第38卷 第24期
Vol.38 No.24

计 算 机 工 程
Computer Engineering

2012年12月
December 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)24—0146—06 文献标识码：A 中图分类号：TP311

# 基于潜在主题的分布式视觉检索模型

陈宜明[1,2,3]，段凌宇[1]，黄 艳[2,3]，李 冰[1,3]，林 杰[1]，黄铁军[1]

(1. 北京大学信息科学技术学院，北京 100871；2. 中国科学院沈阳计算技术研究所，沈阳 110168；
3. 中国科学院研究生院，北京 100049)

**摘 要：**为将文档聚类划分的分布式检索方法直接应用于视觉检索领域，提出一种基于潜在主题的分布式视觉检索模型。给出模型框架，包括图像视觉单词的数据集划分方法和图像子集选择方法，以此优化图像分布式检索性能。实验结果表明，该模型在不降低检索准确率的前提下，能优先选择少量的图像子集进行检索，并提高查询的吞吐量。

**关键词：**分布式检索；视觉检索；词袋模型；图像数据集划分；图像数据集选择；潜在主题

# Distributed Visual Retrieval Model Based on Potential Theme

CHEN Yi-ming[1,2,3], DUAN Ling-yu[1], HUANG Yan[2,3], LI Bing[1,3], LIN Jie[1], HUANG Tie-jun[1]

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;
2. Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China;
3. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

**【Abstract】** In order to use distributed retrieval method for document clustering partition directly in visual search field, this paper proposes a distributed visual retrieval model based on potential theme. This paper gives the model framework, including dataset partition method of image visual words and image subset selection method, to optimize image distributed retrieval performance. Experimental results show that this model searches in a few selected image collections without losing retrieval accuracy and improves the query throughput.

**【Key words】** distributed retrieval; visual retrieval; Bag of Word(BoW) model; image dataset division; image dataset selection; potential theme

**DOI:** 10.3969/j.issn.1000-3428.2012.24.035

## 1 概述

随着互联网技术的发展和应用的普及，在视觉检索(又称基于内容的图像检索)系统面临大规模用户查询时，需要解决单位时间用户查询急剧增加所带来的挑战。运行于单个服务器的视觉检索系统，当面对大规模图像集检索任务，往往由于查询效率有限而无法及时处理庞大的并发请求队列，导致响应延迟或等待，严重影响用户体验。

针对上述问题，本文在视觉检索中引入分布式检索方法。这方面的技术主要有基于子集选择的分布式检索方法[1-2]，文献[2]综合比较了不同子集选择方法的优缺点。基于子集选择的方法主要有集中排序的方法[3]、基于语言模型的方法[4]、基于聚类划分的方法[5]及基于查询的文档划分方法[6]。

分布式检索能够通过减少单个查询所耗费的时间，提高处理并发查询队列的效率。一种广泛使用的分布式检索方法是基于文档均匀划分的分布式检索(Random Document Distributed Retrieval, RDDR)[7]，该方法将文档数据集随机均匀划分成多个子集，每个子集分布到一台服务器上并建立索引。对于每个查询文档，在所有服务器的子集索引上并发查询，以缩短单个服务器上的查询时间，从而加快了单个查询的效率。另一种新颖的分布式检索方法是以文献[5]为代表的基于文档聚类划分的分布式检索(Cluster-based Document Distributed Retrieval, CDDR)方法，该方法使用聚类算法(如 K-means)将相关文档划分到少量的几个甚至一个子集中。对于单个查询文档，首先选择与查询文档相关的子集(子集包含与查询文档相关的文档)，然后在被选择的子集索引上进行检索。与 RDDR 相比，CDDR 仅使用部分服务器处理单个查询请求，其余服务器能够并行处理其他查询请求，进一步提高了检索系统对繁忙查询队列的处理速度。尽管文本检索领域提出了大量关于 CDDR 的工作，但在视觉检索领域几乎没有比较全面的研究方法。

**基金项目：**国家自然科学基金资助项目(60902057)；国家“973”计划基金资助项目(2009CB320902)

**作者简介：**陈宜明(1985－)，男，硕士研究生，主研方向：分布式检索，视觉检索；段凌宇，副教授；黄 艳，副研究员；李 冰、林 杰，博士研究生；黄铁军，教授

**收稿日期：**2012-03-12 **修回日期：**2012-04-22 **E-mail：**lingyu@pku.edu.cn

当前，视觉检索技术普遍采用词袋(Bag of Word, BoW)[8]模型表示图像特征。由于图像视觉特征与文本特征的差异性，直接将 CDDR 算法应用于分布式视觉检索面临新的问题：(1)视觉单词作为一种图像局部特征表达，缺乏语义信息，即语义鸿沟问题。因此，对于图像数据集划分，常用的聚类算法(如 K-means)往往将相关图像分散到较多的图像子集，无法达到依据图像语义划分数据集的目的。(2)背景复杂的查询图像包含大量的噪声视觉单词，采用 BoW 表达很难明确描述图像真实的查询意图。因此，对于图像子集选择过程，仅仅根据视觉单词的统计信息(如单词频率(Term Frequency, TF)、逆向文档词频(Inverse Document Frequency, IDF))无法准确度量查询图像与各个图像子集的相关性，从而导致选择不相关的图像子集进行检索的问题。

为了避免局部特征缺乏语义信息和图像存在背景噪声的问题，引入基于潜在主题的图像表达。假设每个潜在主题对应一个物体或者物体的某部分，每幅图像可以表示成一个或多个潜在主题的组合。其目标是在检索单幅图像时，选择尽量少的相关图像子集，同时，不损失检索准确率。为此，本文提出一种基于潜在主题的分布式视觉检索模型。在数据集划分阶段，将包含同一潜在主题的图像划分到少量的图像子集中。其次，在图像子集选择阶段，给出一种结合局部和全局信息的图像至图像子集的相似性度量方法，用于在视觉单词空间和主题空间分别度量查询图像与各个图像子集之间的相似性。

## 2　基于潜在主题的分布式视觉检索模型

本文贡献主要体现在 3 个方面：(1)本文提出一种基于潜在主题的分布式视觉检索模型。对于查询图像，在保证检索准确率不降低的情况下，仅选择少量相关图像子集检索，有效提升了查询队列的检索效率，及时响应大量的并发用户查询请求。(2)本文的数据集划分策略通过将图像特征投影到潜在主题空间，使用中层语义信息表示图像，将相似图像划分到少数几个甚至一个图像子集中。(3)本文的图像子集选择策略引入基于潜在主题的相似性度量方法，结合基于视觉单词的相似性度量方法，从而有效地确定查询图像的查询意图，降低了查询图像背景中噪声单词的负面影响，解决了图像子集误选的问题，具有较好的鲁棒性。基于潜在主题的分布式视觉检索模型框架如图 1 所示。

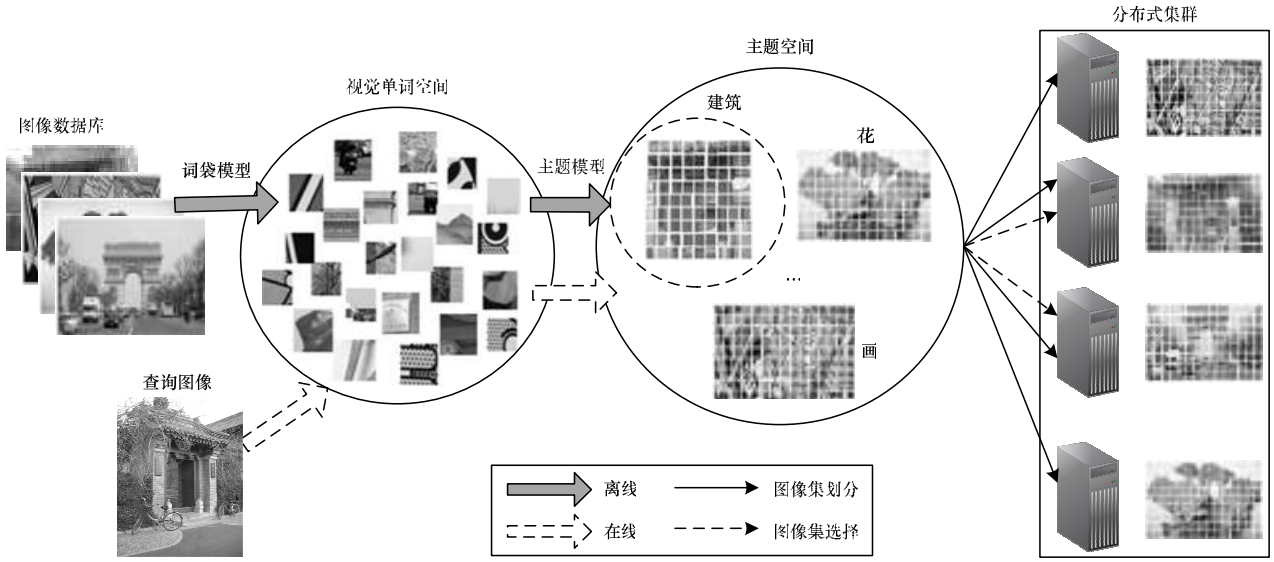



**图 1　基于潜在主题的分布式视觉检索模型框架**

给定图像数据集 $D=\{I_n\}_{n=1}^{N}$ 和查询图像集合 $Q=\{I_q\}_{q=1}^{Q}$，分别包含 $N$ 和 $Q$ 张图像。向量 $\boldsymbol{R}^q=[r_1^q, r_2^q,\cdots,r_N^q]\in\mathbb{R}^N$ 表示第 $q$ 张查询图像 $I_q$ 的相关图像集合；如果 $D$ 中图像 $I_n$ 为第 $q$ 张查询图像 $I_q$ 的相关图像，则 $r_n^q=1$，否则 $r_n^q=0$。

图像数据集划分阶段，将图像数据集 $D$ 划分成 $K$ 个图像子集，每个子集表示成向量 $\boldsymbol{A}^k=[a_1^k,a_2^k,\cdots,a_N^k]\in\mathbb{R}^N$；如果图像 $I_n$ 被分配到第 $k$ 个图像子集，则 $a_n^k=1$，否则 $a_n^k=0$。根据给定查询图像 $I_q$，选择的图像子集表示成向量 $\boldsymbol{S}^q=[s_1^q,s_2^q,\cdots,s_K^q]\in\mathbb{R}^K$；如果图像子集 $\boldsymbol{A}^k$ 被选择，则 $s_k^q=1$，否则 $s_k^q=0$。

分布式视觉检索模型的目标是：给定一幅查询图像，在不降低检索准确率的前提下，选择少量相关图像子集进行检索，被选择的图像子集中应包含尽可能多的相关图像。对于查询图像 $I_q$，假设其选择的各图像子集($s_k^q=1$)所包含的图像($a_n^k=1$)的并集为向量 $\boldsymbol{Q}^q$，向量 $\boldsymbol{Q}^q$ 应满足条件 $\boldsymbol{R}^q\subseteq\boldsymbol{Q}^q$。因此，分布式视觉检索模型定义如下：

$$|\boldsymbol{R}^q|\cong\sum_{k=1}^{K}s_k^q\sum_{n=1}^{N}a_n^k r_n^q \ \text{ s.t. } |\boldsymbol{S}^q|=T \tag{1}$$

其中，常量 $T$ 表示选择的图像子集个数；$|.|$ 表示向量中的非零元素个数。

本文需要求解最优的图像数据集划分策略 $\boldsymbol{A}^k$，以及对每个查询图像的图像子集选择策略 $\boldsymbol{S}^q$。相关图像表达相似的语义信息，如包含相同物体或者物体的某个部分。提出的分布式视觉检索模型通过确定图像是否包含相同

物体或者相同物体的某个部分建立数据集的划分模型和子集选择模型。数据集划分 $a_n^k$ 和图像子集选择 $s_k^q$ 满足如下特性：(1)每个图像子集向量 $\boldsymbol{A}^k$ 对应一个物体或者物体的某个部分；(2)已知查询图像 $I_q$，若 $\boldsymbol{A}^k$ 包含与查询图像 $I_q$ 相同物体或者物体的某个部分，则 $\boldsymbol{A}^k$ 被选择，$s_k^q$=1。

在图像场景分类[9]或者物体识别[10]等领域，潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)模型[11]已被广泛成功应用于挖掘潜在主题，每个潜在主题表达物体或者物体的某个部分。本文首先利用 LDA 模型在大规模的图像库上学习出潜在主题。然后，根据学习获得的图像的主题概率分布和主题的单词概率分布，提出图像数据集划分方法和图像子集选择方法。

### 2.1 基于潜在主题的图像数据集划分方法

图像数据集划分的目标是将相关图像分配到尽可能少的图像子集中。首先，本文利用 LDA 模型学习图像数据集的潜在主题分布；其次，根据潜在主题将图像数据集划分成多个图像子集，每个图像子集具有语义一致性。

已知图像 $I_n$ 用词袋模型表示为集合 $W_n=\{w_j^n\}_{j=1}^{N_n}$，其中，$N_n$ 为图像 $I_n$ 包含的视觉单词数目。LDA 模型将表达由 $I_n$ 生成 $W_n$ 的产生式过程如下：

$$p(D\mid\alpha,\beta)=\prod_{n=1}^{N}\int p(\theta_n\mid\alpha)\Big(\prod_{j=1}^{N_n}\sum_{z_{n,j}}p(z_{n,j}\mid\theta_n)p(w_j^n\mid z_{n,j},\beta)\Big)\mathrm{d}\theta_n \tag{2}$$

其中，$\alpha$ 和 $\beta$ 分别是表示图像的潜在主题概率分布的超参数和单词概率分布的超参数；$\theta_n=\{\theta_n^k\}_{k=1}^K$ 表示图像 $I_n$ 的潜在主题概率分布；$z_{n,j}$ 表示图像 $I_n$ 的第 $j$ 个视觉单词被分配的潜在主题，$z_{n,j}$ 的取值范围为 $z_{n,j}\in\{1,2,\cdots,K\}$。

在获得图像的潜在主题概率分布之后，本文将包含同一潜在主题分布的图像划分到同一图像集合。即已知 $\theta_n$，图像子集向量 $\boldsymbol{A}^k$ 划分方式定义如下：

$$a_i^k=\begin{cases}1 & \theta_{n,k}>\delta \text{ or } k=\arg\max_r\{\theta_{n,r}\}_{r=1}^K\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

其中，$\delta$ 表示阈值常量，根据经验值设定(本文实验中 $\delta=0.5$)。

值得注意的是，CDDR 也是假设按照主题划分，但是本文方法与 CDDR 的数据集划分方式不同。CDDR 采用 K-means 聚类方法划分数据集，每张图像仅属于一个图像子集，而本文方法中每张图像可以属于多个图像集合。一般情况下，单幅图像可能包含多个潜在主题，应该属于多个图像子集。因此，与 CDDR 相比，本文方法更加适用于视觉检索。

### 2.2 基于潜在主题的图像子集选择方法

给定查询图像 $I_q$，图像子集选择的目标是选择与查询图像相关的数据集图像所在的子集。包含相关图像数目越多的图像子集被选择的优先级越高。本文提出一种结合局部与全局信息度量的查询图像至数据子集相似性度量方法。以下分别定义视觉单词空间与潜在主题空间的图像-图像子集相似性度量方法，然后基于局部与全局相似性度量的线性组合设置图像子集的优先级，最后根据优先级选择图像子集向量 $\boldsymbol{S}^q$。

局部相似性度量用于度量查询图像 $I_q$ 和每个图像子集在视觉单词空间中的距离。假设每个图像子集在视觉单词空间中的表示为：向量 $\boldsymbol{U}_k=[p(w_1\mid\boldsymbol{A}_1),p(w_2\mid\boldsymbol{A}_2),\cdots,p(w_M\mid\boldsymbol{A}_k)]$，其中，$M$ 代表视觉词典中单词的个数；$p(w_j\mid\boldsymbol{A}_k)$ 表示图像子集向量 $\boldsymbol{A}_k$ 产生单词 $w_j$ 的概率，$p(w_j\mid\boldsymbol{A}_k)$ 计算公式定义如下：

$$p(w_j\mid\boldsymbol{A}_k)=\frac{c(w_j,\boldsymbol{A}_k)+\sigma}{\mid\boldsymbol{A}_k\mid+\delta\cdot M} \tag{4}$$

其中，$c(w_j,\boldsymbol{A}_k)$ 表示视觉单词 $w_j$ 在图像子集向量 $\boldsymbol{A}_k$ 中出现的次数；$\sigma$ 为平滑因子，防止 $p(w_j\mid\boldsymbol{A}_k)$ 等于零的情况。与文献[10]类似，局部相似性定义为：

$$L_{\mathrm{sim}}(I_q,\boldsymbol{U}_k)=\sum_{w_j\in I_q}\ln p(w_j\mid\boldsymbol{U}_k) \tag{5}$$

$L_{\mathrm{sim}}(I_q,U_k)$ 越大，查询图像 $I_q$ 与第 $k$ 个图像子集向量 $\boldsymbol{A}^k$ 越相关。

全局相似性度量用于度量查询图像和每个图像子集在潜在主题空间中的距离。假设每个图像子集在潜在主题空间的表示为：向量 $\boldsymbol{V}_k=[n_1^k,n_2^k,\cdots,n_M^k]$，其中，$n_j^k$ 表示视觉单词 $w_j$ 被分配到第 $k$ 个图像子集的次数；$n_j^k$ 可以通过视觉单词 $w_j$ 的潜在主题分配 $z_{n,j}$ 计算得到。

$I_q$ 和图像子集 $\boldsymbol{A}^k$ 的全局相似性定义为：

$$G_{\mathrm{sim}}(I_q,V_k)=\frac{\sum_{w_j\in I_q}n_j^k+\alpha}{\sum_{k=1}^K\sum_{w_j\in I_q}n_j^k+\alpha\cdot K} \tag{6}$$

$G_{\mathrm{sim}}(I_q,V_k)$ 越大，查询图像 $I_q$ 与第 $k$ 个图像子集 $\boldsymbol{A}^k$ 越相关。

局部-全局相似性度量本文将局部相似性和全局相似性进行线性组合来度量查询图像与每个图像子集之间的相关性。给定查询图像 $I_q$，全局-局部相似性度量 $Sim_k^q$ 定义如下：

$$Sim_k^q=\rho\cdot L_{\mathrm{sim}}(I_q,\boldsymbol{U}_k)+(1-\rho)\cdot G_{\mathrm{sim}}(I_q,\boldsymbol{V}_k) \tag{7}$$

其中，权重 $\rho$ 的取值范围为 $[0,1]$ 区间(本文实验中 $\rho=0.5$)。$Sim_k^q$ 越大，查询图像 $I_q$ 和图像子集 $\boldsymbol{A}^k$ 越相似，即 $\boldsymbol{A}^k$ 中越可能包含与查询相关的图像集合。

选择图像子集过程中，按照 $Sim_k^q$ 对图像子集排序，查询图像 $I_q$ 的图像子集选择方法 $s_k^q$ 定义为：

$$s_k^q=\begin{cases}1 & RC_k^q\leqslant T\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

其中，$RC_k^q$ 表示第 $k$ 个子集的排序位置。

## 3 实验结果与分析

### 3.1 实验数据集

本文选取 MPEG CDVS 标准数据库中 Zubud 和 ETRI 图像库作为实验数据集。Zubud 和 ETRI 图像库分别标注了查询图像及其相关图像集合。(1)ZuBud 数据集包括对苏

黎世 200 个建筑或场景拍摄的共 1 115 张图像。其中，包含 115 张查询图像，每张查询图像对应平均 5 张~10 张相关图像。(2)ETRI 数据集包括 2 141 张图像，分别来自对 14 个建筑使用 4 个相机从不同角度拍摄的图像。其中，包含 83 张查询图像，每张查询图像平均有 25 个相关图像。

为了验证在大规模数据集下的分布式视觉检索性能，本文选择了一百万的 Flickr 图像数据集[8]作为噪声图像，加入到上述标准数据集的相关图像集合中，增加数据集划分和选择的难度。注意到，Flickr 图像数据集不包含与 ZuBud 和 ETRI 查询图像相关的图像。

### 3.2 实验结果

本文选用 3 种方法与提出的潜在主题式分布式视觉检索模型方法进行对比实验：(1)基于文档均匀划分的分布式检索方法(RDDR)[7]：数据集图像被随机均匀划分到不同的子集中，查询请求需在所有的子集(服务器)上检索。(2)基于文档聚类划分的分布式检索方法(CDDR)[5]：先使用 K-means 聚类方法把图像数据集划分成多个图像子集，对查询图像，根据局部相似性计算查询图像与图像子集的距离，选择与查询图像距离最近的若干图像子集进行检索。(3)整个数据集检索(Exhaustive Search, ES)：没有图像数据集划分过程，查询请求直接在整个图像数据集上检索。

在本文实验中，RDDR 方法、CDDR 方法和基于潜在主题的分布式视觉检索方法都将图像数据集划分成 $K=100$ 个图像子集，分配到 20 台服务器上进行分布式检索，服务器的配置为 2CPU、16 GB 内存。为了验证潜在主题的分布式视觉检索方法的优越性，下面分别给出了分布式检索准确率、图像集划分有效性和图像子集选择有效性的对比试验和分析。

#### 3.2.1 检索准确率实验对比

对于查询图像集，根据返回检索结果列表，本文使用平均准确率(Mean Average Precision, MAP)作为检索性能的度量指标，计算公式如下：

$$mAP@N=\frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{\sum_{r=1}^{N}P(r)rel(r)}{\min(N,\#-relevant-images)}\right) \tag{9}$$

其中，$N\in\{5,10,20,30\}$；$N_q$ 表示图像查询的个数；$r$ 表示排序位置；$N$ 表示返回结果图像的个数；$rel(r)$ 表示第 $r$ 个返回图像是否与查询相关；$P(r)$ 是前 $r$ 个返回结果的准确度；$\#-relevant-images$ 表示与查询相关的图像个数。

本文对比了 4 种方法的平均准确率，除了 ES 以外，其他 3 种方法在图像子集选择时仅选择前 $T$ 个子集。

基于 $T$ 个子集检索的平均准确率越接近于 ES 方法，则表明该方法能够在少量子集中检索时检索准确率的损失越小，即该分布式检索方法越有效。检索准确率结果如表 1 所示，其中，RDDR、CDDR 和本文方法的结果通过在前 $T$ 个图像子集中检索获得，ES 的结果在整个图像数据集上检索获得。对于 $mAP@5$、$mAP@20$ 和 $mAP@30$，CDDR 方法和 RDDR 方法的检索准确率均远低于 ES 方法。而本文方法的检索准确率接近 ES 方法，明显优于 CDDR 和 RDDR 2 种分布式检索方法。值得注意的是，当本文方法仅选择前 5%的图像子集($T$=5)时，检索准确率与 ES 方法相当。从而充分验证了在不损失检索准确率的前提下，本文方法能够选择尽量少的相关图像子集检索，且不明显损失检索准确率，从直观上展示本文方法在检索准确率上的优势。

**表 1 本文方法和 ES、RDDR、CDDR 的检索准确率对比**

| $T$ | $mAP@5$ | | | | $mAP@10$ | | | | $mAP@20$ | | | | $mAP@30$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ES | RDDR | CDDR | 本文 | ES | RDDR | CDDR | 本文 | ES | RDDR | CDDR | 本文 | ES | RDDR | CDDR | 本文 |
| 1 | 0.60 | 0.022 | 0.129 | 0.571 | 0.594 | 0.021 | 0.128 | 0.563 | 0.588 | 0.020 | 0.128 | 0.557 | 0.586 | 0.020 | 0.128 | 0.555 |
| 5 | 0.60 | 0.034 | 0.137 | 0.589 | 0.594 | 0.032 | 0.136 | 0.582 | 0.588 | 0.031 | 0.135 | 0.576 | 0.586 | 0.031 | 0.135 | 0.574 |
| 10 | 0.60 | 0.067 | 0.161 | 0.595 | 0.594 | 0.064 | 0.160 | 0.588 | 0.588 | 0.062 | 0.159 | 0.582 | 0.586 | 0.061 | 0.159 | 0.581 |
| 15 | 0.60 | 0.101 | 0.183 | 0.595 | 0.594 | 0.096 | 0.181 | 0.589 | 0.588 | 0.093 | 0.180 | 0.583 | 0.586 | 0.093 | 0.179 | 0.581 |

#### 3.2.2 图像数据集划分实验对比

为了验证本文图像数据集划分方法的有效性，本文实验对比了基于潜在主题的划分、CDDR 的基于 K-means 划分和 RDDR 的随机均匀划分的 3 种图像数据集划分方法。

对于图像数据集划分，若查询图像的相关图像被划分到越少的图像子集，则该划分方法越有效。为了直观描述图像数据集划分方法的有效性，本文定义一个评价指标：平均比例(Average Proportion of Relevant Image, APR)。首先，计算每个查询图像在各个图像子集中相关图像的数量与其真实相关图像数量的比例(Proportion of Relevant Image, PR)，然后，按照 PR 值从高到低对图像子集进行排序，最后统计所有查询图像排序为 $r$ 的图像子集包含相关图像的平均比例。排序位置为 $r$ 的图像子集 APR 计算公式定义如下：

$$APR(r)=\frac{1}{Q}\sum_{k=1}^{K}\sum_{q=1}^{Q}rel_{k,q}(r)PR_{k,q} \tag{10}$$

其中，$Q$ 表示查询图像的个数；$r$ 表示排序位置；$PR_{k,q}$ 表示 $A^k$ 包含的查询图像 $I_q$ 的相关图像数量与 $I_q$ 的真实相关图像数量的比例；$rel_{k,q}(r)$ 表示图像子集 $A^k$ 按照 $PR_{k,q}$ 排序的位置，若图像子集 $A^k$ 排序在第 $r$ 位，则 $rel_{k,q}(r)=1$。

不同数据集划分方法的平均比例比较如图 2 所示。

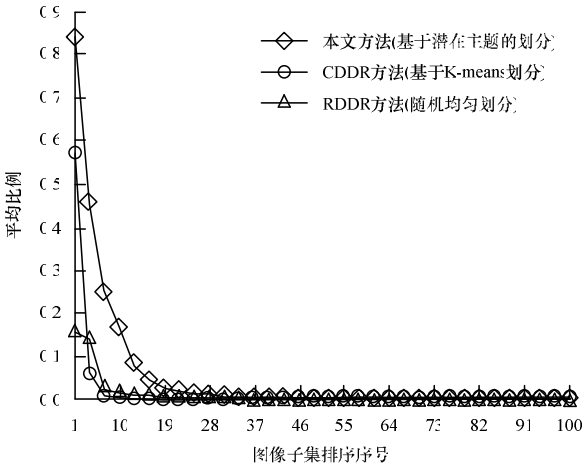



**图 2 不同数据集划分方法的平均比例比较**

APR 曲线越陡峭说明相关图像被划分到越少的图像子集，图像集划分的效果越好。本文划分方法的 APR 曲线的斜率最大，基于 K-means 划分方法的 APR 曲线比随机均匀方法的 APR 曲线陡峭。可见，本文划分方法的图像划分效果最好。具体来说，本文方法能够将 99%的相关图像聚集到 5 个图像子集中，而基于 K-means 划分的 CDDR 方法只能将 70%的图像聚集在 5 个图像子集中。RDDR 随机均匀划分的曲线接近水平，说明相关图像被均匀地分散到所有的图像子集中，不利于分布式视觉检索。

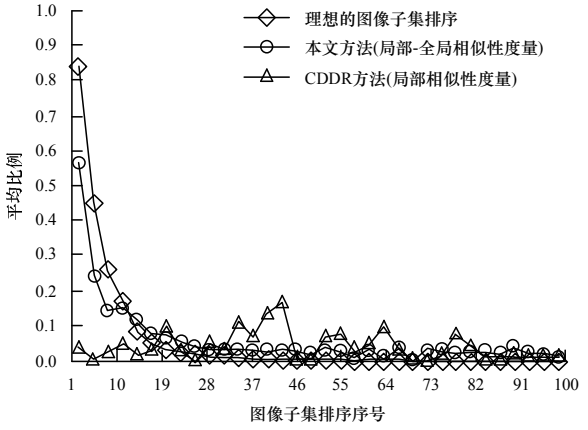
3.2.3 图像子集选择实验结果对比

为了进一步验证本文图像子集选择方法的有效性，实验对比了基于局部相似性度量的 CDDR 方法和本文提出的基于局部-全局相似性度量的子集选择方法。其中，图像集合的划分采用本文提出的基于潜在主题的图像数据集划分方法，并将图像数据集划分成 100 个图像子集。

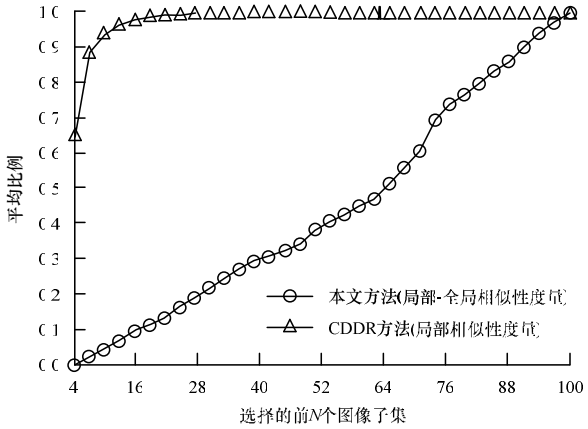
为了评价子集选择方法的有效性，实际上是评价图像子集-查询图像相似性度量有效性。即已知查询图像，按照相似性度量值对子集排序，若 PR 值越高的子集排序越靠前，则该相似性度量越有效。为了定量描述相似性度量能否将 PR 值高的图像子集排在前面，实验将每个查询图像的基于相似性度量产生的图像子集进行排序，然后对于所有查询图像，计算排序序号为 $r(1\sim100)$ 的图像子集的 $APR(r)$值。具体地，已知查询图像 $I_q$，根据 $I_q$ 与各图像子集的相似性度量值对图片子集 $\boldsymbol{A}^k$ 进行排序，再根据排序结果得到 $rel_{k,q}(r)$，然后将所有查询图像的 $rel_{k,q}(r)$ 和 $PR_{k,q}$ 代入式(10)，得到基于相似性度量的 $APR(r)$曲线。

另一方面，为了进一步验证根据相似性度量的子集排序能否接近理想的子集排序，实验比较了理想的子集排序 $APR(r)$曲线与相似性度量的 $APR(r)$曲线之间的差异。其中，理想的子集排序根据图像子集的 $PR_{k,q}$ 从高到低对图像子集进行排序得到。实验按照 $PR_{k,q}$ 对图像子集排序得到 $rel_{k,q}(r)$，再根据式(10)针对所有查询图像得到理想的子集排序的 $APR(r)$曲线。

不同图像子集选择方法的平均比例比较如图 3 所示。




(a)不同子集选择方法各个子集的平均比例比较




(b)不同子集选择方法前 $N$ 个子集的平均比例比较

**图 3 不同图像子集选择方法的平均比例比较**

如图 3(a)所示，本文方法比基于局部相似性度量的 APR 曲线陡峭，因此，相对于基于局部相似性度量的 CDDR 方法，本文方法能够更好地将 PR 值高的图像子集排在前面。而且，基于局部相似性度量的 APR 曲线与和理想图像子集排序 APR 曲线差距较大。可见，基于局部相似性度量的子集选择效果较差，而本文子集选择方法近似于理想图像子集排序的 APR 曲线。

此外，为了进一步证明本文提出的图像子集选择方法的有效性，实验根据相似性度量选择前 $N$ 个图像子集，然后统计被选择的前 $N$ 个图像子集的并集中相关图像数量与真实相关图像数量的比例。如图 3(b)给出的并集 APR 曲线所示，按照局部相似性度量，即使选择了前 30 个图像子集，其图像并集中只包含了 24.08%的相关图像。而根据本文提出的局部-全局相似性度量，只需选择前 10 个图像子集就包含有 96.66%的相关图像，此时局部相似性度量只有近 6.95%的相关图像。

不同图像子集选择方法的查询结果对比如图 4 所示。其中，左侧第 1 列为查询图像，每个查询图像右侧对应一组检索结果，该检测结果从上到下分别对应 3 种不同分布式检索方法，即本文方法、RDDR 方法和 CDDR 方法，由检测结果可知，本文方法要优于 RDDR 方法和 CDDR 方法。

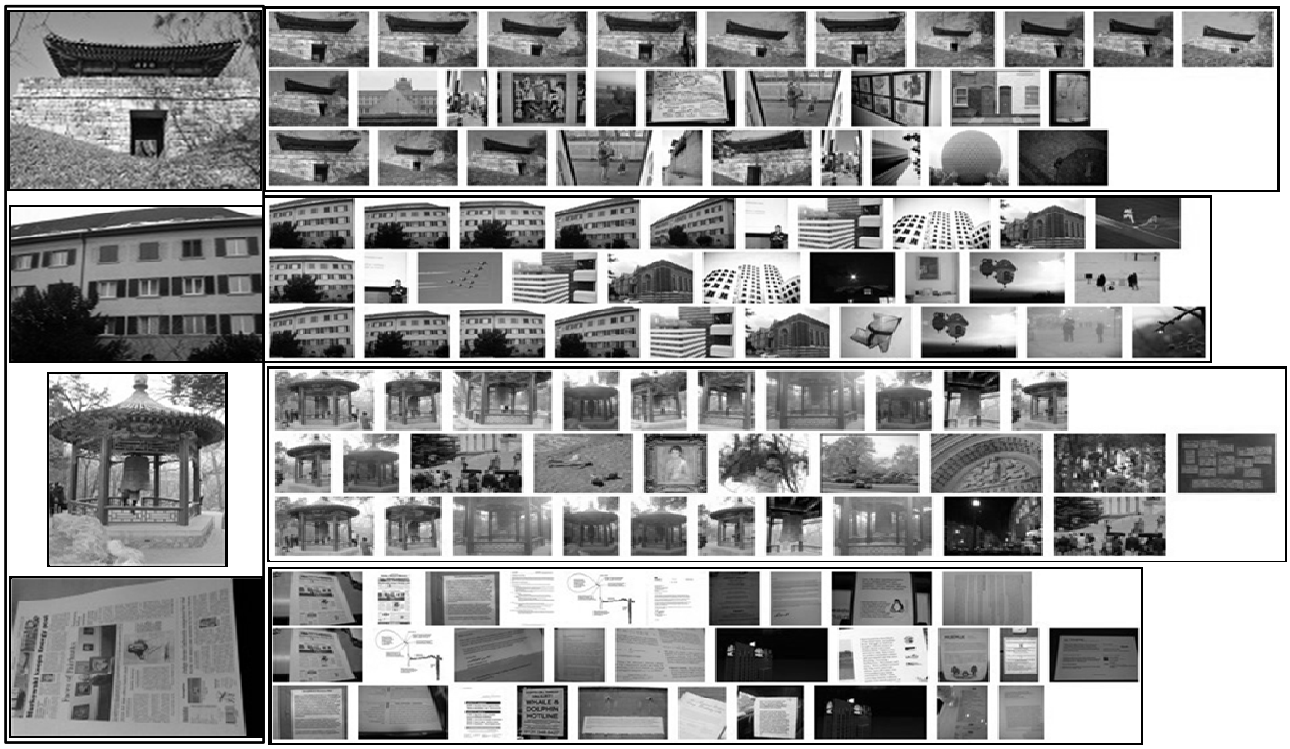


**图 4　不同图像子集选择方法的查询结果对比**

## 4　结束语

本文针对大规模、高吞吐量的视觉检索应用，提出一种基于潜在主题的分布式视觉检索模型，包括图像数据集划分方法和基于查询的数据子集选择方法。对于单个查询图像，本文将检索操作限定在少量的图像子集(服务器)，降低了检索单个查询图像的资源消耗，并且不降低检索准确率，从而提高了分布式视觉检索系统处理并发查询的能力。当前仍然存在的问题是，对于动态更新的图像数据集，新入库图像的视觉语义信息不能保证该图像一定隶属于之前的某一图像子集，因此，导致该图像无法被检索到。为此，今后将研究一种自适应的图像子集更新算法。

### 参考文献


[1] Anagha K, Callan J. Document Allocation Policies for Selective Searching of Distributed Indexes[C]//Proc. of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2010.

[2] Allison L P, James C F. Comparing the Performance of Collection Selection Algorithms[J]. ACM Transactions on Information Systems, 2003, 21(4): 412-456.

[3] Milad S. Central-rank-based Collection Selection in Uncooperative Distributed Information Retrieval[C]//Proc. of the 29th European Conference on IR Research. Berlin, Germany: Springer-Verlag, 2007.

[4] Luo Si, Rong Jin, Jamie C, et al. A Language Modeling Framework for Resource Selection and Results Merging[C]//Proc. of the 11th International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2002.

[5] Xu Jinxi, Croft W B. Cluster-based Language Models for Distributed Retrieval[C]//Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 1999.

[6] Puppin D, Silvestri F, Laforenza D. Query-driven Document Partitioning and Collection Selection[C]//Proc. of the 1st International Conference on Scalable Information Systems. New York, USA: ACM Press, 2006.

[7] Ricardo B Y, Castillo C, Flavio J, et al. Challenges in Distributed Information Retrieval[EB/OL]. (2010-11-21). http://www.citeulike.org/user/ChaTo/article/966384.

[8] 于林森，张田文，张开越. 图像检索中的相似性判别及索引方法综述[J]. 小型微型计算机系统, 2007, 28(2): 356-360.

[9] Bosch A, Zisserman A, Munoz X. Scene Classification Using a Hybrid Generative/Discriminative Approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(4): 712-727.

[10] Philbin J, Sivic J, Zisserman A. Geometric Latent Dirichlet Allocation on a Matching Graph for Large-scale Image Datasets[J]. International Journal of Computer Vision, 2010, 95(2): 138-153.

[11] David M B, Andrew Y N, Michael I J. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(3): 993-1022.


编辑　刘　冰