基于多视图典型相关分析的垃圾网页检测*

高爽^{1,2}, 张化祥^{1,2†}, 房晓南^{1,2}

(1. 山东师范大学信息科学与工程学院, 济南 250014; 2. 山东省分布式计算机软件新技术重点实验室, 济南 250014)

摘要: 首先将垃圾网页特征分为两个不同的视图,即基于内容特征的视图和基于链接特征的视图,利用典型相关分析及其相关改进方法进行特征提取,生成两组新的特征;再对新生成的两视图特征采用不同组合方式产生单视图数据,并用这组数据作为训练数据构建分类算法。实验结果表明,将垃圾网页看成两视图数据,并应用多视图典型相关分析技术,可有效提高垃圾网页的识别精度。

关键词: 垃圾网页检测; 典型相关分析; 多视图分类; 特征抽取

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2013)03-810-04

doi:10.3969/j.issn.1001-3695.2013.03.043

Multi-view canonical correlation analysis based Web spam detection

GAO Shuang^{1,2}, ZHANG Hua-xiang^{1,2†}, FANG Xiao-nan^{1,2}

(1. School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China; 2. Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250014, China)

Abstract: Firstly this paper divided the features of Web spam pages into the content feature based view and the link feature based view. And it employed canonical correlation analysis and promotion methods for feature extraction to generate two new feature sets for each Web page. Then it implemented different combinations of the two new feature sets of Web pages to produce a single view for Web pages, which used to construct classification algorithms. Experimental results show that considering Web page data as two view data and applying multi-view canonical correlation analysis techniques can effectively improve the recognition accuracy of Web spam.

Key words: Web spam detection; canonical correlation analysis(CCA); multi-view classification; feature extraction

0 引言

网络已经成为最主要的信息来源,人们通过信息检索(IR)查找相关信息。用户在使用搜索引擎检索信息时,往往只选取排名靠前的几条结果,某些网站为了达到商业目的,利用一些专门为其他网站提供提高排名服务的盈利组织(如SEO)^[1]误导和欺骗用户,严重影响了用户获取有用信息。由此可见,对垃圾网页进行有效检测是一个亟待解决的问题。

目前垃圾网页主要分为基于内容的垃圾网页、基于链接的垃圾网页和页面隐藏^[2]。基于内容的垃圾网页通过恶意制作网页的内容(如插入与流行的查询条件相关的关键词)来提高搜索排名,通常使用基于语言模型^[3,4]的方法进行检测。基于链接的垃圾网页通过创建一个联系紧密的链接结构影响排名算法,常用基于信任传播的方法进行检测^[5]。页面隐藏通过向搜索引擎和用户发送内容不同的网页来实现,可以利用比较索引版本和用户实际看到的网页页面的方法进行检测。

上述垃圾网页检测方法都是利用网页的内容特征或链接特征进行分类,也有一些研究方法提出将两部分特征同时用于分类^[1-3]:将两部分特征进行了串行合并^[1,2],但简单合并不适用于直接用于分类;在基于内容的垃圾网页检测方法中加入链接

信息进行分类^[3],但实际用于分类的仍然是所链接网页的内容特征。现有垃圾网页数据集(如 webspam2007、webspam2006)通常提供基于内容和基于链接的两类特征,这两方面的特征不同,所以将垃圾网页检测作为一个多视图问题考虑更加恰当。

将垃圾网页作为一个多视图问题,可以利用典型相关分析方法对基于内容和基于链接的特征进行特征提取。典型相关分析(canonical correlation analysis, CCA)^[6]通过最大化两组特征间的相关性,找出两个线性变换的投影矩阵,使变换后的两组数据相关性最大化。CCA已被应用于人脸识别、手写体识别等多个领域。CCA变换为线性变换,很多问题需要非线性变换,因此通过在CCA方法中加入核函数,提出了核典型相关分析(kernel canonical correlation analysis, KCCA)^[7]。局部保持典型相关分析(locality preserving canonical correlation analysis, LPPCCA)^[8]则在CCA方法中考虑样本局部信息,通过最大化每个样本 k 个近邻间的相关性,将全局问题变换为局部问题,同时也使非线性问题转换为线性问题。LPPCCA方法保持了样本的局部信息,但忽略了样本的标记信息。为提高有监督学习的分类性能,判别典型相关分析(discriminative canonical correlation analysis, DCCA)^[9]在CCA中加入了有关类的信息,该方法

收稿日期: 2012-06-18; **修回日期:** 2012-07-19 **基金项目:** 国家自然科学基金资助项目(61170145); 国家教育部高等学校博士点专项基金资助项目(20113704110001); 山东省自然科学基金和科技攻关计划资助项目(ZR2010FM021, 2008B0026, 2010G0020115)

作者简介: 高爽(1988-),女,山东济南人,硕士研究生,主要研究方向为机器学习、模式识别;张化祥(1966-),男(通信作者),教授,博导,主要研究方向为机器学习、模式识别、Web挖掘等(huaxzhang@163.com);房晓南(1979-),男,讲师,博士研究生,主要研究方向为机器学习、数据挖掘。

通过最大化类内相关性、最小化类间相关性得到投影矩阵。

本文研究垃圾网页的有监督学习分类问题,分别针对 Web 网页两视图数据采用上述几种典型相关分析方法:垃圾网页的分类是一个非线性问题,可使用 KCCA;研究表明^[2],在网络节点链接图中,同类页面节点更倾向于聚集在一起,为此考虑节点的局部特性使用 LPCCA;将垃圾网页检测问题作为有监督学习问题,使用加入类信息的 DCCA 方法。此外采用不同分类器算法进行实验比较。本文除了对两视图特征进行典型相关分析外,还在实验中对部分特征与整体特征的融合进行了实验比较。

1 典型相关分析方法

1.1 典型相关分析 CCA

给定 n 个样本的两组特征 $X = [x_1, x_2, \dots, x_n] \in R^{p \times n}$ 和 $Y = [y_1, y_2, \dots, y_n] \in R^{q \times n}$, CCA 方法用来寻找两组投影矩阵 $w_x \in R^{p \times d}$ 和 $w_y \in R^{q \times d}$, 使得投影后的矩阵 $w_x^T X$ 与 $w_y^T Y$ 之间的相关性最大^[10]。相应目标函数可以表示成如下一个等式约束的线性规划问题:

$$\begin{aligned} & \max_{w_x^T, w_y^T} w_x^T X Y^T w_y \\ \text{s. t. } & w_x^T X X^T w_x = 1, w_y^T Y Y^T w_y = 1 \end{aligned} \quad (1)$$

利用拉格朗日乘子法可将式(1)转换成求解广义特征值问题,将特征值按从大到小的顺序排序,取前 d ($d = \min(p, q)$) 个非负的特征值对应的特征向量作为投影矩阵 w_x, w_y 。在使用 CCA 及其推广方法时,需要注意避免小样本问题,否则会导致求解特征值时的矩阵不可逆。

1.2 核典型相关分析 KCCA

CCA 方法只能解决线性问题,于是加入了核函数产生了 KCCA。首先将两组特征用核函数进行投影,由非线性问题转换为线性问题。两组特征集 X 和 Y 隐式的非线性映射为 $\varphi: x \mapsto \varphi(x)$ 和 $\psi: y \mapsto \psi(y)$, 利用映射后的样本进行典型相关分析。于是 KCCA 的解向量为 $w_\varphi = \sum_{i=1}^n \alpha_i \varphi(x_i) = \varphi(X) \alpha$ 和 $w_\psi = \sum_{i=1}^n \beta_i \psi(y_i) = \psi(Y) \beta$, KCCA 的目标函数为

$$\max_{\alpha, \beta} w_\varphi^T \varphi(X) \psi(Y)^T w_\psi = \max_{\alpha, \beta} \alpha^T K_X K_Y \beta \quad (2)$$

其中: $K_X = \{k_x(x_i, x_j)\}_{i,j=1}^n \in \mathcal{R}^{n \times n}$, $K_Y = \{k_y(y_i, y_j)\}_{i,j=1}^n \in \mathcal{R}^{n \times n}$, k_x 和 k_y 为高斯核函数,核函数的参数取值为与某视图中样本的均方距离同数量级的实数。因此相应优化函数为

$$\begin{aligned} & \max_{\alpha, \beta} \alpha^T K_X K_Y \beta \\ \text{s. t. } & \alpha^T K_X K_X \alpha = 1, \beta^T K_Y K_Y \beta = 1 \end{aligned} \quad (3)$$

利用拉格朗日乘子法,可以将式(3)中对偶向量的求解表示为求广义特征值问题;同时为了避免过拟合问题,还需加入正则化的方法^[7]。

1.3 局部保持典型相关分析 LPCCA

LPCCA 方法考虑了局部信息,同时也解决了非线性问题^[10]。首先找出每个视图的特征集中某样本的 k 个近邻。根据样本的近邻,对两视图特征分别定义如下相似度矩阵:

$$S_{ij}^x = \begin{cases} \exp(-\|x_i - x_j\|^2 / \sigma_x^2) & x_i \in ne(x_j) \text{ or } x_j \in ne(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$S_{ij}^y = \begin{cases} \exp(-\|y_i - y_j\|^2 / \sigma_y^2) & y_i \in ne(y_j) \text{ or } y_j \in ne(y_i) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

其中: $x_i \in ne(x_j)$ 表示 x_i 为 x_j 的近邻;参数 σ_x, σ_y 取值为样本

的均方距离。在典型相关分析中加入相似度矩阵后,可以用如下优化问题描述:

$$\begin{aligned} & \max_{w_x, w_y} w_x^T \cdot \sum_{i=1}^n \sum_{j=1}^n S_{ij}^x (x_i - x_j) S_{ij}^y (y_i - y_j)^T \cdot w_y \\ \text{s. t. } & w_x^T \cdot \sum_{i=1}^n \sum_{j=1}^n S_{ij}^x (x_i - x_j) S_{ij}^x (x_i - x_j)^T \cdot w_x = 1 \\ & w_y^T \cdot \sum_{i=1}^n \sum_{j=1}^n S_{ij}^y (y_i - y_j) S_{ij}^y (y_i - y_j)^T \cdot w_y = 1 \end{aligned} \quad (6)$$

LPCCA 方法在一定程度上可以看做是 CCA 方法变形的推广。

1.4 判别典型相关分析 DCCA

由于将垃圾网页检测看做有监督学习,有必要对类信息加以利用,因此以下介绍一种加入判别信息的典型相关分析方法——DCCA 方法。对于 n 个样本的两组特征 X 和 Y , DCCA 方法旨在寻找两个投影矩阵 w_x 和 w_y , 使两个映射后的特征具有最大的类内相关性和最小的类间相关性。其优化问题为

$$\begin{aligned} & \max_{w_x, w_y} w_x^T (C_w - \eta C_b) w_y \\ \text{s. t. } & w_x^T X X^T w_x = 1, w_y^T Y Y^T w_y = 1 \end{aligned} \quad (7)$$

其中: C_w 为类内相关性矩阵, C_b 为类间相关性矩阵, η 是一个平衡因子。求得上述问题的特征值后,取前 d 个最大特征值对应的特征向量, d 除了满足等于 $\min(p, q)$ 还要满足小于等于类别数 c ^[9]。

2 特征抽取后的数据组合方法及分类

以往多视图典型相关分析的研究都是对两个不同视图进行相关性最大化处理^[8,9];除此之外,还可考虑分别将两个视图的局部特征与整体特征进行融合,并利用投影后的两个视图的局部特征分别进行分类。对于第一种情况,将求解广义特征值后得到的投影矩阵 w_x, w_y 进行特征组合,常用到的有并行

组合 $w_x^T X + w_y^T Y$ 和串行组合 $\begin{pmatrix} w_x^T X \\ w_y^T Y \end{pmatrix}$ 两种^[10]。对于第二种情况,将两视图特征串行排列起来组成整体特征 $B = [b_1, \dots, b_n] = \begin{pmatrix} x_1, \dots, x_n \\ y_1, \dots, y_n \end{pmatrix} \in R^{(p+q) \times n}$, X 与 B 进行相关性分析得到投影 w_x, w_b , Y 与 B 进行相关性分析得到投影 w_y, w_b , 仅使用对 X, Y 投影后的特征进行分类,即使用 $w_x^T X, w_y^T Y$ 分别进行分类。

除了采用上述的特征提取方法对垃圾网页的内容及链接特征进行处理外,在实验中还对不同分类器性能进行了比较,使用了最近邻分类器(1NN)、支持向量机分类器(SVM)、朴素贝叶斯分类器(naïve Bayes, NB)和决策树分类器(decision tree, DT)。在典型相关分析的研究文献[8~10]中常用最近邻分类器,同时也常将朴素贝叶斯分类器用于比较^[9]。在垃圾网页检测文献[3,5,11]中常用 SVM 分类器,而在与其他分类器(如 NB 和 DT)的比较文献[3,11]中都取得了很好效果。

3 实验及结果分析

3.1 Webspam 数据集

本文实验采用 Webspam-UK2007 数据集。该数据集网页由 Yahoo 研究的网络算法实验室下载,包含标记、链接图和 HTML 页面三部分。所有网页均是英国域名(.uk),共包含 114 529 个主机名,105 896 555 个网页。由一个志愿者团队对网站是否为垃圾网页进行标注,标注时默认当一个主机名为 spam 或 non-spam 时,其下的所有网页均有相同的分类结果。

每个主机由多人标记,分类结果取平均。同时数据集中提供了已提取好的特征包含基于内容和链接的特征,实验时使用这部分特征。类标和特征都是在主机名下获取。其中训练集已标记数据共有 4 275 个,测试集中已标记数据共有 2 204 个。具体如下如表 1 所示。

表 1 Webspam-UK2007 数据集

数据集	已标记主机个数	spam	non-spam	undecided
训练集	4 275	222	3 776	277
测试集	2 204	122	1 933	149

实验时只取其中标记为 spam 和 non-spam 的样例,此类数据在训练集中共有 3 998 个,测试集中共有 2 055 个。特征数据集中只包含了 109 361 个主机的基于内容的特征,所以真正可以用于实验的数据如表 2 所示。

表 2 实验使用数据集情况

数据集	已标记主机个数	spam	non-spam
训练集	3 849	208	3 641
测试集	1 948	113	1 835

训练集共有已标记数据 3 849 个,测试集共有 1 948 个。特征集中基于内容的特征数为 96 个,基于链接的特征数为 41 个。

3.2 性能评估

使用 F1-Measure 作为评价指标。F1-Measure 是信息检索领域及垃圾网页检测^[2,11]中常用的评价方法。求 F1-Measure 的值,需首先定义表 3 的内容。

表 3 F1-Measure 混合矩阵

	prediction non-spam (positive)		prediction spam (negative)	
	tp	tn	fp	fn
label non-spam (true)				
label spam (false)				

表中, tp 表示分类正确的非垃圾网页, tn 表示分类正确的垃圾网页, fp 表示分类错误的非垃圾网页, fn 表示分类错误的垃圾网页。

在垃圾网页检测问题中,召回率是在所有标记为垃圾网页的页面中正确分类网页所占比率;准确率是在所有预测为垃圾网页的页面中实际标记为垃圾网页的页面所占比率。召回率、准确率及 F1-Measure 计算公式如下:

召回率 (Recall-true positive rate) 为 $R = fn / (fp + fn)$

准确率 (Precision) 为 $P = fn / (tn + fn)$

F1-Measure 为 $F = 2RP / (R + P)$

3.3 特征提取方法和分类器的比较

对 Webspam-UK2007 中的数据,分别使用不同的典型相关分析特征提取方法和分类器进行实验比较,得到表 4 所示结果。其中 content + link 表示不对特征集作任何处理,直接将内容特征及链接特征串行组合。此外,针对数据集中的数据不平衡问题对实验数据进行了相应调整。为了保证求解投影矩阵时矩阵非奇异,对特征也进行了相应的选择。KCCA 方法的参数与 LPCCA 中的近邻数 k 通过实验验证得出。

表 4 不同的特征提取方法和分类器的 F1-Measure 值

分类器	content + link	CCA		KCCA		LPCCA		DCCA	
		并行	串行	并行	串行	并行	串行	并行	串行
INN	0.481 8	0.616 0	0.583 7	0.627 5	0.618 4	0.586 2	0.591 9	0.547 1	0.554 0
SVM	0.675 9	0.566 5	0.609 9	0.596 2	0.627 6	0.609 8	0.660 2	0.547 8	0.634 6
NB	0.405 1	0.287 7	0.372 1	0.678 9	0.690 1	0.660 4	0.670 8	0.388 2	0.597 9
DT	0.626 6	0.556 1	0.561 1	0.642 9	0.645 6	0.633 7	0.629 5	0.614 7	0.553 6

注:加粗数据为同一分类器中的最高值

表 4 中实验结果显示:对于不同的分类器,各种不同典型相关分析方法的分类效果也有所不同。在最近邻分类器中,使用典型相关分析方法后 F1-Measure 值都有所提高;使用支持向量机分类器时,经过特征提取后,分类效果有所下降,LPCCA 方法下降幅度较小,由此可以说明对于垃圾网页检测,除了解决其分类的非线性问题外,保持局部特性也很重要,在文献[2]中提到正常网页常聚集在一起,垃圾网页常聚集在一起,所以 LPCCA 能够取得较好的分类效果;使用朴素贝叶斯分类器时,KCCA 的分类效果最好,其次是 LPCCA 和 DCCA,但 DCCA 方法并行组合时分类精度比较低,原因可能是在 DCCA 方法中求解投影矩阵时矩阵的秩受到类别数目的限制,垃圾网页检测问题是一个二分类问题,使用 DCCA 后特征会降到一个比较低的维数,维数过低反而会影响到分类结果,而串行组合实际上是对特征进行先降维后升维处理,所以分类效果相对较好;在决策树分类器中,KCCA 方法效果优于其他方法。

总体来讲,KCCA 方法在垃圾网页检测问题上取得了最好的分类效果,说明使用核技术能够提高垃圾网页的分类精度,同时很少会出现过拟合现象。支持向量机分类器在 content + link 情况下也有很高的 F1-Measure 值,验证了文献[3,11]的实验结果中支持向量机分类器的优势,但是在使用了典型相关分析方法之后,使用其他分类器会取得更好的效果。

3.4 两视图数据关联分析前后分类性能比较

利用内容与链接两视图学习最大关联矩阵,并用此关联矩阵对内容特征进行投影实现特征抽取,分别用原内容数据及特征抽取后的内容数据训练分类器,比较 F1-Measure 值,得到表 5 所示的实验结果。其中 content 表示不作任何处理的内容特征用于比较。

表 5 基于原内容特征与特征抽取后内容特征的 F1-Measure 值

分类器	content	CCA	KCCA	LPCCA	DCCA
INN	0.549 5	0.570 2	0.607 6	0.681 8	0.526 8
SVM	0.630 5	0.612 4	0.564 7	0.621 4	0.611 7
NB	0.434 8	0.378 0	0.666 7	0.426 0	0.596 9
DT	0.626 6	0.606 1	0.602 6	0.658 3	0.606 8

注:加粗数据为同一分类器中的最高值

表 5 的实验结果表明,使用了 LPCCA 方法的最近邻分类器获得了最好的分类效果;支持向量机分类器在 content 情况下比使用典型相关分析方法进行特征提取后的分类结果好,这一点与表 4 的实验结果相同;在朴素贝叶斯分类器中,使用 KCCA 方法的 F1-Measure 值也比较高;对于内容特征整体而言,决策树分类器比其他分类器的精度要高。本实验又一次验证了支持向量机分类器的优势,但与使用了典型相关分析的其他分类器相比此优势并不明显。

表 6 实验表示只用链接特征及投影后的链接特征训练分类器的结果,其中 link 表示不作任何处理的链接特征用于比较。

表 6 基于原链接特征与特征抽取后链接特征的 F1-Measure 值

分类器	link	CCA	KCCA	LPCCA	DCCA
INN	0.435 2	0.477 1	0.641 0	0.550 5	0.511 6
SVM	0.576 3	0.594 1	0.576 5	0.599 2	0.599 2
NB	0.179 1	0.670 8	0.662 7	0.660 5	0.305 7
DT	0.637 2	0.580 1	0.670 8	0.624 5	0.569 0

注:加粗数据为同一分类器中的最高值

表 6 中实验结果表明,最近邻分类器中,KCCA 方法 F1-

Measure 值较高;支持向量机分类器的优势在链接特征中并不明显,使用 LPCCA 和 DCCA 方法后分类精度较高;朴素贝叶斯分类器在使用了典型相关分析方法之后,F1-Measure 值会大幅提高,其中 CCA 方法的分类效果最好。使用 KCCA 的决策树分类器也取得了最高的 F1-Measure 值。

这一部分的实验更加验证了把垃圾网页检测作为多视图分类问题的必要性,例如在内容特征中分类效果较好的支持向量机在链接特征中并没有很大优势,对链接特征使用典型相关分析方法之后支持向量机分类效果也有所提高,另外对于不同视图的特征所使用的典型相关分析方法也有所不同。由于内容特征的维数比链接特征多,在将两视图数据融合时受内容特征的影响较大,这也解释了支持向量机分类器在表 4 中 content + link 情况下 F1-Measure 值较高的原因。

4 结束语

将垃圾网页检测作为多视图问题来考虑,用典型相关分析及其推广方法对垃圾网页内容和链接特征进行特征提取,解决了如何处理垃圾网页特征的问题。对分类器选择的研究也为今后基于多视图的垃圾网页检测提供了依据。垃圾网页检测除了使用基于多视图的特征提取方法外,还可以从分类的角度运用多视图手段。考虑到现实中大多数的网页都是未被标记的,所以可以利用基于多视图的半监督协同训练来分类。

参考文献:

- [1] BECCHETTI L, CASTILLO C, DONATO D, *et al.* Web spam detection: link-based and content-based techniques [C]//Proc of European Final Workshop Integrated Project Dynamically Evolving, Large Scale Information Systems(DELIS). 2008:99-113.
- [2] CASTILLO C, DONATO D, GIONIS A, *et al.* Know your neighbors: Web spam detection using the Web topology [C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007:423-430.
- [3] BÍRÓ I, SIKLÓSI D, SZABÓ J, *et al.* Linked latent dirichlet allocation in Web spam filtering [C]//Proc of the 5th International Workshop on Adversarial Information Retrieval on the Web. New York: ACM Press, 2009:37-40.
- [4] LOURDES A, JUAN M R. Web spam detection: new classification features based on qualified link analysis and language models [J]. *IEEE Trans on Information Forensics and Security*, 2010, 5(3): 581-590.
- [5] ABERNETHY J, CHAPELLE O, CASTILLO C. Graph regularization methods for Web spam detection [J]. *Machine Learning*, 2010, 81(2):207-225.
- [6] HARDOON R D, SZEDMAK R S, SHAW-TAYLOR R J. Canonical correlation analysis: an overview with application to learning method [J]. *Neural Computation*, 2004, 16(12):2639-2664.
- [7] MELZER T, REITER M, BISCHOF H. Appearance models based on kernel canonical correlation analysis [J]. *Pattern Recognition*, 2003, 36(9):1961-1971.
- [8] SUN Ting-kai, CHEN Song-can. Locality preserving CCA with applications to data visualization and pose estimation [J]. *Image and Vision Computing*, 2007, 25(5):531-543.
- [9] SUN Ting-kai, CHEN Song-can. A novel method of combined feature extraction for recognition [C]//Proc of the 8th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2008:1043-1048.
- [10] 孙廷凯. 增强型典型相关分析研究与应用 [D]. 南京:南京航空航天大学, 2006.
- [11] WANG Wei, ZENG Guo-sun, TANG Dai-zhong. Using evidence based content trust model for spam detection [J]. *Expert Systems with Applications*, 2010, 37(8): 5599-5606.
- [1] BECCHETTI L, CASTILLO C, DONATO D, *et al.* Web spam detection: link-based and content-based techniques [C]//Proc of European Final Workshop Integrated Project Dynamically Evolving, Large Scale Information Systems(DELIS). 2008:99-113.
- [2] CASTILLO C, DONATO D, GIONIS A, *et al.* Know your neighbors: Web spam detection using the Web topology [C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007:423-430.
- [3] BÍRÓ I, SIKLÓSI D, SZABÓ J, *et al.* Linked latent dirichlet allocation in Web spam filtering [C]//Proc of the 5th International Workshop on Adversarial Information Retrieval on the Web. New York: ACM Press, 2009:37-40.
- [4] LOURDES A, JUAN M R. Web spam detection: new classification features based on qualified link analysis and language models [J]. *IEEE Trans on Information Forensics and Security*, 2010, 5(3): 581-590.
- [5] ABERNETHY J, CHAPELLE O, CASTILLO C. Graph regularization methods for Web spam detection [J]. *Machine Learning*, 2010, 81(2):207-225.
- [6] HARDOON R D, SZEDMAK R S, SHAW-TAYLOR R J. Canonical correlation analysis: an overview with application to learning method [J]. *Neural Computation*, 2004, 16(12):2639-2664.
- [7] MELZER T, REITER M, BISCHOF H. Appearance models based on kernel canonical correlation analysis [J]. *Pattern Recognition*, 2003, 36(9):1961-1971.
- [8] SUN Ting-kai, CHEN Song-can. Locality preserving CCA with applications to data visualization and pose estimation [J]. *Image and Vision Computing*, 2007, 25(5):531-543.
- [9] SUN Ting-kai, CHEN Song-can. A novel method of combined feature extraction for recognition [C]//Proc of the 8th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2008:1043-1048.
- [10] 孙廷凯. 增强型典型相关分析研究与应用 [D]. 南京:南京航空航天大学, 2006.
- [11] WANG Wei, ZENG Guo-sun, TANG Dai-zhong. Using evidence based content trust model for spam detection [J]. *Expert Systems with Applications*, 2010, 37(8): 5599-5606.
- [8] 王尚广,孙其博,杨放春. 基于全局 QoS 约束分解的 Web 服务动态选择 [J]. *软件学报*, 2011, 22(7):1426-1439.
- [9] MAURER M, BRANDIC I, EMEAKAROKA V C, *et al.* Towards knowledge management in self-adaptable clouds [C]//Proc of the 4th IEEE International Workshop of Software Engineering for Adaptive Service-Oriented Systems. 2010.
- [10] VASSILIKI D, CHRISTOS M, YANNIS P, *et al.* Techniques to support Web service selection and consumption with QoS characteristics [J]. *Journal of Network and Computer Applications*, 2008, 31(2):108-130.
- [11] SHVACHKO K, HUANG Hai-rong, RADIA S, *et al.* The Hadoop distributed file system [C]//Proc of the 26th IEEE Symposium on Massive Storage Systems and Technologies. Washington DC: IEEE Computer Society, 2010:1-10.
- [12] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [C] //Proc of the 6th Conference on Operating Systems Design & Implementation. New York: ACM Press, 2004:137-150.
- [13] 谢海军,齐连永,窦万春. 基于 Skyline 和局部选择的启发式服务组合方法 [J]. *东南大学学报:自然科学版*, 2011, 41(3):449-452.
- [14] 倪志伟,尹道明,王力,等. 基于和声搜索算法的知识服务动态组合优化 [J]. *计算机工程与应用*, 2012, 48(32):216-221, 247.

(上接第 809 页)

参考文献:

- [1] LUIS M V, LUIS R M, JUAN C, *et al.* A break in the clouds: towards a cloud definition [J]. *ACM SIGCOMM Computer Communication Review*, 2009, 39(1):50-55.
- [2] SUBASHINI S, KAVITHA V. A survey on security issues in service delivery models of cloud computing [J]. *Journal of Network and Computer Applications*, 2011, 34(1):1-11.
- [3] 刘鹏. 云计算 [M]. 北京:电子工业出版社, 2010:10-34.
- [4] GEEM Z W. Effects of initial memory and identical harmony in global optimization using harmony search algorithm [J]. *Applied Mathematics and Computation*, 2012, 218(22):11337-11343.
- [5] 韩红燕,任立群. 求解批量流水线调度问题的和声算法 [J]. *计算机应用研究*, 2011, 28(1):87-90.
- [6] PAN Quan-ke, SUGANTHAN P N, TASGETIRENC M F, *et al.* A self-adaptive global best harmony search algorithm for continuous optimization problems [J]. *Applied Mathematics and Computation*, 2010, 216(3):830-848.
- [7] XU Shou-huai, ZHANG Wei-ning. Knowledge as a service and knowledge breaching [C]//Proc of IEEE International Conference on Services Computing. Washington DC: IEEE Computer Society, 2005:87-94.