文章编号:1671-5896(2017)01-0092-06

基于类空间密度的文本分类特征加权算法

贾隆嘉[a,b],孙铁利[a,b],杨凤芹[a,b],孙红光[a,b]
(东北师范大学 a. 计算机科学与信息技术学院; b. 智能信息处理吉林省高校重点实验室,长春 130117)

摘要: 特征加权是一种依据特征在分类中起到的作用为特征赋予相应权重的过程,是为了提高分类性能而为特征标记权重的策略。基于类空间密度提出了两个新的特征加权算法: tf* ICSDF 和 ICSDF-based。实验中,在 RCV1-4 和 20 Newsgroups 数据集上,采用支持向量机分类器将提出的方法进行了验证。实验结果显示,该方法相比传统的特征加权方法(prob-based、tf* icf 和 icf-based)可以有效地提升文本分类性能。

关键词: 特征加权; 类空间密度; 文本分类; 机器学习
中图分类号: TP31 **文献标识码:** A
DOI:10.19292/j.cnki.jdxxp.2017.01.015

Class Space Density Based Weighting Scheme for Automated Text Categorization

JIA Longjia[a,b], SUN Tieli[a,b], YANG Fengqin[a,b], SUN Hongguang[a,b]
(a. School of Computer Science and Information Technology; b. Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China)

Abstract: Term weighting is a weighting process for terms which is based on the term's effect to the classification. Term weighting is a strategy that assigns weights to terms in order to improve the performance of text categorization. We propose two new class space density based term weighting scheme, tf-ICSDF and ICSDF-based. In the experiments, we investigate the effects of the proposed scheme on the RCV1-4 and 20 Newsgroups datasets using the SVM (Support Vector Machine) as classifiers. The results show that the proposed scheme outperform other traditional term weighting schemes, such asprob-based, tf-icf and icf-based.

Key words: term weighting; class space density; text categorization; machine learning

0 引 言

文本分类特征加权方法中涉及到 3 个重要组成部分: 词频、文档频率和类别频率。目前多数研究中所采用的与类别频率相关的参数主要为逆类别频率(icf)[1-4]。逆类别频率的基本思想与逆文档频率类似,某一给定特征的逆类别频率可以通过如下方法计算: 由训练集中的总类别数目除以包含当前特征的类别数目,然后,将得到的结果(商值),取其对数得到。然而,该方法仅考虑包含特征的类别数目,对于每个类别有多少篇文档包含当前特征没有进一步分析。笔者综合分析了逆文档频率的优点及缺点,并基于两个以逆类别频率为基础的方法: tf* icf 和 icf-based[4],通过结合类空间密度(CSD: Class Space Density),进而引入逆类别空间密度频率(ICSDF: Inverse Class Space Density Frequency),提出了两个新的特征加权算法 tf* ICSDF 和 ICSDF-based。将提出的方法和 3 个当前较为有效的特征加权方法(prob-

收稿日期: 2016-03-07
基金项目: 长春市科技局基金资助项目(14KP009); 吉林省科技厅基金资助项目(20130206041GX); 吉林省发改委基金资助项目(2015Y56; [2013] 779)
作者简介: 贾隆嘉(1988—),男,长春人,东北师范大学博士研究生,主要从事数据挖掘文本分类研究,(Tel) 86-13624498811 (E-mail) jialongjia@163.com; 通讯作者: 孙红光(1970—),女,长春人,东北师范大学副教授,博士,主要从事视频分析、机器视觉与智能信息处理研究,(Tel) 86-13154303182(E-mail) sunhg889@nenu.edu.cn。

based、tf* icf 以及 icf-based) 应用于 RCV1-4 和 20Newsgroups 两个数据集的文本表示模型后，采用支持向量机分类器作为分类方法，将结果在微平均 F_1 与宏平均 F_1 两个方面进行对比，实验结果显示，笔者所提出的特征加权算法可以有效地提高文本分类的性能。

1　相关特征加权算法

近年来，各种特征加权算法被广泛应用于文本分类领域。Debole 等[3] 采用训练集中已知的类别信息，提出了有监督方法。在特征加权的过程中，通过引入 3 个特征选择方法(卡方检验、信息增益及信息增益率) 代替 idf。Liu 等[4] 提出了一种基于概率的有监督特征加权方案(prob-based) ，用以解决文本分类中失衡数据集的分类问题。Wang 等[2] 提出了 tf* icf 和 icf-based 方法。

1.1　prob-based

Liu 等[4] 提出的 prob-based 特征加权方法，主要是应用于失衡数据集的分类。公式(1) 中提出了两个参数: a/b 与 a/c; 其中 a/b 可以被认为是某一特征对特定类别的一个相关指标，比值越高，则代表当前特征与此类别的相关性越高; 相应地，a/c 反映的是: 如果一个特征相比其他特征在某一类别中出现的文档数越多，则代表该特征与此类别更相关。通过结合上述两个参数，替代了传统的tf* idf特征加权方法中的 idf，提出了新的特征加权方法，公式如下[4]

$$P_{\text{prob-based}}(t_i) = \frac{T_{\text{tf}t_i}}{\max(T_{\text{tf}_{t_i}})}\log\left(1 + \frac{a}{c}\frac{a}{b}\right) \tag{1}$$

其中 $\max(\text{tf}_{t_i})$ 代表当前特征所在文档中最大的特征词频，a 代表在正类别文档中含有当前特征的文档数，b 代表在正类别中不含有当前特征的文档数，c 代表负类别中含有当前特征的文档数。

1.2　tf* icf 和 icf-based

许多研究发现: 在文本分类中，词频(tf) 是一个非常重要的参数，单独使用 tf 加权可以获得较好的效果[5-7] 。因此，人们集中精力于替代文档频率(df) 的部分。Wang 等[8] 认为在一篇文档中，特征的区分能力不仅与 tf 相关，而且还与特征在各类别之间的分布有关。他们提出了“icf 假设”: 特征在各类别中出现的越少，特征的区分能力越强。在“icf 假设”中，icf 在加权时侧重于在类别级上出现稀少的特征。

Wang 等[8] 通过将 icf 引入到特征加权中，通过改进 tf* idf 与 prob-based 分别提出了 tf* icf 和 icf-based方法，公式如下

$$T_{\text{tf}} - T_{\text{icf}}(t_i) = T_{\text{tf}}(t_i)\log\left(1 + \frac{|C|}{C_{\text{cf}}(t_i)}\right) \tag{2}$$

$$I_{\text{icf-based}}(t_i) = T_{\text{tf}}\log\left(2 + \frac{a}{\max(1,c)}\frac{|C|}{C_{\text{cf}}(t_i)}\right) \tag{3}$$

其中 $|C|$ 代表当前训练集中的类别数，$C_{\text{cf}}(t_i)$ 代表当前特征在训练集中出现的类别数目; a、c 所代表的含义与式(1) 中相同。

拥有高类别频率的特征将不会是对区分文档有效的特征，这样的特征应该被赋予较低的权重。因此，采用逆类别频率计算时，意味着出现在多个类别中的特征将不会是一个“好特征”。逆类别频率将会为在类别空间中出现在许多个类别中的特征，赋予相对低的特征权重。上述所提方法(prob-based、tf* icf和 icf-based) 都通过不同的方式在计算特征权重时采用了类别信息。其中 tf* icf 和 icf-based 使用了逆文档频率度量类别信息，逆类别频率虽然将拥有高类别频率的特征赋予了较低的权重，但计算时仅考虑了特征的类别频率，而未考虑在此类别频率下，特征在每个类别中出现的文档频率[9,10] 。笔者针对此问题提出新的改进方法。

2　算法的描述与实现

2.1　算法的研究动机

在绝大数多的文本数据集中，不同特征在多个类别中出现次数(文档频率) 存在明显差异，如果仅考虑此方面因素，则出现次数不同的特征应具有不同的权重。然而当前使用较为广泛的逆类别频率不能对

上述情况中的不同特征给予区分，而是赋予相同的权重。为解决此问题，笔者首先引入类空间密度，进而引入逆类别空间密度频率。通过使用新的参数对特征加权方法中的 icf 参数更新，达到了对上述情况中不同的特征分别赋予不同特征权重的效果。

2.2 算法的实现

首先定义了 3 个相关变量：类文档密度(CDD：Class Document Density)、类空间密度(CSD：Class Space Density)以及逆类别空间密度频率(ICSDF：Inverse Class Space Density Frequency)。

CDD 定义为某一特征的类文档密度，表示在当前类别中含有某一特征的文档数目与当前类别的文档总数的比，表示为

$$C_{\mathrm{CDD}\,c_k}(t_i) = \frac{n_{c_k}(t_i)}{N_{c_k}} \tag{4}$$

其中 $n_{c_k}(t_i)$ 表示在类别 c_k 中含有特征 t_i 的文档数目；N_{c_k} 表示在类别 c_k 中的文档总数。

CSD 定义为某一特征的类空间密度表示为

$$C_{\mathrm{CSD}}(t_i) = \sum_{k=1}^{|C|} C_{\mathrm{CDD}\,c_k}(t_i) \tag{5}$$

其中 $|C|$ 代表类别总数。根据类空间密度 CSD，类似逆文档频率，可以得到某一特征的逆类别空间密度频率 ICSDF

$$I_{\mathrm{ICSDF}}(t_i) = \log\frac{|C|}{C_{\mathrm{CSD}}(t_i)} \tag{6}$$

下面举例说明 ICSDF 方法与 icf 方法对不同特征在多个类别中出现次数(文档频率)存在明显差异时，两种方法对特征权重的不同赋予情况。

假设有一个训练集包含两个类别：C_1、C_2，特征 t_1 同时出现在这两个类别中。考虑以下两种情况。

第 1 种情况中，每个类别分别包含 8 篇文档，特征 t_1 在类别 C_1 的 5 篇文档中出现、在类别 C_2 的 4 篇文档中出现。在每个类别含有相同数量文档的情况下，将其认定为均衡类别的情况，如图 1 所示。

在图 1 中，特征 t_1 在 C_1 中的类别密度 $C_{\mathrm{CDD}_1}(t_1) = \frac{n_{c_1}(t_1)}{N_{c_1}} = \frac{5}{8} = 0.625$，特征 t_1 在 C_2 中的类别密度 $C_{\mathrm{CDD}_2}(t_1) = \frac{n_{c_2}(t_1)}{N_{c_2}} = \frac{4}{8} = 0.5$。因此特征 t_1 的类别空间密度：$C_{\mathrm{CSD}}(t_1) = \sum_{k=1}^{2} C_{\mathrm{CDD}}k(t_1) = 0.625 + 0.5 = 1.125$。相应的逆类别空间密度：$I_{\mathrm{ICSDF}}(t_1) = \log\frac{|C|}{C_{\mathrm{CSD}}(t_1)} = \log\left(\frac{2}{1.125}\right) = 0.249\,9$。

第 2 种情况中，两个类别 C_1、C_2 分别包含 12 篇文档与 4 篇文档，特征 t_1 在类别 C_1 的 9 篇文档中出现、在类别 C_2 的 3 篇文档中出现。在每个类别含有不同数量文档的情况下，将其认定为失衡类别的情况，如图 2 所示。

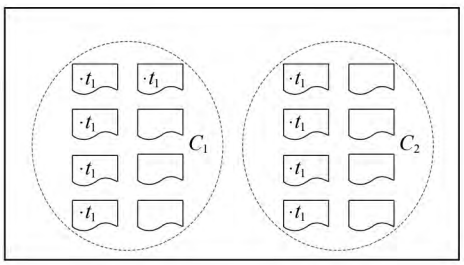

图 1 均衡数据集类空间密度示意图

Fig. 1 Sketch map for class space density of balanced data set

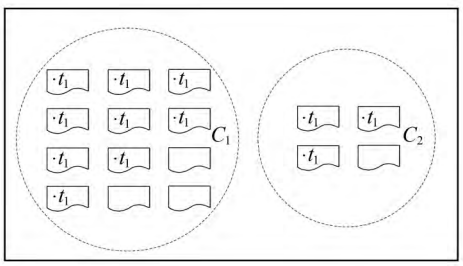

图 2 失衡数据集类空间密度示意图

Fig. 2 Sketch map for class space density of imbalanced data set

在图 2 中，特征 t_1 在 C_1 中的类别密度 $C_{\mathrm{CDD}_1}(t_1) = \frac{n_{c_1}(t_1)}{N_{c_1}} = \frac{9}{12} = 0.75$，特征 t_1 在 C_2 中的类别密度

$C_{\mathrm{CDD}_2}(t_1)=\dfrac{n_{c_2}(t_1)}{N_{c_2}}=\dfrac{3}{4}=0.75$。因此，特征 t_1 的类别空间密度：$C_{\mathrm{CSD}}(t_1)=\sum_{k=1}^{2}C_{\mathrm{CDD}_k}(t_1)=0.75+0.75=1.5$。相应的逆类别空间密度：$I_{\mathrm{ICSDF}}(t_1)=\log\dfrac{|C|}{C_{\mathrm{CSD}}(t_1)}=\log\left(\dfrac{2}{1.5}\right)=0.1249$。

在图 1 与图 2 中，如果不采用逆类别空间密度 ICSDF 计算，而采用逆类别频率 icf 计算，则在两种情况中都将赋予 icf 相同的特征权重，计算方法如下：$I_{\mathrm{icf}}(t_1)=\log\left(1+\dfrac{|C|}{C_{\mathrm{cf}}}\right)=\log\left(1+\dfrac{2}{2}\right)=0.301$。产生上述结果的原因主要是由于 icf 仅考虑某一特征在相应类别中是否出现，而不考虑特征出现的次数。换句话说，假设特征 t_1 在类别 C_1 中的所有文档中都出现，而在类别 C_2 中仅一篇文档中出现，在这种情况下，特征 t_1 对于类别 C_1 有非常大的区分能力，本应该被赋予较高的特征权重，但由于 icf 只考虑特征在类别中出现与否，特征 t_1 对在两个类别中出现频率都是较高或较低的、不具备区分能力的特征赋予相同的权重。

因此，从上述结果可以看出，ICSDF 方法可根据不同情况有效地区分不同的特征，并依据特征在各个类别中出现的文档频率，为特征赋予不同权重。

通过采用 ICSDF 方法更新 icf 方法，可以得到相应改进的特征加权方法。

将 tf-icf 方法改进后，得到 tf-ICSDF

$$T_{\text{tf-ICSDF}}(t_i)=T_{\mathrm{tf}}(t_i)\left(1+\log\frac{|C|}{C_{\mathrm{CSD}}}\right) \tag{7}$$

将 icf-based 方法改进后，得到

$$I_{\text{ICSDF-based}}(t_i)=T_{\mathrm{tf}}(t_i)\log\left(2+\frac{a}{\max(1,c)}\frac{|C|}{C_{\mathrm{CSD}}}\right) \tag{8}$$

笔者所提出方法的描述如下：

算法　逆类别空间密度频率(ICSDF)

输入：fea：经过处理的训练集向量空间模型矩阵；
　　gnd：训练集矩阵中每一个样例对应的标签列向量；
输出：ICSDF(t_i)：训练集中的每个特征对应的逆类别空间密度频率。
通过 fea 获得训练集的样例数目 N、特征维数 M；
通过 gnd 获得训练集的类别数目 $|C|$；
for $i=1$ to N
统计各个类别中所含文档数量 N_{c_i}；
end for
for $i=1$ to M
for $j=1$ to $|C|$
依次统计每个类别中含有当前特征的文档数目 $n_{c_j}(t_i)$；
end for
end for
for $i=1$ to M
for $j=1$ to $|C|$
$C_{\mathrm{CDD}_{c_j}}(t_i)=n_{c_j}(t_i)/N_{c_j}$
end for
end for
for $i=1$ to M
$C_{\mathrm{CSD}}(t_i)=\sum_{j=1}^{|C|}C_{\mathrm{CDD}_{c_j}}(t_i)$
end for
for $i=1$ to M
$I_{\mathrm{ICSDF}}(t_i)=\log\dfrac{|C|}{C_{\mathrm{CSD}}(t_i)}$
end for

3　实验与结果分析

3.1　数据集

1) RCV1-4 数据集。RCV1 是路透社提供的由新闻报道组成的语料库。实验中，采用了由其中 4 个类别(“C15”、“ECAT”、“GCAT”和“MCAT”)构成的子集，称为“RCV1-4”，4 个类别分别包含的文档数量为 2,022、2,064、2,901 和 2,638；构成的数据集共包含 9 625 篇文档，29 992 个特征。

2) 20Newsgroups 数据集。20Newsgroups 是由 Ken Lang 收集的在文本分类领域被广泛采用的数据集。该数据集中共有 19 997 篇文本文档，被近乎平均地分到 20 个类别中(类别 16 中含有 997 篇文档)，相比 Reuter-21578 数据集，它是个典型的均衡数据集。在本实验中，笔者选取最常用的“bydate”版本，共有 18 846篇文档，分为 11 314 篇(大约占总文档数目的 60%)训练文档和 7 532 篇(大约占总文档数目的

40%)测试文档。其中20个类别分别是 "alt. atheism"、"comp. graphics"、"comp. os. ms-windows. misc"、"comp. sys. ibm. pc. hardware"、"comp. sys. mac. hardware"、"comp. windows. x"、"misc. forsale"、"rec. autos"、"rec. motorcycles"、"rec. sport. baseball"、"rec. sport. hockey"、"sci. crypt"、"sci. electronics"、"sci. med"、"sci. space"、"soc. religion. christian"、"talk. politics. guns"、"talk. politics. mideast"、"talk. politics. misc"、"talk. religion. misc"。

3.2 分类器及评价标准

为验证笔者所提出的特征加权方法的有效性，笔者采用支持向量机分类器在微平均 F_1(各个文档性能指标的算术平均)和宏平均 F_1(每个类别性能指标的算术平均)两个方面与 prob-based、tf* icf 以及 icf-based特征加权算法进行了全面比较。

3.3 实验结果

1) RCV1-4 数据集上的实验。图3与图4分别展示了支持向量机分类器采用 tf* icf、icf-based、prob-based、tf* ICSDF、ICSDF-based 5个特征加权方法在 RCV1-4 数据集上运行的微平均 F_1 值及宏平均 F_1 值。

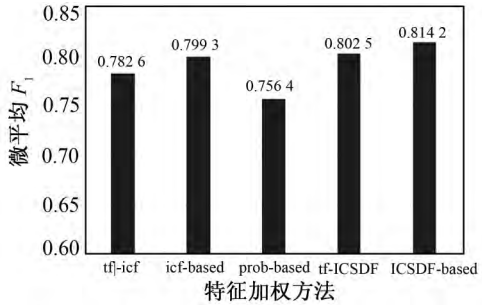

图3 RCV1-4 数据集上微平均 F_1 值对比图

Fig. 3 A comparison on Micro F_1 on RCV1-4 dataset

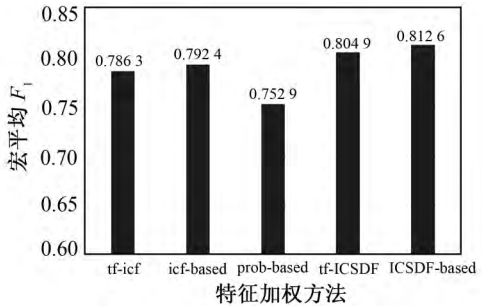

图4 RCV1-4 数据集上宏平均 F_1 值对比图

Fig. 4 A comparison on Macro F_1 on RCV1-4 dataset

从图3可以看出，ICSDF-based 特征加权方法与支持向量机分类器在 RCV1-4 数据集上分类的微平均 F_1 值高于其他特征加权算法与支持向量机分类器结合时的分类性能。另外，tf* ICSDF 特征加权方法获得结果第2高微平均 F_1 值(0.802 5)。从图4可以看出，宏平均 F_1 最高值(0.812 6)由 ICSDF-based 特征加权方法获得，并且 tf* ICSDF 特征加权方法获得了第2高(0.804 9)。

2) 20 Newsgroups 数据集上的实验。图5与图6分别展示了支持向量机分类器采用 tf* icf、icf-based、prob-based、tf* ICSDF、ICSDF-based 5个特征加权方法在 20 Newsgroups 数据集上运行的微平均 F_1 值及宏平均 F_1 值。

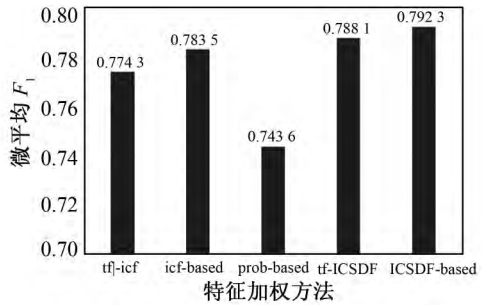

图5 20 Newsgroups 数据集上微平均 F_1 值对比图

Fig. 5 A comparison on micro F_1 on 20 Newsgroupsdataset

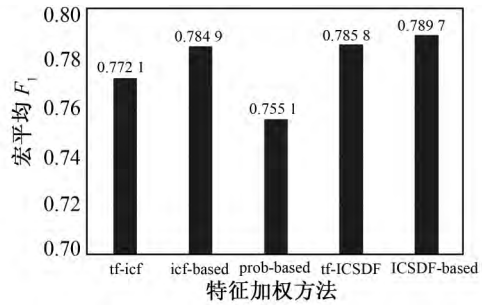

图6 20 Newsgroups 数据集上宏平均 F_1 值对比图

Fig. 6 A comparison on macro F_1 on 20 Newsgroups dataset

从图5中可以看出，ICSDF-based 特征加权方法与支持向量机分类器在 20 Newsgroups 数据集上分类的微平均 F_1 值高于其他特征加权算法与支持向量机分类器结合时的分类性能。另外，tf* ICSDF 特征加权方法获得了第2高的微平均 F_1 值(0.788 1)，icf-based 获得的微平均 F_1 值(0.783 5)排在第3位。从

图 6 可以看出，宏平均 F_1 最高值(0.789 7)由 ICSDF-based 特征加权方法获得，tf* ICSDF 与icf-based两个特征加权算法获得的微平均 F_1 值仅次于 ICSDF-based 特征加权方法，分别为 0.785 8 与0.784 9。

4 结　语

笔者分析了逆类别频率(icf)方法的优缺点：icf 方法与传统无监督特征加权方法(tf、tf* idf 等)的主要区别在于该方法引入了类别信息用于对特征进行加权。但 icf 方法在统计特征的类别频率时，只考虑特征在某个类别中是否出现，而不考虑在此类别中特征出现的次数(文档频率)，使在不同类别中出现次数不同的特征不能得到区分。ICSDF 方法可根据特征在各个类别中出现的文档频率不同，为特征赋予不同的权重。通过将 tf-icf 和 icf-based 中 icf 方法替换为 ICSDF 方法，提出了两个新的特征加权方案：tf-ICSDF和 ICSDF-based。实验采用支持向量机分类在 RCV1-4 数据集和 20Newsgroups 数据集上进行了验证，将提出的方法与基于 icf 的 3 种特征加权方法(prob-based、tf-icf 和 icf-based)在微平均 F_1 值和宏平均 F_1 值两方面进行了全面的对比分析。实验结果表明：笔者提出的基于类空间密度的特征加权算法可以有效地提升文本分类性能。

参考文献：

[1] QUAN X, WENYIN L, QIU B. Term Weighting Schemes for Question Categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 1009-1021.

[2] WANG D, ZHANG H. Inverse-Category-Frequency Based Supervised Term Weighting Scheme for Text Categorization [J]. Journal of Information Science and Engineering, 2013, 29(2): 209-225.

[3] DEBOLE F, SEBASTIANI F. Supervised Term Weighting for Automated Text Categorization [C] // Text Mining and Its Applications. [S.l.]: Springer Berlin Heidelberg, 2004: 81-97.

[4] LIU Y, LOH H T, SUN A. Imbalanced Text Classification: A Term Weighting Approach [J]. Expert Systems with Applications, 2009, 36(1): 690-701.

[5] LAN M, TAN C L, SU J, et al. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721-735.

[6] SEBASTIANI F. Machine Learning in Automated Text Categorization [J]. ACM Computing Surveys (CSUR), 2002, 34(1): 1-47.

[7] LEOPOLD E, KINDERMANN J. Text Categorization with Support Vector Machines, How to Represent Texts in Input Space [J]. Machine Learning, 2002, 46(1/3): 423-444.

[8] WANG D, ZHANG H. Term Frequency-Function of Document Frequency: a New Term Weighting Scheme for Enterprise Information Retrieval [J]. Enterprise Information Systems, 2012, 6(4): 433-444.

[9] REN F, SOHRAB M G. Class-Indexing-Based Term Weighting for Automatic Text Classification [J]. Information Sciences, 2013, 236(1): 109-125.

[10] LANG K. Newsweeder: Learning to Filter Netnews [C] // Proceedings of the 12th International Conference on Machine Learning. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1995: 331-339.

(责任编辑：刘东亮)