Journal of Computer Applications
计算机应用,2013,33(8):2280-2282,2292
ISSN 1001-9081
CODEN JYIIDU
2013-08-01
http://www.joca.cn

文章编号:1001-9081(2013)08-2280-03
doi:10.11772/j.issn.1001-9081.2013.08.2280

新的短文本特征权重计算方法

马雯雯[1*],邓一贵[1,2]
(1.重庆大学 计算机学院,重庆 400044; 2.重庆大学 信息与网络管理中心,重庆 400044)
(*通信作者电子邮箱 ma-wen1024@163.com)

摘 要:短文本固有的特征稀疏和样本高度不均衡等特点,使得传统长文本的加权方法难以直接套用。针对此问题,提出一种针对短文本的特征权重计算方法——综合类别法。该方法引入反文档频和相关性频率的概念,综合考虑了样本在正类和负类中的分布情况。实验结果表明,相对于其他特征权重方法,该方法的微平均和宏平均值均在90%以上,能增强样本在负类中的类别区分能力,改善短文本分类的查准率和查全率。

关键词:短文本;特征权重;不均衡样本;文本分类
中图分类号:TP311 **文献标志码**:A

New feature weight calculation method for short text

MA Wenwen[1*], DEGN Yigui[1,2]
(1. *College of Computer Science, Chongqing University, Chongqing* 400044, *China*;
2. *Center of Information and Network, Chongqing University, Chongqing* 400044, *China*)

Abstract: The inherent sparse features and unbalanced sample of the short text make it difficult for short text to use traditional weight of long text mechanically. To resolve this problem, an approach of short text feature weight named Integrated Category (IC) was proposed. This approach introduced the concept of inverse document frequency and relevancy frequency, and integrated the distribution of sample in positive category and negative category. The experimental results show that, compared with other feature weight methods, the micro-average and macro-average of this method are above 90%, and it can enhance the sample categories distinguishing ability in negative category, and improve the precision and recall of short text categorization.

Key words: short text; feature weight; unbalanced sample; text categorization

0 引言

最近几年,随着即时通信技术的发展和互联网应用的普及,QQ聊天、BBS、微博、新闻评论等短文本数据呈现指数级增长,并已成为一种重要的信息传播方式。大量短文本的出现有利也有弊:一方面,短文本分类在问答社区、广告分类等应用中有着重要的商业价值和应用背景;另一方面,它也是网络不良信息如恶性舆情传播和垃圾信息等的重要载体。因此,针对这些短文本的分类、聚类、话题发现与跟踪已成为数据挖掘和信息安全领域的一个重点[1],正逐步受到人们的关注。

短文本一般只有100个字左右,长度短、噪声数据多,所包含的有用信息非常少,造成可供抽取的信息匮乏,因而对相关研究提出了更高的要求。虽然现有文本分类领域已取得较大成功,却难以在短文本上直接引用[2]。另外,短文本分类同样面临着文本分类领域的两大难题:向量维度高和特征稀疏。

迄今为止,提高短文本分类性能的方法主要分为两种:一种是利用外部数据源或知识库(如WordNet、Wikipedia等)对短文本中的词进行特征扩展[3-4],该方法简单直观,解决了特征词稀疏的问题,并能取得较好的效果,但分类准确率有待提高;另一种是从分类算法[5]入手,通过改进算法提高分类器的性能来改善分类准确率(Precision)和召回率(Recall),但从短文本的特征权重计算的角度出发来解决这个问题的相关文献较少。王细薇等[6]基于特征的共现关系,将改进的短文本特征权重公式和特征扩展算法相结合,来提高短文本的分类精度。文献[7]采用的是另一种思路:先用隐含语义分析对短文本进行降维和去噪,然后用独立成分分析的方法从中抽取最具表现力的特征。这些都是通过与其他方法结合使用来改善短文本分类的效果,不能说明单独使用特征提取方法的有效性。

1 文本的特征权重

为了将文本转换为计算机可以理解的形式,需要将文本表示成向量,以便分析和计算。向量化是文本处理的基础,即为文本中的词赋予一定权重,表示一个词对表征文本意思的重要程度,权重越大,则该词对文本的表征意义越大。只有文档向量很好地保留了原有的文档信息,文本的分类、聚类才可能有令人满意的结果。

信息检索领域的词条权重方法主要分为两类[8]:一类是无监督的 tf(term frequency)、$tf*idf$(term frequency-inverse document frequency)方法,即不考虑词条的类别信息;另一类是有监督的方法,如 $tf*\chi^2$(tf * chi-square)、$tf*ig$(tf * information gain)、$tf*gr$(tf * gain ratio)等[9]。无监督词条权重计算主要借鉴特征选择方法,因为在特征选择的同时,词条也被赋予不同的值来衡量其对文本分类的贡献程度。但这些

收稿日期:2013-02-25;**修回日期**:2013-03-30。 **基金项目**:重庆市自然科学基金资助项目(cstc2011jjA40023)。
作者简介:马雯雯(1986-),女,陕西西安人,硕士,主要研究方向:计算机网络与信息安全; 邓一贵(1971-),男,四川简阳人,高级工程师,博士,主要研究方向:计算机网络与信息安全、移动代理。

特征权重计算方法主要是针对长文本的,并不能直接适用于短文本。

这是由于短文本的一个突出特点是样本分布高度不均衡,即数据集中某些类的样本数远大于其他类,导致小类别文本被淹没在大量其他类别的文档中而难以识别。但在待处理的海量文本数据中,有时系统真正关心的只是一小部分,例如,在网络舆情分析和热门敏感话题发现与跟踪问题中,有价值的数据在现实环境中占的比例很小。称样本少的类为正类,样本多的类为负类。

然而现有的特征权重方法对所有类别都是"平等"看待,实际上,将传统针对长文本的特征权重方法用于短文本时,文本分类的结果更倾向于负类而忽视正类[10],这个问题在短文本分类中表现尤为明显,导致短文本分类的准确率和查全率不高,不能满足实际应用。

2　综合类别特征选择方法

本文将短文本数据集中的当前类视为"正类"(Positive Category,PC),除当前类以外的其他类称为"负类"(Negative Category,NC)。相关元素信息如表 1。

表 1　数据元素表

特征	正类中包含/不包含 t 的样本数	负类中包含/不包含 t 的样本数
t	tp	tN
$\bar{t}$	$\bar{t}P$	$\bar{t}N$

t 和 $\bar{t}$ 分别表示特征词条 t 在样本中出现和不出现的情况,tp 表示正类中包含特征 t 的样本频;$\bar{t}P$ 表示正类中不包含特征 t 的样本频;tN 表示负类中包含特征 t 的样本频;$\bar{t}N$ 表示负类中不包含特征 t 的样本频。

依据普通文本的思想,本文有以下 3 个相关猜想:1)对给定词条 t_i,当包含 t_i 的文本出现次数较多,即文档频率 df(document frequency)较高时,这些文本的类别表达能力较强;2)当 t_i 在正类中出现的频率比它在负类中出现的频率高,说明它具有较好的类别区分能力,称之为反类频 icf(inverse document frequency);3) t_i 在负类中不出现的频率与它在正类中出现频率的比值越大,说明 t_i 和类别的相关性 rf(relevancy frequency)越大。

基于上面的分析,本文提出一种针对短文本的特征选择方法。由于该方法综合考虑了样本在正类和负类中出现的情况,因此,将其命名为综合类别法(Integrate Category,IC),计算公式如下:

$$IC(t_i,c) = df * icf * rf \tag{1}$$

1) df:文档频,给定词条 t_i 的 df 值由 $\log(tp+1)$ 计算得出,表示正类中出现词条 t_i 的文档频率,其中 tp 是正类中包含词条 t_i 的样本数。

2) icf:反类频,由 $idf = \log\left(\frac{|C|}{cf}+1\right)$ 计算得出。$|C|$ 是总类别数,cf 是包含词条 t_i 的类别数。

3) rf:相关性频率,由 $rf = \frac{\bar{t}N}{\bar{t}P+1}$ 计算得出。rf 与 t 在负类中不出现的频率成正比,与 t 在正类中出现的频率成反比。rf 值越大,表示词条 t 和类别的相关性越大。

在正类样本数一定的情况下,数据集中的样本分布大致可以分为以下三种情况:1)负类样本数大于正类,且较为均衡地分布在不同类别中;2)负类样本小于正类,且均衡较为地分布在不同类中;3)负类样本大于正类,但在负类中分布不均衡,即出现在少量类别中。为了更直观地表示上述关系,给出 10 个类别,共包含 100 个短文本的 3 个词条的简单示例,模拟正、负类样本在短文本数据集中的分布情况,见表 2。

表 2　正类(PC)和负类(NC)词条关系比较

term	PC	NC		tp	cf	$df*icf$	$IC(df*icf*rf)$
	Ca_1	Ca_2	Ca_3				
t_1	10	40	40	10	3	7.3183	13.3060
t_2	10	2	2	10	3	7.3183	63.8688
t_3	10	50	0	10	2	8.9425	40.6478

从表 2 可以看出,包含 t_1 和 t_2 的样本在正类和负类中的分布差别很大,其中包含 t_1 的样本大多数都在负类中,而包含 t_2 的样本在负类中的比例较小,但二者的 $df*icf$ 值相同;考虑到 t 在正类和负类中分布的类别相关性 rf 之后,t_2 的分值大幅增加。相比之下,t_1 的分值增幅较小,从而 t_1 与 t_2 的分值差距加大。t_3 在正类中的频数与 t_1、t_2 相同,但它在负类中的分布范围较小,集中出现在一个类中,所以 t_3 的 $df*icf$ 值比 t_1 稍大,而 $df*icf*rf$ 值介于 t_1 与 t_3 之间,符合前文的猜想。

特征 t 对整个语料的全局特征权重由式(2)获得:

$$IC_{\max}(t) = \max_{1\leqslant i\leqslant n}\{IC(t,c_i)\} \tag{2}$$

IC 方法的优点在于它既考虑了词条在单个样本中的分布,又考虑了文本的类别信息;词条在正类和负类中相关性的评估使得特征词在不平衡数据集中更具区分性。同时,取全局特征的最大值有利于得到最有类别区分度的特征。

IC 算法描述如下:

1)对训练语料库中的文档进行预处理:分词、去停用词;

2)计算每个特征词条和类别的 $IC(t,c_i)$;

3)根据第 2)步结果计算所有类别的 $IC_{\max}(t)$;

4)按 IC 值降序排列,取前 M 个作为特征词予以保留,M 是特征空间的维数。本文中 M 值取 750。

3　实验分析

为验证本文短文本的特征权重算法的性能,本文用基于实例学习的 K 最近邻(K-Nearest Neighbor,KNN)分类算法(用 C#3.0 实现)对短文本进行分类实验。KNN 是一种传统的模式识别算法,该算法简单直观,分类准确率高[11],并且加入新的训练文本时不需要重新训练,从而减少了训练时间,被广泛应用于文本自动分类研究。K 表示类别 $C(X)$ 中 X 的最近邻的个数。分别令 K 取 3 ~ 35[12] 的不同奇数值进行测试,并将最优 K 值的结果用于比较。

3.1　实验设置

3.1.1　数据集

多分类问题最终可以分解为二分类问题,所以,本文的重点就转换为提高正、负两个类别中正类的分类性能。目前尚无通用的短文本数据集,本文通过新浪微博提供的 API 抓取微博数据作为短文本语料。选取 4 个字符以上的微博共 7660篇,平均每个文本 35 个字。取关于 2013 年春晚的 50 条微博作为正类,关于火车票的 3650 条微博作为负类。

3.1.2　实验设计

由于数据集样本高度不均衡,因此在进行分类结果的对比实验时,采用 5 折交叉验证(5-fold cross-validation)。即:将数据集平均分为 5 份,取其中 4 份作为训练集、1 份作为测试

集,进行循环测试。得到的结果求其平均值,以消除实验结果的偶然性,并且保证训练集和测试集没有交集。由此获得的对比结果可以认为是可信的。

3.2 评价方法

目前使用较多的分类性能评估指标有精确率(Precision)、查全率(Recall)和 F1 测试值。对样本分布不均衡的短文本分类来说,精确率和查全率会忽视小类别(正类)的影响。微平均和宏平均是两种对分类结果进行全局评价的方法:微平均(Micro-average)是每一个实例文档的性能指标的算术平均值;宏平均(Macro-average)是先计算各个类别的分类结果,再对所有类别求平均值。具体定义如下:

$$\text{微平均 } MicroF1 = \frac{2 \times MicroP \times MicroR}{MicroP + MicroR} \tag{3}$$

$$\text{宏平均 } MacroF1 = \frac{2 \times MacroP \times MacroR}{MacroP + MacroR} \tag{4}$$

其中:*MicroP* 和 *MicroR* 分别表示微平均的精确率和查全率;*MacroP* 和 *MacroR* 分别表示宏平均的精确率和查全率。计算评价指标前,先介绍一下评价指标中涉及到的数据元素,如表 3 所示。

表 3 数据元素对应表

是否检索到	相关	不相关
是	*A*	*B*
否	*C*	*D*

假设文本总数为 N,则:

$$\begin{cases} MicroP = \dfrac{\sum_{k=1}^{N} A_k}{\sum_{k=1}^{N} A_k + \sum_{k=1}^{N} B_k} \\ MicroR = \dfrac{\sum_{k=1}^{N} A_k}{\sum_{k=1}^{N} A_k + \sum_{k=1}^{N} C_k} \end{cases} \tag{5}$$

$$\begin{cases} MacroP = \dfrac{1}{N}\sum_{k=1}^{N} P_k \\ MacroR = \dfrac{1}{N}\sum_{k=1}^{N} R_k \end{cases} ; P_k = \frac{A_k}{A_k + B_k}, R_k = \frac{A_k}{A_k + C_k} \tag{6}$$

3.3 实验结果及分析

考虑到在实际应用中,正类文本所占比例往往很小,为了比较负类文本数据规模对不同特征权重方法的影响,初始取正、负类微博文本各 50 个,再逐步增加负类文本的个数,直至总文本数达到 3 700。图 1 描述了在 *KNN* 分类器下,6 种不同特征权重计算方法的 *MicroF*1 和 *MacroF*1 值。可以看出,随着数据规模的增大,所有特征权重方法的性能值总体呈上升趋势。原因可能是在正类文本数一定的情况下,数据规模较小时,每个类别中的文本不足以“表征”这个类别,导致分类准确率降低。这种状况随着数据规模的增加而有所改善。

图 1 的实验结果还表明,由于宏平均是对每个类别求平均,受小类别影响较大,所以不同特征权重方法的宏平均值总体比微平均值大。

表 3 是重复五次实验,6 种不同特征权重方法的分类结果的最高值的平均值。从中可以看出,有监督词条权重并不总是优于无监督词条权重方法,如 *tf* 和 *tf* * *ig* 表现较差,而在文本分类领域广受欢迎的 *tf* * *idf* 在此次实验中表现并不优异,但是 *df* * *icf* * *rf* 方法分别从词条出现、不出现在正类、负类等角度来衡量词条与类别的关系,在不同数据规模下均能保持较好的分类性能。这是因为,在传统文本分类中,训练集一般具备以下几个特征:类别分布均衡,每个类别中的文档都能较好地代表该类别,类别中各文档在特征空间的排列比较集中。然而,在实际应用中,真实的短文本语料及使用环境往往不满足以上特点,因此,现有特征权重计算方法在短文本上并不能奏效。

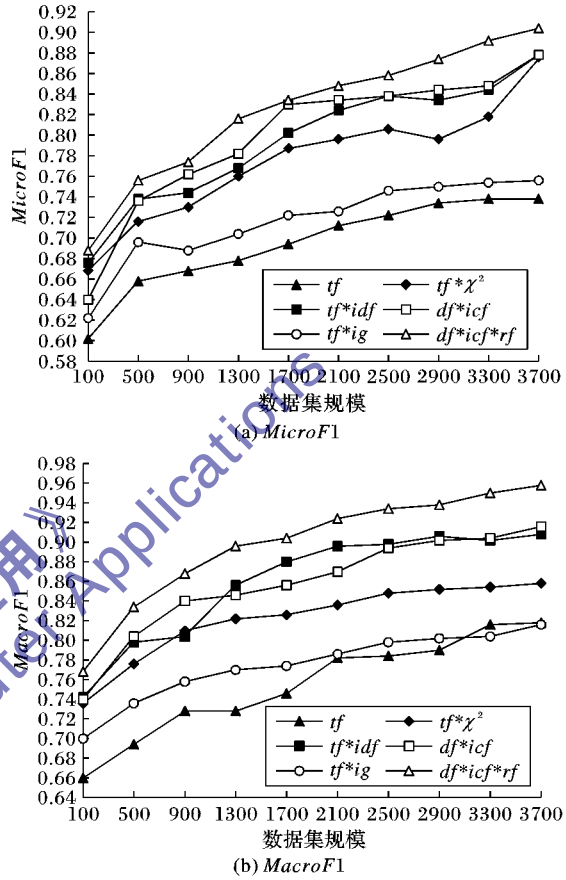

图 1 *KNN* 分类器下不同特征权重计算方法的 *MicroF*1 和 *MacroF*1

表 4 不同特征权重方法的分类结果

性能	*tf*	*tf* * *idf*	*tf* * *ig*	*tf* * χ^2	*df* * *icf*	*df* * *icf* * *rf*
*MicroF*1	0.738 8	0.879 4	0.756 4	0.876 6	0.876 6	0.904 6
*MacroF*1	0.819 6	0.900 8	0.817 0	0.858 4	0.916 2	0.958 4

4 结语

目前短文本分类研究中,特征权重的计算依然沿用传统长文本的方法,但短文本语料往往类别分布不均衡,传统方法不能取得好的分类效果。针对此问题,本文对现有特征权重计算方法作了改进,提出在计算特征权重时,将词条的类别相关性考虑在内,在现实语料环境下测试了该方法的性能。实验结果表明,该方法在一定程度上改善了分类效果;同时发现,短文本的分类性能并没有得到大幅度提高。分析原因是从网络上采集的真实短文本语料,包含大量网络用语和不规范表达,导致系统不能准确识别。下一步工作将引入语义分析的方法,来减小词语表达多样化对分类系统的影响。

参考文献:

[1] DONG H H, HUI S C, HE Y T. Structural analysis of chat messages for topic detection [J]. Online Information Review, 2006, 5(30): 496 – 516.

(下转第 2292 页)

分别抽取两算法成功数中的前 20 个,统计其中各个算法每次所花费的采样节点数。图 5 显示了比较结果,可以看出与 RRT-Connect 算法相比,改进后的算法由于结合任意时间算法,其节点数不断优化减少,最后接近 50 时达到最小值。

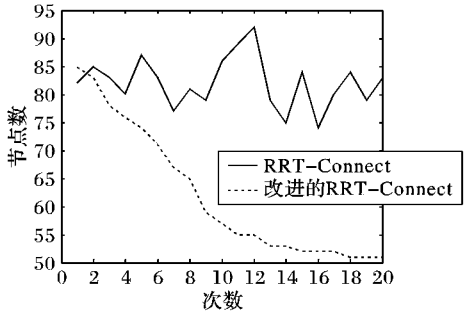

图 5 两算法的搜索节点数比较

5.2 实际的双足机器人实验

如图 6 所示,设定初始的出发点和目标点,控制实验室双足机器人跟踪规划路径。实验结果表明,生成的路径可使双足机器人通过窄道,证明了算法的可行性。

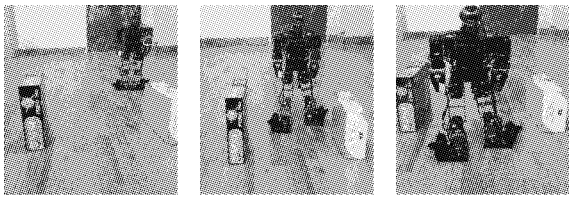

图 6 双足机器人通过窄道

6 结语

本文提出了一种改进的 RRT-Connect 算法,将 RRT-Connect 算法和桥梁检测算法结合起来,解决了 RRT-Connect 算法在狭窄的通道内采样规划路径难的问题;并用任意时间算法进一步改进 RRT-Connect 算法,减少了路径规划的移动代价。但该算法还有一些问题未能得到很好的解决,如桥梁检测算法采样窄道时会产生一些不必要的探索树,这样就降低了算法的性能。这些问题将是下个阶段努力的方向。

参考文献:

[1] 夏泽洋,陈恳,熊璟,等. 仿人机器人运动规划研究进展[J]. 高技术通讯,2007,17(10):1092 - 1099.

[2] 郑慧杰,刘弘,郑向伟. 基于改进群搜索优化算法的群体路径规划方法[J]. 计算机应用,2012,32(8):2223 - 2226.

[3] 张彤,肖南峰. 仿人机器人实时路径规划方法研究[J]. 计算机工程与应用,2009,45(27):4 - 6.

[4] LAVALLE S M, KUFFNER J J, Jr. Rapidly-exploring random trees: progress and prospects [C]// Proceedings of the 4th International Workshop on the Algorithmic Foundations of Robotics: Algorithmic and Computational Robotics: New Directions. Natick, MA, USA: A. K. Peters, 2000: 293 - 308.

[5] LAVALLE S M, KUFFNER J. RRT-Connect: an efficient approach to single-query path planning [C]// Proceedings of the 2000 IEEE International Conference on Robotics & Automation. Piscataway: IEEE, 2000, 4: 995 - 1001.

[6] BRUCE J, VELOSO M. Real-time randomized path planning for robot navigation [C]// Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2002: 2383 - 2388.

[7] FEGUSON D. Replanning with RRTs [C]// Proceeding of the 2006 IEEE International Conference on Robotics & Automation. Piscataway: IEEE, 2006, 5: 1243 - 1248.

[8] ZUCKER M, KUFFNER J J, Jr. Multipartite RRTs for rapid replanning in dynamic environments [C]// Proceeding of the 2007 IEEE International Conference on Robotics & Automation. Piscataway: IEEE, 2007, 4: 1603 - 1609.

[9] ZHEN S, DVAID H, JIANG T T, *et al*. Narrow passage sampling for probabilistic roadmap planning[J]. IEEE Transactions on Robotics, 2005, 21(6): 1105 - 1115.

[10] JEON J H, KARAMAN S, FRAZZOLI E. Anytime computation of time-optimal off-road vehicle maneuvers using the RRT * [C]// Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference. Piscataway: IEEE, 2011: 3276 - 3282.

[11] FERGUSON D, STENTZ A. Anytime RRTs [C]// Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2006: 5369 - 5375.

[12] 夏泽洋,陈恳. 仿人机器人足迹规划建模及算法实现[J]. 机器人,2008,30(3):231 - 237.

[13] (日)梶田秀司. 仿人机器人[M]. 管贻生,译. 北京:清华大学出版社,2007.

[14] 于国晨,刘永信,李晓红. 基于三维线性倒立摆的仿人机器人步态规划[J]. 计算机应用,2012,32(9):2643 - 2647.

[15] 李龙澍,王唯翔,王凡. 基于三维线性倒立摆的双足机器人步态规划[J]. 计算机技术与发展,2011,21(6):66 - 69.

(上接第 2282 页)

[2] DABBLE F, SEBASTIAN F. Supervised term weighting for automated text categorization [C]// SAC '03: Proceedings of the 2003 ACM Symposium on Applied Computing. New York: ACM, 2003. 784 - 788.

[3] ZHANG C, FAN X H, CHEN X N. Hot topic detection on Chinese short text [C]// Advanced Research on Computer Education, Simulation and Modeling: Communications in Computer and Information Science, CCIS 176. Berlin: Springer-Verlag, 2011: 207 - 212.

[4] 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[J]. 计算机应用,2009, 29(3):843 - 845.

[5] 徐易. 基于短文本的分类算法研究[D]. 上海:上海交通大学. 2010.

[6] 王细薇,张凯. 一种改进的基于共现关系的短文本特征扩展算法研究[J]. 河南城建学院学报,2012,21(4):48 - 50.

[7] 胡佳妮,郭军,徐蔚然. 一种基于短文本的独立语义特征抽取算法[C]// 2007 年全国网络与信息安全技术研讨会论文集. [出版地不详]:中国通信学会,2007: 332 - 335.

[8] LAN M, TAN C L, SU J, *et al*. Supervised and traditional term weighting method for automatic text categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(4): 721 - 735.

[9] QUAN X J, LIU W Y. Term weighting schemes for question categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 1009 - 1020.

[10] 林智勇,郝志峰,杨晓伟. 不平衡数据分类的研究现状[J]. 计算机应用研究,2008,25(2):332 - 336.

[11] 崔争艳. 中文短文本分类的相关技术研究[D]. 河南:河南大学,2011

[12] YANG Y M, LIU X. A re-examination of text categorization methods [C]// SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 42 - 49.