图像检索中一种有效的半监督主动相关反馈方案

王文胜[1] 陈利科[2] 邬 俊[2]
(鲁东大学 烟台 264025)[1] (大连海事大学 大连 116026)[2]

摘 要 在交互式图像检索中,基于支持向量机(Support Vector Machines,SVM)理论的主动反馈技术扮演着重要角色。然而,现有的SVM主动反馈方法普遍受到小样本问题、不对称分布问题以及样本冗余性等问题的制约。提出两种新颖策略以应对上述问题:(1)针对相关反馈的技术特点,提出了非对称半监督学习框架,该框架采用不同的学习方法为语义相关类和无关类挑选未标记图像,以有效增强SVM的泛化能力;(2)设计了基于代表性度量的主动采样方法,该方法不仅能够从未标记数据中鉴别出富有信息(most informative)图像,而且确保了待标记图像之间具有较大的差异性。实验结果及对比分析表明,所提方案明显优于其它同类算法。

关键词 图像检索,相关反馈,支持向量机,半监督学习,主动学习
中图法分类号 TP391.41 **文献标识码** A

Efficient Semi-supervised Active Relevance Feedback Scheme for Image Retrieval

WANG Wen-sheng[1] CHEN Li-ke[2] WU Jun[2]
(Ludong University, Yantai 264025, China)[1] (Dalian Maritime University, Dalian 116026, China)[2]

Abstract Active learning plays an important role for boosting interactive image retrieval. Among various methods, support vector machine (SVM) based active learning approaches have been drawn substantial attention. However, most SVM-based active learning methods are challenged by small example problem, asymmetric distribution problem, and redundancy among examples. This paper proposed two mechanisms to tackle above problems: (1) designing an asymmetric semi-supervised learning (ASL) framework that exploits unlabeled data for semantic relevant and irrelevant classes in different ways. Under the influence of ASL, the efficiency of SVM is significantly improved; and (2) developing a representative measure based active selection criterion to identify the most informative images from unlabeled data while the diversity among them is augmented. Experimental results validate the superiority of our scheme over several existing methods.

Keywords Image retrieval, Relevance feedback, Support vector machines, Semi-supervised learning, Active learning

1 引言

如何解决底层视觉特征与高层语义概念之间的"语义鸿沟"问题是基于内容图像检索(Content Based Image Retrieval,CBIR)技术能否走向成熟的关键[1]。为了解决这个问题,相关反馈(Relevance Feedback,RF)技术[2,3]被引入CBIR中,其基本思想是利用人对图像的认知能力,从人机交互过程中挖掘用户的潜在意图,进而达到高层语义与底层视觉特征协同检索的目的。在相关反馈过程中,用户可根据其信息需求对当前检索结果做出相关(正样本)或无关(负样本)的判断,然后系统对用户反馈信息进行学习,从而逐步精化检索结果。近年来,人们对相关反馈技术的研究取得了很大进展,并陆续提出了启发式方法、概率密度估计方法以及基于分类和聚类的机器学习方法等[4-10]。在众多方法之中,支持向量机(Support Vector Machines,SVM)以其优秀的学习性能备受青睐,并被广泛用于相关反馈算法的设计。

在相关反馈过程中,用户对检索结果的标记是一个关键环节。但对用户而言,该过程却是一项十分枯燥、乏味的工作。正是在这样的背景下,基于主动学习的相关反馈技术(简称主动反馈)应运而生。主动学习能够从未标记数据中鉴别出富有信息(most-informative)样本(亦称不确定性样本),一旦用户标记了这些富有信息样本,学习系统就可获得最大的信息增益。换言之,主动反馈的目的是利用少量的标记样本获得高效的检索性能,以期减少用户的标记负担[11]。Tong等[7]提出了SVM主动反馈(SVM-AL)算法,该算法的基本思想是寻找那些能够最大限度减小解释空间的样本。他们证明了靠近SVM边界的样本点可近似地达到这个目的,因此,在他们的工作中,将那些靠近SVM边界的未标记图像当作富有信息样本,并提交给用户进行反馈。随后,Brinker等[8]在SVM-AL的基础之上引入了差异性评估策略,以期减少待标记样本之间的冗余性。然而,由于文献[7,8]均采用传统的监督学习策略,故而很难从少量的标记样本中得到准确的

到稿日期:2010-04-16 返修日期:2010-09-29 本文受国家自然科学基金项目(60773033)和国家高技术研究发展计划(863)项目(2009AA01Z211)资助。

王文胜(1968—),副教授,主要研究方向为多媒体技术,E-mail:wshwang2008@tom.com;陈利科(1979—),博士生,主要研究方向为图像处理及信息安全;邬 俊(1981—),博士生,主要研究方向为多媒体数据挖掘与信息安全。

SVM 分类模型。对于训练样本稀缺问题，学者们普遍采用半监督学习范式对 SVM-AL 进行改进。Wang 等[9]在直推式学习框架下使用未标记图像强化分类边界，以增强 SVM-AL 的检索性能。Hoi 等[10]通过求解一个二次优化问题从标记和未标记的混合数据中为 SVM 构造一个核函数，然后利用该核函数从未标记图像中寻找富有信息样本。但遗憾的是，文献[9,10]方法计算复杂度很高，难以适应相关反馈的实时性需求。此外，上述所有 SVM 主动反馈技术均假设正、负样本具有相似的分布特性，并等同地对待它们。但是图像库中的正、负样本数量相差甚远，且负样本又分属于不同的语义无关类别，因此，应采用不同的学习策略处理语义相关类和无关类。

鉴于此，本文针对主动反馈技术中的 3 个问题：小样本问题、不对称分布问题以及样本冗余性问题，提出一种半监督主动相关反馈方案 SemiSVM-AR(Semi-supervised SVM Active learning with Representative measure)。该算法在非对称半监督学习框架下训练 SVM 分类模型，以应对小样本问题和不对称分布问题。此外，SemiSVM-AR 借助聚类分析手段对未标记图像进行代表性度量，以期减少待标记图像之间的冗余性。实验结果及对比分析表明，本文方案明显优于其它同类算法。

2 背景问题描述及 SVM 反馈技术回顾

给定查询样例 q，图像库可视为由两类样本组成，一类在语义上与 q 相关，称之为正样本；另一类在语义上与 q 无关，称之为负样本。本质上，相关反馈属于统计学习中的二类问题。

设 $DB=\{x_i:i=1\cdots n\}$ 表示图像数据库，它由标记图像样本集 $L=\{(x_i,y_i):i=1\cdots nl\}|y_i\neq 0\}$ 和未标记图像样本集 $U=\{(x_i,y_i):i=1\cdots nu|y_i=0\}$ 组成，其中 $n=nl+nu$，$y_i=1$ 或 -1 分别表示图像 x_i 被用户标记为正样本或负样本，$y_i=0$ 表示未标记样本。给定训练样本，系统利用某种监督学习算法构造分类器，并依据分类器的预测置信度得到相关函数 $f_R(x_i)$，然后再把数据库图像按其所对应的相关值进行排序并返还给用户。随着反馈信息的不断累加，分类器不断被强化，进而检索结果不断被更新。

理论上，系统可使用任何监督学习算法构造分类器。值得注意的是，Gosselin 等[12]在同等实验条件下，分析比较了几种常用的学习算法在相关反馈中的性能，包括 Bayesian、SVM、k 近邻以及 Fisher 判别分析等，他们的实验结果显示，SVM 反馈方法的性能明显优于其它方法。鉴于此，本文亦采用 SVM 作为 CBIR 系统的基分类器。在此，我们首先简单回顾 SVM 的基本原理。SVM 的核心思想是寻找一个最优的超平面，能够将正、负样本分开并保证分类边界最大化，从而降低样本的错分风险。一个线性的超平面可表示为 $\langle w,x\rangle+b=0$，我们可通过求解如下优化问题来寻找最优参数 w 和 b。

$$\max_{\lambda\geq 0}\left(\sum_{i=1}^{nl}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{nl}\lambda_i\lambda_j y_i y_j K(x_i,x_j)\right)\text{ s.t. }\sum_{i=1}^{nl}\lambda_i y_i=0 \quad (1)$$

式中，$\lambda_i(i=1\cdots l)$ 是拉格朗日乘子，$K(x_i,x_j)$ 是核函数，它能够将特征空间的样本点映射到 Hilbert 内积空间：$(x_i,x_j)\rightarrow(\phi(x_i),\phi(x_j))=K(x_i,x_j)$。给定核函数，SVM 的决策函数可表示为

$$f(x)=\sum_{i=1}^{nl}\lambda_i y_i K(x_i,x)+b \quad (2)$$

$f(x)$ 的绝对值反映了样本点 x 到超平面的距离，其值直接反映了 x 与查询概念的相关程度，因此许多 SVM 反馈算法将决策函数 $f(x)$ 作为 CBIR 系统的相关函数 $f_R(x_i)$。此外，$Abs(f(x))$ 越大则 SVM 对 x 的预测置信度就越高；反之，$Abs(f(x))$ 越小则说明 x 的不确定性就越大，其中 $Abs(\cdot)$ 表示绝对值计算函数。在 SVM 主动反馈技术中，$Abs(f(x))$ 经常被用于度量未标记样本的不确定性。

3 基于代表性度量的半监督 SVM 主动反馈方案

本文所提出的 SemiSVM-AR 图像检索方案可概括如图 1 所示，该方案主要包括如下 3 个模块。

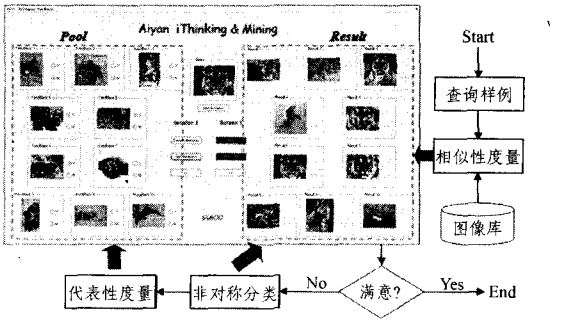

图 1 SemiSVM-AR 图像检索框架

I)初始化检索。用户刚开始检索时，只有查询图像一个训练样本，因此难以构造可靠的分类器。这时可考虑采用传统的相似性度量方法。具体为，利用欧式距离度量数据库图像与查询图像间的相似度，并将相似度最大的前 N 个检索结果返回给用户进行标记，作为初始的训练样本。

II)非对称分类。同时使用已标记样本和未标记样本训练 SVM 分类器，并得到相关函数 $f_R(x_i)$，然后对图像库进行重排序(rerank)。特别是，SemiSVM-AR 采用非对称的方式处理语义相关类和无关类。我们将在 3.1 节详述这种方法。

III)代表性度量。使用聚类分析技术对待标记图像(即学习系统鉴别出的候选富有信息样本)进行差异性评估，从而将代表性的富有信息样本提供给用户进行反馈。具体细节详见 3.2 节。

传统 CBIR 系统的反馈过程直接在检索结果中进行，但是排序比较靠前的图像总要被重复地标记，故而这种方式不能使学习系统获得最大的信息增益。SemiSVM-AR 系统将相关反馈过程从检索结果中分离出来。在每轮反馈过程中(初始化检索步骤除外)，系统将正置信度较高的图像作为检索结果(Result)输出给用户，同时将分类器最不确定的图像置于反馈池(feedback Pool)中等待用户标记，以期减轻用户的标记负担。

3.1 非对称半监督学习框架

如前所述，标准 SVM 难以从小样本数据中获得准确的分类模型。一种可行的做法是借助半监督学习技术改进 SVM 的学习性能。但由于正、负样本间的不对称分布以及实时性需求等问题，大多数半监督学习技术无法直接应用于相关反馈。因此，本文提出一种简单有效的非对称半监督学习(Asymmetric Semi-supervised Learning，ASL)框架，以适应相关反馈的技术特点。

ASL 的核心思想是采用不同的学习策略分别为语义相关类和无关类挑选可信的未标记样本。许多半监督学习方法直接根据分类模型的预测置信度来挑选未标记数据，但该方

法并不适合相关反馈中的学习任务。在反馈过程中(特别是反馈初期),训练样本非常少,故而分类器的性能比较弱。此外,图像库中的正样本远远少于负样本,直接从未标记数据中为语义相关类挑选可信样本是一件极为困难的任务。在本文方案中,ASL 借助 Rocchio 方程式[4]在每轮相关反馈过程中为语义相关类产生一个虚拟的正样本;同时,在处理语义无关类时,该虚拟样本还被用于数据审计。

Rocchio 方程的基本思想是根据用户提供的正、负反馈信息修正查询向量 q,使得修正后的向量 q^* 向着特征空间中正样本集中的区域移动。换言之,q^* 代表的是所有潜在正样本点的质心,故而可将其视为一个置信度极高的正样本。

$$q^*=\alpha q+\beta\sum_{x_i\in L^+}\frac{x_i}{|L^+|}-\gamma\sum_{x_i\in L^-}\frac{x_i}{|L^-|} \tag{3}$$

式中,$L^+\cup L^-=L$,$|\cdot|$ 表示集合的大小,α,β,γ 均为可调常数,用于控制各项之间的比重。关于 Rocchio 方程的参数设置,先前工作已对此进行了详细报道。因此,我们使用经验值 $\alpha=0.3$,$\beta=0.6$ 和 $\gamma=0.3$。此外,整个反馈过程中所有虚拟样本被累计使用,以有效地扩充正样本集。

对于任何给定查询,负样本在整个图像库中所占的比例远远大于正样本。若对图像库进行随机抽样,其得到负样本的概率要远远大于得到正样本的概率。因此,图像库的一个随机子集即是一个候选负样本集。但是,也不排除采样过程中抽到正样本的可能性。为了保证数据质量,本文借助于 Rocchio 方程式对抽样结果进行数据审计。整个语义无关类处理过程可描述如下:

(1)对未标记样本集 U 进行随机采样得到其子集 N_U。

$$N_U=Sampling(U,\sigma_S),|N_U|=fix(\sigma_S|U|) \tag{4}$$

式中,$Sampling(\cdot)$ 表示随机采样操作,$fix(\cdot)$ 表示截断取整,σ_S 表示采样尺度。

(2)借助 Rocchio 方程式对 N_U 进行数据审计。如前所述,q^* 代表的是所有正样本的潜在质心,那么,在特征空间中与之毗邻的样本点是最难确定其归属的样本点。因此我们将这些样本点剔除,以增加数据的可靠性。审计函数定义为

$$\hat{N}=N_U-\{x_i:i=1\cdots r|x_i=\underset{x_i\in N_U}{\operatorname{argmax}}\exp(-\|x_i-q^*\|_2^2)\}$$
$$r=fix(\sigma_C|N_U|) \tag{5}$$

式中,σ_C 表示审计尺度,且 $\sigma_C=\tau\cdot\sigma_S\in[0,1]$,$\tau$ 为常数,用于调节二者之间的比重。值得注意的是,由于候选负样本集 N_U 很容易获得,为了尽可能多地从抽样结果中剔除"问题"数据,本文采用一种比较保守的方法:让审计尺度略大于采样尺度,即 $\tau>1$。此外,被选定的未标记样本仅被暂时当作负样本,而在下一轮反馈过程中,它们又重新被视为未标记数据。通过这种方式,可以抑制误选样本所引起的误差传播。

3.2 基于代表性度量的主动学习

传统的 SVM-AL 仅将不确定性较大的未标记样本返回给用户进行标记,但是,待标记样本之间可能存在较大的冗余性,即待标记样本之间的相似性比较大。在此情况下,由这些样本所张成的解释空间仅占据特征空间中很小的一部分区域,进而可能导致分类器的偏倚性和不稳定性。对于冗余性问题,一种比较合理的方法是通过迭代的方式对所有未标记样本进行差异性评估,以挑选具有代表性的富有信息样本[8]。尽管这种方法被证明是最优的,但其计算复杂度太高,故而无法满足相关反馈的实时性需求。经分析,代表性度量的执行效率主要与待分析数据的规模和所采用的分析手段有关,本文通过采样的方式减小数据规模;同时,采用一种快速聚类分析手段代替传统的差异评估方法,以期快速有效地选择代表性样本。

K-means 算法是一种常用的聚类分析方法。假设待分析数据集为 $\{x_i:i=1\cdots M\}$,K-means 的目标是将数据集分成 K 个簇(设每个簇中心为 $\hat{x}_1\cdots\hat{x}_K$),且保持下列目标函数最小化。

$$D(K)=\sum_{i=1}^{M}\min_{1\leqslant j\leqslant K}(d(x_i,\hat{x}_j))^2 \tag{6}$$

式中,$d(a,b)$ 表示 a 和 b 两点间距离。

由上式不难看出,K-means 需要计算大量的数据点间距离。快速 K-means(FK-means)[13]借助三角不等式确定一个参考下界,从而避免了许多不必要的距离计算。例如,设 x 表示任意数据点,$\hat{x}_b$ 和 $\hat{x}_c$ 表示任意两个簇中心。如果能够预知 $d(x,\hat{x}_c)\geqslant d(x,\hat{x}_b)$,则无需计算 $d(x,\hat{x}_c)$ 的具体值。下述两个引理解释了 FK-means 是如何借助三角不等式达到这一目标的(引理证明可参见文献[13])。

引理 1 给定 $x,\hat{x}_b$ 和 $\hat{x}_c$。如果 $d(\hat{x}_b,\hat{x}_c)\geqslant 2d(x,\hat{x}_b)$,那么 $d(x,\hat{x}_c)\geqslant d(x,\hat{x}_b)$。

引理 2 给定 $x,\hat{x}_b$ 和 $\hat{x}_c$,则 $d(x,\hat{x}_c)\geqslant\max\{0,d(x,\hat{x}_b)-d(x,\hat{x}_c)\}$ 成立。

设 x 表示任意数据点,c 表示 x 所属的簇中心,c' 为其它簇中心。由引理 1 可知,如果 $\frac{1}{2}d(c,c')\geqslant d(x,c)$,那么 $d(x,c')\geqslant d(x,c)$,则无需计算 $d(x,c')$ 的具体值。设 b 表示任意簇中心,b' 表示该中心上轮迭代中的位置。假设上轮迭代中我们已经确定了一个参考下界 l',使得 $d(x,b')\geqslant l'$ 成立。根据引理 2,我们可估计出本轮迭代中所需确定的参考下界 l。

$$d(x,b)\geqslant\max\{0,d(x,b')-d(x,b)\}$$
$$\geqslant\max\{0,l'-d(x,b)\}=l$$

FK-means 通过与参考下界比较的方式避免了大部分的距离计算。标准的 K-means 需要计算的距离数为 nke,其中 n 表示数据点个数,k 表示簇个数,e 表示迭代次数。而 FK-means 所需计算的距离数仅为 n 左右。相比之下,其执行效率有很大幅度提升。鉴于此,本文采用 FK-means 对未标记数据进行代表性度量。

此外,传统的代表性度量方法都是针对所有未标记样本进行分析的。但是,当未标记数据规模很大时,可能会减低 FK-means 的执行效率,因此,本文根据样本点的不确定性对未标记样本集进行下采样,以保证 FK-means 可在一个规模适中的数据集上运行。

综上所述,本文所采用的代表性度量策略可概括如下:

(1)对未标记样本集 U 按其不确定性大小进行降序排序(即 $Abs(f(x))$ 升序,详见第 2 节),并取其前 M 个样本作为待分析数据集。

$$R_{TopM}=\underset{x\in U}{Sort\text{-}in\text{-}Ascent}(Abs(f(x))) \tag{7}$$

式中,集合 R_{TopM} 中存储了 U 中不确定性最大的前 M 个样本。

(2)利用快速 K-means 将 R_{TopM} 聚成 K 个簇。

$$CLU=Fast\text{-}K\text{-}means(R_{TopM}) \tag{8}$$

式中,$CLU=\{cluster_i:i=1\cdots K\}$ 是一个由簇组成的集合。

(3)从每个簇中选择一个代表性样本。

$$Pool=\{x=\underset{x\in cluster_i}{\operatorname{argmin}}(Abs(f(x))):i=1\cdots K\} \quad (9)$$

式中，$Pool$ 表示反馈池，用于存储代表性富有信息样本。

值得注意的是，本文采用聚类分析的目的仅在于降低待标记样本之间的数据冗余，而每个簇的具体含义并不重要，故而无需通过聚类有效性分析等手段确定簇的数目(K 的设置仅取决于反馈池的大小)。折衷考虑有效性和实效性两方面因素，步骤(1)中的 $M=K^2$。

代表性度量确保了训练样本之间具有一定的差异性，从而避免了基分类器的偏倚性和不稳定性等问题。与其他差异性评估方法相比(如文献[8])，本文算法虽然是次优的，但其具有计算简单的特点，能够很好地满足实时系统的需求。

4 实验结果及讨论

实验中所使用的图像全部来自于“Corel Image Gallery”，共包含 50 个语义类，其中每个语义类包含 100 幅图像，总计 5000 幅。我们选用颜色和纹理两种特征来描述图像。颜色特征采用 HSV 空间下的 64 维直方图；纹理特征通过小波变换获得。具体为，首先对图像 YCrCb 空间中 Y 分量进行 3 级小波分解，然后统计每个高频子带的均值和标准差，从而得到一个 18 维的纹理向量。此外，我们采用界定查准率(P@Top$N$)来评价相关反馈算法的性能。其中 P@Top$N$ 表示前 N 个检索结果中被正确检索的图像数与界定范围 N 之间的比值。我们从图像库中随机选取 250 幅图像作为查询图像集来测试相关反馈算法的性能。本文接下来所谈及的查准率均指的是多次测试的平均结果。

为了验证本文算法(SemiSVM-AR)的有效性，我们将其与另外 3 种 SVM 主动反馈方法作对比，它们分别是：标准 SVM-AL 算法[7]、结合差异性评估的 SVM-AL 算法(SVM-AL-Div)[8]和直推式 SVM 框架下的 SVM-AL 算法(TSVM-AL)[9]。所有算法均使用 LIBSVM 软件包[14]解决 SVM 的优化问题，优化过程中使用径向基核函数处理数据的非线性问题。

许多 CBIR 系统要求用户在每轮反馈中标记 20 到 40 幅图像，但该做法并不客观，因为很少有人愿意提供大量反馈信息。SemiSVM-AR 图像检索系统每轮反馈过程中仅向用户提供 10 幅图像(即反馈池的尺寸 $poolsize=10$，亦即 3.2 小节中的聚类数目 $K=10$)，用户只需标记正样本，其它图像由系统自动标记为负样本。我们采用重复验证法来确定采样尺度 σ_S 和审计尺度 σ_C。首先，令 $\tau=1$，即采样尺度与审计尺度相等，$\sigma_S\in\{5\%,10\%,15\%,20\%\}$，图 2(a)给出了本文算法在不同采样尺度下的性能曲线(算法经 1 轮反馈后的 P@Top20)，图 2(b)给出了相应的运算时间。由图 2(a)和(b)可知，当 $\sigma_S=0$ 时，算法性能最弱，这说明使用未标记数据可有效改善算法的检索性能，但随着 σ_S 取值的逐渐增大，算法的性能改善比较缓慢，且运算时间急剧增加。因此，本文将 σ_S 设置为 5%。随后，令 $\sigma_S=5\%$，$\tau\in\{0,0.5,1,1.5,2,2.5,3\}$，图 2(c)给出了本文算法在不同调节因子(即审计尺度)影响下的性能曲线。如图 2(c)所示，$\tau=0$ 时，算法性能最弱，这说明使用数据审计操作有利于检索性能的进一步提升。但 $\tau>1.5$ 时，算法性能趋于稳定，故而本文将 τ 设置为 1.5。

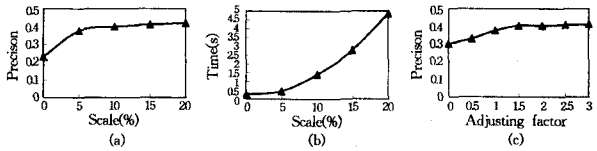

图 2 本文算法在不同的采样尺度和审计尺度设置下的性能曲线

图 3 给出了 4 种算法之间的性能对比(P@Top20，P@Top40，P@Top60 和 P@Top80)。由图 2 可得如下几点结论：(1)两种监督学习方法 SVM-AL 和 SVM-AL-Div 的性能较弱，且 SVM-AL-Div 的性能仅仅略微优于基准线 SVM-AL；(2)相比之下，两种半监督学习方法 SemiSVM-AR 和 SESUSE-AL 的性能明显有所改善；(3)且本文方案 SemiSVM-AR 获得了比传统半监督学习方法 TSVM-AL 更好的检索性能。

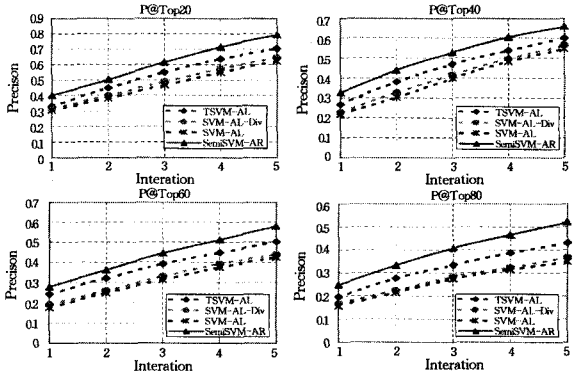

图 3 本文算法与其它几种算法之间的性能比较

结束语 SVM 主动反馈技术是可有效提升 CBIR 性能的重要手段之一，但它受到小样本问题、不对称分布问题和样本冗余性等的制约。本文提出一种新的半监督主动相关反馈方案 SemiSVM-AR，该方案具有如下特点：(1)根据相关反馈自身的技术特点，提出了非对称半监督学习框架，以应对小样本问题和不对称分布问题；(2)借助聚类分析手段对待标记样本进行代表性度量，有效地减少了样本之间的冗余性；(3)所提出的非对称学习框架及代表性度量方法均具有计算简单的特点，与相关反馈技术的实时性需求相一致，具有一定的应用价值。

实验结果表明，本文算法的性能明显优于其它同类算法。在未来工作中，我们欲将所提出算法应用于基于区域的图像检索技术。

参 考 文 献

[1] Datta R, Joshi D, Li J, et al. Image retrieval: ideas, influences, and trends of the new age [J]. ACM Computing Surveys, 2008, 40(2): 1-60

[2] Zhou X, Huang T S. Relevance feedback in image retrieval: a comprehensive review [J]. Multimedia Systems, 2003, 8(6): 536-544

[3] 徐建军，吴玲达. 基于内容图像检索中的相关反馈技术发展[J]. 计算机科学，2004，31(7)：200-202

[4] Rocchio J J. Relevance feedback in information retrieval[M]. Salton G. The SMART System. New York: Prentice-Hall, 1971: 313-323

[5] Ves E, Domingo J, Ayala G. A novel Bayesian framework for relevance feedback in image content-based retrieval system [J]. Pattern Recognition, 2006, 39: 1622-1632

[6] Chen Y X, Wang J Z, Krovetz R. CLUE: cluster-based retrieval of images by unsupervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 14(8): 1187-1201
[7] Tong S, Chang E. Support vector machine active learning for image retrieval [C] // Proc. ACM Int. Conf. Multimedia. Ottawa, Canada: ACM Press, 2001: 107-118
[8] Brinker K. Incorporating diversity in active learning with support vector machines [C] // Proc. Int. Conf. Machine Learning. Washington, DC: AAAI Press, 2003: 59-66
[9] Wang L, Chan K L, Zhang Z. Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval [C] // Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition. Madison, Wisconsin: IEEE Press, 2003: 629-634
[10] Hoi S C H, Rong J, Zhu J K, et al. Semi-supervised SVM batch mode active learning and its applications to image retrieval [J]. ACM Transactions on Information Systems, 2009, 27(3): 1-29
[11] Huang T S, Dagli C K, Rajaram S, et al. Active learning for interactive multimedia retrieval [J]. Proceedings of IEEE, 2008, 96(4): 648-667
[12] Gosselin P H, Cord M. Active learning methods for interactive image retrieval [J]. IEEE Transactions on Image Processing, 2008, 17(7): 1200-1211
[13] Elkan C. Using the triangle inequality to accelerate k-means [C] // Proc. Int. Conf. Machine Learning. Washington, DC: AAAI Press, 2003
[14] http://www.csie.ntu.edu.tw/~cjlin/libsvm

(上接第 194 页)

在 4 个工作线程条件下，11.5Mb，56Mb，115Mb 和 226Mb 对应的平均解析总时间分别是 543ms，2323ms，4312ms 和 10810ms，反映了并行时，执行时间与数据大小接近线性比例，说明该方法有较好的扩展性。

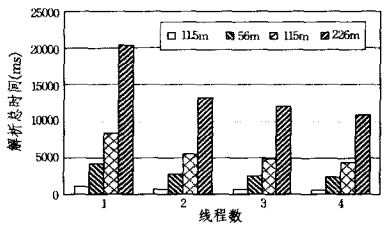

图 4 解析执行时间对比

通过解析执行时间计算加速比。以原有串行处理方法的解析时间 T_1 为基准，在 n 个线程下的解析时间为 T_n，则这时的加速比 $S_n = T_n / T_1$。对不同文档大小，4 个工作线程条件下，加速比接近 2。由于多线程子树构建过程中，存在并发存取问题，在大数据量条件下尤为突出，因此目前未进行数据加载优化，总体加速比未达理想状态，但反映了实际工作情况。

此外，我们测试了各种数据在多线程条件下各阶段时间的分布情况。图 5 给出了在 4 个工作线程下并行解析各阶段的耗时比例。比如 115Mb 数据的并行解析情况，其中并行子树构建的时间占总时间的 83.6%，而子树合并工作只占 13.9%，数据分片操作约占 2.5%。其他不同线程数量条件下的情况类似，子树构建的执行时间占大部分，这说明对其并行化处理的重要性。

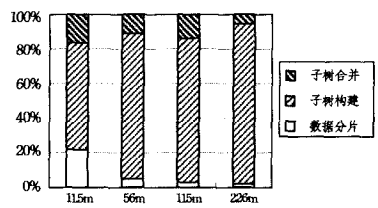

图 5 各阶段耗时比例

结束语 由于 XML 解析性能极大地影响着 XML 应用的整体性能，故 XML 解析优化是个研究重点。随着多核环境的普及，并行化是一个自然而有效的优化选择。相对现有的几种并行 XML 解析方法，本文的方法不需要进行耗时的预处理，而是通过对 XML 文档任意分片，对各数据分片并行处理，构建局部文档子树，最后合并成全局文档树。其关键是对耗时最多的分片解析进行数据并行。通过多核环境下的性能测试验证了整个方法的有效性。

未来工作考虑通过优化设计提高加速比，比如采用数据预读取方法避免并发存取竞争，考虑为每个片段处理串行加载文档的对应部分，以减少文档存取操作对各片段并行解析的影响，实验表明这样串行加载的时间代价小。另一方面，通过合理控制任务粒度，获取较好的负载均衡效果。在工作线程较多的条件下，考虑把 XML 数据划分成更小的片段，采用工作窃取[12]的调度模式使负载均衡容易实现。

参 考 文 献

[1] Matthias N, Jasmi J. XML parsing: a threat to database performance [C] // International Conference on Information and Know-ledge Management, Proceedings. 2003: 175-178
[2] Akhter S, Roberts J, Reinders J, et al. Multi-core programing: increasing performance through software multi-threading [M]. Hillsboro: Intel Press Business Unit, 2004
[3] Lu K, Zhu Y, Sun W, et al. Parallel processing XML documents [C] // Proceedings of International Database Engineering and Applications Symposium. 2002: 96-105
[4] Lu W, Chiu K, Pan Y. A parallel approach to XML parsing [C] // 7th IEEE/ACM International Conference on Grid Computing. 2006: 8-16
[5] Pan Y, Lu W, Zhang Y, et al. A static load-balancing scheme for parallel XML parsing on multicore CPUs [C] // Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07). 2007: 356-365
[6] Pan Y, Zhang Y, Chiu K, et al. Parallel XML parsing using meta-DFAs [C] // 2007 3rd IEEE International Conference on e-Science and Grid Computing. 2008: 237-244
[7] Pan Y, Zhang Y, Chiu K. Simultaneous transducers for data-parallel XML parsing [C] // Proceedings of the 2008 IEEE International Parallel & Distributed Processing Symposium. 2008: 1143-1154
[8] Pan Y, Zhang Y, Chiu K. Hybrid parallelism for XML SAX parsing [C] // Proceedings of the IEEE International Conference on Web Services, ICWS 2008. 2008: 505-512
[9] Bruno N, Koudas N, Sprvastava D. Holistic twig joins: optimal XML pattern matching [C] // Proceedings of SIGMOD. 2002: 310-321
[10] Timothy M, Beverly S, Berna M. Patterns for parallel programming [M]. Massachusetts: Addison Wesley/Pearson, 2004
[11] CWI. An XML benchmark project [CP/OL]. http://www.xml-benchmark.org, 2009
[12] Lea D. A Java fork/join framework [C] // Proceedings of the ACM 2000 conference on Java Grande. 2000: 36-43