

## 基于机器学习的文本分类技术研究进展\*

苏金树<sup>1</sup>, 张博锋<sup>1+</sup>, 徐 昕<sup>1,2</sup>

<sup>1</sup>(国防科学技术大学 计算机学院, 湖南 长沙 410073)

<sup>2</sup>(国防科学技术大学 机电工程与自动化学院, 湖南 长沙 410073)

### Advances in Machine Learning Based Text Categorization

SU Jin-Shu<sup>1</sup>, ZHANG Bo-Feng<sup>1+</sup>, XU Xin<sup>1,2</sup>

<sup>1</sup>(School of Computer, National University of Defense Technology, Changsha 410073, China)

<sup>2</sup>(School of Mechantronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: Phn: +86-731-4513504, E-mail: bfzhang@nudt.edu.cn

Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. *Journal of Software*, 2006,17(9):1848-1859. <http://www.jos.org.cn/1000-9825/17/1848.htm>

**Abstract:** In recent years, there have been extensive studies and rapid progresses in automatic text categorization, which is one of the hotspots and key techniques in the information retrieval and data mining field. Highlighting the state-of-art challenging issues and research trends for content information processing of Internet and other complex applications, this paper presents a survey on the up-to-date development in text categorization based on machine learning, including model, algorithm and evaluation. It is pointed out that problems such as nonlinearity, skewed data distribution, labeling bottleneck, hierarchical categorization, scalability of algorithms and categorization of Web pages are the key problems to the study of text categorization. Possible solutions to these problems are also discussed respectively. Finally, some future directions of research are given.

**Key words:** automatic text categorization; machine learning; dimensionality reduction; kernel method; unlabeled data set; skewed data set; hierarchical categorization; large-scale text categorization; Web page categorization

**摘要:** 文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术,近年来得到了广泛的关注和快速的发展.提出了基于机器学习的文本分类技术所面临的互联网内容信息处理等复杂应用的挑战.从模型、算法和评测等方面对其研究进展进行综述评论.认为非线性、数据集偏斜、标注瓶颈、多层分类、算法的扩展性及Web 页分类等问题是目前文本分类研究的关键问题,并讨论了这些问题可能采取的方法.最后对研究的方向进行了展望.

**关键词:** 自动文本分类;机器学习;降维;核方法;未标注集;偏斜数据集;分级分类;大规模文本分类;Web 页分类

中图法分类号: TP181 文献标识码: A

\* Supported by the National Natural Science Foundation of China under Grant Nos.90604006, 60303012 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20049998027 (国家教育部高校博士点基金)

Received 2005-12-15; Accepted 2006-04-03

随着信息技术的发展,互联网数据及资源呈现海量特征.为了有效地管理和利用这些分布的海量信息,基于内容的信息检索和数据挖掘逐渐成为备受关注的领域.其中,文本分类(text categorization,简称 TC)技术是信息检索和文本挖掘的重要基础,其主要任务是在预先给定的类别标记(label)集合下,根据文本内容判定它的类别.文本分类在自然语言处理与理解、信息组织与管理、内容信息过滤等领域都有着广泛的应用.20世纪90年代逐渐成熟的基于机器学习的文本分类方法,更注重分类器的模型自动挖掘和生成及动态优化能力,在分类效果和灵活性上都比之前基于知识工程和专家系统的文本分类模式有所突破,成为相关领域研究和应用的经典范例<sup>[1]</sup>.

基于机器学习文本分类的基础技术由文本的表示(representation)、分类方法及效果(effectiveness)评估3部分组成.Sebastiani在文献[1]中对文本分类发展历程及当时的技术进行了总结,主要包括:(1)文本关于项(term)或特征的向量空间表示模型(VSM)及特征选择(selection)与特征提取(extraction)两种表示空间降维(dimensionality reduction)策略,讨论了 $\chi^2$ ,IG,MI,OR等用于特征过滤的显著性统计量及项聚类和隐含语义索引(LSI)等特征提取方法;(2)当时较成熟的分类模型方法,即分类器的归纳构造(inductive construction)或模型的挖掘学习过程;(3)分类效果评估指标,如正确率(precision)、召回率(recall)、均衡点(BEP)、 $F_{\beta}$ (常用 $F_1$ )和精度(accuracy)等,以及之前报道的在Reuters等基准语料上的效果参考比较.

然而,互联网中分布传播的海量电子化文本所显现出的种类多样、分布偏斜、关系复杂、更新频繁及标注困难等新的特征,给近年来面向互联网海量信息处理需求的文本分类带来了巨大挑战.文献[1]对分类技术用于解决上述问题时在不同程度上遇到的扩展性差、语料缺乏及精度降低等困难和问题的论述不够,也无法涉及近几年技术的发展以及信息检索、机器学习和数据挖掘等领域权威学术会议及刊物上讨论的重要问题和成果.

本文介绍基于机器学习文本分类技术的最新研究,重点讨论文本分类在互联网信息处理等实际应用中所面临的问题及进展,从相关问题、现状和趋势等方面进行归纳和评论.第1节介绍基础技术的研究动态.第2节讨论现阶段文本分类面向实际应用挑战的主要研究问题及最新进展.最后给出全文的总结和相关技术的展望.

## 1 文本分类基础技术研究动态

近年来,将文本简化为所谓的BOW(bag of words),在特征处理和统计学习算法的基础上获得对文本语义内容及类别信息的估计与预测,已经成为文本分类的标准模式.通过统计理论和语言学(linguistics)两种途径进行的文本表示和分类模型的研究也得到进一步拓宽或发展,相关领域的技术也在文本分类中得到新的应用.

### 1.1 文本表示

VSM仍是文本表示的主要方法,相关研究仍然集中在以什么语义单元作为项及计算项的权重两个问题上.大部分工作仍以词(或 $n$ -gram)作为项,以项的频率为基础计算权重,如 $tf \times idf$ 等<sup>[1]</sup>.值得注意的是,Debole提出了有监督的权重STW,利用项的显著性统计量(如用 $\chi^2$ 等)来平衡其权重<sup>[2]</sup>;文献[3,4]等也使用类似的方法.相对使用 $tf \times idf$ 权重,某些统计量的引入使得SVM及线性分类等方法的分类效果有了不同程度的提高.

除VSM以外,还有人提出基于项概率分布、基于二维视图等模型.Bigi认为,任意文本 $d$ 和类别 $c$ 均可视为所有项的一个概率分布 $P(t_i, d)$ 和 $P(t_i, c), i=1, \dots, |T|$ ( $T$ 为所有项或特征的集合),称为项分布概率表示.通过度量分布间的Kullback-Leibler距离(KLD)相似性的分类方法,获得优于VSM表示下线性方法的效果<sup>[5]</sup>.项分布概率模型本质上仅是在项的权重计算和规格化(normalization)上与VSM不同.Nunzio使用可视的二维表示方法,将所有项的信息压缩到由局部能量和全局能量构成的二维平面上,采用启发式算法进一步计算后,在某些测试集上得到了很高的准确性<sup>[6]</sup>;然而,方法仅是在小数据集上进行了测试,实际应用效果还需要进一步加以验证.

还有一些工作希望通过借鉴自然语言处理的技术考虑被BOW忽略的语义单元间的联系,因此,词义及短语等复杂的项被应用到分类方法的文本表示中.但到目前为止,这些表示方法在分类效果上还没有明显的优势,而且往往需要比较复杂的语言预处理,在分类时影响了分类器的吞吐速度<sup>[7,8]</sup>.到目前为止,非VSM的表示在理论上的合理性及面对实际应用的可扩展性还需要深入验证,适合它们的分类方法比较单一,而且未得到广泛的应用.

## 1.2 表示空间降维

相关研究主要集中在降维的模型算法与比较,特征集与分类效果的关系,以及降维的幅度 3 个方面。

关于降维的模型和算法,很多研究仍按照传统的思路:(1) 用概率统计方法度量并比较项关于类别分布的显著性,如 BNS(bi-normal separation)<sup>[9]</sup>等;(2) 从信息熵角度研究项分布相似性的项聚类方法,如基于全局信息(GI)<sup>[10]</sup>等;(3) 隐含语义分析途径,即通过矩阵的不同分解和化简来获取将向量语义或统计信息向低维空间压缩的线性映射,如差量(differential)LSI<sup>[11,12]</sup>等。一些新颖的研究思路包括:(1) 多步骤或组合的选择方法,即首先用基本的特征选择方法确定初始的特征集,然后以某种标准(如考虑其他项与初始集特征的同现(co-occurrence)等<sup>[13]</sup>)进行特征的补充,或者综合其他因素(如依第 2 种显著性选择标准<sup>[13,14]</sup>或考虑线性分类器系数值大小<sup>[15]</sup>等)进行冗余特征的删减;(2) 尝试借鉴语言学技术进行的研究有从手工输入的特征中学习特征信息<sup>[16]</sup>及基于 WordNet<sup>[17]</sup>的特征提取等方法,但方法所产生的效果都不理想。

必须考虑降维对分类的影响,即关注分类器效果指标随特征数目增加的变化趋势。很多文献中<sup>[9-14,18,19]</sup>比较一致的现象是:合理的降维方法会使多数分类器都呈现出随特征数量增加,效果快速提高并能迅速接近平稳;但若特征数目过大,性能反而可能出现缓慢降低,这表明:降维不仅能大量降低处理开销,而且在很多情况下可以改善分类器的效果。Forman 及 Yang 等人分别从有效性、区分能力及获得最好机会等方面对不同的特征选择方法进行了广泛比较。从结果来看:BNS、 $\chi^2$ 、IG 等统计量及组合方法具有一定的优势;另外,不同分类器倾向于接受不同的特定降维方法<sup>[9,13,18,19]</sup>。常用的特征提取与特征选择算法的效果在不同情况下互有高低或相当<sup>[1,10,20]</sup>。虽然选择方法因为复杂度较低而应用更为广泛,但提取得到的特征更接近文本的语义描述,因此有很大的研究价值。

降维尺度的确定常用经验估算方法,如给定特征数的经验值(PFC)或比例(THR);或者考虑统计量阈值(MVS)或向量空间稀疏性(SPA)等因素。Soucy 给出特征数与文本数成比例(PCS)的方法,并在精度标准下与其他 4 种方法做了比较,得出了 MVS>PCS>SPA>PFC>THR 的结论<sup>[21]</sup>,传统的标准值得重新审视。

## 1.3 机器学习分类方法

分类方法研究的主要目标是提高分类效果,实用的系统还必须兼顾存储和计算能力受限等条件下,学习过程的可扩展性和分类过程的吞吐率(速度)<sup>[22-24]</sup>。近年来,采用多(multiple)分类器集成学习(ensemble learning)的方法被普遍接受;而支持向量机(SVM)仍然代表了单重(single)方法的发展水平。

SVM 的应用是文本分类近年来最重要的进展之一。虽然 SVM 在大数据集上的训练收敛速度较慢,需要大量的存储资源和很高的计算能力<sup>[24-28]</sup>,但它的分隔面模式有效地克服了样本分布、冗余特征以及过拟合(over-fitting)等因素的影响,具有很好的泛化(generalization)能力。有关文献的比较均显示:相对于其他所有方法,SVM 占有效果和稳定性上的优势<sup>[28-32]</sup>。近年来又有很多文献<sup>[1]</sup>中未涉及的一些模型或方法被提出或应用,有的还获得了较好效果,如最大熵模型<sup>[33,34]</sup>、模糊理论<sup>[35,36]</sup>、项概率分布的 KLD 相似性<sup>[5]</sup>、二维文本模型<sup>[6]</sup>以及基于等效半径的方法(SECTILE)<sup>[26]</sup>等(见表 1),但它们仍局限于惯用的相似性度量的分类模式。

Bayes、线性分类、决策树及  $k$ -NN 等方法的能力相对较弱,但它们的模型简单,效率较高,这些方法的修正和改进引起了人们持续的关注。Wu 指出分类器关于数据分布的假设是影响分类效果的重要因素,当模型不适合数据集特点时,性能就可能变得很糟糕。这种模型偏差在弱分类方法中尤为突出,他给出了一种灵活的基于错误矫正的启发式改进策略<sup>[25]</sup>;GIS 方法将样本聚集成不同的实例集(instance set),每个实例集的质心称为推广实例(GI),以 GI 的集合代替样本集合后减少了实例,使得  $k$ -NN 方法的在线速度大为改善,分类效果也有所提高<sup>[37]</sup>;Tsay 利用与 GIS 相反的思路,他增加类别的数目,实质上为原类别选择多个质心,部分地克服了单个质心难以适应样本稀疏的弱点<sup>[38]</sup>;Tan 使用推拉(drag-pushing)策略对 Bayes 和基于质心的方法进行了改进<sup>[39]</sup>;Chakrabarti 的 SIMPL 方法利用 Fisher 线性判别分析将文本表示投影到低维空间后,再进行决策树的构造<sup>[24]</sup>。可以看出,多数分类模型和方法的研究,更侧重在特定测试集上效果基本相当的情况下,获得计算开销上相对 SVM 的优势。

集成学习,也称为多重学习或分类器组合,主要通过决策优化(decision optimization)或覆盖优化(coverage optimization)两种手段将若干弱分类器的能力进行综合,以优化分类系统的总体性能.决策优化对于不同的分类器均采用完整的样本集进行训练,测试时,通过对所有分类器的决策进行投票或评价(如 MV(majority voting),W (weighted)MV 及 WLC (weighted linear combination)等<sup>[1,40]</sup>),确定整个系统输出的类别;Bennett 将特定分类器看作可靠性的指示(reliability indicator);系统利用概率方法综合不同分类器的输出确定最后的决策<sup>[41]</sup>;Xu 和 Zhang 提出一种将 SVM 与 Rocchio 算法进行串行集成方法的思想,即在 Rocchio 算法快速处理全部文本向量后, SVM 对部分感兴趣的类别进行误差校正,用较低的计算代价换取重要类别的精度<sup>[42]</sup>;覆盖优化对同一种学习采用不同的训练子集,形成参数不同的单分类器,这些单分类器决策的某种综合(如 WMV 等)决定每测试样本的分类,如 Bagging 和 Boosting 等方法<sup>[43]</sup>;在 Boosting 方法的迭代过程中,每一轮都关注上一轮的分类错误,用于提升较弱的分类方法并获得了优于 SVM 的结果,AdaBoost.MH 和 AdaBoost.MR 等具体算法都有着广泛的应用<sup>[44]</sup>.

**Table 1** Properties and effectiveness for most of the categorization models or methods

表 1 主要分类模型或方法的性质和效果

Model or method	Examples of algorithm or Implementation	CR	HD	Bi	Best rept eff.	Remark
Probabilistic	Naïve Bayes (NB)				0.773	Easy, highly depend on data distribution
Decision tree (DT)	ID3, C4.5, CART				0.794	Often used as base-lines, relatively weak
Decision rule	DL-ESC, SCAR, Ripper, Swap-1				0.823	
Regression	LLSF, LR, RR <sup>[45]</sup>				0.849	Effective but computing costly
Linear	On-Line				0.822	Weaker but simple and efficient
	Centroid-Based				0.799	
Neural networks	Perceptron, Classi, Nnet				0.838	Not widely used TC
Instance-Based	k-NN				0.856	Inefficient in online classification
SVM	SVM <sup>light</sup> , LibSVM <sup>[46,47]</sup>				0.920	State of arts effectiveness
Ensemble learning	MV, Bagging				N/A	Not widely used and tested yet
	WLC, DCS, ACC, adaboost				0.878	Boosting methods effective and popular
Ensemble learning	STRIVE <sup>[41]</sup>				0.875	Complex in classifier construction
	SVM with Rocchio ensemble <sup>[42]</sup>				+0.019*	*Improvement in a small Chinese corpus
Maximum entropy	Li. KAZAMA <sup>[33,34]</sup>				0.845	Effective but not widely used
Fuzzy	Liu, Widyantoro <sup>[35,36]</sup>				0.892*	*Only accuracy reported
Term prob. distri.	KLD based <sup>[5]</sup>				0.671*	*Better than Rocchio in the same test
Bidimensional	Heuristic approach <sup>[6]</sup>				0.871	Not extensively confirmed
MD and ER based	SECTILE <sup>[26]</sup>				>0.950*	*Only tested in a Chinese corpus, estimated
Wu's Refinement	Rocchio/NB refined <sup>[25]</sup>				0.9/0.926	A little complex in training
Tsay's refinement	Rocchio refined <sup>[38]</sup>				+0.018*	*Improvement, a Chinese corpus
Gener. instance set	GIS-R GIE-W <sup>[37]</sup>				0.860	More efficient than k-NN in testing
Dragpushing	RCC, RNB <sup>[27,39]</sup>				0.859	Easy and computationally efficient
Linear discri. proj.	SIMPL <sup>[24]</sup>				>0.880*	*Estimated form reported data
LS kernel <sup>[48]</sup>	With SVM				0.903	Need expensive matrix processing
Word seq. kernel <sup>[49]</sup>	With SVM				0.915	Complex and time spending in training
String kernel <sup>[50,51]</sup>	With SVM				0.861*	*Estimated form reported data

表 1 中数字角标表示的是: 模型方法; 算法实例或实现; 是否 class ranking 方法(输出测试文本关于每个类的相对形似性参考值或排序); 是否 hard-decision 方法(输出测试文本的类别标记); 是否是二值(binary)方法(方法接受或拒绝当前类,输出±1); (reuters-21578 子集上)报道的最好分类效果(平均的 BEP, F<sub>1</sub> 或精度值,测试条件不同,结果仅供参考); 评注:表 1 的前两部分给出了上述以及文献[1]中涉及的部分方法的主要特征及其在 Reuters-21578 某些子集上(或个别其他语料)上所报道的最好效果指标(平均的 BEP, F<sub>1</sub> 或精度值).由于测试集合和测试条件的差异,指标的数值仅作为方法效果的参考,不能完全作为方法效果间比较的依据.

## 1.4 评估方法

信号检测领域中的 ROC(receiver operating characteristics)曲线,近年来介入到对分类器的效果评估和优化<sup>[41,52-54]</sup>中.对类别  $c$ ,表 2 是其测试结果的邻接表.设  $TPR=TP/(TP+FN)$ , $FRP=FP/(FP+TN)$ ,随着分类器阈值参数的调整,ROC 空间( $TPR,FRP$ )中的曲线不但能直观地反映分类器的性能,曲线下面积 AUC(area under curve)更可以量化分类器接受正例的倾向性.另外,ROC 空间对样本在类别间的分布不敏感,可以反映错误代价(error cost)等指标的变化,具有特别的优势<sup>[52]</sup>.有效地将 ROC 曲线用于分类器的评价、比较及优化,成为近期的一个热点.

**Table 2** The contingency table for category  $c$

表 2 类别  $c$  测试结果邻接表

Category $c$		Expert judgments	
		True	False
Classifier judgments	Positive	$TP$	$FP$
	Negative	$FN$	$TN$

在理论方面,Li 和 Yang 认为关于训练数据的误差及复杂性惩罚使分类器能力间的比较明朗化.通过对常见分类方法进行形式化分析,他们将与分类器获得最优效果条件和标准等价的损失函数(loss function)分为训练损失(training loss)和模型复杂度两部分,从优化的角度给出了一种分类器之间相互比较的方法<sup>[45]</sup>.

方法间的实验比较常在基准语料上进行.Reuters 是重要的基准语料,其中在 Reuters-21578<sup>[55]</sup>版本上进行了最多的测试.常见的语料还包括 OHSUMED,20 Newsgroups,WebKB 及 AP 等<sup>[1,39]</sup>.文献[28]给出了 Reuters-21578 子集的相对难度分析和参考.RCV1(reuters corpus volume 1)是最新整理和发布的较完全的“官方”语料,它改进了之前语料的一些缺点,以适应多层分类、数据偏斜及分类方法扩展性等研究的需要.语料的构建对文本分类研究有着非常重要的促进和参考作用,文献[31]给出了 RCV1 的语料加工技术及部分方法的参考性能.中文分类的公开语料大多处于建设中,特别是经过加工的基准语料相对缺乏,Tan 公开了一个较新的加工中文分类语料 TanCorp 及一些分类方法的参考性能<sup>[39]</sup>.

## 2 主要挑战和研究进展

基于机器学习的文本分类技术经过 20 多年的不断发展,特别是直接从机器学习等领域借鉴最新的研究成果,已能较好地解决大部分具有数据量相对较小、标注比较完整及数据分布相对均匀等特点的问题和应用.但是,自动文本分类技术的大规模应用仍受到很多问题的困扰,如:单是刻画文本间(非线性的)语义联系的问题,都被认为没有很好地得以解决.近年来面临的主要挑战来自于互联网上 Web 等海量信息的处理,其主要特征是:(1) 大规模的类别体系给分类器训练带来扩展性的困难;(2) 建立分类器时所获得的样本相对于海量的未知数据非常有限,模拟样本的空间分布变得困难,这可能带来过拟合(overfitting)及数据偏斜的问题;(3) 文本和类别的更新频繁,在力求对每个类别获得更多的样本时,存在标注瓶颈的问题;(4) 类别间的关系也更加复杂,需要有更好的类别组织方法;(5) Web 文本是一种半结构化(semi-structured)的数据,其结构信息(如链接关系、主题等)可能对分类提供某些帮助.综合来看,我们认为文本分类技术现阶段主要面临非线性、数据集偏斜、标注瓶颈、多层分类、算法的规模扩展性及 Web 页面分类等几个关键的问题.下面主要论述解决这些关键问题可能采取的方法.

### 2.1 非线性问题及核方法

多数文本分类问题的线性可分性<sup>[29]</sup>并未得到理论上的证明,用线性的模型表达复杂的语义内容必然会带来许多误差,非线性的方法仍是处理复杂问题的重要手段.SVM 方法用二元核函数  $K(x,y)$  计算高维空间  $H$  中的内积( $x,y$  是文本表示向量)<sup>[29]</sup>,以应对(降维后的)项空间上不可分的文本分类问题,表达了模型中的非线性变换.SVM 是使用核方法(kernel method)或者核技术(kernel trick)的典型代表,核方法也是 SVM 取得成功的主要因素之一.

在核方法中,通过较复杂的非线性映射 $\phi$ 将项空间的非线性问题变换到高维特征空间 $H$ ,就有可能在 $H$ 中运用线性方法,使问题便于处理和建模;事实上, $\phi$ 的显式构造可能未知或很复杂,但求解过程中却只需利用显式的核函数 $K$ 简单计算 $H$ 中的内积,使得复杂的非线性变换在计算上可行<sup>[56]</sup>。目前,核方法在机器学习领域炙手可热,成为在已有线性算法基础上研究非线性问题的重要途径,如 Zaragoza 将核技巧运用到线性文本分类方法中,此时,仅需将线性决策函数中的内积用核函数 $K$ 进行替换,得到

$$f^-(\mathbf{x}) = \sum_{i=1}^{|\text{Tr}|} \alpha_i^- K(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{|\text{Tr}|} \alpha_i^- \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle,$$

其中: $\text{Tr}$ 是训练样本集合; $\mathbf{x}_i$ 是训练样本的表示( $i=1, \dots, |\text{Tr}|$ ); $\mathbf{x}$ 是待测样本的表示<sup>[57]</sup>。

进一步的研究表明:核方法的效果与核函数的选择密切相关,总是希望它能反映样本相似性的本质。常见的核函数有 RBF, Gauss 及 sigmoid 核等<sup>[29]</sup>。在文本分类中,由于文本空间的特殊性,采用数值核函数获得的分类性能还不能令人满意。因此,新的基于文本语义的核函数成为一个研究重点。文献[48]讨论了基于矩阵分解的隐含语义(LS)核函数;文献[49-51]中使用语法驱动的字符串核及词序列(word sequence)核,直接将文本作为字或词的有序串来计算核;文献[58]讨论了核函数的合成对分类的影响,给出了能够提高分类效果的某些合成条件。核方法的本质是通过核函数引入文本语义相似性的度量,常具有很高的分类准确性(见表 1),但计算开销也较高。

## 2.2 数据集偏斜

通过对机器学习领域的很多研究,发现数据集关于类别的分布往往是偏斜(skewed)或称不均衡的,即类别间样本的数量可能存在数量级的差距,这是导致分类效果很不理想的一个重要因素。在数据偏斜的情况下,样本无法准确反映整个空间的数据分布,分类器容易被大类淹没而忽略小类。在文本分类特别是互联网信息的分类中,大量存在数据偏斜的情况。尤其是在采用二值分类策略时,对某一类,正例的样本可能只占所有样本比例很小的一部分<sup>[59]</sup>。Yang 进行了 SVM, NB 及  $k$ -NN 等方法在样本分布受控情况下的健壮性及分类效果与数据分布之间关系的对比<sup>[30]</sup>,结果表明:SVM 和  $k$ -NN 对样本分布的健壮性要好于 NB 等方法,这印证了 SVM 的泛化性能及 NB 对类别先验概率的依赖性,但所有方法在稀有类别上的准确性均很低。

解决数据偏斜问题的主要对策有:(1) 重取样(re-sampling),可以适当屏蔽大类的信息量或提高小类的分类错误代价<sup>[60]</sup>;(2) 采用新的分类策略,如单类(one-class)SVM 以原点作为未知类别的中心,构造包围训练样本的分隔面,从而将问题转化为等价的受类别分布影响的两类问题<sup>[61]</sup>;文献[62]讨论了在仅有少量正例情况下 SVM 的训练;文献[63]中提出的 NKNN 方法改进了  $k$ -NN 在偏斜数据集上的效果;(3) 采用更好的效果评估方法,如 ROC 曲线或代价曲线等在数据偏斜情况下能够更准确地评估分类器的整体性能<sup>[52,59]</sup>;(4) 在数据偏斜的情况下,特征也很重要,可以分别通过优化特征选择框架或改进特征选择方法获得分类器对小类别特征的重视<sup>[9,64-66]</sup>。目前,所有的方法都还不能将对稀有类别的识别水平(约 0.5 左右或更低的 BEP)整体提高到实际可以接受的程度,相关的研究仍需要进一步的深入。

## 2.3 标注瓶颈

学习算法需要大量的标注样本,但已标注的样本所能提供的信息有限;另一方面,容易获得(如通过互联网)的未标注样本数量相对于标注样本较多,且更接近整个样本空间上的数据分布。提供尽可能多的标注样本需要艰苦而缓慢的手工劳动,制约了整个系统的构建,这就产生了一个标注瓶颈的问题。因此,如何用少量的已标注样本和大量的未标注样本训练出一个好分类器,逐渐引起人们的关注。Nigam 首先利用基于期望最大化(EM)的方法从未标注样本中学习,利用测试样本改进了 Bayes 分类器的分类效果<sup>[67]</sup>;另一种用于未标注文本学习的方法是直推(transductive inference),使得分类器首先通过对已标注样本的学习仅对当前的少量未知样本进行误差最小的预测,而暂不考虑对未来所有实例预期性能的最优性。之后,将这些样本加入到学习过程中来,以改进分类器的效果;Jaochims 使用了直推式支持向量机 TSVM 进行文本分类<sup>[68]</sup>,文献[69]中进行了改进;文献[70]中讨论了直推式 Boosting 文本分类;文献[71,72]采用合作训练(co-training)的方法,使用未标注的样本进行 e-mail 与文本的分类,其思想是从两个视角将样本的特征划分为两个信息充足的子集,分别在两个子集上建立分类器,利用标注样本进行合作学习。另外,文献[73]仅使用正例样本和未标注样本进行学习;文献[74]中利用了 SVM 主动

(active)学习.上述方法在标注样本较少的情况下对提高分类器的性能有很大的帮助(见表3),虽然部分地缓解了标注瓶颈问题,但也以大量迭代为代价.另外,不同的从未标注样本学习方法之间,还没有在同一标准下的比较性工作.

Table 3 Effectiveness of some learning-from-unlabeled methods

表3 一些从未标注样本学习方法的效果

Method/Competitor	Data set	Labeled/ Training set	Unlabeled	Effectiveness/ Effectiveness	Remark
EM/NB <sup>[67]</sup>	20NG	20/20	10000	≈0.36/0.21	Accuracy, estimated from figures
		500/500	10000	≈0.66/0.54	
	WebKB	4/4	2500	≈0.55/0.39	
TSVM/SVM <sup>[68]</sup>		9/9	3957	0.624/0.572	Macro-Average BEP
	Ohsumed	120/120	10000	0.535/0.486	
TBoosting/Boosting <sup>[70]</sup>	RWCP	100/100	1000	0.602/0.479	Macro-Average F1
Co-Train with SVM <sup>[72]</sup>	N/A	9/9	1200	≈0.62/0.77	Accuracy, comparing with startup
Active Learn/Inactive <sup>[74]</sup>	Reuters-21578	22/22	978	≈0.46/0.69	Average BEP

## 2.4 多层分类

通常所讨论的分类问题中,类别间是孤立的,认为它们之间没有相互联系,称之为单层(flat)分类.而在类别较多且关系复杂的情况下,如互联网丰富的Web信息的管理等一大类应用,就需要更好的多层信息组织方式.多层(hierarchical)分类是指多层类别关系下的分类问题<sup>[75-81]</sup>,面对的类别间存在类似于树或有向非循环图的多层分级类别结构,可以更好地支持浏览和查询,也使得部分规模较大的分类问题通过分治的方法得到更好的解决.

多层分类一般采用 big-bang 或自顶向下基于级别两种策略,前者在整个分类过程中使用同一个分类器,即将处于类别树结构上的所有叶节点类别看成平等的类,这本质上还是一种单层分类,不能很好地应用类别间的关系;后者可为不同的级别训练不同的分类器,枝节点的分类器只关心当前的不同分枝<sup>[77]</sup>.Sun 等人讨论了基于类别相似度和类别距离的多层分类效果评估方法,给出了用于说明在不同级别上调度分类器的规范语言<sup>[77-79]</sup>.Ruiz 的博士论文中介绍了早期提出的几种多层分类方法,并给出自己的 HME(hierarchical mixture of expert)模型<sup>[75]</sup>.Huang 等人介绍了用于从 Web 语料中建立多层分类器的 LiveClassifier<sup>[82]</sup>.

多层分类中,类别关系的复杂和相互干扰以及不同类别层次间分类错误的传播都可能对分类器的准确性评估造成影响,仅有 Sun 在文献[83]中考虑了这种影响.对于同一个标签(类别)集合,单层分类设置下的多标签(multi-label,即每个文档可能属于多个类别)分类与多层类别设置下的分类在效果上也需要有一个比较,这些题目目前还都没有得到很深入的研究.

## 2.5 算法的可扩展性

面对互联网海量和复杂的文本内容信息,大规模的文本分类已经成为一个紧迫的需求.大规模的文本分类面对的是庞大的类别数量和训练样本数,这给文本分类带来两个问题:首先,算法的计算时间和存储随类别和样本数量的增长关系;其次,算法是否可以在较大规模下保持有效.目前认为,多层分类是解决算法时间可扩展性的好办法.Yang 的分析和实验表明:分类方法的可扩展性依赖于样本数、类别层次的拓扑结构及类别关于层次的分布,在多层分类中的不同类别层次样本分布满足幂定律(power law)的条件下,SVM 及  $k$ -NN 等算法的复杂性为  $h \cdot O(N_0^{1.5}) \sim h \cdot O(N_0^2)$ .其中,  $h$  和  $N_0$  分别是类别的层次数和首层的文本数,扩展性可以满足对 OHSUMED 全部 14 321 个类别及 233 445 个样本语料的处理<sup>[22]</sup>.然而,Liu 等人指出:大规模的文本分类通常要面对成千上万个类别、较深的类别分级结构及关于类别的偏斜样本分布等状况,目前的算法是否能够有效地扩展到如此大规模的分类依然是一个开放的问题.他首先研究了在大规模分类设置下 SVM 的性能,发现对 Yahoo! Directory 的 24 万多个实际类别下的近 80 万篇文档,其效果远远不能令人满意(几乎所有类的准确率和召回率均下降到 0.3 以下)<sup>[84]</sup>;他的另一个研究结果表明:SVM 大规模的多层分类的计算要远少于单层分类,而对  $k$ -NN 和 NB 的计算量则相反,但三者的分类效果都很差<sup>[23]</sup>.至今,还没有分类方法的准确性随类别规模变化关系的研究,也未见上

述问题的有效解决办法.大规模的文本分类是一个值得开展深入研究的领域.

## 2.6 Web页面分类

传统上所讨论的文本分类一般面向文本内容的本身,在文本的预处理阶段会将文本中所包含的如 HTML 标签(tag)、主题及超链接等结构信息清除<sup>[1]</sup>.然而,在面向互联的信息,特别是 Web 页面的分类中,文本中所包含的这些结构化信息会提供文本归属的丰富信息,如可以考虑测试样本中所含超链接指向文本的类别,借以印证内容分类器的决策<sup>[85,86]</sup>;利用超链接中的锚词(anchor word)或其周围的词语(扩展锚词)作为特征来表达超链接所指向的文本<sup>[86-89]</sup>;利用超链接和 HTML 标签等信息所表现出的结构和拓扑信息来刻画文本间的联系<sup>[86,90]</sup>以及用核函数来表达超链接<sup>[58]</sup>等.这些工作在各自不同的语料上取得的分类效果都较不使用结构信息有所提高.利用结构信息的工作并不都是有效的,如将所链接文本的词当作本地词来处理的方法则降低了分类器精度, Yang 指出,这是由于对话料上超链接与类别间关系模式的假设不当所致<sup>[86,91]</sup>.目前,如何恰当地表示这些结构化信息以及自动地学习它们的统计模式,仍是一个开放的问题.

## 3 总结

本文从文本表示和降维、分类方法以及评估手段等方面总结了基于机器学习的文本分类基础技术近年来的研究进展,重点讨论了近期所面临的一些实际应用需求和数据特点的问题及最新成果,并对将来的一些研究工作进行了展望.

文本分类技术有着广泛的应用,逐渐趋于实用.但随着相关应用的发展及需求的不断提升,仍有很多值得研究的问题,例如:解决大规模分类应用问题的途径和方法;可靠、有效及快速的在线分类;结合自然语言领域的研究,基于语义度量的数据模型和分类方法;缓解样本标注瓶颈以及样本数据分布带来的影响等.随着机器学习和数据挖掘领域理论和技术研究的深入,针对不同实际应用和数据的特征,特别是互联网内容处理和其他一些大规模复杂应用中数据模型、类别规模和性能瓶颈等问题,将成为文本分类相关研究和应用的重点和主要突破的方向.

## References:

- [1] Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys*, 2002,34(1):1-47.
- [2] Debole F, Sebastiani F. Supervised term weighting for automated text categorization. In: Haddad H, George AP, eds. *Proc. of the 18th ACM Symp. on Applied Computing (SAC-03)*. Melbourne: ACM Press, 2003. 784-788.
- [3] Xue D, Sun M. Chinese text categorization based on the binary weighting model with non-binary smoothing. In: Sebastiani F, ed. *Proc. of the 25th European Conf. on Information Retrieval (ECIR-03)*. Pisa: Springer-Verlag, 2003. 408-419.
- [4] Lertnattee V, Theeramunkong T. Effect of term distributions on centroid-based text categorization. *Information Sciences*, 2004, 158(1):89-115.
- [5] Bigi B. Using Kullback-Leibler distance for text categorization. In: Sebastiani F, ed. *Proc. of the 25th European Conf. on Information Retrieval (ECIR-03)*. Pisa: Springer-Verlag, 2003. 305-319.
- [6] Nunzio GMD. A bidimensional view of documents for text categorisation. In: McDonald S, Tait J, eds. *Proc. of the 26th European Conf. on Information Retrieval Research (ECIR-04)*. Sunderland: Springer-Verlag, 2004. 112-126.
- [7] Moschitti A, Basili R. Complex linguistic features for text classification: A comprehensive study. In: McDonald S, Tait J, eds. *Proc. of the 26th European Conf. on Information Retrieval Research (ECIR-04)*. Sunderland: Springer-Verlag, 2004. 181-196.
- [8] Kehagias A, Petridis V, Kaburlasos VG, Fragkou P. A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, 2003,21(3):227-247.
- [9] Forman G. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 2003,3(1):1533-7928.
- [10] Chen W, Chang X, Wang H, Zhu J, Tianshun Y. Automatic word clustering for text categorization using global information. In: Myaeng SH, Zhou M, Wong KF, Zhang H, eds. *Proc. of the Information Retrieval Technology, Asia Information Retrieval Symp. (AIRS 2004)*. Beijing: Springer-Verlag, 2004. 1-11.



- [11] Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier. *Information Processing Letters*, 2003,88(5):203–212.
- [12] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 2005,6(1):37–53.
- [13] Rogati M, Yang Y. High-Performing feature selection for text classification. In: David G, Kalpakis K, Sajda Q, Han D, Len S, eds. *Proc. of the 11th ACM Int'l Conf. on Information and Knowledge Management (CIKM-02)*. McLean: ACM Press, 2002. 659–661.
- [14] Makrehchi M, Kamel MS. Text classification using small number of features. In: Perner P, Imiya A, eds. *Proc. of the 4th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition: (MLDM 2005)*. 2005. 580–589.
- [15] Mladenic D, Brank J, Grobelnik M, Milic-Frayling N. Feature selection using linear classifier weights: Interaction with classification models. In: Jarvelin K, Allan J, Bruza P, Sanderson M, eds. *Proc. of the 27th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-04)*. Sheffield: ACM Press, 2004. 234–241.
- [16] Fernandez J, Montanes E, Diaz I, Ranilla J, Combarro EF. Text categorization by a machine-learning-based term selection. In: Galindo F, Takizawa R, Traummuller R, eds. *Proc. of the Database and Expert Systems Applications (DEXA-04)*. Zaragoza: Springer-Verlag, 2004. 253–262.
- [17] Chua S, Kulathuramaiyer N. Semantic feature selection using WordNet. In: Yao J, Vijay VR, Wang GY, eds. *Proc. of the IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI 2004)*. Beijing: IEEE Computer Society, 2004. 166–172.
- [18] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. In: Fisher DH, ed. *Proc. of the 14th Int'l Conf. on Machine Learning (ICML-97)*. Nashville: Morgan Kaufmann Publishers, 1997. 412–420.
- [19] Gabrilovich E, Markovitch S. Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In: Brodley CE, ed. *Proc. of the 21st Int'l Conf. on Machine Learning (ICML-04)*. Banff: Morgan Kaufmann Publishers, 2004. 41.
- [20] Bekkerman R, Yaniv RE, Tishby N, Winter Y. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 2003,3(2):1183–1208.
- [21] Soucy P, Mineau GW. Feature selection strategies for text categorization. In: Xiang Y, Chaib-Draa B, eds. *Proc. of the 16th Conf. of the Canadian Society for Computational Studies of Intelligence (CSCSI-03)*. Halifax: Springer-Verlag, 2003. 505–509.
- [22] Yang Y, Zhang J, Kisiel B. A scalability analysis of classifiers in text categorization. In: Callan J, Cormack G, Clarke C, Hawking D, Smeaton A, eds. *Proc. of the 26th ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-03)*. Toronto: ACM Press, 2003. 96–103.
- [23] Liu TY, Yang Y, Wan H, Zhou Q, Gao B, Zeng HJ, Chen Z, Ma WY. An experimental study on large-scale web categorization. In: Ellis A, Hagino T, eds. *Proc. of the 14th Int'l World Wide Web Conf (WWW-05)*. Chiba: ACM Press, 2005. 1106–1107.
- [24] Chakrabarti S, Roy S, Soundalgekar M. Fast and accurate text classification via multiple linear discriminant projections. *Int'l Journal on Very Large Data Bases*, 2003,12(2):170–185.
- [25] Wu H, Phang TH, Liu B, Li X. A refinement approach to handling model misfit in text categorization. In: Davis H, Daniel K, Raymond N, eds. *Proc. of the 8th ACM Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD-02)*. Edmonton: ACM Press, 2002. 207–216.
- [26] Wang J, Wang H, Zhang S, Hu Y. A simple and efficient algorithm to classify a large scale of text. *Journal of Computer Research and Development*, 2005,42(1):85–93 (in Chinese with English abstract).
- [27] Tan S, Cheng X, Wang B, Xu H, Ghanem MM, Guo Y. Using dragpushing to refine centroid text classifiers. In: Ricardo ABY, Nivio Z, Gary M, Alistair M, John T, eds. *Proc. of the ACM SIGIR-05*. Salvador: ACM Press, 2005. 653–654.
- [28] Debole F, Sebastiani F. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 2004,56(6):584–596.
- [29] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Nédellec C, Rouveirol C, eds. *Proc. of the 10th European Conf. on Machine Learning (ECML-98)*. Chemnitz: Springer-Verlag, 1998. 137–142.
- [30] Yang Y, Liu X. A re-examination of text categorization methods. In: Gey F, Hearst M, Rong R, eds. *Proc. of the 22nd ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-99)*. Berkeley: ACM Press, 1999. 42–49.

- [31] Lewis DD, Li F, Rose T, Yang Y. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 2004,5(3):361–397.
- [32] Forman G, Cohen I. Learning from little: Comparison of classifiers given little training. In: Jean FB, Floriana E, Fosca G, Dino P, eds. *Proc. of the 8th European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD-04)*. Pisa: Springer-Verlag, 2004. 161–172.
- [33] Kazama J, Tsujii J. Maximum entropy models with inequality constraints: A case study on text categorization. *Machine Learning*, 2005,60(1-3):159–194.
- [34] Li R, Wang J, Chen X, Tao X, Hu Y. Using maximum entropy model for Chinese text categorization. *Journal of Computer Research and Development*, 2005,42(1):94–101 (in Chinese with English abstract).
- [35] Liu WY, Song N. A fuzzy approach to classification of text documents. *Journal of Computer Science and Technology*, 2003,18(5): 640–647.
- [36] Widiantoro DH, Yen J. A fuzzy similarity approach in text classification task. In: *Proc. of the 9th IEEE Int'l Conf. on Fuzzy Systems (Fuzz-IEEE 2000)*, Vol.s 1 and 2. San Antonio: IEEE Computer Society, 2000. 653–658. <http://citeseer.ist.psu.edu/692028.html>
- [37] Lam W, Lai KY. Automatic textual document categorization based on generalized instance sets and a metamodel. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003,25(5):628–633.
- [38] Tsay JJ, Wang JD. Improving linear classifier for Chinese text categorization. *Information Processing and Management*, 2004,40(2): 223–237.
- [39] Tan S, Cheng X, Ghanem MM, Wang B, Xu H. A novel refinement approach for text categorization. In: Otthein H, Hans JS, Norbert F, Abdur C, Wilfried T, eds. *Proc. of the 14th ACM Conf. on Information and Knowledge Management (CIKM-05)*. Bremen: ACM Press, 2005. 469–476.
- [40] Wei YG, Tsay JJ. A study of multiple classifier systems in automated text categorization [PH.D. Thesis]. Chiayi: College of Engineering National Chung Cheng University, 2002.
- [41] Bennett PN, Dumais ST, Horvitz E. The combination of text classifiers using reliability indicators. *Information Retrieval*, 2005,8(1): 67–100.
- [42] Xu X, Zhang B, Zhong Q. Text categorization using SVMs with Rocchio ensemble for internet information classification. In: Lu X, Zhao W, eds. *Proc of the 3rd Int'l Conf on Networking and Mobile Computing (ICNMC-05)*. Springer-Verlag, 2005. 1022–1031.
- [43] Aas K, Eikvil L. Text categorization: A survey. Technical Report, NR 941, Oslo: Norwegian Computing Center, 1999.
- [44] Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 2000,39(2-3):135–168.
- [45] Li F, Yang Y. A loss function analysis for classification methods in text categorization. In: Fawcett T, Mishra N, eds. *Proc. of the ICML 2003*. Washington: AAAI Press, 2003. 472–479.
- [46] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2002. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [47] Joachims T. Making large-scale SVM learning practical. In: Scholkopf B, Burges C, Smola A, eds. *Advances in Kernel Methods—Support Vector Learning*. Cambridge: MIT Press, 1999. 169–184.
- [48] Cristianini N, Shawe-Taylor J, Lodhi H. Latent semantic kernels. In: Brodley C, Danyluk A, eds. *Proc. of the 18th Int'l Conf. on Machine Learning (ICML-01)*. Williams College: Morgan Kaufmann Publishers, 2001. 66–73.
- [49] Cancedda N, Gaussier E, Goutte C, Renders JM. Word sequence kernels. *Journal of Machine Learning Research*, 2003,3(6): 1059–1082.
- [50] Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. *Journal of Machine Learning Research*, 2002,2(2):419–444.
- [51] Leslie C, Kuang R. Fast kernels for inexact string matching. In: Scholkopf B, Warmuth MK, eds. *Proc. of the 16th Annual Conf. on Learning Theory and 7th Kernel Workshop (COLT/Kernel 2003)*. Washington: Springer-Verlag, 2003. 114–128.
- [52] Fawcett T. ROC graphs: Notes and practical considerations for researchers. Technical Report, HPL-2003-4, Palo Alto: HP Laboratories, 2003.
- [53] Yu K, Yu S, Tresp V. Multilabel informed latent semantic indexing. In: *Proc. of the ACM SIGIR-05*. Salvador: ACM Press, 2005. 258–265.

- [54] Lachiche N, Flach P. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning (ICML-01). Washington: AAAI Press, 2003. 416–423.
- [55] Lewis DD. Reuters-21578 text categorization test collection. Distribution 1.0. 1997. <http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt>
- [56] Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks, 2001,12(2):181–202.
- [57] Zaragoza HH, Ralf. The perceptron meets Reuters. In: Proc. of the NIPS 2001 Machine Learning for Text and Images Workshop. 2001. <http://citeseer.ist.psu.edu/456556.html>
- [58] Joachims T, Cristianini N, Shawe-Taylor J. Composite kernels for hypertext categorisation. In: Brodley C, Danyluk A, eds. Proc. of the 18th Int'l Conf. on Machine Learning (ICML-01). Williams College: Morgan Kaufmann Publishers, 2001. 250–257.
- [59] Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. Sigkdd Explorations Newsletters, 2004,6(1):1–6.
- [60] Estabrooks A, Jo TH, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, 2004,20(1):18–36.
- [61] Manevitz LM, Yousef M. One-Class SVMs for document classification. Journal of Machine Learning Research, 2001, 2(1):139–154.
- [62] Brank J, Grobelnik M. Training text classifiers with SVM on very few positive examples. Technical Report, MSR-TR-2003-34, Redmond: Microsoft Research, 2003.
- [63] Tan S. Neighbor-Weighted  $k$ -Nearest neighbor for unbalanced text corpus. Expert Systems with Applications, 2005,28(4):667–671.
- [64] Castillo MDd, Serrano JI. A multistrategy approach for digital text categorization from imbalanced documents. SIGKDD Explorations Newsletter, 2004,6(1):70–79.
- [65] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. SIGKDD Explorations, 2004,6(1):80–89.
- [66] Forman G. A pitfall and solution in multi-class feature selection for text classification. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning (ICML-04). Banff: Morgan Kaufmann Publishers, 2004. 38.
- [67] Nigam K. Using unlabeled data to improve text classification [Ph.D. Thesis]. Pittsburgh: Carnegie Mellon University, 2001.
- [68] Joachims T. Transductive inference for text classification using support vector machines. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML-99). Bled: Morgan Kaufmann Publishers, 1999. 200–209.
- [69] Chen YS, Wang GP, Dong SH. A progressive transductive inference algorithm based on support vector machine. Journal of Software, 2003,14(3):451–460 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/14/451.htm>
- [70] Taira H, Haruno M. Text categorization using transductive boosting. In: Raedt LD, Flach PA, eds. Proc. of the 12th European Conf. on Machine Learning (ECML-01). Freiburg: Springer-Verlag, 2001. 454–465.
- [71] Park SB, Zhang BT. Co-Trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. Information Processing and Management, 2004,40(3):421–439.
- [72] Kiritchenko S, Matwin S. Email classification with co-training. In: Stewart DA, Johnson JH, eds. Proc. of the 2001 Conf. of the Centre for Advanced Studies on Collaborative Research. Toronto: IBM Press, 2001. 8.
- [73] Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Proc. of the 3rd IEEE Int'l Conf. on Data Mining. Melbourne (ICDM-03). IEEE Computer Society, 2003. 179–188.
- [74] Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 2001,2(1):45–66.
- [75] Ruiz M. Combining machine learning and hierarchical structures for text categorization [Ph.D. Thesis]. Ames: Graduate College of University of Iowa, 2001.
- [76] Ruiz M, Srinivasan P. Hierarchical text classification using neural networks. Information Retrieval, 2002,5(1):87–118.
- [77] Sun A, Lim EP, Ng WK. Hierarchical text classification methods and their specification. In: Chan AT, Chan SC, Leong HV, Ng VTY, eds. Cooperative Internet Computing. Dordrecht: Kluwer Academic Publishers, 2003. 236–256.
- [78] Sun A, Lim EP. Hierarchical text classification and evaluation. In: Cercone N, Lin TY, Wu X, eds. Proc. of the 1st IEEE Int'l Conf. on Data Mining (ICDM-01). San Jose: IEEE Computer Society, 2001. 521–528.

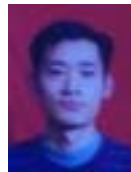
- [79] Sun A, Lim EP, Ng WK. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology*, 2003,54(11):1014–1028.
- [80] Zhou S, Fan Y, Hua J, Yu F, Hu Y. Hierachically classifying Chinese Web documents without dictionary support and segmentation procedure. In: Lu H, Zhou A, eds. *Proc. of the 1st Int'l Conf. on Web-Age Information Management (WAIM-00)*. Shanghai: Springer-Verlag, 2000. 215–226.
- [81] Ceci M, Malerba D. Hierarchical classification of HTML documents with WebClassII. In: Sebastiani F, ed. *Proc. of the 25th European Conf. on Information Retrieval (ECIR-03)*. Pisa: Springer-Verlag, 2003. 57–72.
- [82] Huang CC, Chuang SL, Chien LF. LiveClassifier: Creating hierarchical text classifiers through Web corpora. In: *Proc. of the 13th Int'l World Wide Web Conf.* New York: ACM Press, 2004. 184–192.
- [83] Sun A, Lim EP, Ng WK, Srivastava J. Blocking reduction strategies in hierarchical text classification. *IEEE Trans. on Knowledge and Data Engineering*, 2004,16(10):1305–1308.
- [84] Liu TY, Yang Y, Wan H, Zeng HJ, Chen Z, Ma WY. Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explor. Newsl.*, 2005,7(1):36–43.
- [85] Oh HJ, Myaeng SH, Lee MH. A practical hypertext categorization method using links and incrementally available class information. In: Belkin NJ, Ingwersen P, Leong MK, eds. *Proc. of the 23rd ACM Int'l Conf. on Research and Development in Information Retrieval (SIGIR-00)*. Athens: ACM Press, 2000. 264–271.
- [86] Yang Y, Slattery S, Ghani R. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 2002, 18(2-3):219–241.
- [87] Glover EJ, Tsioutsouliklis K, Lawrence S, Pennock DM, Flake GW. Using web structure for classifying and describing Web pages. In: *Proc. of the Int'l Conf. on the World Wide Web (WWW-2002)*. Honolulu: ACM Press, 2002. 562–569.
- [88] Furnkranz J. Exploiting structural information for text classification on the WWW. In: Hand DJ, Kok JN, Berthold MR, eds. *Proc. of the Advances in Intelligent Data Analysis*. Springer-Verlag, 1999. 487–497.
- [89] Kan MY, Thi HON. Fast Webpage classification using URL features. In: Otthein H, Hans JS, Norbert F, Abdur C, Wilfried T, eds. *Proc. of the 14th ACM Conf. on Information and Knowledge Management (CIKM-05)*. Bremen: ACM Press, 2005. 325–326.
- [90] Shih LK, Karger DR. Using URLs and table layout for Web classification tasks. In: Feldman SI, Uretsky M, Najork M, Wills CE, eds. *Proc. of the 13th Int'l Conf. on the World Wide Web (WWW-2004)*. New York: ACM Press, 2004. 193–202.
- [91] Chakrabarti S, Dom BE, Indyk P. Enhanced hypertext categorization using hyperlinks. In: Haas LM, Tiwary A, eds. *Proc. of the ACM Int'l Conf. on Management of Data (SIGMOD-98)*. Seattle: ACM Press, 1998. 307–318.

#### 附中文参考文献:

- [26] 王健会,王洪伟,申展,胡运发.一种实用高效的文本分类算法. *计算机研究与发展*,2005,42(1):85–93.
- [34] 李陆荣,王健会,陈晓芸,陶晓鹏,胡运发.使用最大熵模型进行中文文本分类. *计算机研究与发展*,2005,42(1):94–101.
- [69] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习. *软件学报*,2003,14(3):451–460. <http://www.jos.org.cn/1000-9825/14/451.htm>



苏金树(1962 - ),男,福建莆田人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络,信息安全.



徐昕(1974 - ),男,博士,副教授,主要研究领域为机器学习,信息安全,自主计算.



张博锋(1978 - ),男,博士生,主要研究领域为信息安全,互联网内容信息分类.