一种新的基于改进聚类检索算法的CBIR系统研究

张志强
(苏州大学计算机科学与技术学院　苏州 215006)

摘　要　随着多媒体技术、网络技术的飞速发展，各种各样的信息爆炸式增长，导致人们对信息检索的要求越来越迫切。本文提出了一个新颖的基于内容的图像检索模型，还提出了一种新颖的适合于图像数据的聚类检索方法，通过自动更新簇特征的权重来提高聚类准确率，从而改进检索效果。通过文中的仿真实验说明，文中提出的检索模型，以及采用的相关算法是可行的，取得了良好的检索效果。

关键词　聚类，CBIR系统，图像检索

New CBIR System Based on the Improved Clustering Image Retrieval

ZHANG Zhi-qiang
(School of Computer Science and Technology, Suzhou University, Suzhou 215006, China)

Abstract　Along with the arrival of multimedia and information time, people can get more and more information. We describe a novel CBIR model which pays attention to both vision feature and semantics. Also, we propose a novel idea of improved clustering which can improve the result of image retrieval. Validity of our thoughts has been testified by the experiments.

Keywords　Clustering, CBIR system, Image retrieval

1　引言

传统的CBIR系统的检索是按照底层特征相似度的大小排序来返回结果，往往排在前面的图像中有一些虽然底层视觉特征相似度很大，但是其语义内容与检索图像却是风马牛不相及。也就是说，底层特征上的相似性只能在一定程度上表示语义上的相似性，这就造成了所谓的语义鸿沟。本文主要针对语义鸿沟问题，提出了一个可行的系统模型。在此模型中采用了改进的聚类算法(Improved Cluster 简称 IC)，并且将语义检索与底层视觉特征的检索相结合，使得图像检索的准确率有所提高。

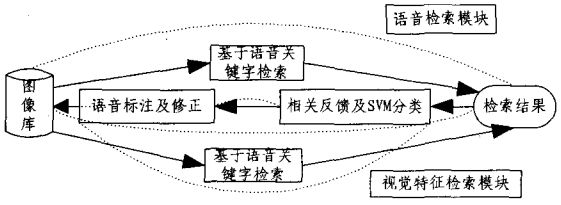

图1　CBIR检索模型

本文所提出的CBIR模型涵盖了图像检索与自动语义标注的内容。分为两种不同的情况，进行基于视觉内容与基于语义关键字的两种不同的检索方式。其基本结构如图1所示。用户进行查询需要提交语义关键字与示例图像，根据由示例图像对没有该关键字标注的数据库图像进行基于视觉特征子模块的检索，根据关键字对已经有该关键字标注的数据库图像进行基于语义关键字的检索，最后将二者的检索结果按照某种策略一并返回给用户，最后是相关反馈以及语义的标注和修改。

张志强　讲师，研究方向为计算机应用与网络。

一种情况是视觉特征检索模块(图2)，针对库中没有进行语义标注的图像：采用先聚类检索然后相关反馈的方法，即上图的视觉特征检索模块。基于改进聚类算法的视觉特征检索模块 Improved Clustering Content-Based Image Retrieval 简称为 IC-CBIR。IC-CBIR 的结构图如图2所示。

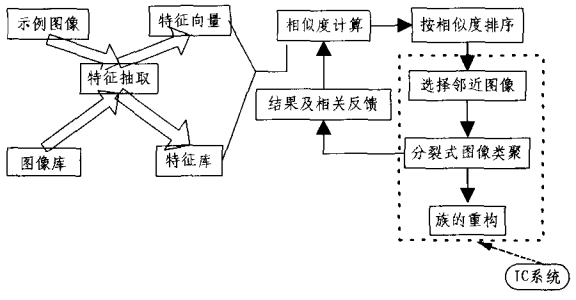

图2　基于视觉特征的检索模块

另外一种情况是语义检索模块(图1)，针对已经有语义标注图像。这种情况的检索比较简单，直接按照某个语义的置信度排序返回结果就可以了，类似于普通的基于关键字的检索。

2　聚类算法

2.1　IC算法概述

将聚类应用于CBIR能够取得良好的效果，改进的聚类检索算法IC是本文的核心思想之一。首先采用基于分裂的方法聚类，利用图表示，借助于经典图算法中的图分割的方法，将邻近图像按照特征相似度分为 n 簇，并计算出各簇的中心图像作为簇的代表。然后计算簇内各个特征对该簇的影响

因子，按照不同特征的影响因子为其赋不同的权重。分别以各簇的簇中心图像为核心，按照调整了特征权重的图像距离公式计算各图像与簇中心的距离。基于最小距离准则对图像进行聚类，从而达到簇的重构，生成语义上更为相似的图像簇。下面是与算法相关的2个定义：

定义1(簇特征影响因子) 簇内各图像在某一个特征上的相似度之和。某一特征的相似度之和越大，该特征的影响因子就越大。与之对应的分别为簇颜色影响因子 S_c、簇纹理影响因子 S_t、簇形状影响因子 S_s。S_c 的求解见公式(1)，S_s，S_s 同理。

$$S_c=(\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}S_c(i,j))/2 \quad (i\neq j,n\text{ 为簇图像数目},S_c(i,j)\text{为}i,j\text{ 在颜色上的相似度}) \tag{1}$$

定义2(簇密度(P)) 用簇密度来描述簇分割的难易程度，由簇内各图像相似度 Image Similarity(IS)之和除以簇图像数目得到。

$$p=(\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}IS(i,j))/2n \quad (i\neq j,n\text{ 为簇图像数目},IS(i,j)\text{为图像}i,j\text{ 的相似度}) \tag{2}$$

2.2 IC聚类检索算法描述

输入：图像特征向量 $V(c,t,s)$，其中，c,t,s 分别表示颜色、纹理和形状特征。

输出：与查询图像 Query Image(QI)语义上相似的图像簇的簇内图像。

步骤：

Step1 令公式(3)中的系数 $a=b=c=1$，计算待查询图像 QI 与其他图像 j 间的图像相似度 $IS(q,j)$。

$$IS(i,j)=a*Sc(i,j)+b*St(i,j)+c*Ss(i,j) \tag{3}$$

Step2 按照第一步求得的 IS 值排序求得待查询图像 QI 的 m 个邻近图像 NI。

Step3 对 NI 用图 $G(V,E)$ 表示，将图像及图像间的关系分别表示为图的顶点 V 和边 E。计算 $IS(i,j)$ 并把它作为边 (i,j) 的权重。

Step4 采用递归的正规化图分割算法 Ncut(文献[1])对图进行分割直到簇密度 P 大于事先规定的某个值或是簇的数目超过了事先给定的界限为止，从而初始聚类结束形成 n 个临时簇(分裂的层次聚类)。

Step5 对每个临时簇 C_k，选取与簇内其他图像相似度之和最大者作为簇中心图像 $r[k]$。

Step6 分别对每个临时簇 C_k 计算其各个特征影响因子 $S_C[K]$，$S_S[K]$，$S_s[k]$，按照簇内各个特征影响因子的大小，调整簇特征权重 $a[k]$，$b[k]$，$c[k]$ 的值。

$$IS(r,j)=a[k]*Sc(r,j)+b[k]*St(r,j)+c[k]*Ss(r,j) \tag{4}$$

Step7 以簇 C_k 的簇中心图像 $r[k]$ 作为聚类中心，按照调整后的权重重新计算 $IS(r,j)$(见公式(4))。基于最小距离准则进行聚类，生成新的簇。

Step8 重新计算新簇的代表图像，与查询图像匹配，返回与查询图像 QI 最相似的簇，代表图像所在的簇作为检索结果。

2.3 图表示以及图分割

一系列图像的集合用赋权无向图 $G(V,E)$ 来表示。其中 $V=\{1,2,3,\cdots,n\}$ 为顶点，代表单张图像；$E=\{(i,j)\mid i\in V,j\in V\}$ 为边，代表图像之间联系，边 (i,j) 赋权值为 $IS(i,j)$ 来表示图像间的相似度 IS。

由于使用了图表示方法，图像在特征空间上的聚类问题自然而然地转变成为图分割问题。将图分割的算法应用于图像检索，可以取得良好的效果，前人已经进行过此类研究。许多学者给出了图分割的近似最优解算法。在此，采用了 J. Shi 的 Ncut 算法，对图进行正规化的递归分割。首先建立赋权无向图并计算节点间的相似度作为节点间的权值，然后递归地对图进行二分，直到达到中止条件。Chen Yixin 的 CLUE 系统同样采用了 Ncut 算法，但是其 CLUE 系统的分割中止条件是簇数目达到一定的界限。这就不可避免地造成对大图的分割，从而导致聚类的结果往往是大小差不多的一些簇，而这些在现实中是不合理的。实际情况是可能存在亲密度关系大而图像数多的簇，同样也有图像数少但是其簇亲密关系小的簇。本文采用簇密度(P)与簇的数目作为算法中止的条件。当簇密度都大于给定的某个值或簇数目达到一定的界限时停止分割，而下一步分割选取的是密度最小的簇。这样便会得到语义上更加合理的结果。图的分割过程如图3所示(椭圆框内的数字为簇图像数目)。在第二步分割时存在两个簇：B和C，由于簇密度 0.36<0.41，故首先分割密度为 0.36 的簇B。

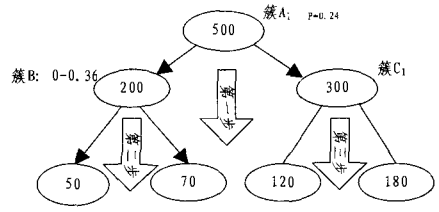

图3 图分割的过程

2.4 中心图像的计算及簇的重构

而对于簇 C 内中心图像 j 的选择，选取与簇内其他图像相似度之和最大者，公式如下：

$$\max_{j\in c_1}\sum_{i\in c_1}Is(i,j) \tag{5}$$

虽然它并不能完全代表该簇的语义，但是这种表示方法还是能够体现簇的大致内容。在找到更好的描述方法之前，本文仍采用此简单办法。

利用图表示和图分割的方法对一系列图像的集合在一定的特征空间内分成了 n 个大小不等的簇，但这些簇不能完全代表语义相似的图像的集合。因为不同的语义簇对不同特征的敏感度不一样，而不敏感的图像特征却往往造成了图像的错误归类。如何选择适当的特征及权重来构造特征空间，对聚类分析显得尤为重要。在这些方面也有很多相关的研究，有基于遗传算法的，也有基于相关反馈的。但是此类方法要么是静态的特征选择，其泛化能力差；要么带有人为的主观色彩，且需要人的大量参与。IC 则不然，实现了动态的特征选择，每一个簇按照在各个特征上的相似度的大小重新调整更加符合语义信息的特征权重，然后以初次聚类的簇代表图像作为新簇的聚类中心，基于最小距离准则，生成新的簇，实现了簇的智能重构，能更好地体现簇图像在语义上的一致性。

3 仿真实验及其结果

该检索算法适用的前提是：利用特征空间的初次聚类能够得到语义上大体一致的图像类别。因为是在聚类的基础上进行的图像检索，所以聚类的好坏直接影响了图像检索的准

(下转第264页)

动的效率。

需要说明的是关联关系不等于因果关系，一个关联规则只是说明在前提为真的情况下，结论为真的概率较大。

4 实现技术

在软件缺陷管理工具中，一元和多元数据分析可按以下方式实现：提取出所有统计类型的缺陷属性，将其名称以列表的方式显示在用户界面上，用户可选择一个或多个属性以及对应的分析指标和结果的显示方式，系统根据用户的选择进行相应的分析并显示分析结果。

上述缺陷数据分析方法的实现涉及到以下两个关键技术。

许多分析结果需要以柱状图、饼图和线图等可视化的形式呈现给用户，虽然可使用开发平台提供的底层图形 API(如 Java 的 2D 和 3D 图形 API)来实现分析结果的可视化，但实现方法较为复杂，且需要开发者熟练掌握图形编程技术。更为常用的方法是使用现有的图形软件包如 JFreeChart，WebChart3D 等来实现图形输出，这些软件包提供了丰富的统计图形绘制功能，可根据分析结果在普通用户界面或 Web 页面上生成各种图形。

缺陷属性关联规则的分析需要调用关联规则挖掘算法，实现这种算法要求具有一定的统计分析和数据挖掘知识，但开发者也可以借助一些已有的工具来实现该算法。例如，新西兰怀卡托大学开发的“Weka”就是一个功能全面的数据挖掘工具，它用 Java 语言实现了几乎所有常用的数据挖掘算法，而且提供了调用各种算法的 API，使开发者可以将该工具中的算法嵌入到自己的应用程序中[5]。

结束语 除本文所介绍的方法以外，软件缺陷数据分析还有许多分析指标和分析方法，如软件可靠性模型的建立，使用线性回归和贝叶斯网等技术预测缺陷数量等。实际上，有关软件开发过程、技术和资源的许多问题都与软件缺陷密不可分，软件组织在项目进展过程中可能会提出各种各样的缺陷度量要求，缺陷数据的分析也常常与缺陷之外的度量数据相联系，因此缺陷分析工具应具有很好的可扩展性标准的数据接口，以便于根据需要增加新的数据分析功能。在今后的工作中，我们将更为全面地研究缺陷分析方法，并开发功能更强、更为简便实用的缺陷分析工具。

参 考 文 献

[1] Dalal S, Hamada M, Matthews P, et al. Using Defect Patterns to Uncover Opportunities for Improvement // Proceedings of International Conference on Applications of Software Measurement (ASM). San Jose, CA, , February 1999: 15-19

[2] Menzies T, Lutz R, Mikulski C. Better Analysis of Defect Data at NASA // Proceedings of SEKE2003, the Fifteenth International Conference on Software Engineering and Knowledge Engineering. San Francisco, July 2003

[3] Bridge N, Miller C. Orthogonal Defect Classification-Using Defect Data to Improve Software Development. Software Quality, 1998, 3(1): 1-8

[4] 毛国君，段丽娟，王实，等. 数据挖掘原理与算法. 北京：清华大学出版社，2005

[5] Witten I H, Frank E. 数据挖掘-实用机器学习技术(英文版). 第2版. 北京：机械工业出版社，2005

(上接第 228 页)

确率。只要说明不同的语义图像簇其聚类的底层特征空间是不同的，即不同图像簇中各个特征的影响因子是不同的，前人通过相关反馈来修改特征权重以及我们此处的实验都证明了这一点。此处改进的聚类检索使得类内相似度高，类间相异度高，达到良好的聚类效果，从而 IC-CBIR 就能够取得较好的检索准确率。

表 1 特征相似度平均值

类别 \ 相似度及权重	颜色 C	文理 T	形壮 S
类 1 鸟(100)	0.387 (25.8%)	0.701 (46.7%)	0.412 (27.5%)
类 2 蓝天海洋(100)	0.754 (46.6%)	0.548 (33.9%)	0.316 (19.5%)
类 3 建筑(100)	0.522 (31.7%)	0.649 (39.4%)	0.475 (28.9%)

表 2 聚类结果分析

类别 \ 相似度及权重	图像数目	普通聚类匹配数目	改进聚类匹配数目
簇 1	100	67	83
簇 2	100	74	69
簇 3	100	59	72

本文以从 http://www.wang.st.psu.edu/IMAGE 下载的包含 10,000 幅测试图像的图像库为实验素材。其中图片规格为 128 * 85 或 85 * 128。首先从图像库中人为地选出 3 类图像，每一类 100 张。为了减弱个人的主观因素，采用多人选取取交集的办法。然后计算类内图像在颜色、纹理、形状三个特征上的相似度。这里类的特征相似度是通过类内所有图像特征相似度的平均值，其结果见表 1，聚类的结果分析见表 2。

由表 1 可以看到，在不同的语义簇内，各个特征所占的比重不是同样大的，这就说明了提取关键特征空间、计算簇特征影响因子的必要性。虽然初次聚类形成簇的特征影响因子与语义上完全正确的簇的簇内特征影响因子并不完全相同，但它们大致接近。由表 2 的聚类准确率可以看到，二次聚类的准确率为 74.7%，而按照统一特征空间的普通聚类的准确率只有 66.7%。这样，新簇的簇内图像语义上更加接近，图像的检索效果也就更好。而且改进的聚类使得一幅图像可以同时属于两个或是多个簇，更加符合语义的模糊性。

结束语 随着多媒体技术、网络技术的发展和需求的带动下，对于它的研究方兴未艾。本文旨在缩小或解决语义鸿沟，提出了一种新颖的 CBIR 检索模型，将语义关键字与底层视觉特征相结合来进行检索。而语义关键字的产生是由用户的相关反馈得来的，具有较高的准确性。这样，用户对图像库检索的次数越多，检索准确率和检索速度也会越好。

参 考 文 献

[1] 温超，耿国华. 基于内容图像检索中的“语义鸿沟”问题. 西北大学学报(自然科学版)，2005

[2] 夏定元. 基于内容的图像检索通用技术研究及应用. 博士学位论文. 2004

[3] 张莉，周伟达，焦李成. 核聚类算法. 计算机学报，2002

[4] 黄元元. 基于视觉特征的图像检索技术研究. 博士学位论文. 2003