基于 SVM 和纹理的笔迹鉴别方法

刘 宏[1,2] 李锦涛[1] 崔国勤[1] 唐 胜[1,2]
[1](中国科学院计算技术研究所数字化室 北京 100080)
[2](中国科学院研究生院 北京 100039)

摘 要 针对与书写内容无关的笔迹,提出利用快速 Gabor 小波提取笔迹图像的整体纹理特征、用支持向量机(SVM)进行训练和识别的方法. SVM 是解决两类问题的算法,而笔迹鉴别是一个多类问题,通过"一对多"的方法将多类问题转化为两类问题. 在 87 人笔迹库上的实验结果表明,文中基于 SVM 和纹理的笔迹鉴别方法是有效的.

关键词 笔迹鉴别;文本独立;Gabor 小波;SVM 分类器
中图法分类号 TP391

Writer Identification Using Support Vector Machines and Texture Feature

Liu Hong[1,2] Li Jintao[1] Cui Guoqin[1] Tang Sheng[1,2]
[1](*Digital Technology Laboratory*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100080)
[2](*Graduate School of the Chinese Academy of Sciences*, *Beijing* 100039)

Abstract A method is presented for text independent writer identification using multi-class SVM and texture feature of the whole script. 2D Gabor filter is applied to feature extraction, after that multi-class SVM is used to train and test the data. In the WS_ICT writer script gallery collected from 87 persons, the proposed algorithm obtained competitive results.

Key words writer identification; text independence; 2D Gabor filter; support vector machines

1 引 言

生物识别技术利用人的生物特征进行身份鉴别. 生物特征可分为生理特征和行为特征[1]. 笔迹是一种相对稳定的行为特征,广泛应用于政府部门、贸易和法律等领域. 文检专家可以鉴别出笔迹的真伪,但是用计算机来描述笔迹的特征,自动地进行笔迹识别却是一个难题[2].

计算机笔迹鉴别根据采样方式可分为在线和离线两类[2]. 在线笔迹鉴别可以采集书写的序列以及压力和速度等信息,离线笔迹鉴别的对象是写在纸上的字符. 根据与文本内容的关系笔迹鉴别可分为文本依存和文本独立两类[3]:文本依存的方法是针对相同的文字进行笔迹鉴别,可以提取更多的特征;文本独立的方法是从大量字符集中提取特征,与书写内容无关,难度较大,但由于其克服了对样本的依赖性,应用更广泛. 本文主要针对离线的文本独立的笔迹鉴别进行研究.

图像的纹理特征提取方法主要有 Fourier 变换、

原稿收到日期:2002-12-25;修改稿收到日期:2003-04-23. 本课题得到中国科学院计算技术研究所领域前沿青年基金(20026180-16)和国家"八六三"高技术研究发展计划(2001AA114190)资助.刘 宏,女,1975 年生,博士研究生,助理研究员,主要研究方向为图像处理、模式识别. 李锦涛,男,1962 年生,博士,研究员,博士生导师,主要研究方向为多媒体技术、多模式人机接口、虚拟现实. 崔国勤,男,1966 年生,硕士,副研究员,硕士生导师,主要研究方向为模式识别、计算机视觉、最优化理论和方法. 唐 胜,男,1973 年生,博士研究生,主要研究方向为图像处理、模式识别.

多通道 Gabor 滤波器[4-6]、灰度共生矩阵[7]、小波变换[8]等. 其中多通道 Gabor 滤波器被广泛用于纹理分析[9]、人脸识别[4]、印刷体汉字识别[10]和笔迹鉴别[5]等领域,是一种比较成熟的纹理分析方法,本文利用此方法提取笔迹样本的纹理特征.

目前,笔迹鉴别一般采用传统的欧氏距离或最近邻(KNN)等算法[2],它们建立在经验风险最小化基础上,在训练样本足够多的情况下才能保证分类效果. 但在笔迹鉴别应用中,得到的训练样本往往很有限. 为了解决这一问题,本文提出了用支持向量机(Support Vector Machines, SVM)对笔迹特征进行训练和识别的方法.

2 算法流程

图 1 所示为本文方法的算法流程. (1)对笔迹图像进行预处理,得到归一化后的训练和测试样本;(2)对样本进行特征提取,采用快速 Gabor 变换,将多个 Gabor 通道得到的小波系数进一步处理得到笔迹特征数据;(3)用多类 SVM 训练算法对训练数据进行学习,得到多个分类面,并分别存入多个模式库;(4)在测试阶段,将测试数据送入多类 SVM 分类器进行识别,最相似的样本对应的书写人作为候选人.

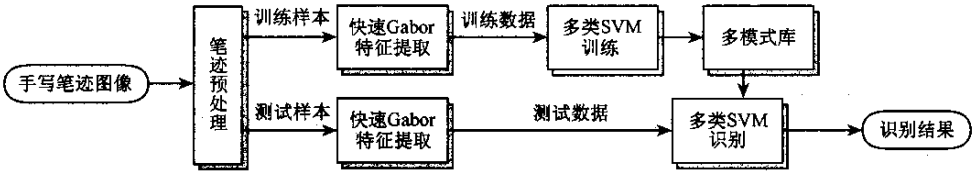


图 1 算法流程图

3 笔迹特征的表示与提取

3.1 笔迹图像的获取

我们采集了 87 个人的笔迹,每人一份,每份笔迹含 300～400 个手写汉字,对内容、书写用笔和纸张均没有限制. 将笔迹文稿按 100 dpi 分辨率 RGB 格式扫描,以 BMP 图像格式存储,组织成 87 人的笔迹库,简记作 WS_ICT. 100 dpi 的分辨率能很好地反映笔迹的整体书写风格,并且在执行速度和存储方面都较优. 而基于特征字的笔迹鉴别,扫描分辨率一般取 300 dpi[3].

3.2 笔迹图像的预处理

为了进行特征提取和分类,需要对笔迹图像进行预处理和归一化操作,以消除书写背景、书写用笔以及不同的行间距、字间距等因素对笔迹特征提取

a 原始笔迹图像(红色格线) b 去除背景,二值化后的图像 c 用于训练的图像 d 用于测试的图像 e 384×384的图像(训练)

f 得到的3个不同书写人的笔迹样本

图 2 笔迹图像的预处理

的影响. 首先利用 RGB 颜色信息进行聚类,根据不同的分布特征设定相应的阈值去除背景和格线,并将图像二值化,如图 2 a, 2 b 所示;然后对图像进行水平方向的投影,投影为零的部分对应行间的空白,不为零的部分对应笔迹行(假设笔迹行是水平的). 采用同样的方法对每一行进行垂直方向的投影[5-6],以压缩掉行间和字间的空白. 将得到的笔迹行一分为二,一份用于训练(如图 2 c 所示),另一份用于测试(如图 2 d 所示),以保证训练数据和测试数据完全不同(如果有足够的笔迹页,可以拿一页做训练,另外的做测试). 对于训练数据以统一的行高 16 个像素对每行进行归一化处理,粘贴到 384×384 的图像中,不足 384 行宽的用该行前部的文字块补充,不足 384 列高的将前面几行文字交叉粘贴,以尽量避免重复(如图 2 e 所示). 将得到的图像平均分割成 9 个 128×128 的图像块用于训练. 用同样的方法将测试数据粘贴成 256×256 的图像,再分割成 4 个 128×128 的图像块用于测试. 图 2 f 所示为通过上述方法得到的三个不同书写人的笔迹样本.

3.3　笔迹特征的提取

每个人都有自己的书写风格,如图 2 f 所示,从整体笔迹图像看,因为它们含有不同的纹理特征,所以很多文献把笔迹识别问题转化成纹理识别问题处理[3,5-6]. 由于笔迹纹理大多具有较强的方向性和频谱特性,即与空间位置信息相关,因此需要采用信号的时频分析.

3.3.1　Gabor 函数

Gabor 函数早在 1946 年由 Gabor 提出,是小波基中的一种,属于一种窗口 Fourier 变换,其窗口是 Gauss 函数[4]. Gabor 函数克服了 Fourier 变换只能反映信号整体特征的缺陷,能同时在时域和频域中较好地兼顾对信号分析的分辨率要求. 20 世纪 80 年代, Daugman 将 Gabor 函数用于计算机视觉领域[11],取得了较好的结果. 二维 Gabor 函数具有方向选择性和带通性,能比较精确地提取图像的局部纹理特征,其滤波函数[4]为

$$g_{vu}(x,y)=\frac{k^2}{\sigma^2}\exp\left(-\frac{k^2(x^2+y^2)}{2\sigma^2}\right)\cdot$$

$$\left[\exp(\mathrm{i}k\cdot(x,y))-\exp(-\frac{\sigma^2}{2})\right],$$

其中,

$$k=\begin{pmatrix}k_x\\k_y\end{pmatrix}=\begin{pmatrix}k_v\cos\varphi_u\\k_v\sin\varphi_u\end{pmatrix},$$

$$k_v=2^{-\frac{v+2}{2}}\pi,\ \varphi_u=u\frac{\pi}{K}.$$

v 的取值决定了 Gabor 滤波的波长,u 的取值表示 Gabor 核函数的方向,K 表示总的方向数. 参数 σ/k 决定了高斯窗口的大小,本文取 $\sigma=\sqrt{2}\pi$. 通过选取不同的参数 v 和 u,可以得到一组 Gabor 滤波器,形成一组非正交基[9]. 用这组基展开信号,可以得到信号在不同的频率和相位下的频域信息.

3.3.2　Gabor 滤波器的设计

Gabor 小波基是非正交的,说明用所有通道(v 和 u 的不同组合)对信号进行滤波得到的结果存在冗余信息[9]. 文献[9]已经证明,就纹理识别而言,参数选择没必要覆盖整个频域,可以通过选择不同的方向和频率参数得到有效的局部纹理特征. 规范的汉字笔画有很强的方向性,主要分布在水平、垂直、对角线方向. 针对手写笔迹,本文经实验比较,取 4 个频率($v=0,\cdots,3$),8 个相位(方向,即 $K=8$,$u=0,\cdots,7$),共 32 个 Gabor 核函数.

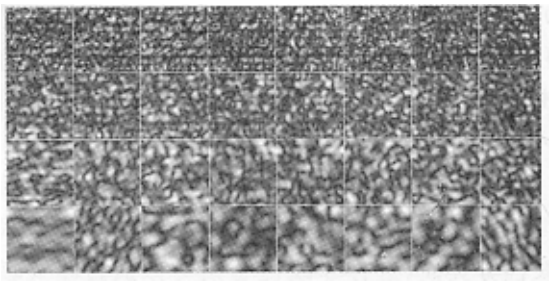
3.3.3　笔迹纹理特征的表示

给定一幅图像 $f(x,y)$,其 Gabor 小波变换[4]可以定义为

$$W_{vu}(x,y)=\iint f(x_1,y_1)\cdot$$

$$g_{vu}(x-x_1,y-y_1)\mathrm{d}x_1\mathrm{d}y_1.$$

实际上是用每个 Gabor 核函数 g_{vu} 与样本进行卷积,如图 3 所示,得到原图像在不同频率和相位下的 32 组小波系数. 分别提取 32 组数据的均值和方差作为特征数据,这样每个样本得到一个多维的特征向量用于分类.



经过幅度增强,注意在原始数据上提取特征

图 3　Gabor 变换后的图像

在空间域中直接进行卷积操作很费时,根据卷积定理[12],时域中的卷积相当于频域中的相乘,即

$$\mathcal{F}\{f(x,y)*g(x,y)\}=F(s,t)G(s,t).$$

针对这一特性,我们采用了快速 Gabor 变换,卷积操作可转化为

$$f(x,y)*g(x,y)=\mathcal{F}^{-1}\{F(s,t)G(s,t)\}.$$

即先利用快速 Fourier 变换将样本和 Gabor 核函数转换到频域,相乘后,再进行 Fourier 反变换变换回

时域. 需要注意的是,Gabor 核函数由实部和虚部构成,需要分开处理,然后取二者的模作为最后的结果. 快速 Gabor 变换的使用极大地提高了算法的效率.

4 多类 SVM 分类器

SVM 是在统计学习理论的 VC 维理论和结构风险最小原理基础上,根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷的方法,目的是通过对有限样本的学习,得到最好的推广能力[13-15]. SVM 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势.

SVM 根据两类样本数据寻求最优分类面,分类面不但能使两类样本无错误地分开,而且要使两类的分类间隔最大[13]. 针对线性可分的 N 个训练数据(x_i, y_i), $i=1,\cdots,N$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1,+1\}$. 其中,y_i 是样本 x_i 所属类的标志,d 是样本的维数. 我们将第 3 节提取的特征向量和其对应的类别属性作为 SVM 的输入. 最优准则是使两类的分类间隔最大,可以通过解带约束的二次规划 QP 问题得到最优线性分类面,对应的分类函数为

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{N} y_i \alpha_i (x \cdot x_i) + b\right\} \qquad (1)$$

其中,$f(x)$的符号用于判定 x 的类别. $(x \cdot x_i)$是内积操作,决策超平面等价于寻找所有非零 α_i,每个非零 α_i 对应的数据 x_i 即是最优超平面的一个支持向量,b 可以由任一支持向量带入式(1)求得. 由式(1)可知,SVM 计算的复杂程度取决于支持向量的数目,而不是特征空间的维数. 对于线性不可分的情况,可以引入松弛因子,在求最优解的限制条件中加入对松弛因子的惩罚函数[13].

当样本非线性可分时,通过非线性变换将输入特征空间变换到高维特征空间,在高维空间中求最优线性分类面,这种非线性变换可以通过定义适当的内积函数(核函数)实现. 此时,分类函数变为

$$f(x) = \operatorname{sgn}\left\{\sum_{i=1}^{N} y_i \alpha_i K(x, x_i) + b\right\}.$$

SVM 的构造主要依赖核函数 $K(x, x_i)$ 的选择,不同的核函数导致不同的分类结果. 实验中采用了下面三种核函数,并依据实验结果进行了比较.

Gauss 函数 $K(x, x') = \exp\left(-\dfrac{|x-x'|^2}{2\sigma^2}\right)$.

多项式函数 $K(x, x') = (x \cdot x' + 1)^p$.

Sigmoid 函数 $K(x, x') = \tanh(kx \cdot x' - \delta)$.

由以上论述可知,SVM 通过寻求最优分类面来解决两类分类问题,但笔迹鉴别是多类问题,我们可以通过构造一个决策函数,将多类问题转化为两类问题处理[14]. 这种转化一般有两种模式:一是‘一对多’,将 n 类问题转化成 n 个两类问题,即第 k 个分类器由第 k 类样本和剩下不属于第 k 类的所有样本构成两类;另一个是‘一对一’,将每一类和其余的 $n-1$类中的每一类构成两类,得到 $n(n-1)/2$ 个分类面. 本文采用‘一对多’的方法,首先得到 n 个分类函数,然后构造解决多类问题的分类器算法

$$L(x) = \arg\max_k \{f_k(x)\}, k = 1,\cdots,n;$$

其中,

$$f_k(x) = \sum_{j=1}^{k_l} y_{kj} \alpha_{kj} K(x_{kj}, x) + b_k;$$

x 是输入的测试数据,$f_k(x)$是第 k 类分类函数. 本文在识别时将测试数据相对于每个分类器的分类函数值 $f_k(x)$按从大到小的顺序排序,提供排在前几位的函数值对应的书写人作为候选人,而不是仅提供值最大的.

5 实验结果和分析

我们在 WS_ICT 笔迹库上对训练样本($87 \times 9 = 783$ 个)和测试样本($87 \times 4 = 348$ 个)进行 32 个通道的快速 Gabor 变换,然后将每个通道得到的数据的均值(或均值和方差,用于最小距离分类器和 KNN)作为特征向量,这样,每个样本得到一个 32 维(或 64 维)的特征向量,分别将得到的训练数据和测试数据存入训练和测试笔迹库. 训练阶段将 783 个特征向量(32 维,仅取均值)送到多类 SVM 分类器中学习,得到 87 个最优分类面. 测试时将 348 个特征向量逐个送入训练好的多类 SVM 分类器进行识别,选取前 8 个(样本类别数的 1/10)函数值对应的书写人作为候选者.

为了与多类 SVM 分类器的识别结果进行对比,同时给出了用均值和方差作为特征向量、用加权欧氏距离作为相似度函数[5-6](针对汉字的特点,我们对相似度函数进行了改进,不同的方向和频率分别选取不同的权重)、分类器采用最小距离分类器和最近邻分类器进行测试的结果. 其中,最小距离分类器将每类 9 个训练数据的均值作为该类代表点,以消除噪声对笔迹识别的影响;而最近邻分类器将所有样本作为代表点,它们都采用 348 个测试数据. 从表 1 和图 4 的结果可以看出,最近邻方法的识别率最低,特别是当径向基函数和二次多项式函数作

为 SVM 核函数时,采用 SVM 分类器的识别结果要好于传统的方法. 而采用 Sigmoid 函数作为核函数的 SVM 分类器的识别率最低,说明核函数的选取对识别结果有重要影响.

表 1　不同的分类器用于笔迹鉴别的识别结果　%

分类器	参数选择	前 1 位	前 2 位	前 3 位	前 4 位	前 5 位	前 6 位	前 7 位	前 8 位
多类 SVM	径向基函数	66.67	81.32	87.07	89.08	90.52	90.80	91.09	93.10
	Sigmoid 函数	56.03	68.68	75.00	76.72	79.60	81.90	81.90	83.05
	二次多项式	67.53	79.89	84.48	88.02	89.09	91.09	91.38	93.10
最小距离	取均值和方差	69.54	77.01	80.75	80.75	87.64	88.79	89.66	92.24
K 近邻法	KNN($K=1$)	68.97	77.87	81.03	82.18	83.05	84.77	85.34	85.63

实验中我们用每人 4 个测试数据分别做测试,但实际应用中,一幅测试样本往往不足以判断书写人的身份,笔迹专家往往也是由测试人的多幅测试样本进行身份鉴别的. 所以我们对以上的测试模式进行改进,用每类 4 个测试数据的均值作为该类的测试数据,即 87 个测试数据,结果如图 5 和表 2 所示. 采用径向基函数或二次多项式函数作为 SVM 核函数的识别效果相当,比最小距离分类器要好,其中,采用径向基函数($\sigma=10$)为核函数的多类 SVM 分类器的识别率最高能达到 97.70%.

本文算法在 Intel Pentium Ⅳ, 1.7 GHz 的 PC 机上用 VC++ 6.0 实现,每一样本的预处理和特征提取时间约为 1.5～2.2 s,每一样本的测试时间为 0.3～0.5 s.

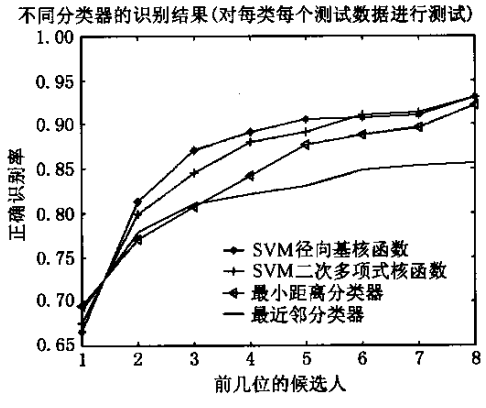


图 4　不同分类器识别结果的比较

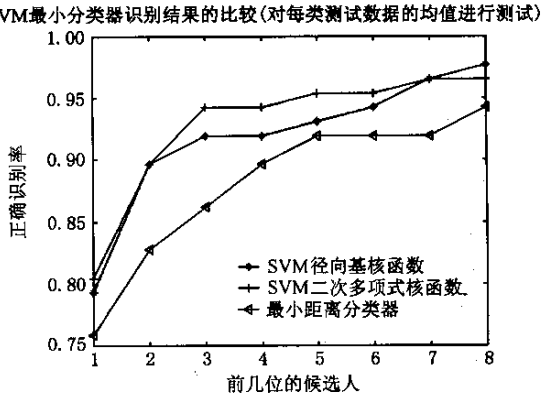


图 5　对每类 4 个测试数据取均值后进行测试的结果

表 2　用每人测试数据的均值代替单个样本的识别结果　%

分类器	参数选择	前 1 位	前 2 位	前 3 位	前 4 位	前 5 位	前 6 位	前 7 位	前 8 位
多类 SVM	径向基函数	79.31	89.66	91.95	91.95	93.10	94.25	96.55	97.70
	二次多项式	80.46	89.66	94.25	94.25	95.40	95.40	96.55	96.55
最小距离	取均值和方差	75.86	82.76	86.21	89.66	91.95	91.95	91.95	94.25

6　结束语

笔迹鉴别大多采用传统的统计模式识别方法,这些方法在样本数足够多时其性能才有保障. 但实际应用中,训练样本的数目往往是有限的,以前的很多方法达不到理想的推广效果. 本文提出了用多类 SVM 和 Gabor 纹理进行笔迹鉴别的方法. 在 87 人的笔迹库上,识别率最高能达到 97.7%.

在实际应用中,笔迹专家除了考虑整体风格外,还通过分析特征字、特征笔画以及运笔等来获取更细微的特征. 下一步将研究基于特征字的笔迹鉴别以及如何将整体特征和局部特征有效地结合,以进一步提高笔迹鉴别的准确性和可靠性.

参　考　文　献

[1] Jain A K, Bolle R, Pankanti S. Biometrics Personal Identification in Networked Society [M]. Boston: Kluwer Academic Publishers, 1999

[2] Plamondon R, Lorette G. Automatic signature verification and writer identification—The state of the art [J]. Pattern Recognition, 1989, 22(2): 107~131

[3] Liu chenglin, Dai Ruwei, Liu Yingjian. Modified Wigner distribution and application to writer identification [J]. Journal of Computers, 1997, 20(11): 1018~1023(in Chinese)
(刘成林, 戴汝为, 刘迎建. 简化的 Wigner 分布及其在笔迹鉴别中的应用 [J]. 计算机学报, 1997, 20(11): 1018~1023)

[4] Laurenz Wiskott, Fellous Jean-Marc, Norbert Kruger, *et al*. Face recognition by elastic graph matching [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 775~779

[5] Said H E S, Tan T N, Baker K D. Personal identification based on handwriting [J]. Pattern Recognition, 2000, 33(1): 149~160

[6] Zhu Y, Wang Y, Tan T N. Biometric personal identification based on handwriting [A]. In: Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, 2000. 801~804

[7] Ohanian P P, Dubes R C. Performance evaluation for four classes of textural features [J]. Pattern Recognition, 1992, 25(6): 819~833

[8] Mallat S G. A theory for multiresolution signal decomposition: The wavelet representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7): 674~693

[9] Bovic A C, Clark M, Geisler W S. Multichannel texture analysis using localized spatial filters [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(1): 55~73

[10] Hamamoto Y, *et al*. Recognition of handprinted Chinese characters using Gabor features [A]. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 1995. 819~823

[11] Daugman J G. Complete discrete 2-D Gabor transform by neural network for image analysis and compression [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1988, 36(7): 1169~1179

[12] Kenneth R Castleman. Digital Image Processing [M]. Beijing: Publishing House of Electronics Industry, 1998. 154~155(in Chinese)
([美]Kenneth R Castleman. 朱志刚, 林学 , 石定机, 等译. 数字图像处理 [M]. 北京: 电子工业出版社, 1998. 154~155)

[13] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer, 1995

[14] Weston J, Watkins C. Multi-class support vector machines [R]. Royal Holloway: University of London, CSD-TR-98-04, 1998

[15] Christopher J C Burges. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121~167

中国计算机学会
第一届全国 Web 信息系统及其应用会议(WISA2004)
征文通知

由中国计算机学会电子政务与办公自动化专业委员会主办,华中科技大学、东南大学与武汉大学承办的'Web 信息系统及其应用(WISA04)'会议将于 2004 年 10 月中旬在湖北省武汉市召开.

一、征文范围(包括但不限于)

Web 信息挖掘与检索	语义 Web 与智能 Web	Web 与网格计算
Web 信息系统环境与基础	Web 应用框架和体系结构	Web 与信息系统安全性
Web 信息系统开发工具	Web 系统度量与分析技术	Web 站点逆向工程与维护技术
Web 测试与 Web 应用的质量保证	多媒体数据管理	工作流模型
组件与中间件技术	代理技术及信息管理	自动文本索引与分类技术
决策支持与分析技术	电子政务与电子商务框架及应用	电子政务与办公自动化发展现状与趋势

二、来稿要求

① 本次会议只接受 Email 投稿.

② 所有来稿请先投中文稿(优秀论文待审后再译成英文稿),一般不超过 6000 字,为便于出版论文集,来稿须附中英文摘要、关键词、资助基金与主要参考文献,注明作者及主要联系人姓名、工作单位、详细通信地址(包括 Email 地址)与作者简介.

三、联系地址

① 大会网址:http://cse.seu.edu.cn/pcegoa/confs/wisa2004/

② 论文投稿地址:(210096)南京 东南大学计算机科学与工程系 徐宝文 周晓宇(zhouxy@seu.edu.cn)

③ 研讨会与专题讨论会联系地址:(100872)北京 中国人民大学信息学院 孟小峰(xfmeng@ruc.edu.cn)

四、重要日期

① 征文截至日期:2004 年 4 月 30 日

② 录用通知发出日期:2004 年 6 月 10 日