

一种基于二进制语义特征向量的协同过滤方法

赵 芸^{1,2},徐 兴¹

(1 浙江科技学院 信息与电子工程学院,杭州 310023;2 浙江工业大学 信息工程学院,杭州 310032)

摘 要:提出了一种图像检索新方法,通过用高层语义描述图像,用二进制特征向量判断语义特征是否存在,由向量计算若干图像相似的可能性(即概率),再根据概率值判断图像的相似度值。通过收集每一次检索时用户的反馈信息,采用协同过滤技术挖掘以往用户的反馈信息,同时给数据库中每幅图像分配一个领域,若检索中图像代码相同则提高它们的相似度。通过比较查全率和查准率,证明该方法优于传统的基于内容的图像检索方法。
关键词:协同过滤;二进制语义特征向量;相关反馈
中图分类号: TP391 .3 **文献标识码:** A **文章编号:** 1671-8798(2008)01-0026-05

Collaborative filtering method based on binary semantic feature vectors

ZHAO Yun^{1,2}, XU Xing¹

(1 . School of Information and Electronic Engineering, Zhejiang University of Science and Technology , Hangzhou 310023 , China;
2 . College of Information Engineering, Zhejiang University of Technology , Hangzhou 310032 , China)

Abstract: Proposes a novel image retrieval method which is based on high-level semantic feature vectors . The binary feature vectors in the experiment representing the presence or absence of a certain semantic feature . The probability of the comparability of images is computed by the feature vectors, which is used to estimate the value of the comparability . Collect the data of the user feedback, work out the useful information from the prior user feedback by collaborative filtering method . Partition the images in the database into several domains with search codes that increse the value of the comparability of the images with the same code . The result of precision ratio and recall ratio comparison shows the method of this paper outperforming conventional content-based image retrieval method .

Key words: collaborative filtering; binary semantic feature vector; relevance feedback

近十多年来,图像检索领域的研究受到了人们的广泛关注,并取得了许多研究成果。现有的许多研究是针对图像的底层内容特征,包括颜色、纹理、形状等,用这些特征描述图像。由于图像的隐性特

收稿日期: 2007-11-05
基金项目: 浙江省自然科学基金项目(107379);浙江科技学院科研基金项目(2006-16)
作者简介: 赵 芸(1981—),女,浙江杭州人,助理工程师,硕士研究生,主要从事模式识别、数据库信息处理、计算机控制等研究。

征较多,底层特征分析较复杂,从而导致检索速度不高、正确度不尽人意。近几年,研究者将研究重点转向了非内容的检索,即在检索中不关心图像的底层内容特征,而只关心语义特征。不论在基于内容的检索系统还是非内容的检索系统中,向量模型始终是描述图像的最通用方法。在利用语义特征匹配图像的检索系统中,用户期望的图像和被查询图像的相似度值是通过比较用户期望图像和被查询图像的语义特征向量模型计算得出的,而用户很难将自己的需要描述为一个实际术语的集合。正因为这个局限,C. Buckley 和 G. Salton^[1]提出了相关反馈技术,即通过以往用户的反馈信息自动调整当前的查询结果。所有图像的处理和感知的任务可以由一群用户来完成,即通过积累用户反馈来实现检索的任务。在应用相关反馈技术的检索系统中,用户的意见对检索的成功与否起到关键性作用,系统可以从用户的反馈中获得可靠的信息,从而得知用户对反馈结果的满意度。系统通过对积累的反馈信息的长期学习,不断修正检索能力。

Y. Rui 等^[2]将用户反馈与基于内容的图像检索系统相结合得到较好的检索效果。周向东等^[3]提出了一种信息过滤方法,结合检索样本图像特征和用户相关评价信息对反馈记录进行分析。S. Uchihashi 和 T. Kanade^[4]提出了用人类的描述取代图像的底层特征的图像检索方法,使人类的理解力得到充分利用。B. Scholkopf^[5]利用支持向量机(Support Vector Machine)挖掘和利用反馈中的正例。

比较上述研究成果可知,现有的图像检索系统存在三个缺点:一是反馈信息的重复率高,从而导致不能提供足够的数据用以训练复杂的视觉模型;二是用户很难准确地以系统定义的方式来描述自己期望的图像;三是利用图像的内容特征进行检索,限制了模型匹配的成功率。针对这些问题,本文提出了一种新的检索方法,即通过使用协同过滤技术累加和分析用户反馈,忽略图像的底层内容特征,采用二进制语义特征向量来描述图像,从而可从用户的反馈信息中获得真正有用的信息并最大限度地加以利用。

1 有关术语

在本文中,笔者给数据库中的每一幅图像建立一个二进制的高层语义特征向量。该向量的每一个元素表示一个高层语义特征。图像具备某一特征,该特征值为1;不具备某一特征,该特征值为0。

假设数据库中有 n 幅图像 $X = \{I_1, \dots, I_n\}$ 。 I 表示第 i 幅图像的特征向量。 Q 表示查询图像的特征向量。 I 和 Q 都是二进制向量,若向量中的元素为1,表示该图像具有所对应的特征;若为0,则该图像不具有对应的特征。每幅图像的特征向量是唯一的,特征向量的每个分量表示一个特征,每幅图像都有若干个特征,用 I^m 表示图像 I_i 的第 m 个特征。当且仅当2个图像至少共有同一个特征时,这2个图像是相似的,用符号“ \sim ”表示两图像相似,如 $I \sim Q$ 表示图像 I_i 和查询图像相似。当用户认为反馈的图像与期望图像相似,并希望系统通过分析该反馈图像能给出更接近期望图像的结果,则该反馈图像称为正例;反之,当用户认为反馈图像与期望图像不相似或无关,则该图像称为反例。在当次查询中,若用户选择了图像 I ,则变量 $X_i = 1$;若用户未选择,则 $X_i = 0$ 。

2 相关反馈计算方法

2.1 ZITNICK 的协同过滤方法

协同过滤技术由 D. Goldberg 于1992年最早应用于邮件系统中。协同过滤方法在电子商务的个性化服务系统中得到了广泛和深入的研究,如网上图像浏览系统 Active Web Museum 就采用了协同过滤方法。周向东等^[6]将协同过滤技术用于多媒体对象的相似性检索中。

根据有关术语的描述已知,若用户选择了图像 I ,则变量 $X_i = 1$;若未选择,则 $X_i = 0$ 。数据库集合中被赋予值的变量的集合称为证明集, $X_E = \{X_{a1}, \dots, X_{aE}\}$;其他未被赋值的变量集合称为隐性集合,即 $X_H = X - X_E$ 。在现实中, X_E 远小于 X_H 。在这样的定义下,非内容的图像检索问题转化成了概率 $P(X_i = 1 / X_{a1} = 1, \dots, X_{aE} = 1)$ 的计算问题,其中 $X_i \in X_H$ 。C. Zitnick^[7]提出通过扩大 Rényi 的二次熵方法,可以得到

$$P(X_i = 1 / X_E) \sim \sum_j f_j \quad (1)$$

式(1)中, f_0, \dots, f_n 是由 $\{X_1, \dots, X_n\}$ 构成的函数。是拉格朗日系数,

$$P_i = \begin{bmatrix} P(X_i = 1 / X_{a1} = 1) \\ \dots \\ P(X_i = 1 / X_{aE} = 1) \end{bmatrix} \quad (2)$$

$$P = \begin{bmatrix} P(X_{a1} = 1 / X_{a1} = 1) \dots P(X_{aE} = 1 / X_{a1} = 1) \\ \dots \\ P(X_{a1} = 1 / X_{aE} = 1) \dots P(X_{aE} = 1 / X_{aE} = 1) \end{bmatrix} \quad (3)$$

由等式(2)和等式(3)可以计算出 P_i 和 P , 由 P_i 和 P 又能计算出 $P(X_i = 1 / X_{a1} = 1, \dots, X_{aE} = 1)$ 的值。

假设数据库中有图像 I_1, \dots, I_E 与用户的期望图像相关, 那么可以将图像按照概率的大小依次排序, 概率越高被选中的几率越大, 系统应给该图像越多的被用户选择的机会。C. Zitnick 给出了详细的计算过程, 具体可参阅文献[7]。

2.2 语义特征向量

在查询过程中, 用户选择与期望图像相关的一幅或几幅图像反馈给系统, 系统分析当前用户反馈

信息中的语义特征, 从而调整查询结果, 直到用户满意为止。

根据有关术语介绍, 将图像描述为特征向量, 假设图像有 3 个特征, 如:

$$I_i = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \text{ 和 } I_j = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

当用户反馈一个正例, 即 $Q \sim I$ 时, 系统通过图像的特征计算 Q 与 I_i 相似的概率, 再以这个概率为条件计算 Q 与数据库中其他图像相似的概率, 即求概率 $P(Q \sim I / Q \sim I)$,

$$P(Q \sim I / Q \sim I) = \frac{P(Q \sim I \mid Q \sim I)}{P(Q \sim I)} = \frac{1 - P\left(Q = \begin{bmatrix} 0 \\ \times \\ 0 \end{bmatrix}\right) P\left(Q = \begin{bmatrix} \times \\ 0 \\ 0 \end{bmatrix}\right) P\left(Q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right)}{P(Q \sim I)} = \frac{1 - [1 - P(Q)] [1 - P(Q)] - [1 - P(Q)] [1 - P(Q)] + [1 - P(Q)] [1 - P(Q)] [1 - P(Q)]}{P(Q \sim I)} \quad (4)$$

为了简便起见, 用 $P(Q)$ 代替 $P(Q = 1)$, 表示图像 Q 的第 k 个特征存在的概率。于是根据概率公式可以算出等式(4)中的每一个 $P(Q)$ 和 $P(Q \sim I)$ 。例如:

$$P(Q \sim I) = P\left(Q = \begin{bmatrix} 1 \\ \times \\ \times \end{bmatrix}\right) P\left(Q = \begin{bmatrix} \times \\ \times \\ 1 \end{bmatrix}\right) P\left(Q = \begin{bmatrix} 1 \\ \times \\ 1 \end{bmatrix}\right) = \frac{4}{8} + \frac{4}{8} - \frac{1}{4} = \frac{3}{4} \quad (5)$$

其中 '×' 表示可以取 0 或 1 中的任意一个。

当用户反馈多个正例时, 要求的就是 $P(Q \sim I / Q \sim I_1, Q \sim I_2, \dots, Q \sim I_n)$, 同样可以根据等式(4)和式(5)的方法计算结果。

3 实验

3.1 实验系统

首先将协同过滤技术应用到检索系统中, 采用 C. Zitnick 的方法。实验过程中发现, 用户选择的若干正例, 可能其中几个正例与期望图像的某一类特性相似, 而其他正例可能与期望图像的这类特性完全相反而只与另几类特征相似。这样的训练样本极易使查询在两个相差较大甚至完全相反的方向之间震荡。对于这个问题, 笔者提出的解决方案是采用语义特征向量描述图像, 用上文提出的二进制语义特征向量的方法描述图像并计算查询样本和被查

询图像间的相似度。另外, 如果一个图像反复被选为反馈图像, 那么它所含的与期望图像相似的特征应该多于被选次数较少的图像, 因此在计算相似概率时为图像加入了被选系数。被选次数较多则被选系数高, 次数少则系数低。其次, 系统按照前文的叙述计算数据库中图像的概率, 大于某一概率值的一批图像为相关图像, 可以呈现给用户, 其他的图像则不呈现给用户, 这个概率值称为基准概率, 系统可根据实际应用情况调整基准概率。

假设数据库中所有图像组成的集合为 Img , 用户选择的图像组成集合 Slt 。用户希望得到的查询结果图像的数量 num , 每幅图像的被选系数用参数 n 表示, 基准概率用 P 表示。将图像分为各个领域, 赋予每个图像一个领域代码。因此, 所提出的算法可描述为:

```

step 1  num  10, 基准概率  50%, 被选系数 n  0。
step 2  系统随机给出 num 个图像;
step 3  IF    用户期望图像就在这 num 个图像中, THEN 查询结束,
        ELSE 用户从 num 个图像中选择若干图像作为反馈, 生成图像集合 Slt,
        AND 用户为系统设置一个期望得到图像的数量 num,
        ENDIF;

```

step 4

IF 集合 Slt 中无图像, THEN 返回 step 2,
ELSE 将 Slt 中每一幅图像的 n 自动加 1,
ENDIF;

step 5

按语义特征向量法计算概率 $P(\mathbf{I}_{\text{img-Slt}} \sim \mathbf{Q} | \mathbf{I}_{\text{lt}} \sim \mathbf{Q})$,
AND $P(\mathbf{I}_{\text{img-Slt}} \sim \mathbf{Q} | \mathbf{I}_{\text{lt}} \sim \mathbf{Q}) = P(\mathbf{I}_{\text{img-Slt}} \sim \mathbf{Q} | \mathbf{I}_{\text{lt}} \sim \mathbf{Q}) \times 1.05^n$,
AND IF 集合 Slt 中有 80 % 以上的图像属于同一个领域,
THEN 集合 Img-Slt 中属于该领域的图像的概率乘以 1.05,
ENDIF,
AND 保存记录;

step 6

将集合 (Img-Slt) 中概率大于 的图像按照概率大小排序,

IF 集合 (Img-Slt) 中概率大于的图像数量多于 num 幅,
THEN 前 num 幅图像送还给用户,
ELSE IF 集合 (Img-Slt) 中有概率大于 的图像,
THEN 将集合 (Img-Slt) 中所有概率大于 的图像送还给用户,
ELSE = - 10 % , AND 返回 step 6,
ENDIF,
ENDIF;

step 7 返回 step 3。

3.2 实验结果与分析

以用户希望得到雪山风景图为例,检索界面如图 1 所示。

图 1 系统检索界面

Fig.1 retrieval interface

实验系统检索效果评价指标采用查全率和查准率。将用户检索的次数和查全率查准率进行对比,可以看出,查全率和查准率都明显高于普通方法,查准率在第 2 次检索后就超过了普通方法 10 次检索的最大值。实验证明,该方法可以降低反馈信息的

重复率,降低对用户知识结构的要求,降低信息传达的误差,从而提高了图像匹配的成功率。图 2 和图 3 分别给出了本文方法与传统方法在查全率和查准率两个评价指标上的对比情况。

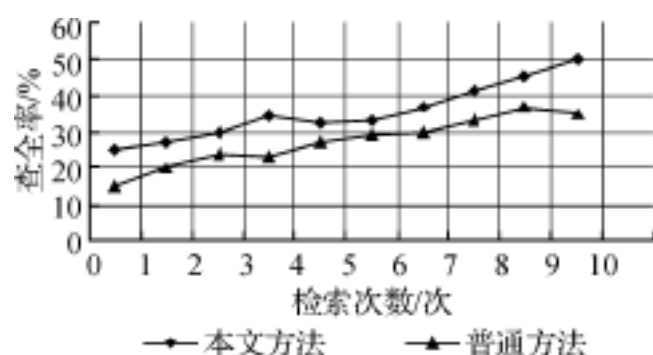


图 2 查全率对比

Fig 2 recall ratio comparison

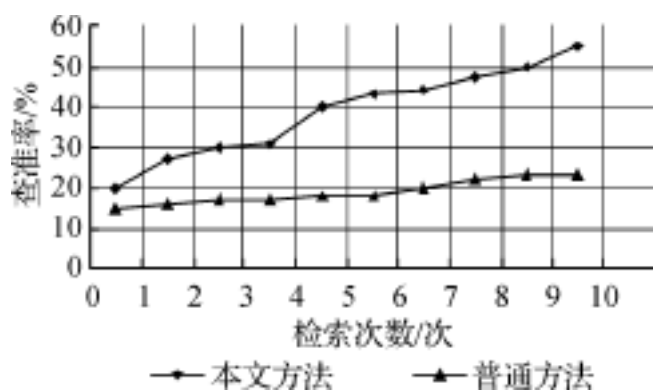


图 3 查准率对比

Fig.3 precision ratio comparison

4 结 语

本文提出的方法采用二进制语义特征向量来描述图像,因为忽略图像的底层内容特征,因此可减轻系统逐一分析复杂的图像底层内容特征的计算量。采用协同过滤技术累加和分析用户反馈,使在不同时间执行检索任务的用户可以互通反馈信息,提高反馈信息的利用率。实验结果证明了该方法可以达到较好的检索效果。

在以后的研究中,笔者将力图找到一种更好的方法使系统在首次接受查询任务时就可以有依据地

选择数据库中的图像呈现给用户。另外,本文的方法是在已有一群用户使用该图像检索系统并给出了足够的反馈数据的假设上进行的,当系统正式投入使用时,先期运行的任务应由谁来完成是在以后的研究中值得考虑的问题。

参考文献:

- [1] BUCKLEY C, SALTON G . Optimization of relevance feedback weights [C] . Washington , United States: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval , 1995:351-357 .
- [2] RUI Y, HUANG T S, ORTEGA M, et al . Relevance feedback: a power tool for interactive content-based image retrieval[J] . IEEE Trans Circuits System Video Technology , 1998(5) :644-655 .
- [3] 周向东,施伯乐 .图像数据库检索中的信息过滤反馈方法[J] .计算机学报,2004 ,27(11) :1505-1513 .
- [4] UCHIHASHI S, KANADE T . User-powered ' content-free ' approach to image retrieval[C] . Tsukuba, Japan: Proceedings of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society 2004 , 2004:24-32 .
- [5] SCHOLKOPF B, PLATT J, SHAWE-TAYLOR J . Estimating and support of a high-dimensional distribution[J] . Neural Computation , 2001,13: 1443-1471 .
- [6] 周向东,施伯乐,张琪,等 .基于长期学习的多媒体数据库相似性检索[J] .软件学报,2004 ,15(1) :86-93 .
- [7] ZITNICK C . Computing conditional probabilities in large domain by maximizing Renyi 's quadratic entropy [D] . Pittsburgh: Carnegie Mellon University , 2003 .