

# Prob1 Doc

洪方舟

2016013259

Email: [hongfz16@163.com](mailto:hongfz16@163.com)

2018 年 5 月 26 日

## 测试系统介绍

### RTtee 接口标准化

### 测试方法

#### Problem1

经过前期预实验发现，测试结果在较少采样量下无法呈现一致的规律，因此实验一中对于每一种维数的特征，每一个查询集大小，重复测试 20000 次，取平均结果。为了观察到较为准确的结果，本实验中对维数的采样范围为 (4, 24) 内所有整数，对查询集的采样范围为 (1000, 5000)，每隔 500 取一点。实验中对于每个维数，每个查询集大小测试两项数据，一项是在固定的查询范围下（每一维度上查询范围均一致）的磁盘读取次数（本实验中使用节点访问次数来近似获取该数据），另一项是平均每查询出 100 个结果所需要的磁盘读取次数。由于本实验所使用的 RTtee 的实现中没有返回节点访问次数的接口，因此需要对原实现做相应修改。

#### Problem2

对于每种不同的特征，记录多组数据，包括每次查询返回的结果数量、准确率 (Precision)、召回率 (Recall)，通过下式计算 F 值 (F-Measure)

$$F = \frac{(\alpha^2 + 1)Precision \times Recall}{\alpha^2(Precision + Recall)}$$

在本实验中取  $\alpha = 1$ 。以查询返回结果数量为 x 轴，F 值为 y 轴，做出曲线分析。

### Problem3

对于每次查询出来的结果集合，对每个结果的特征向量与查询图片的特征向量的欧式距离进行排序，取距离最近的一百张图片，采取如下的打分规则，对于第  $i$  张图片，如果该图片与查询图片属于同一类，则在总得分加上  $(100 - i)$  分，否则减去  $(100 - i)$  分。多次测试取平均得分作为当前测试特征的得分。

## Problem 1

结果如上面立体直方图所示，左侧图片  $z$  轴坐标为固定每一维度查询范围的磁盘读取次数，右侧图片  $z$  轴坐标为每次查询中平均取回 100 个结果所需磁盘读取次数。下面分别对两个结果进行分析

### 固定查询范围

通过立体直方图可以直观的看出，在相同的特征维数下，随着查询集增大，平均磁盘访问次数增多。分析原因如下：相同维数下，固定查询范围，则对于某个特定的查询点，周围点的密度会随着查询集的增大而增大（该性质由测试过程中随机选取图片插入 RTree 的过程保证，由于随机选择，点的分布相对较为均匀，因此随着插入的点的数量增大，点的密度将会随之增大），因而查询返回的结果数量将会增大，从而需要访问磁盘的数量会增多。

对于相同查询集大小的情况下，随着维数增加，磁盘访问次数并没有呈现出一致的总体变化规律，但是在较低维的范围内（4-8 维），随着维数增加，磁盘访问次数增加。分析原因如下：较低维，例如 4 维特征并不能很好的表示相应的图片，导致相应的特征空间中相同类型的图片并没有很好的聚集在一起，而是整体分布较为松散，因此在固定查询范围的情况下，随着维数增大，也即特征空间中类别聚集程度增大，每次取出的结果数量增大，自然所需要的磁盘访问次数也会增大；而当维数增大到一定的程度，例如 8 维及更高的情况下，再增加维数对于特征空间中聚集程度的增加已经没有非常明显的帮助了，因此在较高维数的范围内没能观察到随维数增加磁盘查询次数增多的现象。

### 平均取回 100 个结果

通过立体直方图可以直观地看出，在相同的查询集大小下，随着特征维数增高，平均每取回 100 个查询结果所需要的磁盘访问次数增多。分析原因如下：随着特征维数的增高，RTree 中矩形框重合的数量增多（该性质可作如下简要证明，假设特征为  $n$  维，两个在每一维上取值均满足平均分布的  $n$  维矩形框不重合的概率  $P = p^n$ ，其中  $p$  为在一直线段上随机取两直线段重合的概率，且有  $p < 1$ ，因此随着特征维数的增高， $n$  维矩形框重合的概率增大），在对于每个节点的查询中平均需要访问更多的子节点，因此磁盘访问次数增多。

在维数相同的情况下，从测试结果来看，无法观察到平均取回 100 个结果所需要访问磁盘的次数与查询集大小的关系，分析认为上述两个量无明显关联。原因如下：随着查询集增大，特征空间中查询图片周围点的密度增大，因而固定查询范围的情况下，磁盘访问次数会增多，但是在平均取回 100 个结果的情况下，将查询范围的影响排除，其他因素如 RTree 节点中矩形重合情况，在相同维数下基本维持一致，因此认为平均取回 100 个结果所需要访问磁盘的次数与查询集大小无明显关联。