文章编号：2096-1472(2016)-10-09-06

基于Spark和改进的TF-IDF算法的用户特征分析

张舒雅，王占刚
（天津工业大学计算机科学与软件学院，天津 300387）

摘　要：使用朴素贝叶斯分类算法，结合Spark内存计算框架，对用户观看视频及次数信息进行分析，建立用户性别和年龄区间的分类模型；然后利用特征项的权重优化模型，考虑到每个特征项在各个类别中的权重对分类结果的影响，提出了一种基于特征项与类别间相关性的TFC−IDFC权重计算方法，并与传统的TF−IDF权重计算方法进行比较，通过正确率和F1值两个指标，证明考虑到特征项与类别的相关性所提出的TFC−IDFC权重使得分类模型的分类能力更好。

关键词：Spark；用户特征；贝叶斯；分类；TF−IDF
中图分类号：TP391　　**文献标识码**：A

User Characteristic Analysis Based on Spark and the Improved TF-IDF Algorithm

ZHANG Shuya,WANG Zhangang
(*School of Computer Science and Software Engineering*,*Tianjin Polytechnic University*,*Tianjin* 300387,*China*)

Abstract: Applying Naive Bayes classification algorithm and integrating Spark memory computing framework,the paper analyzes the information about users' video watching and builds up a classification model of genders and age ranges. Then the paper optimizes the model through the weights of characteristic items.Considering the influence of weight of each characteristic item in its own category on the classification results,the paper proposes a TFC-IDFC weight computing algorithm based on the correlation between characteristic items and categories.Through the comparison in the accuracy rate and the F1 value with traditional TF-IDF weight computing algorithms,this TFC-IDFC weight computing algorithm is proved to provide the model with better classification ability.

Keywords:spark;user characteristics;Bayes;classification;TF-IDF

1 引言(Introduction)

随着互联网的快速发展，用户的数量飞速增加，用户属性更加多元化，大数据的应用与创新成为一个重要的关注点。通过用户的网络行为，分析用户的特征，无论在理论研究中还是实际应用中，都是一个热门话题。大数据用户特征分析，整合海量用户数据，将用户标签化，使得计算机能够程序化处理与人相关的信息，通过机器学习算法、模型能够“理解”人。深度分析用户特征，在理论研究上可以更好地挖掘事件关联及预测事件；对于企业而言，无论是搜索引擎、推荐系统、广告投放等各种应用领域，都可以进一步提高获取信息的精准度和效率。

而Spark作为一种基于内存计算的分布式计算框架，正受到越来越多大数据研究者的关注。它提供了一个更快、更通用的数据处理平台，通过将大量数据集计算任务分配到多台计算机上，并且将中间过程的输出结果保存在内存中，不再需要读取和写入HDFS，以提供高效内存计算，因此Spark可以更好的应用于大数据挖掘和机器学习等算法[1−3]。同时Spark引入了弹性分布式数据集(RDD，Resilient Distributed Dataset)。RDD是不可变的、容错的、分布式对象集合，用户可以利用RDD的操作函数并行地操作该集合，以提高计算速度。

目前国内用户特征分析的研究主要是对社交网络、微博评论、日志数据等进行特征分析，少部分人则对视频数据进行分析。张岩峰等人通过用户在微博上的言论、行为和社交圈等公开数据信息，提出了对该用户的个性化维度进行分类分析的方法[4]；张宏鑫等人从海量移动终端日志数据中挖掘用户特征，提出了一种基于日志数据的用户特征分析方法[5]；李冰利用用户观看新闻类视频数据，并通过用户行为分析和建模处理，挖掘用户在类别、国别、年代、热度值、评分等维度的兴趣偏好[6]；冯婷婷通过用户浏览视频的行为，利用支持向量机、逻辑回归等分类器进行性别推理[7]。

国际上，Das S等人通过终端用户的特征标签，提出了基于权重的逻辑回归算法的监督和半监督学习的用户特征分析[8]；Kim H L等人提出通过分析用户标签，实现以用户兴趣为中心的聚类[9]；Gulsen E等人利用网络日志数据，使用url、DMOZ和文本内容三个特征数据集，预测性别[10]。

目前利用用户观看视频信息分析用户特征的研究成果还

比较少。本研究利用用户观看视频及次数信息，基于朴素贝叶斯分类算法[11-15]和Spark内存计算框架，训练用户性别与年龄区间的分类模型，其中年龄区间分为19岁以下、19—30岁、31—40岁、41—50岁和50岁以上，通过计算每个特征项在各个类别中的权重优化模型，提高分类结果的正确率。

2 用户特征分析算法(User characteristics analysis algorithm)

2.1 贝叶斯定理

贝叶斯定理是一则关于随机事件A和B的条件概率的定理。$p(A|B)$表示事件B发生的条件下，事件A发生的概率，其基本公式为

$$p(A \mid B)=\frac{p(AB)}{p(B)} \tag{1}$$

贝叶斯定理为

$$p(B \mid A)=\frac{p(A \mid B)p(B)}{p(A)} \tag{2}$$

2.2 朴素贝叶斯分类算法

朴素贝叶斯分类算法的基本思想是：对于给定的待分类样本，求得该样本出现的条件下各个类别出现的概率，取得最大概率的类别，就认为该样本属于这个类别。

朴素贝叶斯分类的定义如下：

(1)设$x=\{a_1,a_2,\cdots,a_m\}$为一个待分类项，每个a是x的一个特征项。

(2)有类别集合$C=\{c_1,c_2,\cdots,c_n\}$，计算各类别的先验概率，计算公式为

$$p(c_i)=\frac{numSamplesLabel(i)}{sumSamples} \tag{3}$$

式中，$p(c_i)$为类别c_i的先验概率，$numSamplesLable(i)$为类别c_i的样本数，$sumSamples$为样本总数。

(3)计算每个特征项在各个类别下的条件概率，分为两种模式：

①多项式模型

$$p(a_j \mid c_i)=\frac{numFreqs(j,i)+lambda}{numFeatures+numFreqsLabel(i)} \tag{4}$$

式中，$p(a_j|c_i)$为特征项a_j在类别c_i下的条件概率，$numFreqs(j,i)$为特征项a_j在类别c_i中出现的次数，$numFreqsLabel(i)$为类别c_i中所有特征项的总次数，$numFeatrues$是特征项数，$lambda$是平滑因子。

②伯努利模型

$$p(a_j \mid c_i)=\frac{numSamples(j,i)+lambda}{numSamplesLabel(i)+2} \tag{5}$$

式中，$p(a_j|c_i)$为特征项a_j在类别c_i下的条件概率，$numSamples(j,i)$为类别c_i中包含特征项a_j的样本数，$numSamplesLable(i)$为类别c_i的样本数，$lambda$是平滑因子。

为防止分子为0，以上平滑因子$lambda$均取值为1。

(4)计算$p(c_1|x),p(c_2|x),\cdots,p(c_n|x)$

每个特征项是条件独立的，根据贝叶斯定理公式(2)推导为

$$p(c_i \mid x)=\frac{p(x \mid c_i)p(c_i)}{p(x)} \tag{6}$$

$$p(x \mid c_i)p(c_i)=p(a_1 \mid c_i)p(a_2 \mid c_i)\cdots p(a_m \mid c_i)p(c_i)=p(c_i)\prod_{j=1}^{m}p(a_j \mid c_i) \tag{7}$$

(5)因为$p(x)$对于所有类别为常数，只需要将$p(x|c_i)p(c_i)$最大化，即$p(c_k|x)=max\{p(x|c_1)p(c_1),p(x|c_2)p(c_2),\cdots,p(x|c_n)p(c_n)\}$，则$x\in c_k$。

2.3 基于特征项权重的改进

在传统的基于空间向量模型的特征项表示中，特征项对样本类别的代表能力，通常会利用某一种计算方法赋予相应的权重，用来表示他们区分样本类别的重要程度。特征项的权重可以综合反映出该特征项对识别样本的贡献程度以及区分样本类别的能力，选择不同的特征项权重计算方法将对分类模型的分类效果产生较大程度的影响。在传统的TF-IDF权重计算方法的基础上，将特征项和类别的相关性因素加入到权重计算方法中，提出一种改进的特征项权重计算方法。

2.3.1 传统的TF-IDF权重计算方法

TF(Term Frequency)是特征项频率，表示某个特征项在样本集中出现的次数除以样本集中所有特征项出现的总次数。TF的计算公式为

$$TF(a_j)=\frac{numFreqsFeature(j)}{sumFreqs} \tag{8}$$

式中，$numFreqsFeature(j)$是特征项a_j在样本集中出现的总次数，$sumFreqs$是样本集中所有特征项出现的总次数；考虑到特征项的频率信息，使得每个特征项的权重有了很大差异，但是单纯考虑特征项的频率会对高频特征项产生过大的依赖，并且有可能会抛弃一些仅在某个类别中出现的低频特征项，因此只考虑词频，不足以表示一个特征项对样本类别的有用程度；如果TF值高的特征项比较均匀的出现在样本集中，这样的特征项很难说可以代表哪个类别，因此需要计算IDF值。

IDF(Inverse Document Frequency)是反文档频率，它是用包含特征项的样本数来计算特征项的权重，即一个特征项的样本频率越高，说明该特征项出现在大部分样本中，其代表类别的能力就越弱，也就是说该特征项的重要程度越低。IDF的计算公式为

$$IDF(a_j)=\log\frac{sumSamples}{numSamplesFeature(j)} \tag{9}$$

式中，$sumSamples$为样本总数，$numSamplesFeature(j)$为出

现特征项a_j的样本数。

*IDF*计算方法的核心思想是，只在小部分样本中出现的特征项要比在大部分样本中都出现的特征项重要，也就是说可以增强出现在小部分样本中的特征项的有用程度，削弱出现在大部分样本中的特征项的有用程度。

因此特征项a_j在样本中的权重w_j的计算公式为

$$w_j = TF(a_j) \times IDF(a_j) \tag{10}$$

考虑到特征项a_j在样本集中的权重(TF−IDF)，将公式(7)更新为

$$p(x \mid c_i)p(c_i) = p(c_i)\prod_{j=1}^{m} p(a_j \mid c_i)w_j \tag{11}$$

2.3.2 改进的权重计算方法(TFC-IDFC权重)

通过公式(8)和公式(9)发现TF−IDF权重只考虑特征项在整个样本集中的TF值和IDF值，并没有考虑到特征项与类别的相关性；如果一个特征项在某个类别中频繁出现，却很少出现在其他类别中，该特征项区分类别的能力可以说是很强的，因此将权重计算方法改进为TFC−IDFC权重计算方法。

TFC−IDFC权重是计算特征项a_j在类别c_i中的权重方法，用$w_{j,i}$来表示。TFC表示特征项a_j在类别c_i中出现的次数除以样本集中所有特征项出现的总次数。TFC的计算公式为

$$TFC(a_j \mid c_i) = \frac{numFreqs(j,i) + lambda}{sumFreqs} \tag{12}$$

式中，$numFreqs(j,i)$是特征项a_j在类别c_i中出现的次数，*sumFreqs*是样本集中所有特征项出现的总次数。

*IDFC*是指类别c_i中包含特征项a_j的样本数与除类别c_i以外的其他类中包含特征项a_j的样本数的比值。如果某个特征项在某个类别中的*IDFC*值越高，表明特征项在类别间出现越不均匀，其代表类别的能力越强，也就表示该特征项越重要。*IDFC*的计算公式为

$$IDFC(a_j \mid c_i) = \log\left(\frac{numSamples(j,i) + lambda}{numSamplesFeature(j) - numSamples(j,i) + lambda} + L\right) \tag{13}$$

式中，$numSamplesFeature(j)$为样本集中出现特征项a_j的样本总数，$numSamples(j,i)$为类别c_i中包含特征项a_j的样本数；为了防止分子、分母为0和*IDFC*值为负数，以上*lambda*均取值为1，L取值为1。

改进后的权重计算为

$$w_{j,i} = TFC(a_j \mid c_i) \times IDFC(a_j \mid c_i) \tag{14}$$

不同类别的样本中，如果特征项a_j在类别c_i中出现的频率很高，并且在其他类别中很少出现，则该特征项a_j的TFC−IDFC权重很大，也就是说通常$w_{j,i}$值较大的特征项，表示在该类样本中具有较高的权重。

考虑到特征项a_j在类别c_i中的权重(TFC−IDFC)，将公式(7)更新为

$$p(x \mid c_i)p(c_i) = p(c_i)\prod_{j=1}^{m} p(a_j \mid c_i)w_{j,i} \tag{15}$$

通过公式(15)对测试样本进行分类，对比标注标签和分类结果，评价分类模型。

3 实验过程及结果分析(Experiment process and result analysis)

3.1 实验环境

实验中采用6个节点、31核的分布式集群和spark分布式计算框架、通过RDD的转换操作函数和行动操作函数计算类别先验概率和每个特征项在各个类别下的条件概率，得到分类模型。

3.2 实验数据

实验过程中通过YouTube数据集网站获取了一个100万条、2万维的样本数据集，训练用户性别与年龄区间的分类模型，进而可以利用该模型预测待分类样本数据。其中70万条为训练样本，30万条为测试样本。每条样本为用户观看视频信息的数据，数据内容包括用户观看的视频及次数信息的2万维的特征信息；数据格式定义为label,feature1feature2…feature20000，每一个feature为一个视频，值为观看的次数，label为性别或年龄区间的标注标签。

3.3 实验系统架构

实验中将100万条样本数据分别进行性别标注和年龄区间标注，作为2个样本集，将2个样本集分别进行如下过程：取70万条样本作为训练数据，30万条样本作为测试数据；读取70万条训练样本，利用贝叶斯分类算法和RDD的操作函数，计算类别的先验概率和特征项在类别条件下的条件概率，得到分类模型；再读取30万条测试样本，加载分类模型，对测试样本进行分类；将分类结果与标注标签进行对比，计算正确率和F1值，对分类模型进行评价。

考虑到特征项在样本集中的权重(TF−IDF)和特征项在类别中的权重(TFC−IDFC)，则在计算每个特征项在各个类别下的条件概率时，乘以相应权重因子；实验中的系统架构如图1所示。

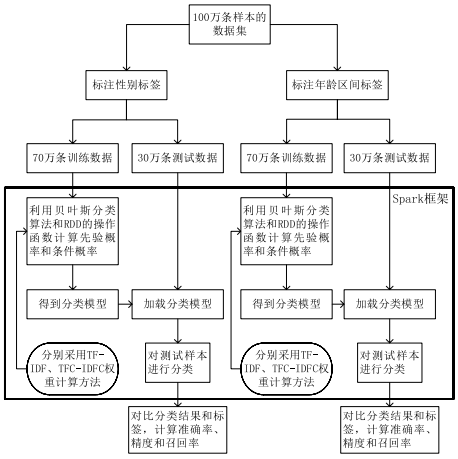

图1 实验系统架构图

Fig.1 Experiment system architecture

3.4 实验过程

从公式(4)和公式(5)可以看出：伯努利模型考虑的是特征项是否在样本中出现，而多项式模型则考虑特征项出现的次数，两种模式侧重不一样。McCallum对比大量的研究发现，当特征项维度比较大的时候，多项式模型的误差总小于伯努利模型[16-25]，由于实验中采用2万维的特征项，并且考虑到用户观看每个视频的次数，因此本文采用多项式模型。实验步骤如下：

(1)对数据进行处理，将数据向量化，并处理成label,feature1feature2…的格式。

(2)对(label,features)格式的样本数据采用combineByKey进行聚合操作，聚合同一个label的features，得到所有label的统计(label，(numSamplesLabel(i)，features_sum_vector))，记为aggregation_one。

(3)对(label,features)格式的样本数据采用combineByKey再次进行聚合操作，聚合同一个label的features，但此次聚合操作中，将features向量中大于0的数置为1进行聚合，得到所有label的统计(label，(numSamplesLabel(i)，features_normal_sum_vector))，记为aggregation_two。

(4)特征维度numFeatures和训练样本总数sumSamples事先定义，分别为2万和70万。

(5)循环aggregation_one，利用$numSamplesLabel(i)$，求得所有label的数组记为labelArray，并根据公式(3)计算每个label的先验概率$p(c_i)$；同时计算features_sum_vector.values.sum，求得类c_i中的总特征数$numFreqsLabel(i)$，遍历features_sum_vector的每一个索引值，求得特征项a_j在类c_i中出现的次数$numFreqs(j,i)$，根据公式(4)计算每个feature在每个label中的条件概率；累加每个label的features_sum_vector，求得的向量为特征项a_j在所有样本中出现的总次数$numFreqsFeature(j)$；累加$numFreqsLabel(i)$求得所有样本中所有特征项的总次数sumFreqs，根据公式(8)计算$TF(a_j)$，根据公式(12)计算$TFC(a_j,c_i)$。

(6)循环aggregation_two，遍历features_normal_sum_vector的每一个索引值，求得类c_i中包含特征项a_j的样本数$numSamples(j,i)$；累加每个label的features_normal_sum_vector，求得所有样本中出现特征项a_j的样本数$numSamplesFeature(j)$，根据公式(9)计算$IDF(a_j)$，根据公式(13)计算$IDFC(a_j,c_i)$。

(7)根据公式(11)和公式(15)计算每个feature在各个label中的加权条件概率。

(8)利用labelArray、类别的先验概率和特征项在类别下的条件概率生成分类模型。

(9)根据生成的分类模型，对测试样本进行测试。

(10)测试过程中对测试样本的标注label和预测label进行计数，分别计算正确率、精度和召回率。

(11)通过正确率和F1值两个指标，对比TF-IDF权重分类模型和TFC-IDFC权重分类模型。

3.5 实验结果

3.5.1 评价标准

实验通过正确率和F1值两个指标进行评估。

对于给定测试样本数据集的分类情况如表1所示。

表1 测试集分类情况

Tab.1 Test set classification

项目名称	实际属于该类	实际不属于该类
预测属于该类	a	b
预测不属于该类	c	d

那么正确率(accuracy)、精度(precision)和召回率(recall)的计算公式分别为公式(16)—公式(18)。

$$accuracy=\frac{a+d}{a+b+c+d} \tag{16}$$

$$precision=\frac{a}{a+b} \tag{17}$$

$$recall=\frac{a}{a+c} \tag{18}$$

正确率是实际类别与预测类别相同的样本的占所有测试样本的比例；精度是实际类别与预测类别相同的样本占预测为该类样本的比例；召回率是实际类别与预测类别相同的样本占实际为该类样本的比例；精度和召回率分别从不同的角度反映了分类器的分类质量，一些情况下这两个指标是互补的，因此可以将两者综合起来考虑，如公式(19)

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{19}$$

通过$F1$值可以综合评价分类模型的性能。

3.5.2 结果分析

基于未加权、TF-IDF权重计算方法和TFC-IDFC权重计算方法在性别分类和年龄区间中的分类情况如表2和表3所示。

表2 未加权、TF-IDF、TFC-IDFC权重对性别的分类情况

Tab.2 Gender classification of unweighted, TF-IDF and TFC-IDFC weight

性别	正确率			*F*1值		
	未加权	TF-IDF	TFC-IDFC	未加权	TF-IDF	TFC-IDFC
男	0.8512	0.8667	0.8871	0.8397	0.8518	0.8741
女	0.8512	0.8667	0.8871	0.8622	0.8788	0.8970

表3 未加权、TF-IDF、TFC-IDFC权重对年龄区间的分类情况

Tab.3 Age classification of unweighted, TF-IDF and TFC-IDFC weight

年龄	正确率			*F*1值		
	未加权	TF-IDF	TFC-IDFC	未加权	TF-IDF	TFC-IDFC
19岁以下	0.8841	0.8983	0.9017	0.6163	0.6324	0.6482
19—30岁	0.8174	0.8313	0.8480	0.7657	0.7718	0.7919
31—40岁	0.7811	0.7950	0.8430	0.7610	0.7746	0.7931
41—50岁	0.8292	0.8450	0.8663	0.6134	0.6259	0.6678
50岁以上	0.8872	0.9033	0.9180	0.7674	0.7837	0.8100

由表中数据可以看出，考虑到每个特征项的权重可以提高分类模型的分类能力，但是仅仅依靠特征项在样本集中的权重，即TD-IDF权重，来优化分类模型的效果并不明显；其原因是该权重并不能反映出在各个类别下的差异；因此基于特征项与类别之间的相关性而改进的TFC-IDFC权重计算方法，理论依据是当一个特征项较为平均的出现在样本集中，其代表某个类别的能力不如频繁的出现在某个类别的样本中，使得优化后的分类模型的分类效果较为理想。性别和年龄区间的分类模型的*F*1值结果比较图如图2和图3所示。

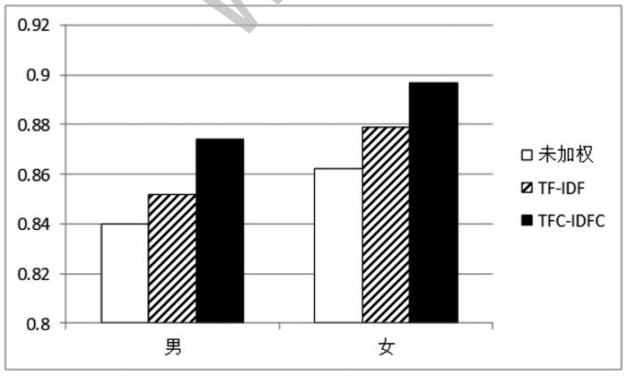

图2 性别分类模型*F*1值结果

Fig.2 *F*1 value of gender classification model

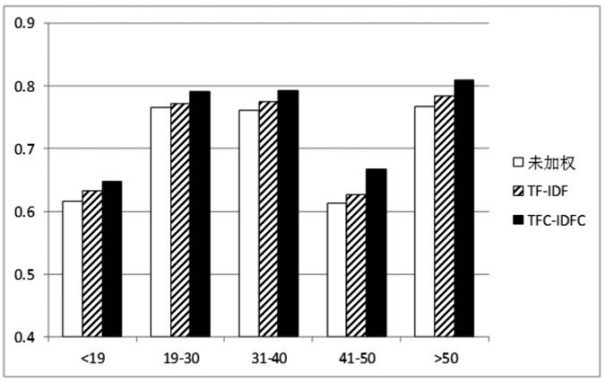

图3 年龄区间分类模型*F*1值结果

Fig.3 *F*1 value of age classification model

4 结论(Conclusion)

本文利用用户观看视频的数据，将朴素贝叶斯分类算法应用到Spark计算框架，训练用户的性别和年龄区间的分类模型、加载模型，对测试样本进行分类，比较分类结果与标注标签，分析模型性能，整个过程耗时大约三分钟。在实验过程中，未考虑特征项权重的分类效果不是很理想；在朴素贝叶斯分类算法中加入传统的TF-IDF权重计算方法，分类效果仅有小幅度提升；其原因是TF-IDF权重考虑的是特征项与整个样本集的相关性，并没有考虑到特征项与类别的相关性，给出的特征项权重并不准确，因此文本提出了一种改进的基于特征项与类别间相关性的TFC-IDFC权重计算方法。该算法的思想是某个特征项在各个类别间出现的越不均匀，则区分类别的能力越强，权重越大，也就意味着每个特征项在各个类别中的权重在很大程度上影响了朴素贝叶斯分类算法的分类能力；另外，通过实验也证明改进的TFC-IDFC权重计算方法，增加对类别重要程度较大的特征项的权重，有利于提高分类模型的质量和分类结果的正确率，分类效果有了明显的提高。

大数据背景下的用户特征分析是当前的研究热点，用户的一切网络行为都是值得挖掘的对象。在最短的时间内，能够对用户更加准确分析是我们的研究目标。采用更多维度、更大量级的数据信息，对用户更多特征的分析将是下一步的研究重点。

参考文献(References)

[1] Zhang F,et al.A Distributed Frequent Itemset Mining Algorithm Using Spark for Big Data Analytics[J].Cluster Computing,2015,18(4):1493-1501.

[2] Semberecki P,Maciejewski H.Distributed Classification of Text Documents on Apache Spark Platform[C].International Conference on Artificial Intelligence and Soft Computing. Springer International Publishing,2016:621-630.

[3] Meng X,et al.Mllib:Machine Learning in Apache Spark[J]. JMLR,2016,17(34):1-7.

[4] ZHANG Yanfeng,et al.A Micro-Blog User Personality Classification Analysis[J].Computer Engineering and Science,2015,37(2):402-409.

[5] ZHANG Hongxin,et al.Visualization of Crowd Characteristics Based on Mobile terminal log data[J].Journal of Software,2016,27(5):1230-1245.

[6] LI Bing.Design and Implementation of Personalized Video Recommendation System based on Hadoop[D].Beijing University of Technology,2015.

[7] Feng T,et al.Tags and Titles of Videos you Watched Tell Your Gender[C].ICC 2014 IEEE International Conference on Communications,2014:1837-1842.

[8] Das S,et al.End-User Feature Labeling: Supervised and Semi-supervised Approaches Based on Locally-Weighted Logistic Regression[J].Artificial Intelligence,2013,204(9):56-74.

[9] Kim H L,et al.Mining and Representing User Interests:The Case of Tagging Practices[J].Systems Man & Cybernetics Part A Systems & Humans IEEE Transactions on,2011,41(4):683-692.

[10] Gulsen E,et al.Big Data Feature Selection and Projection for Gender Prediction Based on User Web Behaviour[C].Signal Processing and Communications Applications Conference (SIU),2015 23th.IEEE,2015:1545-1548.

[11] Luo X,et al.Improvement of Automatic Chinese Text Classification by Combining Multiple Features[J]. IEEJ Transactions on Electrical and Electronic Engineering,2015,10(2):166-174.

[12] Lee C H.A Gradient Approach for Value Weighted Classification Learning in Naive Bayes[J].Knowledge-Based Systems,2015,85(C):71-79.

[13] Bi W,Kwok J T.Bayes-Optimal Hierarchical Multilabel Classification[J].IEEE Transactions on Knowledge and Data Engineering,2015,27(11):2907-2918.

[14] Kim H K,Kim M.Model-Induced Term-Weighting Schemes for Text Classification[J].Applied Intelligence,2016:1-14.

[15] Vicente M,Batista F,Carvalho J P.Twitter Gender Classification Using User Unstructured Information[C].Fuzzy Systems(FUZZ-IEEE),2015 IEEE International Conference on.IEEE,2015:1-7.

[16] McCallum A,Nigam K.A Comparison of Event Models for Naive Bayes Text Classification[C].AAAI-98 Workshop on Learning for Text Categorization,1998,752:41-48.

[17] Peralta D,et al.Evolutionary Feature Selection for Big Data Classification:A MapReduce Approach[J].Mathematical Problems in Engineering,2015,12(05):301-305.

[18] LIANG Hong,XU Nanshan,LU Lingang.Sina Micro-blog Users Characteristics Analysis[J].Computer Engineering and Applications,2015,51(7):141-148.

[19] Bozkurt O O,Taygi Z C.Audio-Based Gender and Age Identification[C].Signal Processing and Communications Applications Conference,2014:1371-1374.

[20] Pentreath N.Machine Learning with Spark:Create Scalable Machine Learning Applications to Power a Modern Data-Driven Business Using Spark[M].Packt Publishing,2015.

[21] Hu W,et al.Tagpref:User Preference Modeling by Social Tagging[C].Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing.IEEE Computer Society,2013:111-118.

[22] Sun X,Lin H.Topical Community Detection from Mining User Tagging Behavior and Interest[J].Journal of the American Society for Information Science & Technology,2013,64(2):321-333.

[23] Wang Z,et al.Analysis of User Behaviors by Mining Large Network Data Sets[J].Future Generation Computer Systems,2014,37(7):429-437.

[24] Han Y,Xia K.Data Preprocessing Method Based on User Characteristic of Interests for Web Log Mining[C]. Instrumentation and Measurement, Computer,Communication and Control(IMCCC),2014 Fourth International Conference on.IEEE,2014:867-872.

[25] Bai S,et al.Predicting Big Five Personality Traits of Microblog Users[C].2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence(WI)and Intelligent Agent Technologies(IAT).IEEE Computer Society,2013:501-508.

作者简介：

张舒雅(1989-)，女，硕士生.研究领域：大数据挖掘.

王占刚(1975-)，男，博士，副教授.研究领域：大数据，计算机检测应用，计算机网络安全.