KNN	Macro-F	0.479	0.516	0.537	0.558	0.584	0.602	0.612	0.634
	Micro-F	0.488	0.52	0.545	0.564	0.595	0.615	0.627	0.649
	训练时间	28 265	51 485	79 578	105 219	132 813	161 391	182 250	210 797
	测试时间	18 359	22 547	27 328	31 625	36 500	41 184	45 813	50 578
TF-IDF	Macro-F	0.708	0.788	0.842	0.865	0.861	0.884	0.895	0.906
	Micro-F	0.708	0.793	0.844	0.866	0.863	0.885	0.897	0.908
	训练时间	32 781	59 266	91 250	117 828	143 046	173 046	238 313	255 781
	测试时间	14 078	14 094	14 015	14 531	14 125	14 110	13 984	14 547

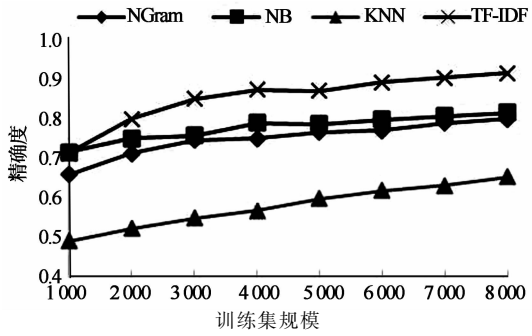

图 2 各分类模型精确度与训练集规模间关系

Fig. 2 Model precision under different training dataset scales

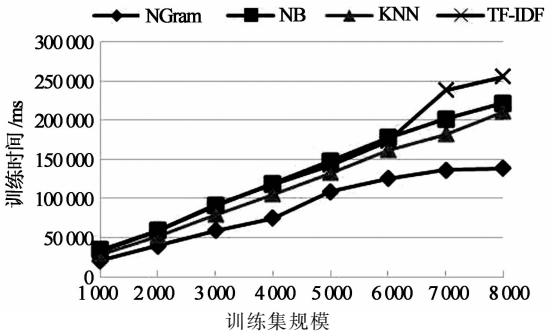

图 3 各模型运行时间与训练集规模训练时间关系

Fig. 3 Model training time under different training dataset scales

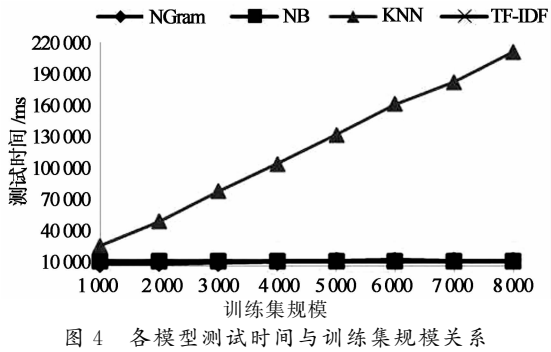

图 4 各模型测试时间与训练集规模关系

Fig. 4 Model testing time under different training dataset scales

3.3 文本长度与性能

为了考察各模型在不同文本长度情境下性能的表现,将训练数据的每个类别都按文档大小进行划分,抽取 5 个子集,分别为 1 000 字节以下,1 000－2 000 字节,2 000－3 000 字节,3 000－5 000 字节,5 000 字节以上.在保证 5 个子集的类别分布和规模分布一致的前提下,随机抽取样本构成训练数据,实验数据如表 6 所示.该实验共进行 3 次,取 3 次实验结果的平均值作为最终结果,如表 7 所示.图 5 展示了不同文本长度情境下,模型性能(微平均下 F 值)

的情况;图 6 和图 7 则展示了分类训练时间和测试时间与文本长度的关系.

表 6　文本长度实验数据集示例
Tab. 6　Experimental dataset sample of the text length:article

	历史	军事	文化	读书	社会	总计
训练集	100	100	100	100	100	500
测试集	30	30	30	30	30	150

表 7　各模型性能与训练集文本长度关系
Tab. 7　Model performance under different text length of training dataset　　ms

文本长度		1 000−−	1 000−2 000	2 000−3 000	3 000−5 000	5 000++
N-Gram	Macro-F	0.407	0.742	0.761	0.771	0.775
	Micro-F	0.44	0.74	0.767	0.773	0.773
	训练时间	4 687	7 625	11 015	14 875	26 094
	测试时间	1 587	1 891	2 078	2 344	2 860
NB	Macro-F	0.466	0.782	0.806	0.727	0.724
	Micro-F	0.507	0.780	0.807	0.727	0.792
	训练时间	9 885	16 095	23 505	27 411	39 786
	测试时间	2 687	2 703	2 735	2 703	2 719
KNN	Macro-F	0.366	0.399	0.462	0.367	0.393
	Micro-F	0.353	0.433	0.473	0.373	0.307
	训练时间	2 218	7 343	12 984	18 015	34 375
	测试时间	3 046	3 313	3 453	3 594	3 844
TF-IDF	Macro-F	0.507	0.841	0.756	0.816	0.837
	Micro-F	0.54	0.84	0.767	0.82	0.84
	训练时间	9 745	20 923	23 458	30 651	48 401
	测试时间	2 739	2 739	2 802	2 781	2 739

对文本长度情境而言,从表 7 以及图 5～图 7 可以看出,随着文本长度的增加,除 KNN 模型外,其它 3 个分类模型的准确性能在初期快速提升,其后趋缓而渐趋平稳,但 KNN 快速提升后却逐渐下降.实验还发现,在文本长度短时(小于 1 000 字节),TF-IDF 模型要好于其它 3 个模型.从时间性能来看,各模型都随文本长度增加而近似呈线性增长趋势.其中 NB 和 TF-IDF 相对较逊一筹.

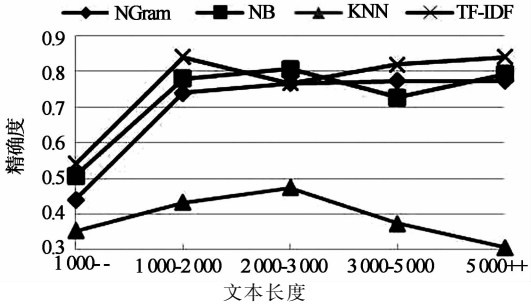

图 5　各分类模型精确度与文本长度间关系
Fig. 5　Model precision under different text length of training dataset

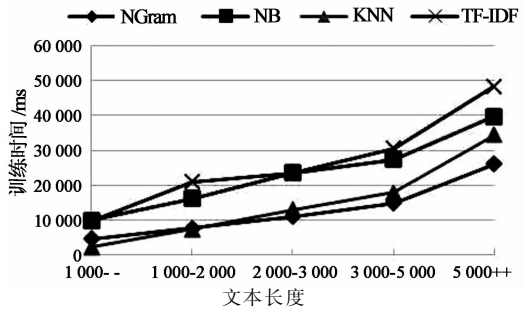

图 6　各模型训练时间与文本长度关系
Fig. 6　Model training time under different text length

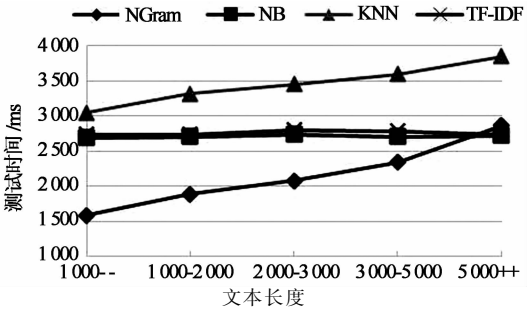

图 7　各模型测试时间与文本长度关系
Fig. 7　Model testing time under different text length

3.4　类别偏斜与模型性能

为了考察各模型在类别是否偏斜情境下性能的表现,我们设计了类别均衡和非均衡两组实验,实验数据集如表 8 所示,训练集和测试集规模相同,非均衡训练集中读书和社会两个类别远高于其它类别数据,而均衡训练集中各类别文本数相同.每组实验共进行 3 次,随机抽取样本构成训练数据.取 3 次实验结果的平均值作为最终结果,如表 9 所示.

表 8　文本类别偏斜实验数据集示例
Tab. 8　Experimental dataset sample of skewed categories

	历史	军事	文化	读书	社会	总计
非均衡训练集	500	500	500	3 000	3 000	7 500
均衡训练集	1 500	1 500	1 500	1 500	1 500	7 500
测试集	200	200	200	200	200	1 000

对类别偏斜情境而言,观察可知,各模型整体性能及各类别分类性能在均衡训练集情境下的表现均优于在非均衡训练集下的性能表现.在非均衡训练集情境下,对各小类而言,包含文本数较多的读书和社会两个类别的分类性能要优于其它类别,与文献[18]对训练集类别分布对文本分类影响的研究结果类似.4 个分类模型中,对于类别均衡数据而言,TF-IDF 表现最佳,对于类别非均衡数据而言,NB 表现最佳.

表 9 文本类别偏斜实验各小类分类情况(F1 值)
Tab. 9 Each category precision of skewed experiment

		历史	军事	文化	读书	社会	整体
N-Gram	均衡	0.816	0.779	0.688	0.686	0.933	0.786
	非均衡	0.650	0.688	0.186	0.736	0.871	0.675
NB	均衡	0.826	0.875	0.801	0.841	0.988	0.866
	非均衡	0.617	0.807	0.579	0.835	0.946	0.768
KNN	均衡	0.735	0.471	0.483	0.721	0.854	0.677
	非均衡	0.376	0.295	0.266	0.641	0.733	0.525
TF-IDF	均衡	0.884	0.947	0.842	0.940	0.962	0.916
	非均衡	0.574	0.776	0.360	0.750	0.864	0.698

4 结 论

本文所考察的几个模型是当前文本分类领域应用最为广泛、最为经典的. 在实践当中,各模型的表现各异,而在理论上又很难分析和评价其优劣. 为此本文构建了多组实验来考察不同情境下各模型的表现,形成的结论及模型选择建议如下:

1)几个模型在运行效率方面没有明显的区别,训练和测试时间都与数据集的规模和文本长度呈线性关系. 实验结果看,NB 模型和 TF-IDF 虽然稍逊,但实际应用中,此差别并不明显,可以忽略.

2)不管在何种情境下,KNN 的表现都最差,因此不推荐采用此模型.

3)从训练集规模来看,几种模型都是随着规模增大而分类精确性稳步增加,训练集的大小不构成模型选择的关键依据.

4)N-Gram、TF-IDF、NB 三个模型的分类精度受文本长度的影响差别不大. 都随文本长度的增加而精度得到提升,且都在长度低于 1 kb(约 400 汉字)时提升较明显,而此后提升速度放缓. KNN 则未能因文本长度的增加而提升其精确度.

5)对类别分布是否偏斜的情况,NB 模型表现出良好的稳定性,比 N-Gram 和 TF-IDF 都要好. 因此,在类别偏斜严重的情况下,推荐采用 NB 模型.

参考文献

[1] NASSIRTOUSSI A K, AGHABOZORGI S, WAH T Y, *et al*. Text mining for market prediction: A systematic review[J]. Expert Systems with Applications, 2014, 41(16): 7653-7670.

[2] 袁军鹏, 朱东华, 李毅, 等. 文本挖掘技术研究进展[J]. 计算机应用研究, 2006,23(2):1-4.
YUAN J P, ZHU D H, LI Y, *et al*. Survey of text mining technology[J]. Application Research of Computers, 2006,23(2):1-4. (In Chinese)

[3] 谭文堂, 王桢文, 殷风景, 等. 一种面向多文本集的部分比较性混合模型[J]. 湖南大学学报: 自然科学版, 2013, 40(11): 101-107.
TAN W T, WANG Z W, YIN F J, *et al*. A partial comparative mixture model for multi-collections documents[J]. Journal of Hunan University: Natural Sciences Edition, 2013, 40(11): 101-107. (In Chinese)

[4] ZHU F, PATUMCHAROENPOL P, ZHANG C, *et al*. Biomedical text mining and its applications in cancer research[J]. Journal of Biomedical Informatics, 2013, 46(2): 200-211.

[5] XU X, CHENG X, TAN S, *et al*. Aspect-level opinion mining of online customer reviews[J]. Communications, China, 2013, 10(3): 25-41.

[6] 胡凌云,胡桂兰,徐勇,等.基于 Web 的新闻文本分类技术研究[J]. 安徽大学学报:自然科学版,2010,34(6):66-70.
HU L Y, HU G L, XU Y, *et al*. Research of text classification technology based on Web news pages[J]. Journal of Anhui University: Natural Sciences Edition, 2010,34(6):66-70. (In Chinese)

[7] 蔡华丽,刘鲁,王理. 突发事件 Web 新闻多层次自动分类方法[J]. 北京工业大学学报:自然科学版, 2011,37(6): 947-954.
CAI H L, LIU L, WANG L. Automated multiple hierarchical classification of web news of unexpected events[J]. Journal of Beijing University of Technology, 2011,37(6): 947-954. (In Chinese)

[8] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9): 1848-1859.
SU J S, ZHANG B F, XU X. Advances in machine learning based text categorization[J]. Journal of Software, 2006, 17(9): 1848-1859. (In Chinese)

[9] 卢苇, 彭雅. 几种常用文本分类算法性能比较与分析[J]. 湖南大学学报:自然科学版, 2007, 34(6): 67-69.
LU W, PENG Y. Performance comparison and analysis of several general text classification algorithms[J]. Journal of Hunan University: Natural Sciences Edition, 2007, 34(6): 67-69. (In Chinese)

[10] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys (CSUR), 2002, 34(1): 1-47.

[11] BROWN P F, DESOUZA P V, MERCER R L, *et al*. Class based N-gram models of natural language[J]. Computational Linguistics, 1992, 18(4): 467-479.

[12] 杨健, 汪海航. 基于隐马尔可夫模型的文本分类算法[J]. 计算机应用, 2010, 30(9): 2348-2350.
YANG J, WANG H H. Text classification algorithm based on hidden Markov model[J]. Journal of Computer Applications, 2010, 30(9): 2348-2350. (In Chinese)

[13] 熊志斌, 刘冬. 朴素贝叶斯在文本分类中的应用[J]. 软件导刊, 2013, 12(2):49-51.
XIONG Z B, LIU D. Application of naive Bayes in documents[J]. Software Guide,2013, 12(2):49-51. (In Chinese)

[14] 张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类[J]. 计算机工程, 2005, 31(8).
ZHANG N, JIA Z Y, SHI Z Z. Text categorization with KNN algorithm [J]. Computer Engineering, 2005, 31(8). (In Chinese)

[15] 张玉芳, 彭时名, 吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程, 2006, 32(19): 76-78.
ZHANG Y F, PENG S M, LV J. Improvement and application of TFIDF method based on text classification [J]. Computer Engineering, 2006, 32(19): 76-78. (In Chinese)

[16] 李明江. 结合类词频的文本特征选择方法的研究[J]. 计算机应用研究,2014,31(7): 2024-2026.
LI M J. Research on method of feature selection in text combined with word frequency in class[J]. Application Research of Computers, 2014,31(7): 2024-2026. (In Chinese)

[17] 陈建华. 中文文本分类特征选择方法研究[D]. 西北师范大学, 2012.
CHEN J H. Research of feature selection method for Chinese text classification[D]. Northwest Normal University, 2012. (In Chinese)

[18] 张启蕊, 张凌, 董守斌, 等. 训练集类别分布对文本分类的影响[J]. 清华大学学报: 自然科学版, 2005,45(S1).
ZHANG Q R,ZHANG L,DONG S B, *et al*. Effects of category distribution in a training set on text categorization[J]. Journal of Tsinghua University: Science and Technology, 2005,45(S1). (In Chinese)