基于长期学习的多媒体数据库相似性检索*

周向东+, 施伯乐, 张 琪, 张 亮, 刘 莉

(复旦大学 计算机与信息技术系,上海 200433)

A Long-Term Learning Based Similarity Retrieval of Multimedia Database

ZHOU Xiang-Dong+, SHI Bai-Le, ZHANG Qi, ZHANG Liang, LIU Li

(Department of Computing and Information Technology, Fudan University, Shanghai 200433, China)

+ Corresponding author: Phn: +86-21-55073942, E-mail: xidzhou@etang.com, http://www.cit.fudan.edu.cn

Received 2002-11-20; Accepted 2003-03-04

Zhou XD, Shi BL, Zhang Q, Zhang L, Liu L. A long-term learning based similarity retrieval of multimedia database. *Journal of Software*, 2004,15(1):86~93.

http://www.jos.org.cn/1000-9825/15/86.htm

Abstract: An approach is presented for multimedia similarity query using an on-line analysis of feedback sequence logs. The approach is based on user's feedback sequence accumulation and on-line collaborative filtering to predict the semantic correlation between the media objects in database and query sample. Edit distance is used to evaluate the similarity between current retrieval's feedback sequence and the prefixes of the records in the feedback logs. A prototype image retrieval system is implemented. Integrated with the retrieval method based on the generalized Euclidean distance, the performance of similarity query can be improved apparently. Experiments over 11 000 images demonstrate that this method outperforms the conventional ones.

Key words: similarity query; user's relevance feedback; sequence analysis; collaborative filtering; multimedia database

摘 要: 基于内容的相似性检索是多媒体数据库研究的重要内容之一.近年来,利用用户相关反馈技术改善检索性能的研究成为新的热点.但是,在传统的相关反馈方法中,系统积累的反馈历史数据未得到充分利用.为了进一步提高检索系统的性能,提出了一种对相关反馈序列日志进行协同过滤在线分析的相关反馈检索方法.该方法使用编辑距离对用户的反馈序列进行相似性度量,并根据协同过滤的思想对数据库中的媒体对象与当前检索的语义相关性进行预测,从而改善检索的效果.实现了一个图像数据库检索原型系统.对 11 000 幅图像数据库进行的实验表明,与传统相关反馈技术相比,该方法对检索性能有明显的改善.

关键词: 相似性检索;用户相关反馈;序列分析;协同过滤;多媒体数据库

中图法分类号: TP311 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.69933919 (国家自然科学基金)

作者简介: **周向东**(1969－),男,河南封丘人,博士,讲师,主要研究领域为多媒体数据库,信息检索;**施伯乐**(1935－),男,教授,博士生导师,主要研究领域为数据库理论与应用;**张琪**(1979－),女,硕士生,主要研究领域为数据库,信息检索;**张亮**(1963－),男,博士,教授,主要研究领域为支持多媒体应用的数据库技术及信息集成;**刘莉**(1978－),女,硕士生,主要研究领域为数据库,信息检索.

多媒体对象往往具有丰富的语义信息和复杂的视觉特征,使得基于文本标注的传统多媒体数据库检索系统面临着一系列问题,如标注工作量巨大、标注的主观性和不一致性、用户难以描述检索目的等.基于向量模型的多媒体数据库相似性检索(similarity query)首先从多媒体对象中抽取多维属性或视觉内容特征(如颜色、纹理、形状等),组成特征向量后存入数据库中.检索时,系统根据用户提交的检索样本在数据库中返回一定数量的与之最相似的多媒体对象,即系统把检索对象映射到特征空间中,使得检索变成了特征空间中的相似向量查找,比如,使用加权的欧式距离[1]或一般的加权欧式距离[2,3]进行特征向量的相似性度量,从而使检索的自动化程度得到很大的提高.但是,由于目前的计算机视觉技术还不能稳定地建立起多媒体对象的语义信息与其视觉特征间的对应关系,使得基于视觉特征对比基础上的多媒体数据库相似性检索在检索的准确性上还难以满足实际应用的要求.

为了弥补检索性能上的不足,用户相关反馈技术这样一种通过系统与用户进行交互、动态地调整检索目标和相似性度量函数的检索机制被引入到相似性检索中.用户相关反馈通常是一个人机交互的循环过程,就是在检索过程中由用户对检索结果进行评价,指出哪些检索结果是与检索目的相关的(正例)/或不相关的(负例),然后根据这些用户评价信息调整检索样本或相似性度量函数,进行新一轮的检索,如此反复,直至用户得到满意的检索结果或者系统的检索精度达到了稳定状态[1~4]为止.

相关反馈技术是当前多媒体数据库检索研究中最为活跃的领域.早期的相关反馈方法主要依据一些启发式思想进行检索样本与参数的调整,如修改查询向量使其向相关检索对象的分布中心移动[4],根据反馈信息调整距离度量公式中各分量的权重等[1].Ishikawa 等人[2]在 MindReader 系统中使用最优化方法进行查询参数优化,Rui 等人[3]在此基础上给出了利用最优化方法求解最优参数的统一的相关反馈框架.近来,机器学习方法如支持向量机(SVM)[5]等也与相关反馈方法相结合,以进一步提高检索精度.在上述相关反馈方法中,检索系统并不保存以前用户的反馈信息,即在新的检索过程中,以前用户对数据库中多媒体对象的语义评价的反馈信息已被丢弃,系统并未利用它来改进新检索的效果.

当前利用系统积累的反馈信息改进检索性能的研究已引起了人们的关注[6,7].由于忽略了对反馈历史记录中隐含的数据库对象与当前检索的语义相关性的发掘,已知工作存在着反馈信息的利用缺乏针对性以及对检索精度的改进不明显等问题.本文给出一种通过发掘反馈历史记录中隐含的数据库对象与当前检索的语义的相关性来改善检索性能的方法.该方法把用户在检索中所进行的反馈操作序列(用反馈例编号表示)作为反馈日志记录入数据库中,进行新的检索时使用协同过滤方法对用户反馈的序列模式进行分析,进而有针对性地对数据库中的多媒体对象与当前检索的语义相关性进行预测,通过与使用一般的加权欧式距离的检索方法相结合,明显地提高了检索的性能.本文给出的检索方法既保持了一般基于内容检索的特征,又针对当前检索的个性化特点,隐含地结合了媒体对象的语义相关性信息.我们实现了一个关于图像数据库相似检索的原形系统,实验显示,与传统的相关反馈方法相比,本文提出的检索方法能够明显地改善相似检索的效果.

本文首先给出本文工作的背景,然后对相关工作进行介绍.第 2 节给出用户反馈的序列模式概念与反馈日志分析方法及相应算法.第 3 节是关于本文的检索方法.第 4 节报告和分析实验结果.最后是本文的研究结论.

1 相关工作

用户相关反馈通常分为检索点移动和检索参数调整两种类型.检索点移动通常用 Rocchio 公式来描述[4],即检索样本中与正例相关的特征得到增强,与负例相关的特征被减弱,使检索点移向能够带来更好检索结果的位置.Rui[1]给出了一种分层的权重调整方法.该方法的核心就是考察用户(正)反馈样本集合中的特征向量的各个分量在特征空间的各个维上的分布.反馈集合中的各个特征向量在向量空间第 i 维上的分布用其标准差 r_i 度量,r_i 越大,分布得越杂乱,该分量与检索的相关程度也就越小,所以,应减小该分量的权重;反之则应增加该分量的权重.

MinderReader[2]使用一般的欧式距离进行特征向量的相似性度量.文献[2]证明了当反馈例个数大于等于特征向量的维数时,可以得到一般加权欧式距离的最优权矩阵 W.但是在实际检索中,特征向量的维数往往明显高于每一轮反馈中用户给出的反馈正例的数量,所以,只有当检索正例累积到一定程度时,一般欧式距离才能有

效地改进检索效果.本文通过对反馈日志进行分析,利用所得到的多媒体对象的语义相关性来扩充用户的正反馈例集合,使检索精度得到了明显的改善.

Bartolin 等人[6]提出了 FeedbackBypass 系统.该系统在用户提交的检索样本与检索中通过相关反馈得到的“最优”参数之间建立起映射关系,并记录入数据库中.当新的检索进行时,系统首先在数据库中进行相似匹配,得到“最优”检索参数,然后使用这些参数进行检索.该方法提高了检索的效率,但并未明显改善检索精度.

H.Muller 等人[7]给出了一个使用大图像特征集的基于文本检索方法的图像检索系统.该系统通过分析用户的反馈日志来调整图像特征相关度公式中的相关度因子.调整的准则是,在同一次检索中,同时出现在两个正例图像中的特征的相关度应提高,同时出现在一个正例与一个负例图像中的特征的相关度需要降低.该系统对用户的反馈日志进行离线分析,缺乏在线的动态调整能力.由于采用特殊的检索策略,该反馈日志分析方法难以扩展到一般的相似性检索系统中.

在多媒体系统的个性化研究方面,A.Kohrs 和 B.Merialdo 给出了一个使用推荐方法的自适应网站图像浏览系统[8,9].该系统使用协同过滤方法对用户给出的图像评价数据进行离线分析,并结合使用图像的内容过滤方法来生成个性化的浏览内容.

2 用户反馈序列日志的在线分析

用户进行的每一次检索都具有其自身的特点.从宏观上看,大量用户的多次检索又具有一定的统计规律性.把相似检索中的用户相关反馈数据记录下来就形成了反馈日志文件,对反馈日志文件的分析有助于改进检索系统的性能.我们认为,通过分析以往检索产生的反馈信息,系统可以对数据库中的多媒体对象与当前检索的相关性进行预测,从而达到改善检索效果的目的.

2.1 多媒体对象语义相关性预测

协同过滤是一种通过已知用户群(或相关对象组)的兴趣或特征来预测新的未知用户(对象)的兴趣或特征的方法,最早由 D.Goldberg 等人[10]在一个邮件过滤系统中用“协同过滤”(collaborative filtering)加以命名.协同过滤的思想基于这样一个常识性假设,即不同的人在某些共同事物上观点相同,则他们在其他事物上观点相同的可能性较大[10,11].协同过滤方法在电子商务的个性化服务与推荐系统中得到了广泛和深入的研究,如关于新闻组与电影推荐的 GroupLens[11]和网上图像浏览系统 ActiveWeb Musuem[8,9]是采用 CF 方法的研究系统.

借鉴上述思想,我们在多媒体相似检索中使用协同过滤方法分析反馈日志文件,预测数据库中的多媒体对象与当前进行的检索之间的语义相关性.其基本原理用向量模型描述如下:

检索过程中,用户在一次检索活动中的相关反馈(只考虑正例)用一组多媒体对象的编号(用户认为语义相关的对象,即正例)来表示,这组编号就看做是一条评价记录,这里称为反馈记录,反馈记录用一个 n 维向量表示:$r_i=\langle I_1,I_2,...,I_n\rangle$.经过用户的不断检索,系统产生的反馈记录日志文件就构成了一个反馈信息数据库.相关反馈信息数据库用一个 $m\times n$ 的矩阵 R 来表示,$R=\{r_{ij}\}$,矩阵中行代表不同的检索所进行的反馈,列代表数据库中的对媒体对象,其中 r_{ij} 代表第 i 次检索对第 j 个多媒体对象的相关评价指标,则 R 的每一行就代表了用户的一次检索过程中对多媒体对象的相关评价(也反映了该次检索的特征).

在检索系统中,令 $I,I=|n|$ 表示数据库中多媒体对象的集合,设 d_i 是第 i 次检索中用户进行了相关评价的对象集合,则 $\bar{r}_i=\frac{1}{|d_i|}\sum_{j\in I_i} r_{ij}$ 表示第 i 次检索中用户对数据库中对象的平均评价指标.当前检索的反馈用向量 a 表示,则对象 j 对于当前检索的相关评价的预期值 $p_{a,j}$ 用如下公式描述[12]:

$$p_{a,j}=\bar{r}_a+\lambda\sum_{i=1}^{m} w(a,i)(r_{i,j}-\bar{r}_i) \tag{1}$$

其中 m 是相关反馈信息数据库中反馈记录的数量,权值 $w(a,i)$表示 a 与数据库中其他反馈之间的相似程度,λ是规格化因子.从公式(1)可知,$w(a,i)$的大小决定了数据库中的对象与本次检索的相关程度,因此如何在日志数据库中查找相似记录非常关键.

2.2 用户反馈的序列模式

通常意义上的日志是指系统对用户所作的操作序列的记录.类似地,我们利用用户在检索中的操作序列来刻画相应的检索(用反馈例的编号序列表示).用户在每一轮相关反馈过程中给出的反馈序列形式如下:

〈检索编号:相关对象 1,相关对象 2,…,相关对象 M〉.

为了客观地反映用户检索的特点,我们把每一轮反馈中用户给出的反馈例都按照它们在该轮检索结果中的排名顺序进行排列.把一次检索中各轮反馈的反馈序列按反馈的轮次顺序组成一个长序列记录下来,就得到了该次检索的反馈日志记录,通过积累这些反馈记录就得到了检索系统的反馈日志数据库.

我们有如下定义:

定义 1(用户反馈的序列模式). 令检索集合为 U,数据库中项目的集合为 I.设检索 $a\in U$,a 包含 l 轮反馈,则第 i 轮反馈包含的数据库对象的集合为 d_{a_i},$|d_{a_i}|=k$,反馈对象在第 i 轮反馈中对于检索样本的相似度评价函数为 $E:d_{a_i}\to R^+$,则对于 d_{a_i} 上的一个排列 p_i,且 $E(p_i(I_1))\geq E(p_i(I_2))\geq\ldots\geq E(p_i(I_k))$:

检索 a 的第 i 轮反馈模式用一个 k 元组表示:

$$W(a_i)=\langle p_i(I_1),p_i(I_2),\ldots,p_i(I_k)\rangle.$$

检索 a 的反馈序列模式用一个 l 元组表示:

$$W(a)=\langle W(a_1),W(a_2),\ldots,W(a_l)\rangle.$$

为了方便起见,以后我们直接用反馈例的符号序列代替元组来表示反馈模式,称为序列模式.

2.3 基于编辑距离的序列模式相似性度量

由于本文提出的用户反馈模式以序列形式表示,因此,参照字符串相似匹配中编辑距离的概念,可以用两个序列的编辑距离来表示它们之间的相似程度.下面给出字符串编辑距离的一些基本概念[12].

任意的字符序列 AB 的距离 $d(AB)$定义为 A 经过一系列字符的插入,替换和删除操作转化为 B 的最小代价(即这一系列编辑操作的代价之和最小),并且 $d(AB)$满足距离公理,则所有的字符序列组成一个度量空间.令γ是关于编辑操作 $a\to b$ 的代价函数,对于不同的编辑操作,γ赋予不同的实数.对于一个编辑序列 $S=\{s_i\}$,$\gamma(S)=\sum^m_{i=1}\gamma(s_i)$,则字符串 A 转换到 B 的编辑距离$\delta(AB)$为

$$\delta(AB)=\min\{\gamma(S)|S\text{ 是 }A\text{ 转换到 }B\text{ 的编辑操作序列}\}.$$

经典的计算编辑距离的动态规划算法的复杂度为 $O(s\times n)$,其中 s 为编辑距离[12].

本文通过计算日志中的反馈序列前缀与当前反馈模式的编辑距离来查找相似序列模式.相关的概念与定义如下:

定义 2. 序列模式 w_i 的 k 前缀是指 w_i 的长度为 k 的前导子序列,用 $[w_i]^k$ 表示.

一次检索通常包含多轮反馈,设新的检索包含 m 次反馈,其中第 i 轮反馈后系统得到的反馈模式用 c_i 表示,用户在完成本次检索后最终给出的总的反馈序列模式就是 c_m,且 $c_i=[c_m]^{|c_i|}$,即每一轮反馈后系统得到的反馈序列模式都是本次检索的总序列模式的一个前缀.

直接用两个序列的编辑距离来度量序列的相似性存在如下问题:

设 a,b 为反馈日志中记录的两个序列模式,$c_i,i=1,\ldots,m$ 为新检索中的第 i 次反馈后系统得到的反馈模式,且 $c_i=[a]^{|c_i|}$,$c_m=a$,即新检索与日志中 a 是相同的反馈模式,又 $b\cap c_m=\varnothing$,即 b 与 c_i 无共同元素,显然 a 就是我们希望从日志中得到的相似记录.在每一轮反馈后我们都用新的反馈模式去日志中进行相似匹配,设第 k 轮反馈时,$|a|>|b|=|c_k|$,则根据编辑距离的定义 $\delta(a,c_k)=|a|-|c_k|$,$\delta(b,c_k)=|b|=|c_k|$.显然,当 $|a|>2|c_k|$ 时,$\delta(a,c_k)>\delta(b,c_k)$,即 b,c_k 的编辑距离比较小,因此系统得出 b,c_k 更相似的结论.然而,事实上,序列 a 才是我们感兴趣的反馈模式,而由于 b 与 c_k 没有共同元素,它们相关的可能性非常小.

因此,我们使用当前反馈序列模式与日志中的反馈序列的相应前缀进行对比,并给出定义 3.

定义 3. 当前检索的反馈模式 c 与日志中的反馈记录 w_i 的相似度定义为

$$sim(c,w_i)=1-\frac{\delta(c,[w_i]^{|c|})}{|c|} \tag{2}$$

其中$|c|$是c的长度,$[w_i]^{|c|}$表示w_i的长为$|c|$的前缀.

设反馈日志库中与当前检索的反馈模式最相似的k条记录组成的候选反馈记录集合为U,U中的反馈记录所包含的数据库对象组成的集合为P称为候选对象集合,令U中包含$p \in P$的反馈记录组成的集合为U_P.根据公式(1),我们给出定义4.

定义 4. P中的数据库对象p与当前检索反馈模式c的语义相关度定义如下:

$$cor(p,c)=\sum_{i=1,w_i\in U_P}^{|U_P|}\frac{sim(c,w_i)}{|w_i \setminus c|}, \tag{3}$$

其中$|w_i \setminus c|$表示w_i带来的新对象的个数.在初始反馈时,c通常比较短,可能使个别与c有相同前缀的超长模式序列带来的大量数据库对象获得非常高的相关度,从而完全影响了整个预测.为了避免这种过早的主导影响可能带来的偏差,我们在公式中除以$|w_i \setminus c|$,以降低过长的序列模式对整个相关预测的主导作用.

2.4 语义相关对象集的生成算法

设检索系统中$I=\{I_1,I_2,\ldots,I_n\}$为多媒体对象集,则系统的相关反馈日志R由m条反馈序列组成,其中反馈序列r_i选中的多媒体对象用集合$d(r_i)$表示,$i=1\ldots m$.当进行相关反馈时,用户的相关反馈序列用c表示.在R中查找与c最相似的k条反馈记录并组成相邻集合$neighbor$,且$r_{i_j} \in neighbor$,其中$j=1,\ldots,k$,令$sim(c,r_{i_j})$表示c与r_{i_j}的相似度.设$neighbor$中出现的多媒体组成的集合为$I'=\bigcup_{j=1\ldots k} d(r_{i_j})$,$p \in I'$对应于$c$的语义相关度$cor_p$由公式(3)可以得到.语义相关对象集的生成算法如下:

算法 1. 相关对象集合生成算法.

输入: $threshold1$是相关反馈记录相似度阈值;

$threshold2$是数据库对象语义相关度阈值;

c为当前反馈模式, R为反馈日志库.

输出: 相关对象的编号和相关度.

```
Function result=GetCorrelation(c,R,threshold1,threshold2)
Correlation=∅
I'=∅
for each r_i∈R
    similar←sim(c,r_i)
    if similar>threshold1
        newp=d(r_i)\I'    //判断新增对象
        for each p'∈newp
            correlation(p')←0    //新增对象相关度初始化
        endfor
        I'←I'∪newp
        for each p_i∈d(r_i)
            correlation(p_i)←correlation(p_i)+similar/|r_i−c|    //根据定义 4
        endfor
    endif
endfor
for each p_i∈I'
    规格化 correlation;
    if correlation(p_i)>threshold2
        result←result∪⟨p_i,correlation(p_i)⟩
    endif
endfor
```

该算法最坏情况下的时间复杂度为$O(n\times(|c|^2+m))$,其中$n=|R|$, $m=\max_{i=1}^{|R|} r_i$.一般情形下,m和$|c|$是比较小的,因此,算法的时间复杂度主要由日志中的反馈记录的数量决定.显然,当反馈日志中的记录数量很大时,采用合适的索引结构是必要的.限于篇幅,本文集中讨论如何提高检索的精度问题.

3 检索方法

Rui[3]给出了使用最优化方法的通用相关性反馈框架.本文在该框架的基础上,使用协同过滤方法对用户提交的反馈样例集合进行扩展,以加快反馈速度、提高检索效率.按照文献[3],检索样本用 q 表示,q 表示其对应于第 i 个特征的向量.用户相关反馈样例集合用$\{x_1,x_2,\ldots,x_N\}$表示,则 x_{ni} 表示第 n 个反馈例关于第 i 个特征的向量. $\pi=\{\pi_1,\pi_2,...,\pi_N\}$ 为用户反馈样例的相关度指标,则对于第 i 个特征的最优检索向量为

$$q_i^{T^*}=\frac{\pi^T X_i}{\sum_{n=1}^{N}\pi_n} \tag{4}$$

其中 X_i 为 $N\times K_i$ 的对应于第 i 个特征的反馈例矩阵,其中 K_i 为第 i 个特征向量的维数.最优权重矩阵为

$$W_i^*=(det(C_i))^{\frac{1}{K_i}}C_i^{-1} \tag{5}$$

其中 C_i 是关于 X_i 的协方差矩阵.

$$C_{i_{rs}}=\frac{\sum_{n=1}^{N}\pi_n(x_{nir}-q_{ir})(x_{nis}-q_{is})}{\sum_{n=1}^{N}\pi_n},\quad r,s=1,...,K_i \tag{6}$$

上述反馈公式中的输入参数为反馈样例的特征向量 $\{x_n\}$ 及其相关度指标 $\{\pi_n\}$.当反馈例个数大于第 i 个特征的维数时,W_i'为全阵,这时相似性度量函数为一般的加权欧式距离.一般情形下,随着用户反馈次数的增加(反馈样例的增多),检索效果在总体上是不断上升的(不是稳定的).但是,这往往需要一个用户反馈的积累过程.

根据算法 1 获得的相关反馈例集合及其相关度指标集分别用$\{p_i\}$和$\{correlation(p_i)\}$表示,则扩展后的反馈例集合 $x=d_c\cup\{p_i\}$,相应的相关度指标为 $\pi=\pi_c\cup\{correlation(p_i)\}$,其中 d_c,π_c 为用户给出的反馈例集合及其相关度指标集.分别把 x,π 代入公式(4)和公式(5),得到新的反馈检索向量与权重,使用新的反馈检索向量和权重计算相似度,得到检索结果.本系统初始时无反馈积累,其功能等同于传统的基于用户相关反馈的检索系统(如文献[3]),随着反馈的积累,检索的效果会有明显的改善.

基于协同过滤的相似检索工作步骤如下:

(1) 提交检索样例 Q;

(2) 如果检索结果不满意,则转步骤(4);

(3) 否则,记录本次检索中的相关反馈信息,结束.

(4) 用户提交相关反馈列表;

(5) 调用算法 1 得到相关对象及其相关度指标集;

(6) 扩展用户相关反馈样例集合及其相关度指标;

(7) 按照公式(4)和公式(5),计算反馈后的多媒体对象的物理特征和各分量的权重,计算相似度;

(8) 按相似度排序,输出检索结果.转步骤(2).

4 性能评价

4.1 实验系统

我们的实验系统是基于色彩与纹理的图像检索系统,图像的物理特征用颜色直方图与小波子段标准差来描述.系统在图像颜色特征上使用 $u=R/(R+G+B)$,$v=G/(R+G+B)$,u 和 v 上各取 16 个 bin,构成一个 32 维的向量,再对图像进行 3 次小波变换得到 10 个子段,取每个子段的标准差合成后得到一个 10 维的向量,即数据库中的每一个图像由两个向量来描述.在用户相关反馈中按照 Rui[3]的反馈模型进行检索样本与权重调整,并在此基础上实现了本文提出的检索方法.

实验中使用的图库包含 11 000 幅图片*.检索的主题固定为 6 个:鸟类、海洋鱼类、城堡、飞机、赛马、火车(经人工分类,每个主题含 100 幅相关图片).实验中用于对比的是 Rui[3]方法,以下称为传统方法.反馈日志文件包含对 10 个主题进行的共 700 次不同检索的反馈记录.

4.2 检索性能评价的标准

多媒体信息检索中一般采用查全率(recall)和查准率(precision)来评判检索系统的性能:*recall*=|*relevant* ∩ *retrieved*|/|*relevant*|,*precision*=|*relevant* ∩ *retrieved*|/|*retrieved*|.其中 *retrieved* 表示系统返回的检索结果集合,*relevant* 表示数据库中实际相关的对象集合.由于我们的数据库中每个主题有 100 幅相关图像,我们统计每次检索后系统输出的前 100 幅图像中实际相关的图像个数,按照上述查全率和查准率的公式,在本文的实验中,二者数值上是相等的,以下统称为检索精度.

4.3 实验结果与分析

4.3.1 检索性能的改进

首先考察本系统的检索性能,使用本文的方法分别和传统反馈方法对每一实验子类各进行 10 次检索,每一次检索进行五六次反馈.各子类的平均最大检索精度对比结果在图 1 中给出.图 2 给出了平均检索性能对比结果.从实验结果可以看出,检索精度比传统反馈平均提高了 6.85%(初始无反馈为 9.93%,传统反馈为 20.05%,本文的方法为 26.9%,如图 2 所示).所以,实验显示本方法具有更好的检索性能.

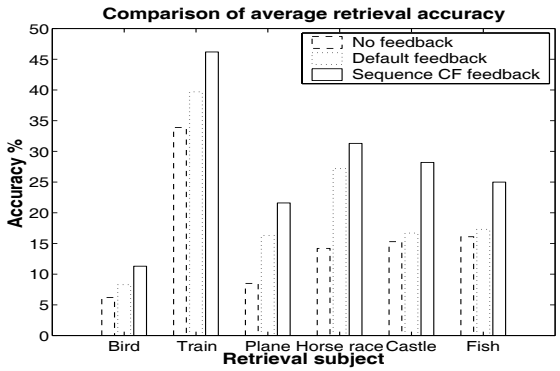

Fig.1 Comparison of the maximum retrieval accuracy
图 1 最大检索精度对比直方图

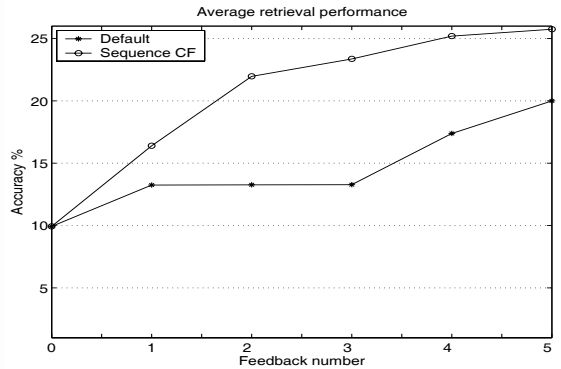

Fig.2 Comparison of the average retrieval performance
图 2 平均检索性能的对比结果

4.3.2 反馈日志记录的数量和检索精度的关系

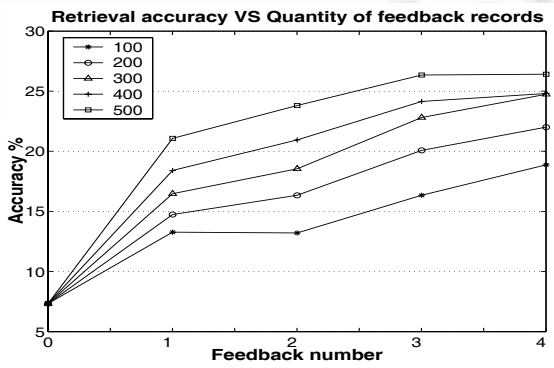

Fig.3 The relationship between relevance feedback records and the retrieval performance
图 3 反馈日志记录的数量与检索性能的关系

为了考察反馈日志的数量对检索精度的影响,我们从 700 条反馈记录中分别随机提取 100~500 条记录组成 5 个大小不等的按 100 递增的反馈日志文件.使用这 5 个反馈日志文件分别对海洋鱼类、飞机和城堡 3 个主题各进行 5 次检索,每次检索进行 4 次反馈.图 3 给出了总的平均检索性能和反馈日志记录数的关系.由于我们是随机选取的反馈例,因此反馈日志文件的增加并不总是意味着所有子类的相关反馈记录数都有显著的增长.但是从平均结果可以看出,检索的精度随着反馈日志记录的条数的增多而同步增长.因此,即随着反馈记录的积累,系统的检索性能会不

* 感谢 Dr. James WANG 提供下载的实验数据 http://jzw.standford.edu/IMAGE/download/corel1m.60k.tar.

断地提高,趋近检索系统的最大检索能力.

5 结 语

相关反馈在基于内容的相似性检索中具有重要意义.如何减少反馈交互次数,提高反馈效率,是相关反馈实用化所必须解决的问题.我们认为,系统曾经发生的大量检索具有统计规律性,可以用统计的方法对相关反馈历史数据进行分析与利用.本文通过对相关反馈的序列模式的协同过滤分析,对数据库中的多媒体对象进行语义相关性预测,用那些与当前检索有较高相关度的数据库对象来扩充反馈例集合,以提高检索的性能.我们通过一个图像检索系统对本文的方法进行评价.实验显示,与传统的反馈方法相比,本文的方法能够明显地提高相关反馈的效率.

References:

[1] Rui Y, Huang TS, Mehrotra S. Content-Based image retrieval with relevance feedback in MARS. In: Proc. of the IEEE Int'l. Conf. on Image Processing. New York: IEEE Press, 1997. II815~818.

[2] Ishikawa Y, Subramanya R, Faloustos C. MinderReader: Query database through multiple examples. In: Gupta A, Shmueli O, Widom J, eds. Proc. of the 24th Int'l. Conf. on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 1998. 218~227.

[3] Rui Y, Huang TS. A novel relevance feedback technique in image retrieval. In: Proc. of the 7th ACM Int'l. Conf. on Multimedia. Orland: ACM Press, 1999. 67~70.

[4] Yu CT, Luk WS, Cheung TY. A statistical model for relevance feedback in information retrieval. Journal of the ACM, 1976,23(2): 273~286.

[5] Tong S, Chang E. Support vector machine active learning for image retrieval. In: Proc. of the 9th ACM Int'l. Multimedia Conf. Ottawa: ACM Press, 2001. 107~119.

[6] Bartolini I, Ciaccia P, Waas F. FeedbackBypass: A new approach to interactive similarity query processing. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th Int'l. Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers, 2001. 201~210.

[7] Muller H, Muller W, Squire D. Learning feature weights from user behavior in content-based image retrieval. In: Proc. of the Int'l. Workshop on Multimedia Data Mining (MDM/KDD2000). Boston, 2000. 67~72. http://www.cs.ualberta.ca/~zaiane/mdm_kdd2000/proceedings.html

[8] Kohrs A, Merialdo B. Improving collaborative filtering with multimedia indexing techniques to create user-adapting Web sites. In: Proc. of the 7th ACM Int'l. Conf. on Multimedia. Orland: ACM Press, 1999. 27~36.

[9] Goldberg D, Nichols D, Oki B, Terry D. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992,35(12):61~70.

[10] Sarwar B, Karypis G, Konstan J, Riedl J. Analysis of recommendation algorithms for E-commerce. In: Proc. of the 2nd ACM Conf. on Electronic Commerce (EC 2000). Minneapolis: ACM Press, 2000. 158~167.

[11] Herlocker J, Konstan J, Borchers A, Riedl J. An algorithmic framework for performing collaborative filtering. In: Proc. of the 22nd Annual Int'l. ACM SIGIR Conf. on Research and Development in Information Retrieval. Berkeley: ACM Press, 1999. 230~237.

[12] Berghel VH, Roach D. An extension of Ukkonen's enhanced dynamic programming ASM algorithm. ACM Trans. on Information System, 1996,14(1):94~106.