搜索评价

评价标准

1. 用户满意度
    1. 信息覆盖面
    2. 响应速度
    3. 界面易用性
    4. 结果相关性、准确性
        - 用户调研
        - benchmark（标准测试）
2. 用户回访率
3. 商品选择成功率

如何评价搜索准确性

- 用户调研
- 标准测试 Benchmark

评价值

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

准确率定义：

$$Precision = \frac{TP}{TP + FP}$$

召回率定义：

$$Recall = \frac{TP}{TP + FN}$$

F1度量定义：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F1是基于准确率和召回率的调和平均定义的

在一些应用中，对准确率和召回率的重视程度不同，例如在商品推销系统中，为了尽可能少打扰用户，更希望推荐内容是用户感兴趣的，此时准确率更重要.而在逃犯信息检索系统中，更希望尽可能少漏掉逃犯，此时召回率比较重要.

将F1一般化可得到F_β的定义：

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

其中$\beta = 1$时退化为标准的F1，$\beta > 1$时对准确率有更大影响，$\beta < 1$时对召回率有更大影响

从信息检索角度考虑ROC

信息检索排序后我们可以返回Top-k的结果，不同的k取值对应不同的Precision和Recall， 基于这一系列的点对，我们便能绘制出ROC，基于此我们可以得到AUC

此外也能考虑： Mean Average Precision 、 Normalized Discounted Cumulative Gain

建立一套标准测试集

1. 选择适当的文档集
2. 常见搜索任务
3. 针对每个搜素任务，对文档的相关性进行标注

- 不同专家的标注存在差异 故引入 Kappa Measure

$$[P(A) - P(E)][1 - P(E)]$$

$P(A)$：标注一致的概率，$P(E)$：随机标注情况下，一致的概率

- 经验性指标
    - 0.8： 一致

- 0.63– 0.8：基本一致
- 0.63：可疑

Kappa Measure Example：

Number of docs	Judge 1	Judge 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

```
P(A) = 370/400 = 0.925
P(nonrelevant) = (10+20+70+70)/800 = 0.2125
P(relevant) = (10+20+300+300)/800 = 0.7878
P(E) = 0.2125^2 + 0.7878^2 = 0.665
Kappa = (0.925 – 0.665)/(1–0.665) = 0.776
```

搜索引擎的在线评价

- A/B testing
  - 大部分用户使用已有的排序方法
  - 选择一小部分用户使用新的排序方法