doi:10.3969/j.issn.1000-436x.2014.12.023

面向互联网的大规模重复图像检索技术研究

王树鹏[1]，陈明[2]，吴广君[1]
(1. 中国科学院 信息工程研究所，北京 100093；2. 郑州轻工业学院 软件学院，河南 郑州 450000)

摘　要：针对互联网上典型的社交媒体应用，提出了一个基于随机投影和分块 DCT 系数的大规模分布式重复图像检索方法。该方法在 Hadoop 集群的基础上，首先利用随机投影映射生成图像签名，再由图像签名高效的检索 HBase 表以获得具有高召回率的候选图像集，最后依赖分块 DCT 系数对候选图像进行进一步过滤来提高检索精度。实验结果表明，对于 1 200 万张微博图像，当 H =2 且 T=150 时，该方法的召回率为 98%，精确率为 93.2%，平均检索时间为 6.7 s。
关键词：社交媒体；随机投影映射；图像签名；分块 DCT 系数；Hadoop 集群
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1000-436X(2014)12-0196-07

Large-scale duplicate image retrieval technical research for the internet

WANG Shu-peng[1], CHEN Ming[2], WU Guang-jun[1]
(1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China;
2. School of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou 450000, China）

Abstract: For the typical social media application on the internet, a large-scale distributed duplicate image retrieval approach based on random projection and the block DCT coefficients was proposed. On the basis of Hadoop, this approach exploited image signatures generated by random projection mapping to retrieve HBase efficiently. And candidate images with high-recall were achieved. Then in order to improve the retrieval precision, the block DCT coefficients were used to further filter candidate images. For 12 million images, experimental results showed that with our approach the recall ratio reached 98%, the precision ratio reached 93.2%, and the average retrieval time was 6.7s when H=2 and T=150.
Key words: social media; random projection mapping; image signature; block DCT coefficients; Hadoop cluster

1　引言

微博的出现标志着个人互联网时代的到来[1]，关于微博多媒体数据(尤其是图像)的研究也已经引起了学术界的广泛兴趣[2~4]。

在开放的互联网上，由于微博图像的复制和转发异常方便，这给用户带来便利的同时，也容易滋生大量安全问题。例如 2013 年 3 月，有媒体援引法国一报告称[5]，美国育种公司孟都山公司的转基因玉米可能致癌，并引用央视截图作为说明以增加可信度，从而引发大量粉丝转载，造成群众性恐慌。然而后经证实，该报道是 2012 年 9 月的旧闻，并且该结论早已被法国生物技术最高委员会和国家卫生安全署先后否认。目前这种利用旧的或者不相关照片在微博上就行话题炒作，散布谣言的行为日益猖獗，因此追踪和验证图像来源就变得非常必要。

追踪和验证图像来源首先需要解决大规模重复图像检索问题。由于数以亿计的用户参与，一张

收稿日期：2013-07-21；修回日期：2013-12-20
基金项目：国家自然科学基金资助项目(61271275，61202067)；国家高技术研究发展计划(“863”计划)基金资助项目(2013AA013205, 2012AA013001, 2013AA013204)；北京市科技计划基金资助项目(Z131100001113034)
Foundation Items: The National Natural Science Foundation of China (61271275, 61202067); The National High Technology Research and Development Program of China (863 Program) (2013AA013205, 2012AA013001, 2013AA013204); Beijing Municipal Science and Technology Project(Z131100001113034)

图像往往会经过缩放、增加水印、转换格式等变化，生成多张内容相同但是形式不同的图像，因此如何在大规模图像集中高效、精确地检索此类图像就成为目前迫切需要解决的问题。现今已有的文献中提出了大量的重复图像发现方法[6~8]，这些方法都能够合理地处理自己设计的场景，但是对于大规模图像检索来说由于计算的复杂性和检索精度的下降，其很难满足实时性和精确性的需要。因此本文的研究目标就是针对大规模微博图像，提出一种高效、精确的分布式重复图像检索方法，该方法在 Hadoop 集群的基础上，分为快速过滤和精确过滤 2 个阶段。在快速过滤阶段利用随机投影生成图像签名，然后根据签名比较来判断图像相似性以获得候选图像集。在精确过滤阶段利用分块 DCT 系数顺序测度方法[6]提取图像特征来进一步判断候选图像是否重复。实验结果表明，在大规模图像集上，当 H =2 且 T=150 时，该方法的召回率为 98%，精确率为 93.2%，平均检索时间为 6.7 s。

2　算法描述

2.1　重复图像定义

一般来讲，“重复”是一种精确术语，即数据的内容完全相同。据统计，完全相同的图像占微博用户上传图像的 77%，此时通过比较语义信息(URL）和特征值(MD5)，可以快速地发现完全相同的图像。然而对于实际的多媒体应用来说，“重复”则更侧重于内容而不是形式，因此其定义还应包括由同一图像或视频经过一组可容忍变换生成的不同副本[9]。不同场景下可容忍变换的定义也不尽相同，经过对微博图像的研究发现，可容忍变换的类型主要有如下 3 种。

1) 尺度变换：包括图像的等比缩放和非等比拉伸。

2) 水印变换：同一张图像上具有不同的水印信息。例如新浪、腾讯等微博服务提供商在用户上传图像时会自动添加各自的水印信息。

3) 存储格式变换：内容相同的图像采用不同的存储格式，如 JEPG、PNG、TIFF 等。

对于如光亮变换、旋转变换等更为复杂的光学和几何操作等由专业图像处理产生的副本，由于其在微博中出现的概率非常小，故不在本文的考虑范围之内。

2.2　快速过滤阶段

大规模图像检索对系统实现提出了较高的实时性和扩展性要求，因此需要一种快速的签名比对方法以满足系统需要，本文利用随机投影算法[10]将图像的分块灰度均值特征转化为二进制图像签名，通过比较签名之间的汉明距离判断 2 个图像的相似性。

随机投影算法能够保持数据之间的相似性，即把高维空间中的向量投影到汉明空间，使高维空间中相似的数据在汉明空间中的二进制编码距离较小。该方法的优点在于二进制签名计算简单快速，而且结构紧凑，占用存储空间小，易于比较和扩展，但是缺陷在于精度不足。

具体的算法流程如下。

1) 提取图像特征

将输入图像统一缩放为 $M \times M$ 的灰度缩略图 K，然后将 K 均匀划分成 n 个图像块，计算每一个图像块 K_i 的平均灰度值，生成图像 K 的块平均灰度特征 $V_K = (v_1, v_2, \cdots, v_n)$。若图像元素的灰度值用 $g(x, y)$ 表示，则式(1）成立。

$$v_i = \frac{n}{M^2} \sum_{x, y \in K_i} g(x, y) \tag{1}$$

2) 生成图像签名

定义 1　随机投影散列 $h(V)$：在 n 维空间中随机选取一个非零向量 $X = (x_1, x_2, \cdots, x_n)$，其中 X 的每一维分量均随机地取自标准正态分布 $N(0,1)$，考虑任意 n 维特征向量 V 与向量 X 之间的夹角，若为锐角，则令 $h_X(V) = 1$，否则 $h_X(V) = 0$。用向量内积∘表示，有式(2)成立，即

$$h_X(V) = \begin{cases} 1, V \circ X \geqslant 0 \\ 0, V \circ X < 0 \end{cases} \tag{2}$$

定义 2　随机投影签名 $sign(V)$：随机产生 f 个 n 维向量 $X_1, X_2, \cdots, X_f$，其中 $f \leqslant n$，然后分别计算 n 维特征向量 V 与 X_i 的内积，若 $V \circ X_i \geqslant 0$，则第 i 位散列值 $h_{X_i}(V) = 1$，否则 $h_{X_i}(V) = 0$，即

$$sign(V) = h_{X_f}(V) h_{X_{f-1}}(V) \cdots h_{X_1}(V) \tag{3}$$

定理 1　随机投影算法生成的二进制签名中 0、1 出现的概率均等且无关[11]。

证明　设特征向量 $V \in R^n$ 为 n 维实空间的向量，则有以下结论。

1) 无关性：令 $S_i = V \circ X_i$，其中 $X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n})$，$\forall x_{i,j} \sim N(0,1)$，则有 $S_i = \sum_{j=1}^{n} v_j \times x_{i,j}$ 且 $v_j \times x_{i,j} \sim N(0, {v_j}^2)$，根据正态分布的可加性，可以

得到随机变量 $S_i \sim N(0, \sum_{j=1}^{n} v_j^2) = N(0, \delta^2)$ ，这表明随机投影算法生成的每个分量都属于期望为 0，方差为 $\sum_{j=1}^{n} v_j^2$ 的正态分布。

对任意2个分量S_p和S_q，其协方差$\mathrm{cov}(S_p, S_q) = E(S_p \times S_q) - E(S_p) \times E(S_q)$，其中，$S_p = \sum_{j=1}^{n} v_j \times x_{p,j}$，$S_q = \sum_{j=1}^{n} v_j \times x_{q,j}$。由于 S_p 和 S_q 属于期望为 0 的正态分布，所以 $\mathrm{cov}(S_p, S_q) = E(S_P \times S_q) = E((\sum_{j=1}^{n} v_j x_{p,j}) \cdot (\sum_{j=1}^{n} v_j \times x_{q,j}))$，由乘法分配律可得 $\mathrm{cov}(S_p, S_q) = E(\sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j x_{p,i} x_{q,j}) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j E(x_{p,i} x_{q,j})$，由于任意的 $x_{p,i}$ 和 $x_{q,j}$ 都相互独立且属于标准正态分布，因此 $\mathrm{cov}(S_p, S_q) = 0$。

2)均等性：由定义 1 可知，$h_X(V) = \mathrm{sig}(S) = \mathrm{sig}(V \circ X) = \begin{cases} 1, V \circ X \geqslant 0 \\ 0, V \circ X < 0 \end{cases}$，其中，sig()代表符号函数，然后由结合 1)中结论 $S \sim N(0, \sum_{j=1}^{n} v_j^2)$ 可以推出式(4)和式(5)，即随机投影签名每一位出现 0、1 的概率均等

$$\Pr(h_X(V) = 1) = \Pr(\mathrm{sig}(S) = 1) = \Pr(S \geqslant 0) = \frac{1}{2} \tag{4}$$

$$\Pr(h_X(V) = 0) = \Pr(\mathrm{sig}(S) = 0) = \Pr(S < 0) = \frac{1}{2} \tag{5}$$

定理 2　随机投影算法生成的 f 位向量签名完全相等的概率 $\Pr(s) = (1 - \frac{\arccos(s)}{\pi})^f$，其中，$s$ 代表 2 个向量的相似度。

证明　设 2 个 n 维向量分别为 U 和 V。

如图 1 可知，若 2 个向量 U 和 V 的夹角为 θ，则随机向量 X_i 只有落在 U，V 的法向量夹角之内才会使得 $h_{X_i}(U) \neq h_{X_i}(V)$，此时对应位签名不等的概率为 $\frac{\theta}{\pi}$。结合定理 1 的无关性，2 个 f 位向量签名完全相等的概率 $\Pr(s) = \left(1 - \frac{\theta}{\pi}\right)^f$，由余弦相似性原理 $s = \frac{U \circ V}{|U||V|} = \cos\theta$，可以得到 $\Pr(s) = \left(1 - \frac{\arccos(s)}{\pi}\right)^f$。

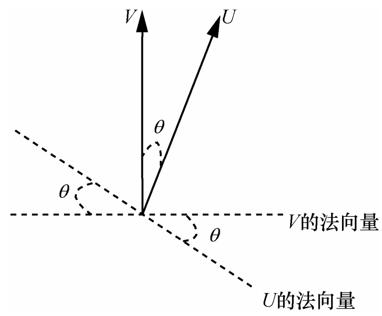

图 1　随机投影

通过定理 2 可以看出，Pr(s)是关于 s 的单调递增函数，即满足高维空间中相似的数据在汉明空间中的二进制编码距离较小。因此相似度越高的向量，其散列值相等的概率也越大，如 s=1 时，$\theta = 0$，则 $\Pr(s) = 1$，即向量签名完全相等。

3) 签名比较

利用上述方法生成的图像签名，本身可能受到噪声的影响，因此为了提高系统的召回率，不仅需要考虑图像签名完全相等的情况，还需要考虑签名相近的情况，在这里认为 2 个图像签名的汉明距离不大于参数 H 时，图像相似。

$$D_{\mathrm{Ham}}(sign(U), sign(V)) = \sum_{i=1}^{f} (h_{X_i}(U) \oplus h_{X_i}(V)) \leqslant H \tag{6}$$

其中，D_{Ham} 代表 2 个图像签名的汉明距离。

2.3　精确过滤阶段

从包含信息的角度来看，原始图像包含了最丰富的内容，但是这些信息是隐藏在图像整体结构中，无法直接表现出来供计算机理解，因此需要提取图像特征来表示图像，但实际上在每一步的特征提取过程中都存在着一定程度的信息损失，从而使其在大规模图像检索过程中难以满足用户的精度需求，这种损失对于感知散列尤为明显，而在大规模数据的背景下，精度不足将导致错误返回的图像数过多，严重影响用户体验。因此一种更为复杂的精确匹配就是必不可少的，本文通过分块 DCT 系数的顺序度量来提高重复图像的检索精度，这一阶段的图像匹配速度较慢，但是由于上一阶段剔除了大量的候选图像，使系统的实时性得到保证。

具体的算法流程[6]如下。

1) 预处理：统一缩放图像尺寸到 64×64，并将其转化为灰度图 K。

2) 分块 DCT 变换：把处理后的灰度图 K 均匀切分成 64 小块，并计算每一小块的平均灰度值，对由

平均灰度值构成的 8×8 矩阵 $\boldsymbol{M}_K$ 进行二维 DCT 变换。

3) 顺序测度：提取变换后矩阵 $\boldsymbol{M}_{DCT}$ 左上角的 32 个 AC 系数作为生成矩阵，然后对其中所有元素进行排序，用排序的结果(即排名)代替原有元素的值以生成新的排序矩阵，如图 2 所示。本文采用顺序测度进行图像匹配的主要原因是：当一个图像包含大量噪声时，分块灰度特征间的典型 L_1、L_2 距离均不具有顽健性，而顺序测度对于灰度变化不敏感，因此具有更好的匹配效果。

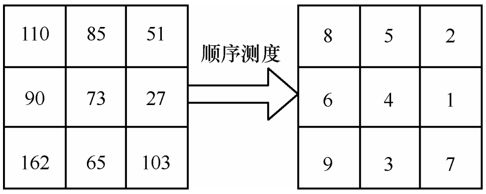

图 2　顺序测度

4) 相似性度量：通过比较排序矩阵的曼哈顿距离与阈值 T 的关系来判断是否重复。

分块 DCT 系数生成的图像特征对于 2.1 节中所列举的变换具有较好的抵抗力。首先，预处理过程中所有图像均被映射为统一大小的缩略图，这样不管图像的尺寸如何变化，其对应的分块灰度均值都保持了不变，因此能够适应尺度变换。其次，离散余弦变换[6](DCT, discrete cosine transform) 是 JPEG(joint photographic experts group)国际标准有损压缩算法的核心部分，利用 DCT 变换可以将图像的能量集中在少数低频 DCT 系数上，这些位于矩阵左上角的低频系数反映了图像的整体信息。而图像的局部失真对图像整体信息的变化影响较小，其分块 DCT 系数中的低频部分几乎不会发生变化，因此对水印变换具有较好的适应性。最后，根据信号原理可知，均值可以更好地减弱随机噪声的影响，因此对于由存储格式变换或者其他原因引起的噪声具有较好的适应性[7]。此外分块 DCT 系数的顺序度量计算简单快捷，可以直接采用随机投影算法的中间结果(每个图像块的平均灰度值)，因此两阶段过滤只需对图像进行一次扫描即可，能够适应大规模图像检索的实时性需求。

3　系统架构

为了保证大规模图像检索的实时性，本文设计一种基于 Hadoop 的分布式系统架构。在该架构中，Hadoop 集群提供分布式并行处理环境，HDFS 作为整个架构的根基用于支撑上层结构，在上层结构中，HBase 主要是用于管理图像信息和图像特征。另外，为了提升 HBase 中获取候选图像集的效率，本文采用布鲁姆过滤器(BF, Bloom filter)[12]剔除部分扩展签名以减少签名定位时间，如图 3 所示。

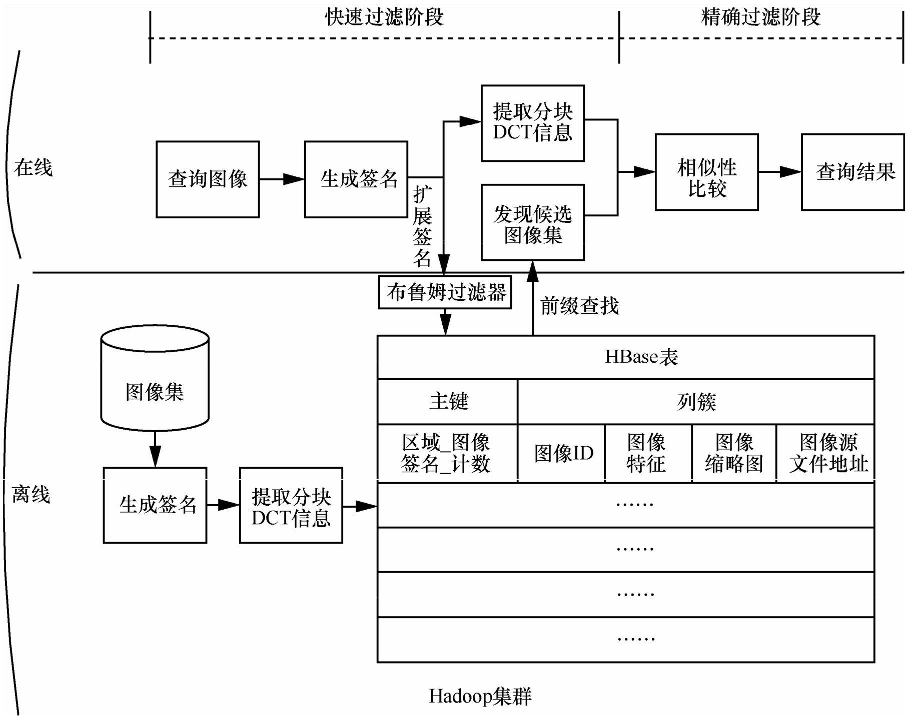

图 3　系统架构

3.1 HBase 的设计

由 2.2 节可知，在快速过滤阶段，为了提高系统召回率，需要将所有与查询图像签名的汉明距离不大于 H 的测试图像作为候选图像集。这里的关键问题是如何能够在可容忍的时间内将查询结果返回，简单的全局扫描因为耗时长，无法满足实际应用需求，因此需要根据应用来设计 HBase 并做出相应的优化。具体的 HBase 表结构如表 1 所示，其中区域代表图像签名的分区号，图像签名采用十六进制表示以提高比较速度，计数是图像个数统计，可以有效避免前缀字段的重复。

表 1 HBase 设计

字段	值
主键	区域_图像签名_计数
$C1$：图像 ID	图像名
$C2$：图像特征	图像的整体 DCT 信息，以字符串的形式存放
$C3$：图像缩略图	源图像的缩略图，以文件形式存放
$C4$：图像源文件地址	源图像在系统中的存放位置，以字符串形式存放

针对 HBase 的设计，将查询图像签名 H 位以内的所有变化组合作为扩展签名，然后根据这些扩展签名进行前缀查找即可，由于不用全局扫描，只是对其中部分数据进行处理，因此能够有效提高查询效率。例如，当 H=2 时，一共需要进行 $C_{32}^{\ 0}+C_{32}^{\ 1}+C_{32}^{\ 2}=529$ 次前缀查找，由此可看出，随着 H 的增大，查找的次数将会急剧增加。

3.2 优化

在进行前缀查找时，首先要确定可能的扩展签名。由于图像签名并不是一种均匀分布，对于同一个签名，可能存在很多图像与之对应，也可能不存在对应的图像。因此为了减少前缀查找次数，本文利用布鲁姆过滤器对扩展签名进行过滤，布鲁姆过滤器是一种用于判断一个元素是否在一个集合中的数据结构，其实质是将集合中的元素通过 k 个散列函数映射到位串向量中，对于每一个元素只需要保存几个比特即可，因此其占用空间较少，可以常驻内存中。在每次前缀查找前，先将该签名与布鲁姆过滤器中的存储情况作对比，若对应位都为 1，说明在该签名上可能存在相应图像，则进行查找，否则说明一定不存在对应的图像，则跳过本次查询，通过布鲁姆过滤器可以有效剔除部分扩展签名，减少磁盘 IO 时间。

4 实验

为了验证本文思想，分别进行如下实验：1) 参数估计，目的是为了提高特征向量的辨识能力，对参数进行最优化选择；2) 算法比较，目的是将本文算法与最新研究进行对比，以验证本文算法的有效性；3) 大规模图像检索，目的是验证大规模图像检索的效果。

为了更接近真实情况，本文从网易、搜狐、新浪等门户网站下载 1 200 万张微博图像作为测试数据。其中，由于微博图像中存在大量重复图像，难于统计，因此对于参数估计和算法比较，本文从中随机选取 23.8 万张图像在单机进行处理以提高结果的准确性，单机的索引结构采用倒排表的形式。对于大规模图像检索的验证，本文以 1 200 万张微博图像为主，通过构建 Hadoop 集群并生成 HBase 表来完成，其中 Hadoop 集群包括 4 台主机，1 台 master，3 台 slaver，主机配置如下：Intel(R) Xeon(R) CPU E5645 @ 2.40 GHz，32 GB 内存，600 GB 硬盘。

此外为了生成图像副本，从 1 200 万张微博图像中随机选取 50 张图像作为查询图像集，然后将查询图像分别进行各种缩放变换和水印变换，其中缩放图像 400 张(缩小 50%、缩小 25%、缩小 12.5%、放大 2 倍、放大 4 倍、放大 8 倍、0.8：0.6 拉伸，1.2：2 拉伸各 50 张)，水印图像 100 张(不同位置的网易水印、新浪水印、腾讯水印、搜狐水印)。通过将生成的图像副本分别与 23.8 万张微博图像和 1 200 万张微博图像合并即可作为不同阶段的测试数据集。

4.1 参数 *H* 的选择

为了提高图像签名的抗干扰能力，本文通过 ROC 曲线来确定参数 H 的选择，ROC 曲线体现了不同阈值 H 下的召回率(recall)和精确率(precision)的关系。如图 4 所示，在相同精确率的情况下，H=2 的召回率最大，此外对未检索到的重复图像进行分析发现，此时影响召回率的主要因素是阈值 T 的选择，继续增大 H，对召回率的影响非常有限，反而导致精确率的下降和检索时间的提高，因此本文选择 H=2。

4.2 阈值 *T* 的选择

阈值 T 的设置影响图像整体 DCT 信息的抗失真能力。当阈值 T 过大时，难以起到区分作用；当阈值 T 过小时，虽然能很好地识别非重复图像，但同时会把部分重复图像当作非重复图像给过滤掉。因此需要根据实验情况选择合适的阈值 T, 如图 5

所示，当阈值 T 在 140~150 之间变化时，精确率和召回率具有一个较好的平衡，而最佳平衡点的取得更为靠近 150，因此选择 T=150。

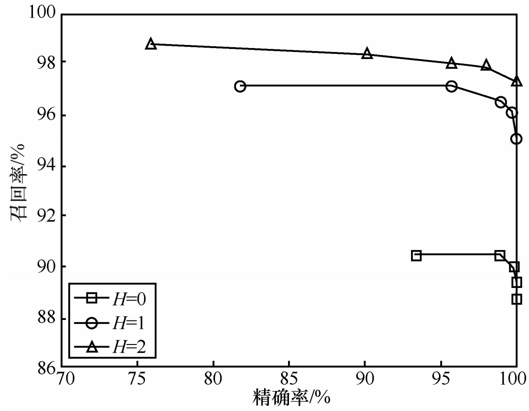

图 4　随阈值 H 变化的 ROC 曲线

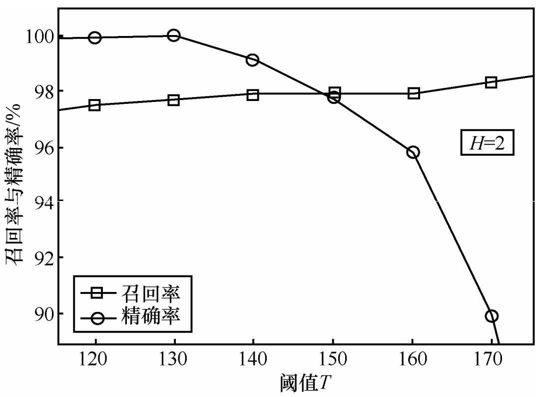

图 5　随阈值 T 变化的 PR 曲线

4.3　算法比较

为了验证本文算法的效果，将其与文献[13]提出的整体 DCT 系数的顺序测度算法进行对比，该算法已被证实对于图像的局部几何失真和缩放变换等具有良好的健壮性。

从图 6 可以看出，本文算法相对文献[13]具有以下 2 个优点：1)在具有相同精确率的情况下，本文算法的召回率最接近于理想值(100%)。2)文献[13]的算法在精确率提升时，召回率急剧下降(尤其是为保证 100%精确时)，而本文算法的召回率虽然也会有所下降，但幅度较小，所以在精确率的表现上更为优秀。这里需要指出，相对于本文算法，整体 DCT 系数的顺序度量对于细节的变化更加敏感，对于精度过滤的阈值 T，如果只是缩放图像，$T \leqslant 100$ 就能满足大部分的需求，但是因为部分水印图像的失真程度更大，不得不放宽阈值，这也是产生错误检索的主要来源，而本文算法的分块特性使其容错性更强。此外在整个实验过程中，当 H=2，T=150 时，在测试图像集中除了生成的 500 张重复图像外，还额外发现 99 张重复图像，这也表明在微博中确实存在大量的重复内容。

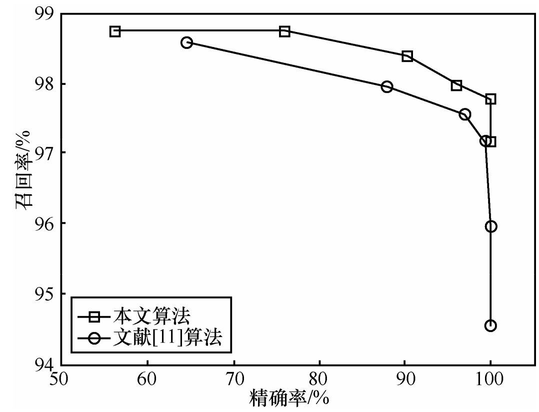

图 6　算法对比

4.4　大规模图像检索

4.4.1　时间效率

在本节实验中，分别测试了不同图像数量下的平均检索时间。如图 7 所示，随着图像数量的增加，平均检索时间接近于线性递增，当图像数量为 1 200 万时，平均检索时间仅为 6.7 s，而且随着集群的横向扩展可以进一步减少平均检索时间。

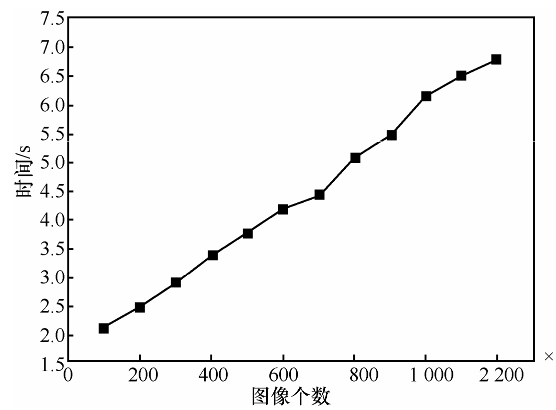

图 7　不同图像数量下检索时间的对比

4.4.2　算法优化

本节分别测试布鲁姆过滤器和多线程(MT, multi thread)对于本文基本算法(BA, basic algorithm)在检索时间上的性能提升。如图 8 所示，对于 1 200 万张图像，基本算法的平均响应时间为 20.3 s，在其基础之上通过添加布鲁姆过滤器，平均检索时间缩短为 16.7 s，效率提高了 17.7%，而且随着数据量的增大，HBase 会将数据分布到各个节点上，在同一节点内

部也会分为不同的区域，因此采用多线程可以有效提高系统的平均检索时间，平均检索时间为 6.7 s。

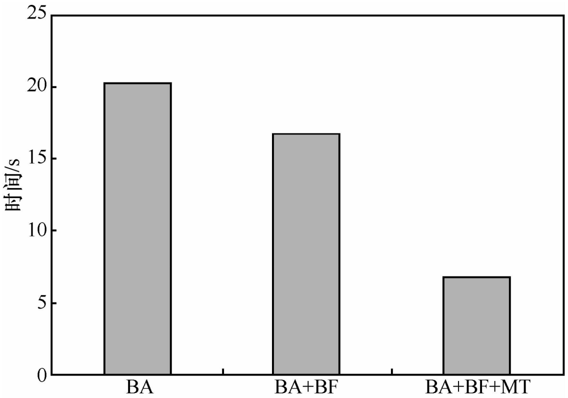

图 8　算法优化

4.4.3　检索结果

在小规模数据集中(23.8 万张测试图像)，当 H=2，T=150 时，单机的召回率为 98%，精确率为 97.8%。随着数据集的扩展，由于手动生成的重复图像总数并未发生改变，且集群模式并不会影响算法的重复图像发现能力，因此大规模图像检索的召回率仍为 98%，精确率为 93.2%，而随着检索到的错误图像逐渐增多，精确率将逐渐下降，这是难以避免的。但是换个角度来看，测试数据从 23.8 万张图像扩展到 1 200 万张图像，图像规模扩展了将近 50 倍，而检索到的错误图像仅从 11 张变为了 36 张，只扩展了 3 倍，这从侧面反映了本文算法具有很高的精确性。

5　结束语

为了追踪和验证微博图像来源，本文提出了一种高效、精确的重复图像发现方法，该方法具有以下几个显著特征：1)利用随机投影算法生成图像签名，使高维空间中的相似数据投影在汉明空间中的二进制编码距离较小，然后在 Hadoop 集群中，通过 HBase 的前缀查找来代替全局扫描以满足系统检索对实时性和扩展性的要求。2)利用布鲁姆过滤器剔除部分不存在的扩展签名，可以有效减少前缀查找次数及磁盘 IO，从而提高系统的平均检索时间。3)从图像整体特征入手，选取分块 DCT 系数中对局部几何失真和缩放不敏感的中低频系数，通过计算其顺序测度作为图像特征向量来保证重复图像的检索精度。实验结果表明，与已有算法相比，本文算法对缩放变换、水印变换和存储格式变换具有较高的精确率和召回率，而且在实时性和扩展性上也能满足人们的需求。未来的研究工作主要是对本文算法的扩充，针对其他重复图像定义，尤其是简单裁剪的重复图像检索进行研究。

参考文献：

[1] http://www.baike.com/wiki/%E5%BE%AE%E5%8D%9A[EB/OL].2013.

[2] SANG J T. Collective search and recommendation in social media[A]. Proc of the 20th ACM International Conference on Multimedia[C]. 2012.1421-1424.

[3] SANG J T, XU C S. Browse by chunks: topic mining and organizing on web-scale social media[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2011, 7S(1):30.

[4] SANG J T, XU C S. Learn to personalized image search from the photo sharing Web sites[J]. IEEE Transactions on Multimedia, 2012, 14(4): 963-974.

[5] http://www.bjnews.com.cn/finance/2013/03/14/252955.html[EB/OL]. 2013.

[6] CHANGICK K. Content-based image copy detection[J]. Signal Processing: Image Communication, 2003, 18(3):169-184.

[7] WANG B, LI Z W, LI M J. Large-scale duplicate detection for Web image search[A]. International Conference on Multimedia Computing and Systems/International Conference on Multimedia and Expo[C]. 2006.353-356.

[8] THOMEE B, HUISKES M J, BAKKER E. Large scale image copy detection evaluation[A]. Proc of the 1st ACM International Conference on Multimedia information Retrieval[C]. 2008.59-66.

[9] JOLY A, BUISSON O, FRELICOT C. Content-based copy retrieval using distortion-based probabilistic similarity search[J]. IEEE Transactions on Multimedia, 2007, 9(2):293-306.

[10] CHARIKAR M S. Similarity estimation techniques from rounding algorithms[A]. Proc of the 34th Annual ACM Symposium on Theory of Computing[C]. Montréal, Canada, 2002.380-388.

[11] YUAN P S, SHA C F, WANG X L, *et al*. C-approximate nearest neighbor query algorithm based on learning for high-dimensional data[J]. Journal of Software,2012, 23(8):2018-2031.

[12] XIE K, WEN J G, ZHANG D F, *et al*. Bloom filter query algorithm[J]. Journal of Software, 2009, 20(1):96-108.

[13] LING H F, XU Z H, ZOU F H, *et al*. A robust image copy detection scheme using ordinal measure of full DCT coefficients[J]. Journal of Computer Research and Development, 2010, 47(10):1812-1822.

作者简介：

王树鹏（1980-），男，山东济南人，中国科学院副研究员，主要研究方向为海量数据存储、网络安全。

陈明（1983-），男，河南驻马店人，郑州轻工业学院讲师，主要研究方向为大数据处理、网络安全。

吴广君（1981-），男，辽宁辽阳人，中国科学院副研究员，主要研究方向为大数据存储与分析、分布式存储。