一种基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT

戴新宇[1], 田宝明[1], 周俊生[2], 陈家骏[1]
(1. 南京大学计算机软件新技术国家重点实验室, 江苏南京 210093; 2 南京师范大学计算机科学系, 江苏南京 210097)

摘 要: 本文针对训练数据较少以及在基于图的分类算法中的文本表示问题, 提出了一种基于潜在语义分析技术和直推式谱图算法的文本分类方法 LSASGT, 该方法将潜在语义分析技术和直推式谱图算法这两种基于谱分析理论的技术有机地结合在一起, 对所有训练数据和测试数据进行统一建模, 挖掘数据中潜在的多种结构信息. LSASGT 引入潜在语义分析技术用于构造文本图表示模型, 在能够反映人的分类标准的潜在语义特征空间中, 描述文本之间的语义相关性; 基于这样的文本表示, 利用半监督的直推式谱图算法进行文本分类. 在基准英文文本分类数据集 Reuters21578 和中文文本分类数据集 TanCorp 上的实验结果表明, 本文给出的 LSASGT 文本分类方法获得了较好的分类结果.

关键词: 直推式谱图; 潜在语义分析; 文本分类; 图构造
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2008) 08-1626-05

LSASGT: an Approach to Text Categorization Based on Latent Semantic Analysis and Spectral Graph Transducer

DAI Xin-yu[1], TIAN Bao-ming[1], ZHOU Jun-sheng[2], CHEN Jia-jun[1]
(1. *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210093, *China*;
2. *Department of Computer Science, Nanjing Normal University, Nanjing, Jiangsu* 210097, *China*)

Abstract: In this paper, an approach to text categorization named LSASGT is proposed, which combines Latent Semantic Analysis(LSA) with Spectral Graph Transducer(SGT) for the task of text categorization. For both LSA and SGT are originated from spectral analysis theory which can mine some latent structure information within all training and testing data, we integrate them tightly in one model. Firstly, according to the characteristic of natural language, LSA is used to represent documents in a latent semantic space in which documents and their semantic relationships can be reflected more pertinently. Then we construct a graph based on the latent concept-based subspace, and apply the graph into SGT for text categorization. The experiments demonstrate that LSASGT can improve classification performance on both English and Chinese datasets of Reuters21578 and TanCorp-12.

Key words: spectral graph transducer; latent semantic analysis; text categorization; graph construction

1 引言

文本分类, 是将一篇文本自动地分配到预先定义的某个类别中. 统计学习方法是当前进行文本分类的主流方法. 代表性的有监督的分类方法包括 K 近邻分类算法(K Nearest Neighbor, KNN)、朴素贝叶斯方法(Naïve Bayes, NB) 和支持向量机算法(Support Vector Machine, SVM)[1,7]. 由于带类别标注的文本数据规模往往较小, 多种半监督学习算法被广泛地用于文本分类并表现出不错的效果, 如 EM 算法[2]、co-training 算法[3,4]以及直推式 SVM[5](Transductive SVM, TSVM). 近年来, 鉴于数据的图表示模型具有较强的描述能力, 以及各种图算法较好的谱图理论基础[8,15], 基于图的直推式半监督学习方法也得到了广泛的研究. 这其中, 直推式谱图算法(Spectral Graph Transducer, SGT) 最具代表性, 并表现出比其它直推式算法如 TSVM 具有更好的分类性能[12].

将基于图的分类算法用于文本分类时, 首先需要利用图表示模型描述文本数据, 其中, 图的结点代表文本, 两结点间边的权重反映文本之间的相似性. 数据的图表示对于基于图的分类算法非常重要. 对于文本数据来说, 一个"好的"图表示模型应该能够在恰当的维度特征空间中准确地描述出文本数据之间的内在语义关系, 这

收稿日期: 2008-01-17; 修回日期: 2008-05-04
基金项目: 国家 863 高科技研究发展计划(No. 2006AA01Z143, No. 2006AA01Z139); 国家自然科学基金(No. 60673043); 江苏省自然科学基金(No. BK2006117)

符合人对文本类别划分的标准. 然而, 传统基于词汇特征的文档向量表示模型往往并不能准确地描述文档之间的语义相关性[6]. 而潜在语义分析(Latent Semantic Analysis, LSA)技术则[10,11]可以在一个低维潜在概念语义空间重新描述自然语言文本, 从而可以更好地反映文本之间的语义相似度.

综合考虑文本分类中的特点及问题, 本文提出一种基于潜在语义分析和直推式谱图算法的文本分类方法(LSASGT). 该方法基于所有训练数据和测试数据, 对潜在语义分析和直推式谱图算法进行统一建模, 首先, 利用LSA 技术将自然语言文本映射到潜在语义空间, 构造更能反映文本间语义相似关系的图表示模型; 然后, 在基于潜在语义的文本图表示模型基础上, 利用直推式谱图分类算法进行文本分类. 该方法一方面利用潜在语义分析技术描述文本间的语义相关性; 另一方面吸收了直推式谱图算法的半监督特性, 在训练数据较少的情况下, 对基于潜在语义空间表示的文本进行类别的划分. 在中英文文本分类数据集上的实验结果表明, LSASGT 方法表现出较好的分类性能.

2　直推式谱图算法

针对带标注训练数据不足的问题, Vapnik 首先提出直推式学习思想[5]. 在直推式学习的框架下, 一种基于数据图表示的直推式谱图算法[12]被提出. 该方法将分类任务转换为图的分割, 分割的目标如公式(1)所示:

在公式(1)中, G^+ 和 G^- 分别表示数据图 G 中包含所有正例($y=+1$)和反例($y=-1$)的子图, A_{ij} 表示图中两数据点之间的相似度.

$$\min_y \frac{cut(G^+, G^-)}{|\{i: y_i=+1\}||\{i: y_i=-1\}|} \qquad (1)$$

其中: $cut(G^+, G^-)=\sum_{i\in G^+, j\in G^-} A_{ij}$

$y_i=+1$, 如果 i 是训练集中的正例

$y_i=-1$, 如果 i 是训练集中的反例

$\boldsymbol{y}\in\{+1,-1\}^n$, n 是训练集和测试集样本数之和

公式(1)中的分母表示所有正例和反例个数的乘积, 以保证求得一个更为平衡的分割[12]. 对于公式(1)最佳分割向量 $\boldsymbol{y}$ 的求解是一个 NP-hard 问题[9]. 谱图方法可以给出近似的全局最优解[12], 它用公式(2)来描述带约束的最小分割问题的求解目标:

$$\min_{\boldsymbol{w}} \boldsymbol{w}^{\mathrm{T}} D\boldsymbol{w} + c(V\boldsymbol{w}-\gamma)^{\mathrm{T}} C(V\boldsymbol{w}-\gamma) \qquad (2)$$

其中 $\boldsymbol{D}$ 和 $\boldsymbol{V}$ 是表示初始数据图 G 的相邻矩阵 $\boldsymbol{A}$(由 A_{ij} 构成)所对应的拉普拉斯矩阵的特征值对角矩阵和特征向量矩阵. γ 是基于训练集中正例和反例比例的约束向量. C 和 c 是两个给定的相关经验值常量参数. 最终的图分割求解变成了对于利用谱图方法对公式(2)最佳分割向量 $\boldsymbol{w}$ 的求解[12].

直推式谱图算法有效地解决了直推式学习算法中普遍存在的退化分割问题, 在语音识别、手写数字识别、文本分类等应用中取得了不错的效果[12].

3　文本的潜在语义分析

本文引言中已经指出, 文本数据的图表示模型, 对基于图的分类算法最终的分类效果具有重要的影响.

传统的文本表示来源于向量空间模型. 在向量空间模型中, 文本被表示成带权向量 $\boldsymbol{d}$. 所有的文本向量, 则构成了一个高维的词条-文档矩阵 $\boldsymbol{X}=(\boldsymbol{d}_1, \cdots, \boldsymbol{d}_i, \cdots, \boldsymbol{d}_n)$.

在传统的基于词特征空间的向量空间模型中, 文本间的相似性取决于文档间的词汇特征的共现率. 然而, 在自然语言文本中普遍存在着同义词和多义词的现象, 多义词的现象导致两篇包含很多共有词汇的文本并不一定很相似, 而同义词现象导致相似文本间可能并没有太多的共现词汇, 这样, 基于词特征空间的文档表示有时并不能很好地反映文档之间的语义相关性. 针对自然语言的这一特点, 潜在语义分析(LSA)[10]利用奇异值分解(Singular Value Decomposition, SVD)技术[16]对高维的词条-文档矩阵进行处理, 在潜在的语义结构子空间中重新表示文本及文本间的相似度.

经过奇异值分解, 可以发现自然语言潜在的语义领域知识[11]. 在潜在的语义特征空间重新描述文本的图表示模型, 文本被描述为维潜在语义空间中的点, 点间边的权重进而能够描述文本之间的潜在语义相关性[10].

4　基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT

潜在语义分析技术和直推式谱图算法均来源于谱分析技术[13,14], 它们利用矩阵的特征值和特征向量信息发现矩阵所描述数据中潜在的结构信息. 在潜在语义分析中, 对词条-文档矩阵进行奇异值分解, 在潜在的语义空间重新表示文本. 而在直推式谱图算法中, 是对文档-文档的相似度矩阵进行谱分析, 利用谱图方法找到文档间的最佳分割. 基于上述考虑, 同时结合文本分类的特点和问题, 本文将这两种方法有机地结合起来, 提出一种基于潜在语义分析的直推式谱图文本分类方法——LSASGT. 下面给出 LSASGT 文本分类方法中需要考虑的一些问题, 以及 LSASGT 方法的详细描述.

首先, 我们为所有标注文本(训练集)和非标注文本(测试集)构造完整的词条-文档矩阵. 这样做的好处是, 一方面, 大量的非标注文本可以为构造潜在的语义子空

间提供更多的语义背景知识,从而更准确地反映标注和非标注文本之间的相似关系;另一方面,由于直推式谱图算法需要构造一个包含所有训练和测试数据的相邻矩阵,这样,就使得 LSA 可以和直推式谱图算法在训练数据和测试数据上进行统一建模,从而进一步保证了本文提出的 LSASGT 分类方法的可行性.

另外,由于原始的 SGT 算法是一个 2 元分类算法,我们采用一对多的(one-versus-all)方法使之可以用于多元分类,为每一个类构造一个分类器,在对某一类别进行分类时,将训练集中属于该类别的样本标注为+1,其它类别的样本标注为-1.

下面是 LSASGT 方法的整个流程:

(1)预处理文本分类数据.对英文文本,进行词根化和删除停用词处理,对中文文本,则进行分词处理;

(2)为所有的训练和测试文本数据构造一个 $t*d$ 的词条-文档矩阵 $\boldsymbol{X}$,t 是作为特征的词条数,d 是包含所有训练集和测试集的总文档数;

(3)对矩阵 $\boldsymbol{X}$ 进行奇异值分解 $[\boldsymbol{T}_0, \boldsymbol{S}_0, \boldsymbol{D}_0] = SVD(\boldsymbol{X})$,选择并保留 $\boldsymbol{S}_0$ 中最大的 k 个奇异值,并得到相应的左右奇异矩阵 $\boldsymbol{T}_k$,$\boldsymbol{S}_k$ 和 $\boldsymbol{D}_k$;

(4)计算经过潜在语义分析后的文本相邻矩阵 $\boldsymbol{A} = \boldsymbol{D}_k\boldsymbol{S}_k^2\boldsymbol{D}_k^T$,其中,$A_{ij} = sim(d_i, d_j)$ 反映两个文本之间的相似度.构造矩阵 $\boldsymbol{A}$ 的 k 近邻矩阵 A',此时 $\boldsymbol{A}'_{ij} = \left\{ \begin{array}{ll} \frac{A_{ij}}{\sum_{d_k \in knn(d_i)} A_{ik}}, & \text{if } d_j \in knn(d_i) \\ 0, & \text{else} \end{array} \right\}$,对 $\boldsymbol{A}'$ 进行对称化处理,得到最终的用于直推式谱图算法的文本相似度矩阵 $\boldsymbol{A}^{sgt} = \boldsymbol{A}' + \boldsymbol{A}'^{\mathrm{T}}$;

(5)计算 $\boldsymbol{A}^{sgt}$ 的拉普拉斯矩阵 $\boldsymbol{L} = \boldsymbol{B}^{-1}(\boldsymbol{B} - \boldsymbol{A}^{sgt})$,其中 $\boldsymbol{B}_{ii} = \sum_j \boldsymbol{A}_{ij}^{sgt}$,$\boldsymbol{B}$ 为对角线矩阵.计算 $\boldsymbol{L}$ 的第 2 小到第 $d+1$ 小的特征值组成的对角矩阵 $\boldsymbol{D}$ 和对应的特征向量矩阵 $\boldsymbol{V}$;

(6)利用谱图方法对公式(2)的最佳分割向量 $\boldsymbol{w}$ 进行求解:

(a)首先估计约束向量 γ,基于训练集 l,$\gamma_+ = \sqrt{\frac{(l+)}{(l-)}}$,$\gamma_- = \sqrt{\frac{(l-)}{(l+)}}$,其中 $(l+)$ 和 $(l-)$ 分别是训练集中正例和反例样本的数目,对应训练集中的正例(反例)文本,$\gamma_i = \gamma_+ (\gamma_-)$,对应测试集中的文本,$\gamma_i = 0$;

(b)计算 $\boldsymbol{G} = (\boldsymbol{D} + c\boldsymbol{V}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{V})$,$\boldsymbol{b} = c\boldsymbol{V}^{\mathrm{T}}\boldsymbol{C}\gamma$,求得矩阵 $\begin{bmatrix} \boldsymbol{G} & -\boldsymbol{I} \\ -\frac{1}{n}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}} & \boldsymbol{G} \end{bmatrix}$ 的最小特征值 λ^*;

(c)求得 $\boldsymbol{w}^* = (\boldsymbol{G} - \lambda^*\boldsymbol{I})^{-1}\boldsymbol{b}$;

(7)计算 $\boldsymbol{z}^* = \boldsymbol{W}\boldsymbol{w}^*$,得到所有文本的预测向量;

(8)通过取符号函数 $y_i = \mathrm{sign}(z_i - \Theta)$ 求得样本的标记是否属于该类别(+1 或-1),通常设分割阈值 $\Theta = \frac{1}{2}(\hat{\gamma}_+ + \hat{\gamma}_-)$.

5 实验

为验证本文提出的 LSASGT 文本分类方法的有效性,本文在中英文数据集上进行了文本分类实验.

5.1 实验数据集及评估指标

选取 Reuters21578[18] 文档数最多的 10 个类别的文档,通过"ModApte Split"方法[14] 划分训练集和测试集,其中训练集 2532 篇文本,测试集文本共 6454 篇.经过对文本进行词根化和去除停用词的预处理,共有 21726 个词条特征项,词特征维权重用 tf* idf 表示.

TanCorp-12[17] 是一个中文的文本分类数据集.14150 篇文档分属 12 个类别.取数据集中每个类别的前 20% 的文档(按文件名排序)作为训练集,剩下的 80% 作为测试集.该数据集中共有 72603 个中文特征词汇.特征权重同样用 tf* idf 表示.

实验结果评估采用在多类别文本分类中常用的 F-值(F-Measure)、宏平均(Macro-F)和微平均(Micro-F)值作为分类性能评估指标.

5.2 实验和分析

本文分别使用有监督的 SVM 算法、直推式谱图算法(SGT)以及本文提出的基于潜在语义分析的直推式谱图分类方法(LSASGT)分别对实验数据做了对比实验.其中,SVM 分类器我们采用 LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/),参数采用该工具包中的默认设置.SGT^{light} (http://sgt.joachims.org) 提供了直推式谱图分类算法,在进行 SGT 算法进行分类时,需要设定一些参数,本实验算法依据文[12]中给出的推荐设置,将 k(构造 k 近邻邻接矩阵中的参数)和 d(选取拉普拉斯矩阵特征值的参数)分别设定为 800 和 100,其它参数则采用 SGT^{light} 中的默认设置.对于潜在语义分析,使用 SVDLIBC 的工具包(http://tedlab.mit.edu:16080/~dr/SVDLIBC/),使用该工具时,需将词条-文档矩阵转换成 st(sparse text)格式可以高效地进行 SVD 处理.

表 1 和表 2 分别给出了在 Reuters21578 和中文 TanCorp-12 数据集上,每个类别的分类结果 F 值以及衡量整体分类性能的微平均和宏平均值.从表 1 可以看出,对于有足够多的训练样本的类别,如 Earn 和 Acq 类别,三种方法均能取得较好的分类效果.而对于训练文档数相对较少的类别,如 Ship 和 Corn 类别,SVM 表现较差.训练数据的稀疏对半监督的 SGT 算法影响并不大,它在所有类别的分类中,均取得了较好的效果.而本文提

出的 LSASGT 文本分类方法则进一步提高了分类性能. 在总共 10 个类别中, 7 个类别的分类结果表现得最好, 反映分类整体性能的微平均和宏平均值也均有较大幅度的提高. 从表 2 可以看出, 就单个类别来说, LSASGT 在 TanCorp-12 数据集的某些类别上并没有显著的提高, 但是衡量整体分类性能的微平均和宏平均值均明显高于其它两种算法.

表 1　Reuters21578 数据集对比实验结果

类别	F 值		
	SVM	SGT	LSASGT
Earn(1080/2861)	88.18%	93.11%	97.57%
Acq(718/1648)	94.39%	74.47%	93.38%
Money-fx(179/534)	55.21%	80.10%	84.33%
Crude(186/385)	7.5%	91.50%	90.66%
Grain(148/428)	55.86%	92.71%	92.98%
Trade(116/367)	66.88%	83.94%	88.56%
Interest(131/345)	17.86%	75.84%	71.25%
Wheat(71/211)	45.95%	62.57%	64.37%
Ship(87.191)	12.29%	75.93%	82.76%
Com(56/180)	12.75%	65.04%	72.82%
Macro-F1	45.70%	79.52%	83.87%
Micro-F1	77.17%	84.05%	91.65%

表 2　TanCorp-12 数据集对比实验结果

类别	F 值		
	SVM	SGT	LSASGT
人才(121/487)	74.00%	90.60%	89.47%
体育(561/2244)	98.69%	99.14%	98.65%
卫生(281/1125)	87.12%	83.60%	85.98%
地域(30/120)	15.30%	69.27%	75.48%
娱乐(300/1200)	73.23%	82.01%	76.23%
房产(187/748)	92.90%	97.52%	94.92%
教育(161/647)	78.82%	75.31%	76.24%
汽车(118/472)	67.24%	75.82%	89.30%
电脑(588/2355)	91.52%	95.41%	97.68%
科技(208/832)	69.98%	59.29%	67.96%
艺术(109/437)	49.81%	57.13%	50.78%
经济(163/656)	66.26%	58.00%	71.12%
Macro-F1	72.07%	78.59%	81.02%
Micro-F1	83.46%	84.92%	87.06%

在表 1 和表 2 提供的实验结果中, 在对原始数据进行 SVD 处理时, 均设定选取前 50 大的奇异值($k=50$). 而对于潜在语义分析, k 值的选取非常重要. 为验证 LSA 产生的潜在语义子空间的维度大小对于分类性能的影响, 在进行 SVD 处理时, 我们选取 $k=10$、20、30、40、50、100、150、200 的潜在语义子空间维度, 在 Reuters21578 和 TanCorp-12 数据集上分别做了实验, 实验结果如图 1 所示. 从图 1 中可以看出, 在中英文数据集上, 当 $k=50$ 时, 基本已经得到较好的结果. 综合效率和性能的考虑, k 设为 50 是一个比较合理的选择.

实验表明, 本文提出的 LSASGT 方法, 在潜在语义分析选择合适的维度的条件下, 在英文和中文文本分类数据集上均表现出较好的分类性能. 从表 1 和表 2 来看, LSASGT 方法在英文数据集大幅提高了分类的结果, 而在中文数据集上提高的幅度相对较小. 我们的分析是, 除了受数据集的规模、质量等方面的因素影响之外, 不同语种文本的语义对词汇的依赖程度可能不同, 从而导致潜在语义分析得到的潜在语义结构信息对文本语义的反映程度有所不同. 论文所提分类方法与不同语言的相关性分析将是进一步研究的问题.

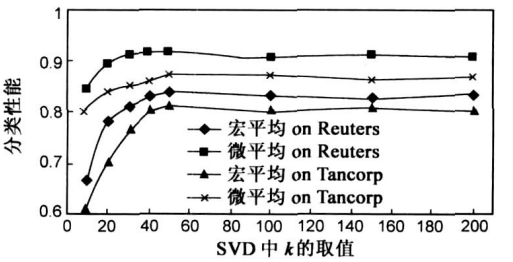

图 1　SVD 中 k 的取值对分类性能的影响

6　总结

本文针对自然语言文本的特点, 提出一种基于潜在语义分析的直推式谱图文本分类方法, 该方法在潜在的低维语义空间中描述文本数据及其之间的语义相关性, 在此基础上利用直推式谱图算法进行文本分类. 实验表明, LSIASGT 方法在基准英文文本分类数据集 Reuters21578 和中文数据集 TanCorp-12 上均取得了较好的效果.

从本质上看, 潜在语义分析技术(LSA)和谱图方法均来源于谱分析技术. 在 LSA 中, 对词条-文档矩阵进行分析, 在潜在语义空间重新表示文本. 而在直推式谱图算法中, 是对文档-文档的相似度矩阵进行分析, 找到文档间的最佳分割. 从本文的实验来看, 这两种基于谱分析的技术在文本分类中, 具有互补的作用. 本文作者将对两者之间的内在联系作进一步的研究.

参考文献:

[1] Fabrizio, Cosiglio, et al. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-47.

[2] Nigam K, McCallum K, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39: 103-134.

[3] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[A]. Proceedings of the 11th Annual Conference on Computational Learning Theory[C]. Madison, 1998. 92-100.

[4] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training[A]. Proceedings of 9th International Conference on Information and Knowledge Management[C]. Virginia, 2000. 86-93.

[5] Vapnik V. Statistical Learning Theory[M]. Wiley, 1998.

[6] Nello C, Shawe J, Lodhi H. Latent semantci kernels[J]. Journal of Intelligent Systems, 2002, 18(2/3): 127- 152.

[7] 李蓉, 叶世伟, 史忠植. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. 电子学报, 2002, 30(5): 745-748.
Li R, Ye S W, Shi Z Z. SVM-KNN classifier——a new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2002, 30(5): 745- 748. (in Chinese).

[8] Sepandar K, Klein D, et al. Spectral learning[A]. Proceedings of International Joint Conference of Artificial Intelligence[C]. Cambridge MA: Mit Press, 2003. 561- 566.

[9] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on PAMI, 2000, 22(8): 888- 905.

[10] Deerwester S, Dumais S, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391- 407.

[11] Kumar CA, Srinivas S. Latent semantic indexing using eigenvalue analysis for efficient information retrieval[J]. International Journal of Applied Mathematics and Computer Science, 2006, 16(4): 551- 558.

[12] Joachims T. Transductive learning via spectral graph partitioning[A]. Proceedings of the 20th International Conference on Machine Learning[C]. Menlo Park, CA: AAAI Press, 2003. 290- 297.

[13] Hagen L, Kahng A. New spectral methods for ratio cut partitioning and clustering[J]. IEEE Transactions on CAD, 1992, 11(9): 1074- 1085.

[14] Bast H, Majumdar D. Why spectral retrieval works[A]. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM Press, 2005. 11- 18.

[15] Fan RK Chung. Spectral Graph Theory[M]. American Mathematical Society, 1997.

[16] 谭铁牛, 刘瑞祯. 基于奇异值分解的数字图像水印方法[J]. 电子学报, 2001, 29(2): 168- 171.
Tan T N, Liu R Z. SVD Based Digital Watermarking Method[J]. Acta Electronica Sinica, 2001, 29(2): 168- 171. (in Chinese).

[17] 谭松波, 王月粉. 中文文本分类语料库 TanCorpV1.0[DB/OL]. http://www.searchforum.org.cn/tansongbo/corpus.htm., 2008-01-17

[18] Lewis D. Reuters-21578 text categorization test collections[DB/OL]. http://www.daviddlewis.com/resources/testcollections/reuters21578/, 1997.

作者简介:

戴新宇 男, 1979 年生于江苏盱眙, 南京大学计算机科学与技术系讲师, 博士. 主要研究方向包括机器翻译、信息检索、信息抽取.
Email: mtlab@nju.edu.cn

田宝明 男, 1984 年生于山西太谷, 南京大学计算机科学与技术系硕士研究生. 主要研究方向包括文本分类, 信息检索等.