

# 一种基于颜色矩和多尺度纹理特征的彩色图像检索方法

杨红菊<sup>1,2</sup> 张 艳<sup>2</sup> 曹付元<sup>1</sup>

(山西大学计算机与信息技术学院 太原 030006)<sup>1</sup> (北京理工大学计算机科学技术学院 北京 100081)<sup>2</sup>

**摘要** 特征提取是基于内容的图像检索的关键步骤,仅基于一种特征的方法只能表达图像的部分属性。由于对图像内容的描述比较片面,缺乏足够的分辨能力,在图像有较大变化的场合常不能取得理想的检索效果。提出了一种基于图像颜色和纹理的图像检索方法,颜色特征采用颜色矩,纹理特征采用小波多尺度高频子带的方差特征。采用组合特征进行图像检索,选择 MPEG 推荐的 ANMRR 方法对两种特征间的权值分配进行分析。实验表明,该方法具有很好的检索性能。

**关键词** 基于内容的图像检索,小波变换,多尺度分析,颜色矩  
**中图分类号** TP391.41 **文献标识码** A

## Color Image Retrieval Approach Based on Color Moments and Multi-scale Texture Features

YANG Hong-ju<sup>1,2</sup> ZHANG Yan<sup>2</sup> CAO Fu-yuan<sup>1</sup>

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)<sup>1</sup>

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)<sup>2</sup>

**Abstract** Feature extraction is one of the key steps in content-based image retrieval, but one feature based approach expresses only partial attributes of an image, which unilaterally describes the image content and is short of enough resolving power. This approach may not achieve ideal result in the case of images varying greatly. An approach was proposed based on image color and texture features, wherein color moments is the color feature and the variances of the multi-scale high frequency sub-bands of wavelet domain. In experiment section, the presented method was compared to the traditional histogram and color moments methods. Combining the two features for image retrieval, we analysed the weight distribution between the two features by using ANMRR evaluation method which MPEG-7 recommends. The experiment shows that the approach has good retrieval performance.

**Keywords** Content-based image retrieval, Wavelet transform, Multi-scale analysis, Color moments

颜色是彩色图像最重要的内容之一,被广泛用于图像检索中。但从图像中提取颜色特征时,很多算法都先要对图像进行量化处理<sup>[1-4]</sup>。量化处理容易导致误检,并且产生的图像特征维数较高,不利于检索。Stricker 和 Orengo 提出了颜色矩的方法<sup>[5]</sup>,颜色矩是一种简单有效的颜色特征表示方法,有一阶矩(均值)、二阶矩(方差)和三阶矩(斜度)等,颜色矩已证明可有效地表示图像中的颜色分布,该方法的优点在于:不需要颜色空间量化,特征向量维数低;但实验发现该方法的检索效率比较低,因而在实际应用中往往用来过滤图像以缩小检索范围。

纹理特征是图像的重要特征之一,其本质是刻画像素的邻域灰度空间分布规律。图像纹理分析是图像分析处理研究中的一个重要组成部分,纹理分析在场景分析、医学图像分析、遥感图像分析等领域有着重要的应用<sup>[12]</sup>。小波分析是一种全新的时、频分析,是信号的时间-尺度分析方法,具有多分辨率分析的特点,近年来越来越多的研究集中于如何借助小波变换进行图像纹理分析<sup>[6,8,11,13]</sup>。

图像的一种特征只能表达图像的部分属性,要更好地描述图像,需要综合多种特征描述图像。为此本文提出一种基于颜色矩和多尺度纹理特征的彩色图像检索方法。首先把图像从 RGB 转化到亮度与颜色分离的 YUV 空间中,求出 YUV 空间中图像的每个分量的三阶矩;其次提取图像的纹理特征;对包含更多能量的 Y 分量图像进行小波塔式分解,从多尺度细节系数矩阵中提取纹理信息,结合颜色与纹理特征进行彩色图像检索。该方法采用的都是统计特征,计算简单,维数不高,利于图像数据库的索引,且不需要对图像进行量化。实验结果表明,该方法具有较好的检索性能。

## 1 图像特征抽取

### 1.1 颜色空间转换和颜色特征抽取

由于 RGB 颜色模型中,每个分量(R,G,B)图像中都包含亮度和饱和度信息,因此两个具有不同亮度的相似颜色利用欧式距离进行相似度计算时,可能有很大的距离。本文采用 YUV 颜色空间。Y 指亮度;U 和 V 分别为 R-Y 和 B-Y 分量,

到稿日期:2008-10-28 返修日期:2009-03-18 本文受“十一五”总装预研项目(513040103)资助。

杨红菊(1975-),女,博士生,讲师,主要研究方向为图像检索等,E-mail:hongju\_yang@126.com;张 艳(1973-),女,博士,主要研究方向为图形图像处理等;曹付元(1974-),男,讲师,主要研究方向为模式识别等。

称为色度,描述了色彩饱和度信息。YUV的优点是它的亮度信号(Y)和色度信号(U,V)相互独立,只需要用U和V两个分量即可表示色彩。RGB到YUV颜色空间的传统转换公式为:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B & (1) \\ U &= 0.492(B - Y) \\ V &= 0.877(R - Y) \end{aligned}$$

图像的颜色矩,即一阶矩(均值)、二阶矩(方差)和三阶矩(斜度)被证明可以很有效地表示图像中的颜色分布,3个颜色矩的数学定义如下<sup>[5]</sup>:

$$\begin{aligned} \mu_i &= \frac{1}{N} \sum_{j=1}^N p_{i,j} \\ \sigma_i &= \left( \frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}} \\ s_i &= \left( \frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}} \end{aligned} \quad (2)$$

其中,  $p_{i,j}$  表示彩色图像第  $i$  个颜色通道分量中灰度为  $j$  的像素出现的概率,  $N$  表示图像中的像素个数。

图像的3个分量Y,U,V图像的前三阶颜色矩组成一个9维直方向向量,即图像的颜色特征表示如下:

$$F_{color} = [\mu_Y, \sigma_Y, s_Y, \mu_U, \sigma_U, s_U, \mu_V, \sigma_V, s_V] \quad (3)$$

## 1.2 多尺度分析与纹理特征提取

### 1.2.1 多尺度分析

1989年, Mallat 提出了多分辨率分析(multi-resolution analysis, MRA)的理论<sup>[7]</sup>。多分辨率分析又称为多尺度分析,主要思想是通过可分离分解滤波器组把一幅图像按照 Mallat 金字塔算法分解为相应多种尺度的小波系数矩阵,每种尺度下包含一个近似系数矩阵和3个不同方向的细节系数矩阵。

本文选用具有紧支集的正交小波基 Daubechies 来构造滤波器。可分离分解滤波器组的结构如图1所述。

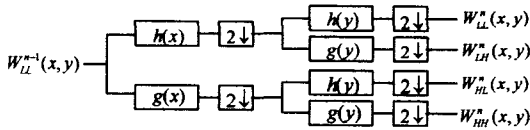


图1 可分离分解滤波器组结构

第  $n-1$  级图像  $W_L^{n-1}(x, y)$  通过滤波器组时就会被分解成4个较低分辨率(较低尺度)的分量,其中包括一个近似系数矩阵(低频图像)  $W_L^n(x, y)$  和3个方向(水平、垂直、对角线)的细节系数矩阵(高频图像)  $W_{LH}^n(x, y)$ ,  $W_{HL}^n(x, y)$ ,  $W_{HH}^n(x, y)$ 。  $h(x)$  和  $g(x)$  分别为低通和高通分解滤波器。  $W_L^n(x, y)$  还可以进一步分解为下一级的4个系数矩阵。图2为  $W_L^{n-1}(x, y)$  图像的一、二级分解示意图。

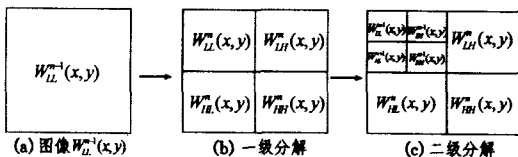


图2 图像的小波分解示意图

多尺度分析方法对图像进行多级分解,有如下优点:效率较高;可以获取不同尺度下图像的信息;变换后的能量大部分集中在低频部分。多尺度方法多用于分析不大于二维的信号。

### 1.2.2 纹理特征提取

对图像进行多尺度后,生成一系列不同尺度的子图像,低频图像往往反映了原图像的主要内容,而高频图像中更多地包含了图像的纹理信息。因此,本文从各级高频图像中获取图像的纹理信息。

具体方法描述如下:

(1) 按照式(1)把一幅彩色图像从 RGB 彩色空间转换到 YUV 颜色空间。

(2) 利用 Daubechies 4 小波基对图像的 Y,U,V 分量分别进行小波3级分解。每个分量图像将得到3个不同尺度的系数矩阵共10个,表示为:  $\{W_{LL}^i, W_{LH}^i, W_{HL}^i, W_{HH}^i, W_{LH}^{i+1}, W_{HL}^i, W_{HH}^i\}$ , 其中包括9个高频系数矩阵。

如果考虑3个分量分解的所有细节系数,从中提取纹理信息,则需要考虑27个系数矩阵。为了用尽可能紧凑的特征来描述图像,则只选用其中的一些系数矩阵进行分析。经过实验发现,各分量图像经小波变换后,能量主要集中在Y分量的小波系数中。

实验过程如下:

子图( $m * n$ 个像素)的能量定义为:

$$E = \frac{1}{mn} \sum_j \sum_i |x(i, j)|^2 \quad (4)$$

以一幅图像I为例,如图3所示。



图3 图像I

计算 YUV 空间中图像的3个分量图像经三级小波分解后10个系数矩阵的能量之和,结果分别为  $1.83 \times 10^9$ ,  $0.17 \times 10^9$ ,  $0.13 \times 10^8$ , 所占的总和的百分比为:90.78%, 8.57% 和 0.65%。本文基于多幅图像进行了相同实验,都有相似的规律。

(3) Y分量集中了原图的绝大多数能量,能够反映图像的大部分纹理信息。所以本文在进行纹理分析时仅仅考虑Y分量图像小波分解后各细节系数矩阵,即提取不同尺度下不同方向9个细节系数矩阵的标准方差来表示原图像的纹理信息。标准方法定义如下:

$$\sigma_k^i = \frac{1}{N} \sum_{x,y} (W_k^i(x, y) - \mu_k^i)^2 \quad (5)$$

其中,  $\mu_k^i$  表示高频细节子图像  $W_k^i(x, y)$  的所有系数的均值,  $N$  为子图  $W_k^i(x, y)$  中系数的个数,  $i$  表示分解级数,  $k$  表示3个方向 LH, HL, HH, 即水平、垂直和对角线方向。

图像的纹理特征可以表示如下:

$$T_{Texture} = \{\sigma_{LH}^3, \sigma_{HL}^3, \sigma_{HH}^3, \dots, \sigma_{HH}^1\} \quad (6)$$

## 2 相似度比较

对于一个特征向量,由于特征内部各个分量的物理意义和取值范围不同,所包含的幅度较大的值会产生很大的偏差。因此需要对它们进行内部归一化,这样在计算相似距离时,可

使各分量具有相同权重。这里采用高斯归一化,其特点是少量超大或超小的元素值对整个归一化后的元素值分布影响不大<sup>[12]</sup>。

将从一幅图像中提取的重要位平面直方图特征向量记为  $F=[f_1, f_2, \dots, f_9]$ , 用  $I_1, I_2, \dots, I_M$  代表图像库中的  $M$  幅图像。图像  $I_i$  对应的特征向量记为  $F_i=[f_{i,1}, f_{i,2}, \dots, f_{i,9}]$ , 根据图像库中的  $M$  幅图像, 就可以得到  $M \times 9$  的特征矩阵  $F=\{f_{i,j}\}$ , 其中  $j=1, 2, \dots, 9, f_{i,j}$  是  $F_i$  的第  $j$  个特征分量。根据式(7), 可以将原序列归一化为  $N(0, 1)$  分布的序列。

$$f_{i,j} = \frac{(f_{i,j} - \mu_j)}{\sigma_j} \quad (7)$$

设查询图像与目标图像分别为  $Q$  和  $I$ , 高斯归一化后的颜色特征向量分别为  $Q=(x_1, x_2, \dots, x_9)$  和  $I=(y_1, y_2, \dots, y_9)$ 。  $Q$  和  $I$  之间的颜色特征距离可以定义为:

$$D_{color}(Q, I) = \sum_{i=1}^9 |x_i - y_i| \quad (8)$$

同样对图像的纹理特征进行内部归一化, 然后用  $L_1$  距离进行比较:

$$D_{texture}(Q, I) = \sum_{i=1}^9 |a_i - b_i| \quad (9)$$

式(9)中的  $a_i$  和  $b_i$  分别对应  $Q$  和  $I$  纹理特征归一化后的各个分量。然后对颜色特征距离和纹理特征距离加权求和, 计算两幅图像间的距离。

两幅图像  $Q$  和  $I$  之间的相似度可以表示为:

$$D(Q, I) = w_1 D_{color}(Q, I) + w_2 D_{texture}(Q, I) \quad (10)$$

其中,  $w_1$  和  $w_2$  分别是对颜色距离和纹理距离的加权值,  $w_1 + w_2 = 1$ 。

### 3 实验与分析

为了评估所提出的方法, 本文使用了包含 1000 幅图像的数据库<sup>[6]</sup>, 共包括 10 类图像, 如 Africa, Beach, Building 等, 每类包括 100 幅图像。

采用 ANMRR (average normalized modified retrieval rank) 对算法进行了性能评价<sup>[10]</sup>。设  $N(q_i) (i=1, 2, \dots, Q)$  表示图像库中与图像  $q_i$  相似的所有图像数目,  $M = \max\{N(q_1), N(q_2), \dots, N(q_Q)\}$ ,  $K = \min\{4N(q_i), 2M\}$ , 设与查询图像相似的图像在检索结果序列中所处的位置为:

$$rank(k) = \begin{cases} k; & k \leq K \\ K+1; & k > K \end{cases} \quad (11)$$

则 ANMRR 定义如下:

$$ANMRR = \frac{1}{Q} \sum_{i=1}^Q \frac{\sum_{k=1}^{N(q_i)} rank(k) - 0.5 - 0.5 * N(q_i)}{K + 0.5 - 0.5 * N(q_i)} \quad (12)$$

由上式可知 ANMRR 的取值越小表明该算法的检索性能越好。除了主观因素, 图像检索的准确率与图像数据库有很大的关系。即使同一算法根据相同的评价方法在不同的数据库中计算出的检索准确率都可能会存在较大的差异。

为了平衡两个权值, 使本文算法具有尽可能高的检索效率, 分别设定了一组  $w_1$  值: 0.1, 0.2, 0.3, ..., 0.9, 进行了实验, 在实验数据库中的每类图像中随机抽取 10 幅图像, 共计 100 个查询图像, 分别就  $w_1$  和  $w_2$  的 9 种组合进行了分析, 得出这 100 次检索结果的 ANMRR 值。表 1 给出了最终结果。

表 1 10 种取值组合下 ANMRR 结果

$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

	0.1	0.9	0.2	0.8	0.3	0.7	0.4	0.6	0.5	0.5
ANMRR	0.4730		0.4669		0.4400		0.4256		0.4128	
	$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$	$w_1$	$w_2$
	0.6	0.4	0.7	0.3	0.8	0.2	0.9	0.1	1	0
ANMRR	0.4048		0.4032		0.4151		0.4376		0.4736	

从表 1 可以看出, 当颜色特征的权值  $w_1 = 0.1$  逐渐增大时, ANMRR 的值逐渐减小, 直到增大到  $w_1 = 0.7$  时, ANMRR 值取得最小值, 从而得到最好的检索结果, 这种现象可以初步反映出颜色矩特征在该图像库的检索中, 比纹理特征能更好地描述图像。此后如果继续增加颜色特征权值, 则 ANMRR 值反而增大, 当  $w_1 = 1$  时, ANMRR 取得最大值, 表示单独用颜色矩特征作为图像特征进行相似度比较, 检索效果不如特征组合法。

上述方法只是对所提出算法的一种综合评价。真实情况是, 图像类别不同, 反映图像主要内容的特征也不同。在实验数据库中的每类图像中随机抽取 20 幅图像作为查询图像进行检索, 得出 20 次查询的 ANMRR 值, 用来研究图像特征与图像类别的关系以及特征之间的权值分配问题。

表 2 显示了部分图像类单独使用颜色矩特征、单独使用多尺度纹理特征、使用特征组合进行检索时检索结果的 ANMRR 值以及在两种特征下距离的最佳加权方式。

表 2 ANMRR 值

	汽车	恐龙	鲜花
颜色矩特征	0.5425	0.0140	0.4856
多尺度纹理特征	0.2966	0.3620	0.0383
最佳特征组合	0.2642	0.0087	0.0130
最佳加权方式	$w_1 = 0.3, w_2 = 0.7$	$w_1 = 0.7, w_2 = 0.3$	$w_1 = 0.2, w_2 = 0.8$
	马	大象	建筑
颜色矩特征	0.3845	0.5193	0.6075
多尺度纹理特征	0.4563	0.6931	0.5234
最佳特征组合	0.2852	0.4939	0.5050
最佳加权方式	$w_1 = 0.7, w_2 = 0.3$	$w_1 = 0.8, w_2 = 0.2$	$w_1 = 0.3, w_2 = 0.7$

可以看出, 有关汽车、鲜花、建筑类的图像包含较多的纹理信息, 所以纹理特征能更好地描述图像, 单独采用纹理特征进行检索, 检索效果比单独使用颜色特征的检索结果好, 当利用组合特征进行图像检索, 距离加权时纹理的权值应高一些; 而恐龙、大象、马类图像, 颜色特征能更好地表示图像, 单独采用颜色特征进行检索, 检索效果比单独使用纹理特征的检索结果好, 因而距离加权时, 颜色特征距离的权值应高一些; 利用特征组合进行检索总体上检索性能要优于单独使用一种特征进行检索。

**结束语** 本文提出了一种基于颜色矩和多尺度纹理特征的彩色图像检索方法, 该方法特征维数低, 计算简单, 不需要对图像进行量化, 比单独使用颜色矩或小波纹理特征进行图像检索, 其效果更好。通过分析特征之间的权值分配方法, 了解到总相似度中各特征的重要程度与图像本身的特点有关。这为下一步研究相关反馈和图像检索中的学习机制奠定了基础。但是本方法不具有旋转、平移等不变性, 同时也没有考虑图像颜色的空间分布信息, 因而如何解决上述的两个问题也将成为下一步研究的关键。

### 参考文献

- [1] Cinque L, et al. Color - Based Image Retrieval Using Spatial - Chromatic Histograms [M]. IEEE Computer Society Washington, DC, USA

- [2] Huang J, et al. Image Indexing Using Color Correlograms [M]. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE): 762-768
- [3] PASS G, et al. Comparing images using color coherence vectors [M]. New York, NY, USA, ACM Press, 65-73
- [4] Swain M J, Ballard D H. Color indexing [J]. International Journal of Computer Vision, 1991, 7(1): 11-13
- [5] Stricker M, Orengo M. Similarity of color images [J]. Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995, 2420: 381-392
- [6] 李明, 吴艳, 吴顺君. 基于小波多通道特征级融合的彩色纹理图像分析 [J]. 光子学报, 2004, 33(8): 999-1003
- [7] Mallat S. A theory for multiresolution signal decomposition: The wavelet representation [J]. IEEE Transactions on PAMI, 1989, 11(7): 674-693
- [8] Kokare M, et al. Texture image retrieval using new rotated complex wavelet filters [J]. Ieee Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2005, 35(6): 1168-1178
- [9] Smith J R, Chang S F. Tools and techniques for color image retrieval [J]. SPIE, 1996, 2670: 426-437
- [10] Muller H, Muller W, Squire D M, et al. Performance evaluation in content-based image retrieval: overview and proposals [J]. Pattern Recognition Letters, 2001, 22(5): 593-601
- [11] 傅蓉, 许宏丽. 基于小波多尺度分析的彩色图像检索方法 [J]. 中国图像图形学报, 2004, 9(1): 1326-1330
- [12] 周明全, 耿国华, 韦娜. 基于内容图像检索技术 [M]. 北京: 清华大学出版社, 2007
- [13] Chun Y D, Seo S Y, Kim N C. Image retrieval using BDIP and BVLC moments [J]. Circuits and Systems for Video Technology, IEEE Transactions on, 2003, 13(9): 951-957

(上接第 241 页)

距离计算, 同时以不选择任何查询扩展算法的原始检索作为对比样本 Base.  $SRTM(W, \theta, L, R, M)$  取  $\theta=0.1, L=5, R=2, M=10$ . 基于不同算法检索结果的查全率与查准率如图 5 所示.

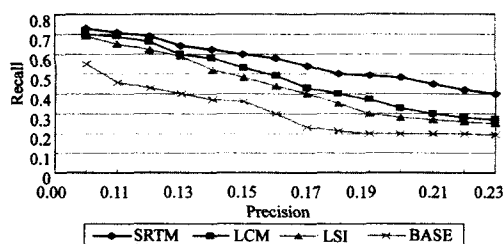


图 5 实验结果对比

由图可以看到, 基于语义关联树的查询扩展算法 (SRTM) 较 LCM 和 LSI 有明显优势, 在查准率大于 0.2 时依然能保持较高的召回率。这一方面是由于引入了基于用户交互的主题选取过程, 使扩展的范围更加精确; 另一方面是因为运用了多个筛选函数, 使扩展词筛选更严格, 避免了噪声的引入。而基于 LSI 的效果甚至比 LCM 模型更差, 这主要是因为 LSI 引入了太多噪声。

假设扩展语料库的规模为  $w$ , 初始查询向量  $W$  的长度为  $i$ , 语义关联树的层数为  $L$ , 每层最大词数为  $M$ , 此时的时间复杂度为  $O(w \times i \times L \times M)$ , 考虑到实际应用中  $W, L, M$  都比较小, 该时间复杂度趋于  $O(w)$ 。因此构建语义关联树获取查询扩展词的时间复杂度较低, 时效性比较高。

实验表明本文提出的语义关联树不但有效降低了词相似度矩阵的工作量, 而且参数可控, 能够根据需求灵活地生成不同的语义空间, 更重要的是它能够弥补传统方法可能引入大量噪声的缺陷, 结合了领域知识也使扩展更有效。

**结束语** 本文提出了一种结合主题选取与局部反馈进行分类查询扩展的方法, 用户输入查询的同时在主题分类语料库中选取与自己意愿最相近的分类, 再结合初次检索回的文档构建语义关联树, 通过筛选函数的严格筛选, 选出最佳的扩展词进行第二次检索, 最终得到扩展结果。通过构造语义关联树有效降低了词相似度矩阵计算的工作量。现阶段用户可根据扩展精度的需求人为设置参数, 灵活高效地生成不同规模的语义空间, 但如何根据不同需求自适应的设置参数还有待于今后进一步研究。根据从网络上下载的 Web 文档按主题进行分类, 建立了主题分类语料库, 但这个语料库还是不完

善的, 今后我们会进行进一步的研究, 取得更准确的领域知识。在网络检索环境下纳入用户个人偏好, 以及搜索结果自动聚类推荐, 提供更准确更个性化的查询扩展结果及检索结果将是下一步研究工作的重点。

### 参考文献

- [1] Furnas G W, Landauer T K, Gomez T K, et al. The vocabulary problem in human-system communication [J]. Communication of ACM, 1987, 30(11): 964-971
- [2] Deerwester S, Dumain ST, Furnas G W, et al. Indexing by latent semantic analysis [J]. Journal of ACM Transaction on Information System, 2000, 18(1): 79-112
- [3] Xu J X, Croft W B. Query expansion using local and global document analysis [C] // Frei H P, Harman D, Schauble P Wilkinson R, eds. Proceedings of the 19<sup>th</sup> Annual International SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1996: 4-11
- [4] Rocchio J J. Relevant Feedback in Information Retrieval. Chapter 14, Prentice-Hall INC, 1997: 313-323
- [5] Xu J X, Croft W B. Improving the effectiveness of information retrieval with local content analysis [J]. ACM Transaction on Information System, 2000, 18(1): 79-112
- [6] Zhang Cheng-qi, Qin Zhen-xing, Yan Xiao-wei. Association-based segmentation for Chinese-crossed query expansion [J]. IEEE intelligent Informatics Bulletin, 2005, 5(1): 18-25
- [7] Qianli Jin, Jun Zhao, Bo Xu. Query expansion based on term similarity tree model [C] // Proceedings of IEEE Natural Language Processing and Knowledge Engineering. 2003: 400-406
- [8] Ponte J, Croft W. A language modeling approach to information retrieval [C] // Proceedings of the 21<sup>st</sup> ACM SIGIR Conference on Research and Development in Information Retrieval. 1998: 275-281
- [9] 赵军, 金千里, 徐波. 面向文本检索的语义计算 [J]. 计算机学报, 2005, 12(28): 2068-2077
- [10] Shen D, Sun J-T, Yang Q, et al. A comparison of implicit links for Web page classification [C] // WWW'06: Proceedings of the 15<sup>th</sup> International Conference on World Wide Web. New York, NY, USA, ACM Press, 2006: 643-650
- [11] 桑艳艳, 刘培刚, 李勇. 基于语义计算的查询扩展的查询优化研究 [J]. 情报学报, 2007, 26(5): 704-710
- [12] Wang Y, Hodges J E. Document clustering using compound words. In IC-AI, 2005: 307-313