利用频繁模式挖掘进行图像标注*)

周 祥 周向东 周浩峰 王智慧 汪 卫 施伯乐
(复旦大学计算机与信息技术系 上海 200433)

摘 要 在基于内容的图像检索与计算机视觉研究领域中,如何将底层的视觉特征与高层的语义信息相联系,即如何有效地根据图像的底层特征提取其表达的语义概念是备受关注的难题之一。特别是当图像包含了多个语义概念时,问题就变得更为棘手了。本文中,我们提出一种基于图像底层特征值频繁模式的语义概念标注方法,针对图像分块的特点实现了一组有效的模式挖掘算法,并设计了标注规则的生成算法。权威的真实数据集上的实验表明我们的方法在对含有多个语义概念的图像进行概念标注时要比之前的一些算法效果更好。

关键词 基于内容的图像检索,语义概念,频繁模式,多概念标注

Using Frequent Pattern Mining for Image Labeling

ZHOU Xiang ZHOU Xiang-Dong ZHOU Hao-Feng WANG Zhi-Hui WANG Wei SHI Bai-Le
(Department of Computing and Information Technology, Fudan University, Shanghai 200433)

Abstract One major challenge in the content-based image retrieval and computer vision research is to relate low-level visual features with semantic concepts, that is, to extract semantic concepts from an image effectively according to its low-level visual features. Especially when images contain more than one concept, the problem will be even more intractable. In this paper, we provide a method to label an image based on the frequent patterns of its low-level visual features. According to the specialty of image segmentation, effectivealgorithms are implemented to mine such patterns and to generate labeling rules. It is shown in the experiments on authoritative and real datasets that our method is more effective than some previously proposed methods for labeling images containing multiple concepts.

Keywords Content-based image retrieval, Semantic concepts, Frequent patterns, Multiple concept labeling

1 简介

基于内容的图像检索是当前多媒体数据库应用中的重要问题,其技术关键是如何自动识别图像所表达的内容。图像内容的分析可分为三个层面:第一层面是指分析图像的视觉特征,如颜色,形状,纹理等。第二层面是指分析图像所包含的语义对象,如"老虎","台灯"等。第三层面是指分析图像所表达的主题,如"浴室","厨房"等。在第二,三层面所分析的对象被称为高层语义概念。

在图像内容的第一层面上,已经产生了一些很好的方法进行分析(如文[1~3])。但是如何将图像底层的视觉特征与高层的语义概念相联系,即如何有效地根据图像的底层特征提取其表达的语义概念是当前备受关注的热点。特别是针对包含多个语义概念的图像的处理方法是其中的难点。

本文所要解决的问题就是当一个图像包含多个语义概念的情况下,基于图像底层的视觉特征值,自动识别出图像所包含的高层语义概念,我们称之为图像的(多概念)标注问题。

1.1 相关工作

文[4~8]提出了各自的方法以识别图像包含的高层语义概念。它们大都是有监督的机器学习方法,主要的思想都是构建一个精确的分类器(以每一个语义概念作为一个分类),然后利用分类器进行语义信息标注。这类方法具有一定的分类准确性,但是其构建的分类器的可解释性不强。

频繁模式挖掘是近年来研究的热点,主要目的是从数据集中找到所有频繁出现的项集。它是关联规则挖掘的前提,后者是为了发现大量数据中项集之间的有趣的关联。

对于频繁项集挖掘,Apriori[9]是最有影响力的一种算法。Apriori算法的主要代价在于计算项集在事务中出现的次数,后来的很多算法都基于此进行了改进,提高了算法的效率[10,11]。2000年提出的FP-growth算法[12]采取分制策略,将数据集压缩到一棵fp-tree上,在不产生候选项集的情况下生成所有的频繁项集。

利用图像视觉特征中的频繁模式构建标注规则对图像进行标注,一方面可以直观地建立起底层特征到高层语义之间的联系,规则的解释性较强,分类器的语意更明确。另一方面,频繁模式挖掘是基于离散的模型,相较连续模型降低了问题的复杂度,减小了过度拟合的可能性。同时相比原始信息频繁模式中包含了更多的语义内容,利于提高标注准确性。文[13]提出了用在正例中频繁而在反例中罕见的模式(emerging pattern),构建分类器,但这种方法只是针对一个样例只有一个类标签的情况。但对图像而言,一般一幅图像会出现多个语义概念。文[14]讨论了当图像包含多个语义概念时的情况,提出决定性特征模式(decisive feature pattern),它对为多个语义概念所共有的特征模式进行了进一步过滤,使得每个概念的决定性特征模式对于识别这个概念而言更具区分性。

1.2 本文的主要思路及贡献

本文就图像的多概念标注的问题在一个经过专家人工标

*)基金项目:国家自然科学基金项目(60403018)。周 祥 硕士研究生,主要研究方向为数据挖掘;周向东 副教授,主要研究方向为多媒体数据管理等;周浩峰 博士,主要研究方向为数据挖掘;王智慧 博士研究生,主要研究方向为数据库,数据挖掘;汪 卫 教授,博导,主要研究方向为数据挖掘,安全数据库,XML等;施伯乐 首席教授,博导,主要研究方向为数据库与知识库等。

注过的训练图像集上，寻找对于标识某一语义概念而言唯一的或者是关键的决定性表征模式，并依此根据效果制导的原则自动生成标注规则，对新的测试图像中包含的语义概念进行识别。

本文根据当前具有代表性的图像处理技术，在基于分块后的图像上，扩展了决定性特征模式的概念，设计了有效的挖掘算法，并基于找到的特征模式设计了自动生成标注规则的算法，使得算法更具现实意义，并通过实验验证，我们的算法与其他同类算法相比，在效果方面表现优良。

1.3 论文的结构

第2节给出了问题的定义，第3节就给定的问题提出了解决的步骤，并对每个步骤进行了详述。第4节通过在真实数据集上的性能分析验证了算法的实效性。最后是算法的小结与对未来的展望。

2 问题定义

设有训练图像集为 Sample$=\{S_1, S_2, \cdots, S_n\}$，$n$ 为图像个数。$S_i=\{S_i.F, S_i.C\}$，$i=1, 2, \cdots, n$，$S_i.F$ 表示第 i 幅图像的视觉特征，$S_i.C$ 表示第 i 幅图像上的标注集合(图像所包含的语义概念的集合)。而每幅图像的视觉特征是分块进行提取的。令 $S_i.F=\{S_i.B_1, S_i.B_2, \cdots, S_i.B_{mi}\}$，$i=1, 2, \cdots, n$ 表示第 i 幅图像被分成 mi 个块。$S_i.B_j=\{S_i.B_j.f_1, S_i.B_j.f_2, \cdots, S_i.B_j.f_v\}$，$j=1, 2, \cdots, m_i$，表示第 i 幅图像在第 j 个分块上提取出的图像底层视觉特征值集合，而 v 则是特征的维数，$S_i.B_j.f_k$，$k=1, 2, \cdots, v$ 表示第 i 幅图像的第 j 个分块在第 k 维特征上的取值。我们用 $c=\underset{S\in Sample}{Y} s.c$ 表示概念(全)集。通过在 Sample 上学习后对于测试图像集中的图像 $S=\{S.F, S.C\}$，能自动给出标注 Cl，使得 Cl 与 $S.C$ 尽可能相符。

之所以假定图像的视觉特征是在分块后提取的，是因为能够标识语义概念的特征应该在包含这一语义概念的图像区域中表现得比较突出。本文使用一种有效的分块方法[3]，其分出的每一个块中图像的视觉特征比较一致。更重要的是从直观上讲，特征一致的块很可能只包含一个语义概念，而包含两个或更多不同的语义概念的可能性比较小。当然它也可能只是包含一个语义概念的一部分。本文基于 Duygulu[5] 提供的一个经过分块处理后的比较完整也较为权威的真实图像集开展研究和进行实验。

3 解决方法

3.1 解决步骤

我们采用如下步骤来解决上述问题：1)对图像的视觉特征值进行离散化。2)挖掘标识语义概念的特征模式。3)基于这些模式，构建标注图像的规则。4)根据标注规则对新的图像进行标注。

3.2 数据离散化

离散化的目的是降低问题的复杂度以及避免过度拟合。由于每一维上不同的特征值太多，因此直接进行频繁模式挖掘代价太大。而且在同一维的特征上两个相近的值之间没有质上的差异，可以视为同一项。离散化有很多成熟的方法，其中比较常用并且通常效果比较好的是基于最小熵的方法。本文将图像的标注作为其特征值的类标签，采用文[15]中的算法寻找分割点，将每一维上连续的特征值分割成若干互不相交的集合，每个集合内的特征值用同一离散值表示。该算法依据熵增益最大化的原则递归寻找分割点，直到熵增益不可能大于某一参数时停止。

3.3 特征模式的挖掘

数据被离散化之后我们就可以在其上寻找那些对于标识语义概念至关重要的模式了。我们这里所说的模式是指特征值的组合，比如{黄色，黑条纹，王字形}。

定义 1(正例，反例) 称 $\text{Sample}_+(c)=\{S|S\in \text{Sample}\wedge c(S.C)\}$ 为概念 c 的正例。称 $\text{Sample}_-(c)=\{S|S\in \text{Sample}\wedge c(S.C)\}$ 为概念 c 的反例。

显然 $\text{Sample}_-(c)\cup\text{Sample}_-(c)=\text{Sample}$，$\text{Sample}_+(c)\cap\text{Sample}_-(c)=\Phi$。

定义 2(覆盖，支持度，频度，频繁模式) 模式 P 在图像集 Sample 上的覆盖为：

$$\text{cover}(P, \text{Sample})=\{S \mid (S\in \text{Sample})\wedge(\exists_j, S.B_j\supseteq P)\}$$

cover (P, Sample)表示的是出现模式 P 的图像集合。

模式 P 在图像集 Sample 上的支持度为：

$$\text{support}(P, \text{Sample})=|\text{cover}(P, \text{Sample})|$$

support(P, Sample)表示图像集中出现模式 P 的图像个数。

模式 P 在图像集 Sample 上的频度：

$$\text{sup}(P, \text{Sample})=\frac{\text{support}(P, \text{Sample})}{|\text{Sample}|}$$

显然 sup (P, Sample)表示模式 P 在图像集中出现的频率，意即频繁程度。

给定 α，当 sup (P, Sample)$>$a 时我们就称 P 为 Sample 上的频繁模式。

一个很自然的想法是，$\text{Sample}_+(c)$ 上的频繁模式应该能够表征 c，但是如果这个模式在大多数图像上都出现就没有意义了，比如，关于野生动物的图像中一般都会有树木，草地，那么{绿色}这一模式对于“老虎”这一概念来说一定会是频繁的，但它不是一个很好的表征“老虎”这个概念的模式。

定义 3(置信度) 称 $\text{conf}(P, c)=\frac{\text{support}(P, \text{Sample}_+(c))}{\text{support}(P, \text{Sample}_+(c))+\text{support}(P, \text{Sample}_-(c))}$ 为模式 P 对于概念 c 的置信度。我们为置信度设一阈值来滤除那些在大多数图像中都出现的模式。

定义 4(表征模式) 给定两个阈值 α，β，称满足下列条件的模式 P 为概念 c 的表征模式：

$$\text{sup}(P, \text{Sample}_+(c))\geqslant\alpha \text{ 且 } \text{conf}(P, c)\geqslant\beta$$

于是我们要寻找每一个概念的表征模式用以识别这些概念。

注意，按照我们对支持度的定义，在计算 support (P, Sample+(c))时，即使 P 在同一个图像的不同分块中都出现，但在计数时只会累计一次。这是由于我们的概念是专家对未经处理的图像进行标注的而不是对分块，以图像为计数单位的目的是为了减少当 P 在不表达 c 的分块中出现和 c 由多个分块表达时带来的干扰。所以在挖掘表征模式的时候我们要把一幅图像作为一个由若干子事务组成事务，每个子事务对应图像中的一个分块，分块中的每个特征值就是一个项，一个模式就是一个项集。而计算项集的支持度时，无论项集是否在同个事务的多个子事务中都出现，只要出现就只累计一次。正是由于这一计算支持度时的特殊性以及数据结构的复杂化，使得我们无法应用 fp-growth 算法进行挖掘。所以我们

仍然采用改进后的 Apriori 算法[11]来挖掘表征模式。

以下我们就来描述一下如何寻找给定概念 c 的表征模式。

算法 1 寻找给定概念 c 的表征模式 CPMine(Sample, c, α, β)

输入：训练图像集 Sample，概念 c，两个域值 α, β
输出：c 的表征模式集 PS(c)，$\forall P \in PS(c)$，有 $sup(P, Sample_+(c)) \geq \alpha$ 且 $conf(P,c) \geq \beta$
1. 计算 $Sample_+(c)$ 与 $Sample_-(c)$
2. 调用 Apriori ($Sample_+(c)$, $|Sample_+(c)| * \alpha$) 找出满足 $sup(P, Sample_+(c)) \geq \alpha$ 的模式集 PS(c)
3. 对 PS(c) 中的每一个模式 P 计算 conf(P, c)，删除 PS 中满足 $conf(P, c) \leq \beta$ 的模式 P
4. 输出 PS(c)

当两个语义概念 c 与 c′经常同时出现时（比如白云与太阳），由于 c 与 c′的正例重合比例很大，则实际表征 c 的模式会在挖掘 c′的表征模式时被挖掘出。（比如，{圆形，红色}可能会被作为白云的表征模式挖掘出）显然，应该要求 c 的表征模式在 c 出现但 c′不出现的图像集上也是频繁的。

定义 5 （决定性表征模式） 给定参数 α，我们称概念 c 的表征模式 P 为 c 的决定性表征模式当且仅当：

$\forall c' \in C, c' \neq c$，若 $sup(c', Sample_+(c)) \geq \alpha$，则有：

$$sup(P, Sample(c \wedge \neg c')) \geq \alpha$$

其中 C 为概念集，$Sample(c \wedge \neg c')$ 表示 Sample 中包含概念 c 但不包含概念 c′的图像集，而

$$sup(c', Sample_+(c)) = \frac{|\{S | S \in Sample_+(c) \wedge c' \in S.C\}|}{|Sample_+(c)|}$$

表示的是 c′在 c 的正例上出现的频繁程度。

我们用 DFP(c) 表示概念 c 的所有决定性表征模式的集合。下面给出如何找出概念集 C 中所有概念的决定性表征模式的算法。

算法 2 寻找概念集中所有概念的决定性表征模式 CDPMine(Sample, C, α, β)

输入：训练图像集 Sample，概念集 C，域值 α, β
输出：$DFP(C) = \{DFP(c) | c \in C\}$
1. 对于 C 中的每一个概念 c，用 CPMine(Sample, c, α, β) 挖掘它的表征模式集 PS(c)
2. 令 $DFP(C) = \phi$
3. 对于 C 中的每一个概念 c
4. 对 PS(c) 中的每一个模式 P
5. 计算 $C'(c, P) = \{c' | c' \in C - \{c\}$，且 $P \in PS(c')\}$
6. 令 $C'(c, P) = C'(c, P) - \{c' | sup(c', Sample_+(c)) < \alpha\}$
7. 对 $C'(c, P)$ 中的每一个概念 c′，若 $sup(P, Sample(c \wedge \neg c')) < \alpha$ 则令 $PS(c) - \{P\}$
8. $DFP(c) = PS(c)$，$DFP(C) = DFP(C) \cup DFP(c)$
9. 输出 DFP(C)

3.4 标注规则

我们主要以两个指标来衡量标注结果：查准率与查全率。

定义 6（查全率，查准率） 在测试图像集 T 上进行多概念标注，即对 $S = \{S.F, S.C\} \in T$，我们给出 S 的标注集 $S.C_1$。

概念 c 的查准率定义为：

$$precision(c) = \frac{|\{S \in T | c \in S.C \wedge c \in S.C_1\}|}{|\{S \in T | c \in S.C_1\}|}$$

概念 c 的查全率定义为：

$$recall(c) = \frac{|\{S \in T | c \in S.C \wedge c \in S.C_1\}|}{|\{S \in T | c \in S.C\}|}$$

对于概念集 C 我们用其中的每个概念的平均查准率和查全率来衡量标注的结果。

我们本着提高在训练图像集上标注后的查全率与查准率的原则制定标注规则。若用模式 P 是否在图像的某个分块中出现来判断其对应的概念 c 是否在该图像中出现：

$$\exists S.B_i, P \in S.B_i \rightarrow c \in S.C_1 \qquad \text{规则(1)}$$

则其在 Sample 上标注后的查全率为：

$$\frac{|\{S \in Sample | c \in S.C \wedge \exists S.B_i, P \in S.B_i\}|}{|\{S \in Sample | c \in S.C\}|} = sup(P, Sample_+(c))$$

而查准率为：

$$\frac{|\{S \in Sample | c \in S.C \wedge \exists S.B_i, P \in S.B_i\}|}{|\{S \in Sample | \exists S.B_i, P \in S.B_i\}|} = conf(P, c)$$

定义 7（重要度） 称 $Imp(P, c) = sup(P, Sample_+(c)) * conf(P, c)$ 为模式 P 对于概念 c 的重要度，我们用重要度表示 P 对标识 c 的重要性。

当然一个语义概念不可能用单一的模式来标识。DFP(c) 中的每一个模式对于识别 c 来说都会起到作用，而它们各自所起作用的大小就可以由 Imp(P, c) 来表示。

定义 8（关联度） 称 $R(S, c) = \sum_{P \in DFP(c) \wedge \exists S.B_i, 有 P \in S.B_i} Imp(P, c)$ 为图像 S 与概念 c 的关联度。

R(S, c) 应该超过怎样的阈值，才能断定图像 S 包含概念 c，我们依然从提高在 Sample 上标注后的查全率与查准率入手。

令 $R(Sample, c) = \{R(S, c) | S \in Sample\}$。对 R(Sample, c) 中的每一个值 x，用：

$$R(S, c) \geq x < c \in S.Cl \qquad \text{规则(2)}$$

在 Sample 上进行标注（详见下节）选择标注效果最好的 x 作为真正的阈值 τ。这样我们便实现了无监督地产生标注规则的方法。

下面我们描述生成标注规则的算法：

算法 3 生成识别概念集 C 的规则集 GenRule(Sample, C, DFP(C))

输入：训练图像集 Sample，概念集 C，所有概念的决定性表征模式集 DFP(C)
输出：规则集 Rules
1. 对 C 中的每一个概念 c
2. 对 DFP(c) 中的每个模式 P，计算 Imp(P, c)；
3. 对 Sample 中的每幅图像 S，计算 R(S, c)；
4. 令 $e = 0, \tau = 0$, $Rules = \phi$
5. 对 R(Sample, c) 中的每一个值 x
用规则(2)在 Sample 上标注
计算 $et = \frac{2 \cdot precision(c) \cdot recall(c)}{(precision(c) + recall(c))}$
if $et > e$ then $e = et, \tau = x$
6. $Rules = Rules \cup R(S, c) \geq \tau \rightarrow c \in S.C_1$
7. 输出 Rules

算法中用 et 来表示标注的综合效果（因为查全率与查准率是相互牵制的，查全率只关心标对的图像个数，所以标得越多越有可能提高，而查准率还关心标错的图像个数，标得越多，错的概率也更大）。

3.5 图像标注

以下我们给出标注新图像的算法

算法 4 图像标注算法

输入：测试图像 S，概念集 C，Rules，DFP(C)
输出：S 上的标注 C_1
1. $C_1 = \phi$
2. 对 C 中的每个概念 c：
在 Rules 中取得 c 的标注规则中对应的 τ 值
3. 计算 R(S, c)，若 $R(S, c) \geq \tau$ 则 $C_1 = C_1 \cup \{c\}$
4. 输出 C_1

4 实验

4.1 数据集

我们采用 Duygulu[5] 提供的一个 5000 幅图像的真实数据集作为实验数据集，其中的图像已按文[3]提出的方法进行了分块，每幅图像分成 1～10 个块，每个块提取了 36 维的视觉特征值。每幅图像上人工标注了 1～5 个概念，共有 371 个概念在图像集中出现。

4.2 比较算法

我们用 DFPM 表示我们的标注算法，并与以下几个算法进行了比较：文[4]提出的“co-occurrence”模型将图像上的标注赋予每个分块，并根据概念与分块同时出现的频率来自动标注，我们用 COM 表示。文[5]提出的翻译模型(“the translation model”)，将图像分块聚类，采用典型的翻译机模型将分块聚类翻译成概念集，我们用 TRM 表示。文[6]提出的跨媒体相关模型(“cross-media relevance model”)学习了同一幅图像中的所有分块及概念的联合概率分布，我们用 CMRM 表示。我们还自己实现了一个与本文提出的算法类似的方法，只是它将图像上的标注全赋予分块，将每个分块视为一个事务，针对分块挖掘决定性表征模式，并对分块进行标注，最后将图像中所有分块的标注作为图像的标注，我们用 DFPM-PB 表示。

4.3 实验及结果

我们在 5000 幅图像中随机选出 500 幅图像作为测试样例集，剩余的 4500 幅图像作为训练样例集。

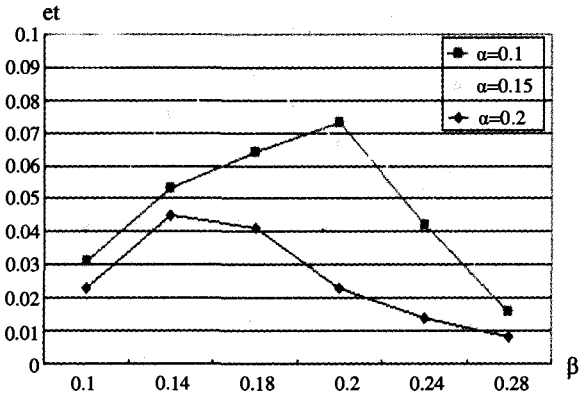

图 1 α,β 对标注效果的影响

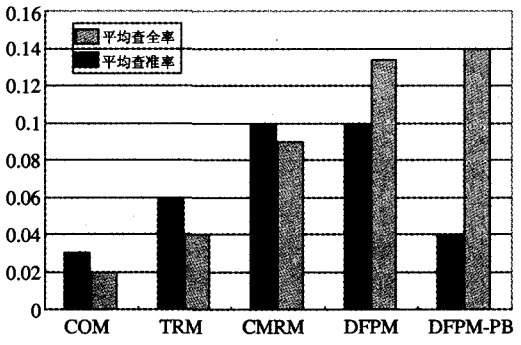

图 2 各算法的查率与查准率真比较

首先，我们来观察算法中的两个参数 α 与 β 对标注效果的影响，图 1 给出了实验结果。我们以 $et=2mr*mp/(mr+mp)$(其中 mp 表示平均查准率，mr 表示平均查全率)表示标注的综合效果，可以看到，当频度阈值和置信度阈值较低时，由于找到的模式区分概念的能力较差，导致标注效果较差，此时找到的模式也太多太滥。当两个阈值逐渐上升时，由于找到的模式更具代表性和区分性，标注效果也因而上升，但当两个阈值过大时，由于找到的模式非常少，导致标注时基本没有匹配到的模式，从而很多概念极少被标注，甚至不被标注，因此标注效果很差。

随后我们在数据集上进行了比较实验，图 2 显示了各个算法在测试图像上标注后，所有概念的平均查全率与查准率，图 3 显示的是各算法的综合效果 et(我们的算法中 α 取 0.155，β 取 0.185)。显然 DFPM 要比 COM，TRM 与 DFPM-PB 表现优异许多，比 CMRM 表现较好。而 DFPM 要比 DFPM-BP 效果好很多，这就说明了，我们将整幅图像作为一个有结构的事务进行挖掘而不将一个分块作为一个事务挖掘是合理而有效的。

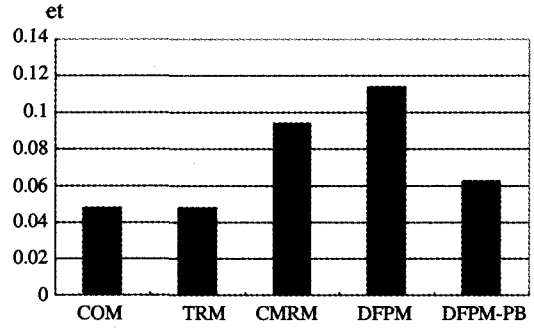

图 3 各算法的综合效果 et 的比较

最后我们将算法在产生规则时确定阈值 ι 的方法分别改成随机选择 R(Sample,c)中的一个值和选择 R(Sample$_-$(c),c)中的最大值以及选择 R(Sample$_+$(c),c)中的最小值进行实验，分别记为 DFPM-random，DFPM-max，DFPM-min。图 5 给出了实验结果，可以看到其他几种选择 τ 值的方法效果都不近人意，DFPM-max 将 τ 值取得太高导致查全率很低，但由于很多概念一个都没被标到图像上，使其查准率为 0，所以平均查准率不升反降。而 DFM-min 则将 τ 值取得太低，从而牺牲太多查准率换取查全率得不偿失。可见我们根据在训练样例集上标注的效果选取 τ 值的方法是合理而有效的。

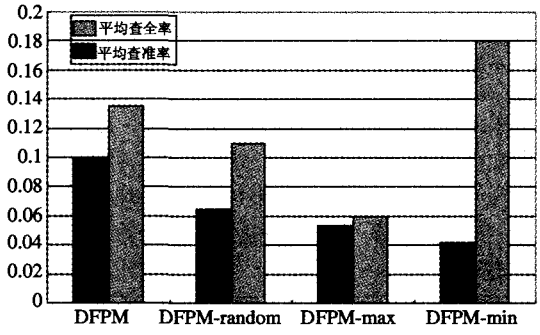

图 4 τ 的取值方法对标注效果的影响

小结 本文针对现存的图像数据库，基于图像分块的处理技术，利用频繁模式挖掘寻找识别语义概念的决定性表征模式，并依此根据效果制导的原则设计了无监督地产生标注规则的方法，实验表明我们的算法比同类算法标注效果更佳，我们的模型和制定标注规则的方法是合理有效的。

本文只是对一些简单的语义概念进行自动标注。是否可以通过在简单语义概念上的数据挖掘来对图像表达的主题进行标注？是否能挖掘出语义概念的子概念以使标注更准确？我们都将在今后工作中探索。

参 考 文 献

1 Ma W Y, Zhang H. Handbook of multimedia computing. In: chapter of Content-based Image Indexing and Retrieval, London: Routledge, 1998. 19～20

2 Brandt S, Laaksonen J, Oja E. Statistical shape features in content-based image retrieval. In: Proc. ICPR, 2000

(下转第 196 页)

同理，已知交通线路停靠站点，站点和附近单位有临近关系，则可推出交通线路经过单位。那么可以在单位类中添加一个“经过公交线路”的属性。对出行者来说，只要提供单位名称，即可查询附近有哪些交通线路可以乘坐。如例1中长风公园站临近单位有长风公园，可以推理出如果出行者的目的地是长风公园，可以考虑乘坐67路公交车，并且可以在长风公园类的属性“经过公交线路”中添加进一条记录为67路。

5 城市交通本体在语义查询系统中的应用

由上述方法设计的本体及获取的知识库可以应用在城市交通语义查询系统。系统设计模型如图5所示。该系统由以下几个模块构成：

(1)城市交通元数据库。保存城市交通元数据信息和属性。

(2)城市交通本体库。是领域概念化模型的形式化表示，包括概念、属性、关系、实例等，以附加文件形式保存，或和城市交通元数据一体化存储。

(3)推理引擎。主要是本体推理。根据用户概念化的信息需求描述和出行方式的显式规则选择，结合本体库对用户需求进行分析，生成针对视图的查询计划，然后把结果返回给用户界面。

(4)查询。包括数据查询和本体查询。本体查询对城市交通本体所包含的概念、属性、关系和实例进行检索和查询，将查询结果返回。

结论和进一步的工作 本文从本体论的角度出发探讨城市交通本体建立和获取公理的方法，建立了比较完善的城市交通本体，包括类和公理集合，其中类定义了属性和关系集。形成了城市交通本体的知识库，公理用来对已经获取的知识进行分析，从而确保知识的一致性和对未知知识的推理。

由于城市交通是门类众多、变化迅速的知识领域，下一步我们需要不断补充完整和添加新的概念、属性和关系集；其次，也是很重要的一点，公理库中要添加更多的公理，以保证知识的完备性和一致性。继而我们需要开展城市交通本体的应用研究，将交通本体和知识推理应用在城市交通查询系统中，并进一步研究城市交通本体与交通信息服务结合的可行性和有效性。

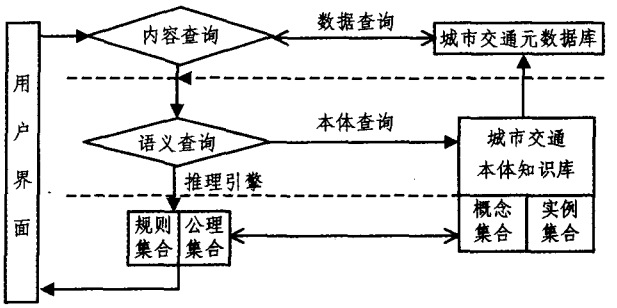

图5 语义查询系统设计模型

参考文献

1 Smith B L, William T S. Development of integrated intelligent transportation system [A]. In: Transportation Research Record 1675 [C], Washington D C: TRB, 1999. 85～90

2 Cycl. http://www.cyc.com/tech.html#cycl

3 Worldnet. http://www.cogsci.princeton.edu/～wn/

4 DOPE. http://aduna.biz/dope/

5 宋炜. 语义网络教程. 北京：高等教育出版社，2004

6 Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse: [PhD thesis]. University of Twente, Enschede, 1997

7 邓志鸿，唐世渭，张铭，等. Ontology研究综述. 北京大学学报(自然科学版)，2002，38(5)：730～738

8 周肖彬，曹存根. 基于本体的医学知识获取. 计算机科学，2003，30(10)：35～39

9 Staab S, Stude R. Knowledge processes and ontologies [J]. IEEE Intelligent Systems, 2001(1): 26～34

(上接第173页)

3 Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 25(9): 1075～1088

4 Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: Proc. of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999

5 Duygulu P, Barnard K, de Freitas N, et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Proc. of the Seventh European Conference on Computer Vision, Copenhagen, Denmark: Springer, 2002. 97～112

6 Jeon J, Lavrenko V, Manmatha R. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: Proc. of the 26th Annual International ACM SIGIR Conference, Toronto, Canada: ACM, 2003. 119～126

7 Wang J Z, Li J. Learning-based linguistic indexing of pictures with 2-d mhmms. In: Proc. ACM Multimedia, 2002

8 Joachims T. Making large-scale svm learning practical. advances in kernel methods - support vector learning, MIT-Press, 1999

9 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: VLDB'94, 1994. 487～499

10 Agrawal R, Srikant R. Fast algorithms for mining association rules: [IBM Research Report RJ9839]. IBM Almaden Research Center, San Jose, California, June 1994

11 Park J S, Chen M S, Yu P S. An effective hash based algorithm for mining association rules. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, SIGMOD Record, ACM Press, 1995, 24(2): 175～186

12 Han Jiawei, Pei Jian, Yin Yiwen. Mining Frequent Patterns without Candidate Generation. ACM SIGMOD, 2000, 29(2): 1～12

13 Dong G, Zhang X, Wong L, et al. Caep: Classiffication by aggregating emerging patterns. Discovery Science, 1999

14 Wei W A, Zhang A. Evaluation of low-level features by decisive feature patterns [content-based image retrieval]. Multimedia and Expo., ICME'04, 2004, 2: 1007～1010

15 Fayyad U M, Irani K B. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. of the 13th International Joint Conference on Artificial Intelligence. Chambery, France: Morgan Kaufmann, 1993. 1022～1029