doi:10.3969/j.issn.1000-436x.2012.12.006

用于文本相似度计算的新核函数

王秀红[1,2,3]，鞠时光[4]

(1. 江苏大学 科技信息研究所，江苏 镇江 212013；2. 江苏大学 理学院，江苏 镇江 212013；
3. 加州大学戴维斯分校 农业与环境科学学院，加利福尼亚州 戴维斯 95616；4. 江苏大学 计算机科学与通信工程学院，江苏 镇江 212013)

摘 要：为了提高文本相似检测的综合表现，在文本文档相似特征的基础上构造了新的核函数 S_Wang 核函数。结合文本相似计算过程中的实际情况，将待比对的文本表示成向量，考虑通过 2 个向量间的乘积和欧氏距离来描述向量之间的相似程度，从而构造了适合文本相似度计算的新核函数，并根据 Mercer 定理证明了所构造函数可以作为核函数。实验验证了新构造的核函数在文本文档相似度计算中的表现，实验结果表明 S_Wang 核其相似度计算精度和综合指标均分别优于 Cauchy 核、潜在语义核（LSK）以及 CLA 复合核。S_Wang 核适用于文本相似度计算。

关键词：信息检索；文本相似度；核函数；S_Wang 核; 潜在语义核；Cauchy 核；CLA 复合核

中图分类号：TP312 **文献标识码：**A **文章编号：**1000-436X(2012)12-0043-06

Novel kernel function for computing the similarity of text

WANG Xiu-hong[1,2,3], JU Shi-guang[4]

(1. Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013, China;
2. Faculty of Science, Jiangsu University, Zhenjiang 212013, China;
3. College of Agricultural and Environmental Sciences, University of California-Davis, Davis 95616, USA;
4. School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: To enhance the performance of detecting similar documents, a novel kernel function named S_Wang kernel was constructed. Based on the actual situation of computing text similarity, the S_Wang kernel was newly bu lt with consideration of the Euclidean distance and angle between vectors that represented the text documents to be compared. It was proved that the function could be constructed as a kernel function according to Mercer theorem. Experimental verification of the performance of the kernels in the text document similarity calculation was provided. The results show that the S_Wang kernel is significantly better than the precision and F1 performance of other kernels like Cauchy kernel, Latent Semantic Kernel (LSK) and CLA kernel. S_Wang kernel is suitable for text similarity computation.

Key words: information retrieval; text similarity; kernel function; S_Wang kernel; LSK; Cauchy kernel; CLA kernel

1 引言

核方法的思想是：将在低维空间中一个非线性可分的问题向高维空间转化，即映射到高维空间，使其在高维空间中变得线性可分，然后在特征空间中使用线性学习机建立优化超平面，利用高维特征空间中的内积来对低维空间的问题进行分类，从而解决问题。而转化最关键的部分在于找到输入空间中 x 到高维空间中 $f(x)$ 的映射方法，如何找到这个映射 f 没有系统的方法。事实上，该映射函数往往不易找到，且不一定能显式表达。这个办法带来的困难就是计算复杂度的增加，且直接在这个特征空间作内积计算会面临一个维数灾难问题。核函数的基本作用就是接受 2 个低维空间里的向量输入值 x

收稿日期：2011-12-26；修回日期：2012-09-18

和 z，能够计算出经过某个变换后在高维空间里的向量内积值，而无需寻找那个从低维空间到高维空间的具体映射。关于核函数的定义如下。

设 $x, z \in X$, X 属于 R（n）空间，非线性函数 F 实现输入空间 X 到特征空间 H（内积空间或 Hilbert 空间: $H, <\cdot,\cdot>$）的映射（$?$ ：X? H），其中，H 属于 R（m），$n<<m$。

$$k(x,z) = \langle f(x), f(z) \rangle \tag{1}$$

其中，$<,>$为内积，$k(x,z)$ 为核函数。

核函数的应用很好地解决了计算复杂度和维数灾难的问题。根据泛函的有关理论，只要一种核函数 K 满足 Mercer 条件，它就对应某一变换空间中的内积，满足 Mercer 条件的任意对称函数都可以作为核函数。针对具体的问题，选择和构造适合该问题的核函数是解决该领域具体非线性分类问题的关键所在，具有非常重要的意义。

高维空间中的映射函数的内积即反映了 2 个输入数据之间的距离（即相似性）。文档相似检测本质上是计算 2 篇文档的相似程度。每一个文档均可表示成一个向量，文档相似检测问题就转化为计算 2 个输入向量的相似度问题。2 篇文档相似与不相似是一个在低维空间中非线性可分的问题。

构造核函数使其通过接受待比对文档低维空间的输入值，就能算出高维空间的内积值，有效地解决文档相似计算的问题是本文研究的出发点和目标。目前已有的一些核函数[1]，如径向基核函数 (RBF, radial basis functions) Homogeneous kernels 具有 $k(x,x')=k(\|x-x'\|)$的形式。

用于文本处理的核方法有：将文本视为词包（bag of word）的核，如点积或多项式核(dot product or polynomial kernel)，该类核的表示能力不太强；将文本视为概念体集合（set of concept）的核，如 GVSM 核、Kernel LSI（或 kernel PCA），该类核用于计算的精准率（精度）较低；此外还有将文本视为字符串的核如字符串核（string kernel）、将文本视为词或概念串的核如词序核（word sequence kernel）和将文本视为树结构的树核（tree kernel）等，此类核用于计算时其召回率较低。核方法在文本相似计算方面已有广泛的应用，如：在文本分类上的应用[2]、基于 Web 的核函数用于短文本片段的相似计算[3]、利用树核的短文本相似度计算[4]、利用信息几何学的方法计算文本的相似度[5]以及利用复合核函数计算检索结果和提问相似度从而对检索结果进行融合的 CLA 核[6]等。事实上，现有的核函数在用于文本相似度计算的时候，其相似度计算的精准率和召回率均有待提高。本文研究的目的在于构造适合文本文档相似度计算这一特定问题的新核函数，使其在文本文档相似计算方面具有更好的精准率和综合表现。

2 模型的建立

2.1 新核的构造

关于核函数的构造方法，目前已有一些相关研究[7~10]。而如何针对具体的问题构造适合解决该问题的核函数是不同领域应用核函数的关键所在。本文旨在根据文本文档自身的特点和相似判别的具体要求，构建适合此类相似比对的核函数。首先是利用词包法先将文档表示成向量。

2.1.1 新核的构造思想

举例说明如下：假设 2 篇文档 X 和 Z，统计词后具有如表 1 所示的词项。

表 1 文档中词的集合

X	Z
A	B
B	C
C	D
F	G
P	L
M	D
B	

则 2 篇文档出现的所有词为：A，B，C，D，F，G，L，M，P 共有 9 个概念(此处 N=9)。则在映射 f_2 下将待比对的 2 篇文本文档表示为向量 $\boldsymbol{x}$ 和 $\boldsymbol{z}$，如表 2 所示。

构造思路步骤如下。

1) 如果词典中某一词 t_i 在某一篇文档中未出现，即对应的向量维数位置值为 0，则认为该词对 2 篇文档相似的贡献值为 0，如果待比对的 2 篇文档没有共同的词，则认为该 2 篇文档的相似度为 0，于是考虑利用 2 个行向量对应维数相乘 $\boldsymbol{x}\boldsymbol{z}^{\mathrm{T}}$ 形式来计算其相似度，作为构造核函数的分子。

2) 如果某一词 t_i 在 2 篇待比对的文档中词频统计结果差值 $|tf(t_i,x)-tf(t_i,z)|$ 越大，表明 2 篇文档越不相似，该词 t_i 使相似程度的计算结果越小，

表 2　　文档的实值向量表示

文档 *term*	$term_1$	$term_2$	$term_3$	$term_4$	$term_5$	$term_6$	$term_7$	$term_8$	$term_9$
词典（*N*）	A	B	C	D	F	G	L	M	P
x	1	2	1	0	1	0	0	1	1
z	0	1	1	2	0	1	1	0	0

于是考虑用$\| x-z \|^2$表明 2 篇文档之间由于词语不同产生的欧氏距离，且将其置于构造核函数的分母上。

3) 如果 2 篇文档完全相同，则 $x=z$，此时有 $\| x-z \|=0$，且 $xz^{\mathrm{T}}=1$；当 2 篇文档完全相同的时候，其相似度计算值应为 1，于是考虑构造核函数的分母形式为 $xz^{\mathrm{T}}+\| x-z \|^2$。

4)$s(s>0)$为宽度参数，用来控制函数的径向作用范围，调节由于词语不同导致 2 篇文档距离对相似度的影响。

最终构造出核函数形式为

$$k(x,z)=\frac{x^{\mathrm{T}}z}{x^{\mathrm{T}}z+\dfrac{\| x-z \|^2}{s}} \tag{2}$$

2.1.2　理论证明

下面证明 2.1.1 节设计的函数可作为核函数。

统计学习的理论指出，根据 Hilbert-Schmidt 原理，只要一种运算满足 Mercer 条件，则可作为变换空间的内积使用，即可作为核函数。

引理 1（Mercer 定理）：令 X 是 R^n 上的一个紧集，$k(x,z)$ 是 $X\times X$ 上连续实值对称函数，则

$$\iint_{X\times X} k(x,z)f(x)f(z)\mathrm{d}x\mathrm{d}z \geq 0, \forall f\in L_2(x) \tag{3}$$

称式(3)为 Mercer 条件。

式(3)等价于 $k(x,z)$ 是一个核函数，即 $k(x,z)=(f(x)\cdot f(z))$，$x,z\in X$，其中，f 为某个从 X 到 Hilbert 空间 H 的映射 $f:|\to f(x)\in H$，$(\cdot)$ 是 Hilbert 空间 L_2 上的内积。

下面证明所构造的函数可以作为核函数，看其是否满足 Mercer 条件。

证明

1)令 $k_1(x,z)=x^{\mathrm{T}}z$，$k_2(x,z)=\dfrac{\| x-z \|^2}{s}$，则式(1)可变为 $k(x,z)=\dfrac{k_1(x,z)}{k_1(x,z)+k_2(x,z)}$。

2）显然 $k_1(x,z)=x^{\mathrm{T}}z$ 是线性核函数，它满足当 X 是 R^n 上的一个紧集时，$k_1(x,z)$ 是 $X\times X$ 上的连续实值对称函数，因文档向量 x 和 z 所有元素值均为非负，所以 $k_1(x,z)$ 为非负。

3）$k_2(x,z)=\dfrac{\| x-z \|^2}{s}$（$s$>0）是 Homogeneous kernel(RBF) 径向基核函数，只依赖于距离的大小。它满足当 X 是 R^n 上的一个紧集时，$k_2(x,z)$ 在 $X\times X$ 上为连续实值对称函数，且因 s>0，所以为非负。

4）当 $x-z$ 为 0，即 2 篇文档 X 和 Z 完全相同时，$k_2(x,z)$=0，而此时必然有 $k_1(x,z)=x^{\mathrm{T}}z=1\neq 0$。当 2 篇文档完全不同时，$k_2(x,z)$=1，而此时必然有 $k_1(x,z)=x^{\mathrm{T}}z=0$。可见式（2）的分母不可能为 0。

综上所述，当 X 是 R^n 上的一个紧集时，$k(x,z)=\dfrac{x^{\mathrm{T}}z}{x^{\mathrm{T}}z+\dfrac{\| x-z \|^2}{s}}$ 是 $X\times X$ 上的连续实值对称函数，且为非负。则由 Mercer 定理有：$\iint_{X\times X} k(x,z)f(x)f(z)\mathrm{d}x\mathrm{d}z\geq 0,\forall f\in L_2$。于是有 $k(x,z)$ 可以作为一个核函数，即 $k(x,z)=(f(x)\cdot f(z))$，$x,z\in X$。

证毕。

为了方便描述，本文将构造的新核命名为 S_Wang 核。

2.2　文档的向量表示

为了使待比较的文本文档表示成输入向量形式，需要将文档进行向量表示。

2.2.1　文档的词包表示

文档的词包表示是将文本文档表现成向量的一种方法[10]。假设有 2 篇文档 X 和 Z 需要计算其相似程度。将每一篇文档视为词包，将其表示为一个行向量，其每一个维数都与词典的一个词相关联。此处的词典为需要比对的 2 篇文档包含的所有规范化词的一个集合，大小为 N。则通过某种映射，文档 d 可以表示为一个行向量。

$$f_1: d\to f_1(d)=(tf(t_1,d),tf(t_2,d),\cdots,tf(t_N,d))\in R^N$$

其中，$tf(t_i,d)$)是词 t_i ($i=1,2,\cdots,N$)在文档 d 中出现的频率，通常为 1。

2.2.2 进一步考虑词语义信息的文档词包表示

词包表示未考虑词的语义信息，为了解决此问题需在词包表示法的基础上构建语义核。不同的词对主题的重要程度不同。用一个词在文档中出现的频率来量化这个词所带信息的重要程度，即 IDF（inverse document frequency）规则，具体表示为

$$w(t)=\ln\left(\frac{l}{df(t)}\right) \tag{4}$$

其中，l 为文集中存在的文档个数，$df(t)$是包含词 t 的文档个数，$w(t)$为逆文档频率 IDF 规则定义词 t 的衡量词权重的绝对尺度。从而赋予不同词以不同的权重 $w(t)$，建立新的考虑词义信息的映射，进一步将文档 d 映射到带有语义信息的特征空间，选矩阵 $\boldsymbol{R}$ 的元素是 $\boldsymbol{R}_{tt}=w(t)$ 的对角矩阵，从而映射变为

$$d\to f_2(d)=f_1(d)R$$

进一步地，考虑词的语义信息，文档可表示为

$$f_2:d\to f_2(d)=(w(t_1)tf(t_1,d),\\ w(t_2)tf(t_2,d),\cdots,w(t_N)tf(t_N,d))\in R^N$$

3 算法描述与复杂度分析

3.1 算法描述

具体算法如下：

输入：现有文档 Z 和待检测的文档 X。

输出：现有文档 Z 和待检测的文档 X 的相似度值 S。

Step1 分别将现有文档 Z 和待检测的文档 X 映射到特征空间，用词包表示

$$f_1:\boldsymbol{z}\to f_1(\boldsymbol{z})=(tf(t_1,\boldsymbol{z}),tf(t_2,\boldsymbol{z}),\cdots,tf(t_N,\boldsymbol{z}))\in R^N$$

$$f_1:\boldsymbol{x}\to f_1(\boldsymbol{x})=(tf(t_1,\boldsymbol{x}),tf(t_2,\boldsymbol{x}),\cdots,tf(t_N,\boldsymbol{x}))\in R^N$$

Step 2 建立文档—词矩阵

$$\boldsymbol{D}=\begin{pmatrix} tf(t_1,d_1) & \cdots & tf(t_N,d_1)\\ \vdots & \ddots & \vdots\\ tf(t_1,d_l) & \cdots & tf(t_N,d_l)\end{pmatrix}$$ 和核矩阵 $\boldsymbol{K}=\boldsymbol{DD}'$

Step3 将现有文档 Z 和待检测的文档 X 进一步映射到带有语义信息的特征空间。

$$f_2:\boldsymbol{z}\to f_2(\boldsymbol{z})=(w(t_1)tf(t_1,\boldsymbol{z})\\ w(t_2)tf(t_2,\boldsymbol{z}),\cdots,w(t_N)tf(t_N,\boldsymbol{z}))\in R^N$$

$$f_2:\boldsymbol{x}\to f_2(\boldsymbol{x})=(w(t_1)tf(t_1,\boldsymbol{x}),\\ w(t_2)tf(t_2,\boldsymbol{x}),\cdots,w(t_N)tf(t_N,\boldsymbol{x}))\in R^N$$

Step 4 利用新设计的核 S_Wang 计算 2 篇文档向量的内积，得现有文档 Z 和待检测的文档 X 之间相似程度 S。

$$S=k(f_2(\boldsymbol{x}),f_2(\boldsymbol{z}))=k(\boldsymbol{x},\boldsymbol{z})=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}+\dfrac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{s}}$$

去除文本长度的不同对结果的影响，单位化为

$$\%=\frac{S}{\sqrt{SS}}$$

3.2 算法复杂度分析

从计算的角度看，文档的向量表示 $f_2(\boldsymbol{d})$ 的计算复杂度为 $O(1\times N)$；新构造的用于内积计算的核函数 $k(\boldsymbol{x},\boldsymbol{z})=\dfrac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}+\dfrac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{s}}$ 的算法复杂度为 $O(N)$，所以总的计算复杂度为 $O(1\times N^2)$。

4 实验

4.1 实验语料

为了与以往的研究进行对比，本文实验语料选择本领域通用的语料，数据测试集为 TREC 文本集合和主题：50 个 TREC 检索主题（251-300）以及 524 000 多个文档，包括 AP88、CR93、FR94、FT91-94、WSJ90-92 以及 ZF 等。对语料进行了随机划分，训练/测试的比例是 3:1。实验评估了 4 种核函数，包括潜在语义核 (LSK)[12]、Cauchy Kernel(Basak, 2008)[13]、CLA 复合核[6]以及本文提出的 S_Wang 核函数。使用的软件有 MATLAB7.0 和 Lemur，实验中线性学习器采用 LibSVM[14]。实验是在 8 个文档水平（top 5、top 10、top 15、top 20、top 25、top 30、top 50、top 100）上进行。这里所说的文档水平是指经过融合排序后的结果，称排在最前面的 n 个文档(top n)为一个文档水平用来实验验证有效性的对象。

4.2 实验评价指标

采用典型的信息检索评价指标：精准率（precision）、召回率（recall）和综合评价指标，具体算法为

$$precision=\frac{true\ positive}{true\ positive+\ flase\ positive} \tag{5}$$

表 3　4 种核函数在 8 个文档水平上的相似度计算精度、召回率以及 F_1 实验分值

核函数	指标	top 5	top 10	top 15	top 20	top 25	top 30	top 50	top100
Cauchy Kernel s =1	P	0.290 2	0.271 9	0.250 5	0.240 1	0.226 7	0.207	0.181 1	0.154 1
	R	0.809	0.796 7	0.781 3	0.770 2	0.766	0.753 4	0.747 7	0.726 5
	F_1	0.427 2	0.405 4	0.391 4	0.365 8	0.349 9	0.324 8	0.291 6	0.254 3
LSK	P	0.241 4	0.218 8	0.202 6	0.190 2	0.182 1	0.170 9	0.158 2	0.139 5
	R	0.984 5	0.977 8	0.963 3	0.951 2	0.939	0.921 2	0.901 9	0.881 6
	F_1	0.387 7	0.357 5	0.334 8	0.317	0.305	0.288 3	0.269 2	0.240 9
CLA kernel	P	0.325 3	0.291 2	0.273 4	0.255 9	0.241	0.219 9	0.198 3	0.157 7
	R	0.879 2	0.860 3	0.844 5	0.826 7	0.818 9	0.810 3	0.794 5	0.775 3
	F_1	0.474 8	0.435 6	0.413 2	0.390 8	0.372 4	0.345 9	0.317 4	0.262 1
S_Wang Kernel s =1	P	0.356 1	0.326	0.304	0.281	0.264 5	0.241	0.209 3	0.169 8
	R	0.907 9	0.89	0.867 8	0.849	0.833 3	0.826	0.803 2	0.784 4
	F_1	0.511 6	0.477 2	0.450 3	0.422 2	0.401 5	0.373 1	0.332 1	0.279 2

$$Recall = \frac{true\ positive}{true\ positive + flase\ negative} \tag{6}$$

$$F_b - measure = \frac{(1+b^2)\,precision \times recall}{b^2\,precision + recall} \tag{7}$$

考虑到结果融合中召回率和精准率同等重要，本文综合评价指标中的参数 ß 取 1，得 F_1 的指标。

4.3　实验设计与结果分析

实验评估了多种核函数。将检索的召回率和精准率视为同等重要，故在文献[6]的式(14)中，CLA 核的系数 $d_1 = 0.5$，最终得到不同核函数的相似度计算表现，实验结果如表 3 所示。其中，P 表示相似度计算精准率，R 表示相似度计算召回率。

通过对不同的核函数的相似度计算精准率进行分析，分析结果如图 1 所示。

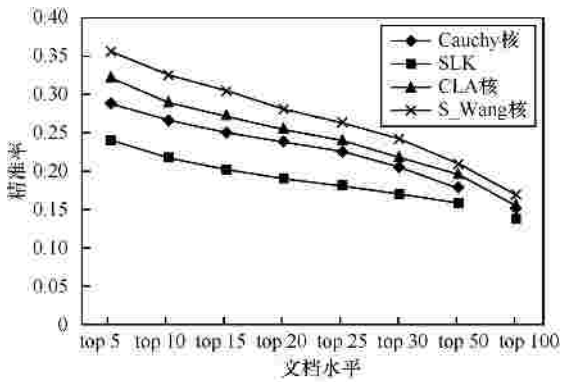

图 1　不同核函数进行相似计算，在不同的文档水平上的精准率表现

从图 1 可以看出 8 个不同文档水平上进行文本相似度计算的精准率表明：随着考察文档水平数的增加，各核的相似度计算精度逐渐减小，且各核之间的表现相差渐小。此外，虽然潜在语义核 LSK 因考虑到不同词间的潜在语义关系而具有较高的召回率，但其精准率在 4 个核中表现最差，在 top 5 文档水平上达到 0.241 4，在 top 100 文档水平上只达 0.139 5；CLA 核在精准率上略好于 Cauchy 核，在 top 5 文档水平上达到 0.325 3，在 top 100 文档水平上达 0.157 7 ;而 S_Wang 核在 8 个文档水平 top 5、top10、top 15、top 20、top 25、top 30、top 50 及 top 100 上其精度分别达到 0.356 1、0.326、0.304、0.281、0.264 5、0.241、0.209 3 及 0.169 8，均分别大于其他 3 个核函数在对应文档水平上的相似度计算精准率。

分别将 4 个核函数在相似度计算中的精准率、召回率和综合表现 F_1 在 8 个文档水平上平均后，比较不同核函数的相似度计算表现，其结果如图 2 所示。

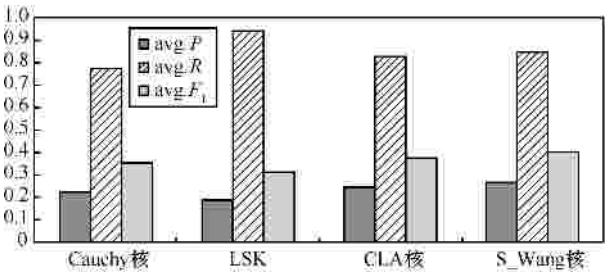

图 2　4 种核函数在 8 个文档水平上的平均表现

从图2可以看出，S_Wang核用于文本文档相似度计算时表现突出。S_Wang核的平均召回率为0.845 2，虽不及LSK核，但均高出Cauchy Kernel（s =1）和CLA复合核；S_Wang核在精准率上明显高于其他核方法，其平均精度达0.268 96，分别比Cauchy Kernel （s =1）高出18.12%，比潜在语义核（LSK）高出43.09%，比CLA复合核高出9.63%。S_Wang核的综合表现F_1优势明显，高达0.405 9，比Cauchy Kernel（s =1）、潜在语义核（LSK）和CLA复合核分别提高了15.54%、29.87%和7.8%。

5 结束语

本文构造的新核函数S_Wang，用于文本文档的相似检测计算，实验结果表明：S_Wang核的文本相似计算表现优于其他核函数，具有更好的相似度计算性能。适用于文本的相似度计算，如文本信息过虑、文档相似检测、论文抄袭剽窃检测、文本分类等领域。此外，该新核函数在生物信息基因序列比对方面也具有一定的应用前景。

参考文献：

[1] Kernel functions for machine learning applications[EB/OL]. http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html, 2011.

[2] MARTINS A, FIGUEIREDO M, AGUIAR P. Kernels and similar y measures for text classification[A]. Proceedings of ConfTele'2007[C]. New York, USA, 2007. 1-4.

[3] SAHAMI M, HEILMAN T D. A Web-based kernel function for measuring the similarity of short text snippets[A]. Proceedings of WWW'06[C]. New York,USA, 2006. 1-10.

[4] TIAN Y, LI H, CAI Q, *et al*. Measuring the similarity of short texts by word similarity and tree kernels[A]. Proceedings of Conf YC-ICT'10[C]. New York, USA, 2010.69-72.

[5] HOFMANN T. Learning the Similarity of Documents: an Information-Geometric Approach to Document Retrieval and Categorization[M]. Cambridge, MA: The MIT Press, 2000.

[6] 王秀红,鞠时光.基于混合核函数的分布式信息检索结果融合[J].通信学报,2011,32(4):112-118.
WANG X H, JU S G. Result merging method based on combined kernels for distributed information retrieval[J]. Journal on Communications, 2011, 32(4):112-118.

[7] 王国胜.核函数的性质及其构造方法[J]. 计算机科学, 2006, 33(6):171-174,178.
WANG G S. Properties and construction methods of kerne in support vector machine[J]. Computer Science, 2006, 33(6):171-174,178.

[8] 任双桥,魏玺章,黎湘等. 基于特征可分性的核函数自适应构造[J]. 计算机学报, 2008, 31(5):803-809.
REN S Q, WEI X Z, LI X, *et al*. Adaptive construction for kernel function based on the feature discriminability[J]. Chinese Journal of Computers, 2008, 31(5):803-809.

[9] 王华忠,俞金寿.核函数方法及其模型选择[J]. 江南大学学报(自然科学版), 2006, 5(4):500-504.
WANG H Z, YU J S. Study on the kernel based methods and its model selection[J]. Journal of Southern Yangtze University(Natural Science Edition), 2006, 5(4):500-504.

[10] 吴涛,贺汉根,贺明科.基于插值的核函数构造[J].计算机学报, 2003, 26(8):990-996.
WU T, HE H G, HE M K. Interpolation based dernel funct n's construction[J]. Chinese Journal of Computers, 2003, 26(8):990-996.

[11] SHAWE-TAYLOR J, CRISTIANINI N. Kernel Methods for Pattern Analysis[M]. Lodon: Cambridge University Press, 2004.

[12] CRISTIANINI N, SHAWE-TAYLOR J, LODHI H. Latent semantic kernels[J]. Journal of Intelligent Information Systems, 2002, 18(23): 127-152.

[13] BASAK J. A least square kernel machine with box constraints[A]. 19th International Conference on Pattern Recognition[C]. New York: IEEE Press, 2008. 1-4.

[14] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

作者简介：

王秀红（1975-），女，江苏南通人，博士，江苏大学硕士生导师，主要研究方向为信息检索、信息分析、模式识别、专利情报、农业信息系统工程。

鞠时光（1955-），男，江苏南通人，博士，江苏大学教授、博士生导师，主要研究方向为信息安全、信息检索。