信息检索
Information Retrieval

第三章 信息检索的评价

为什么要进行检索系统的评价

- 有很多的检索模型、排序算法……，哪一个更好的？
- 哪一个更好：
  - 排序函数（内积、余弦相似度……）
  - 词的权重计算（TF、TF-IDF）
  - ……
- 系统返回一个结果列表，用户沿着列表多久能找到所需要的结果文档？

信息检索系统评价中的困难之处

- 检索系统的效果与检索结果的相关性有关
- 相关性
  - 是一个连续的量，非二值的量
  - 即使是二值的量，也很难确定
- 从人的角度出发，相关性是：
  - 主观的：依赖于用户的主观判定
  - 时变的：依赖于用户当前的需求
  - 认知的：依赖于用户的认知和行为
  - 动态的：随着时间发生变化的

关于评价

- 评价无处不在
  - 工作、生活、娱乐、招生
- 评价是检验学术进步的唯一标准，也是杜绝学术腐败的有力武器

IR中评价什么

- 效率(Efficiency)—可以采用通常的评价方法
  - 时间开销
  - 空间开销
  - 响应速度
- 效果(Effectiveness)
  - 返回的文档中有多少相关文档
  - 所有相关文档中返回了多少
  - 返回的正确结果是否靠前

其他指标：

覆盖率（Coverage）
访问量
数据更新速度
……

如何评价效果？

- 相同的文档集合，相同的查询主题集合，相同的评价指标，不同的检索系统进行比较
  - The Cranfield Experiments, Cyril W. Cleverdon, 1957~1968（上百篇文档集合）
  - SMART System, Gerald Salton, 1964~1988（数千篇文档集合）
  - TREC (Text Retrieval Conference), Donna Harman，美国标准技术研究所，1992~（上百万篇文档）

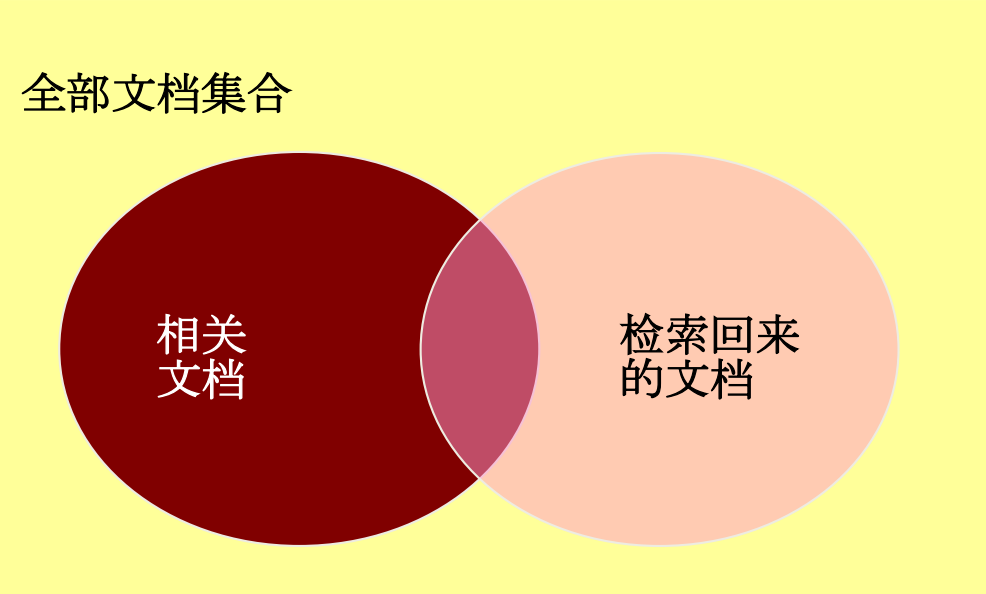
如何评价效果？

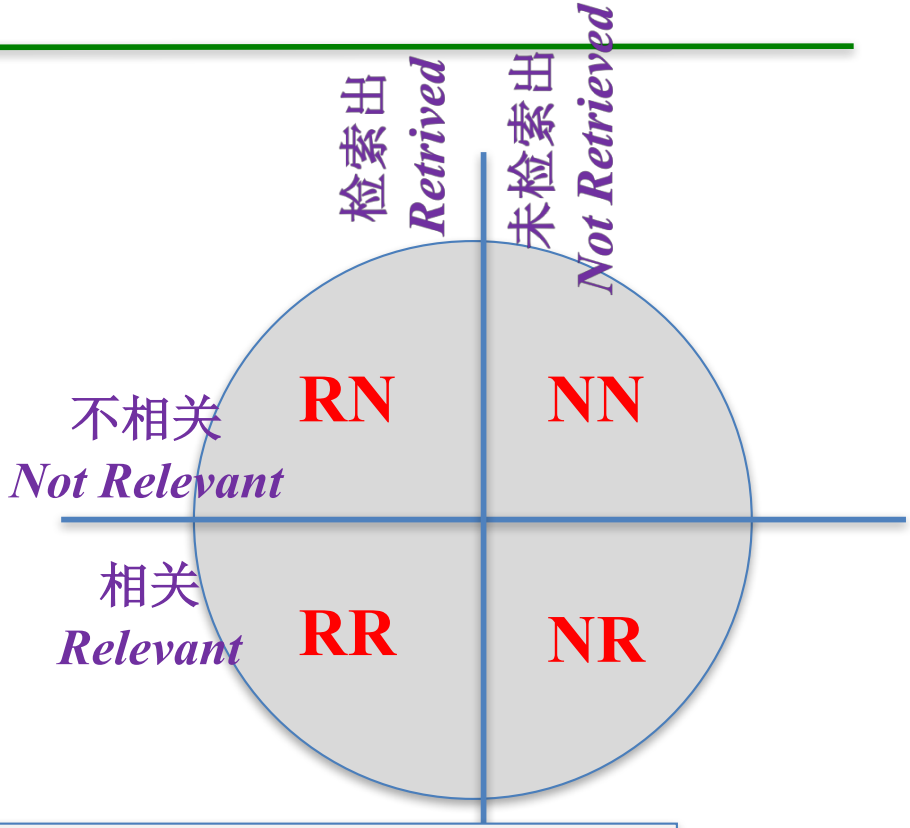
- 一个文档集合C
  - 系统将从该集合中按照查询要求检出相关文档
- 一组用户查询需求{q_1, q_2, q_3, ……, q_n}
  - 每个查询需求q_i描述了用户的查询需求
- 每个用户查询需求的标准相关文档集{R_1, R_2,…, R_n}
  - 该集合可由人工方式构造
- 一组评价指标
  - 这些指标反映系统的检索性能
  - 通过比较系统实际检出的结果文档集和标准的相关文档集，对它们的相似性进行量化，得到这些指标值

主要内容

- 准确率、召回率、F值
- 单值概括（不考虑召回率）
  - MAP，R-Precision，准确率直方图，Precision@N，RR&MRR，Bpref，NDCG
- 相关评测
- 一致性检验

准确率、召回率、F值





正确率（**Precision**）：$\frac{RR}{RR+RN} = \frac{\text{检索回来的相关文档数}}{\text{检索回来的文档总数}}$

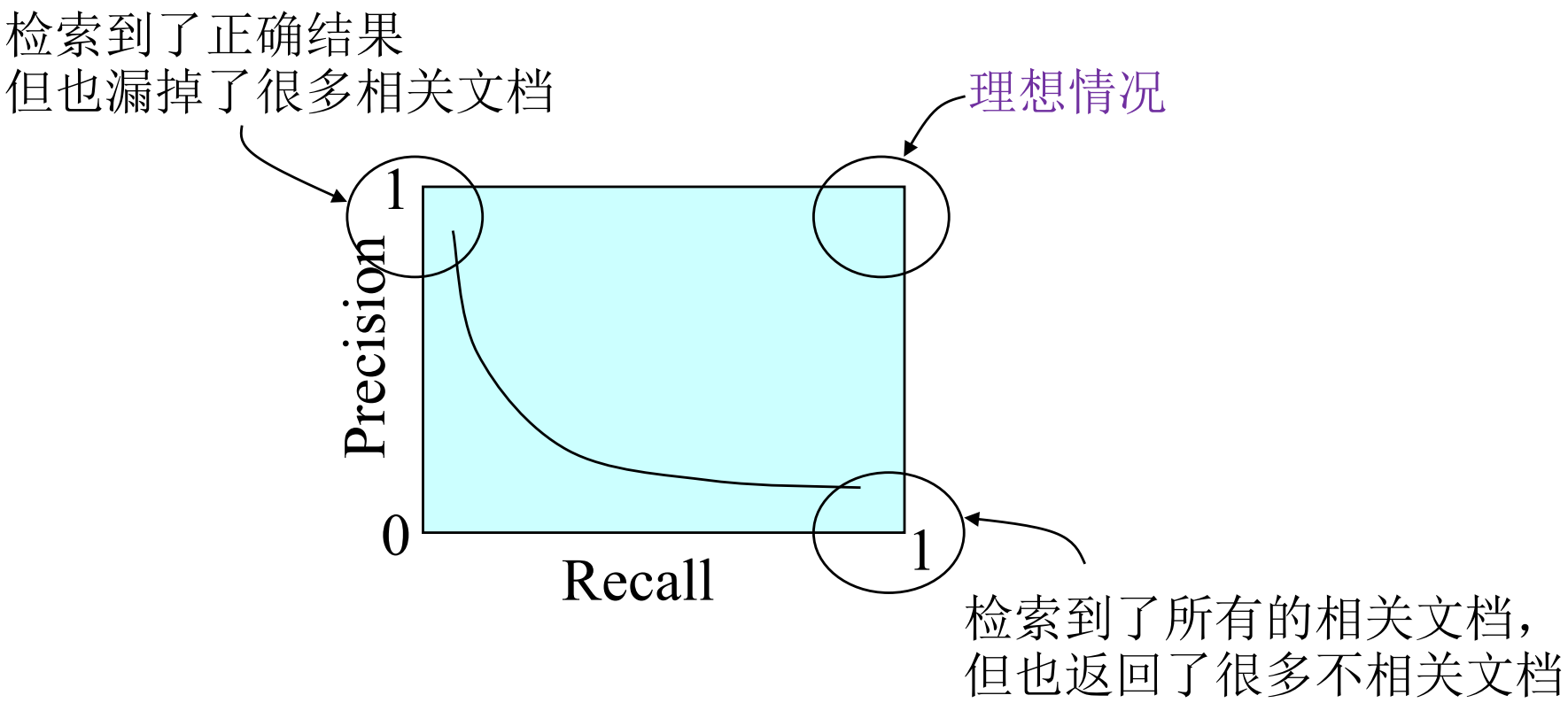
召回率（**Recall**）：$\frac{RR}{RR+NR} = \frac{\text{检索回来的相关文档数}}{\text{相关文档总数}}$

准确率、召回率、F值

- 召回率的确定很困难
  - 大规模文档集中，针对一个查询确定相关文档的数量很困难
  - “Pooling”方法
    - 假设
      - 绝大多数的相关文档都收录在这个文档池中
      - 没有进行判断的文档即未被认为是不相关的
    - 操作过程
      - 针对某一检索问题，所有参与其检索试验的系统分别给出各自检索结果中的前K个文档（例如$K=100$），将这些结果文档汇集起来，得到一个可能相关的文档池“*pool*”

准确率、召回率、F值

- 准确率和召回率的权衡分析



准确率、召回率、F值

- 计算准确率/召回率的分数

① 对于一个给定的查询，生成一个排好序的检索结果

② 调整阈值（threshold），这样可以得到不同的排好序的文档集合，在此基础上可以得到不同的准确率/召回率结果

③ 对结果集中的每一个文档进行相关性标注

④ 针对排序结果中每一个相关文档的位置计算准确率/召回率

准确率、召回率、F值

- 计算准确率/召回率的分数—例1

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

相关文档总数= 6

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

准确率、召回率、F值

- 计算准确率/召回率的分数—例2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

相关文档总数= 6

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/3=0.667

R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

R=6/6=1.0; p=6/14=0.429

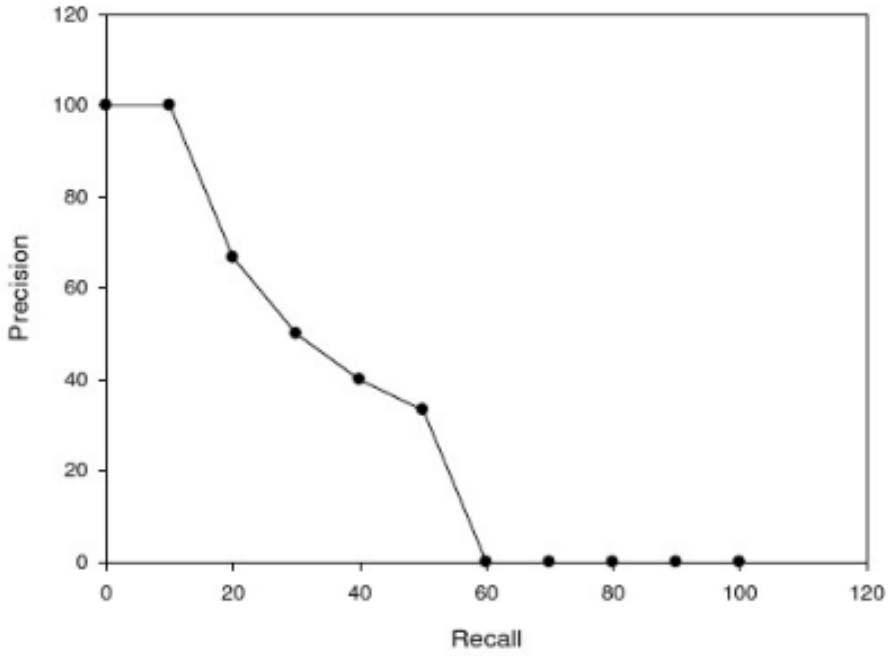
准确率、召回率、F值

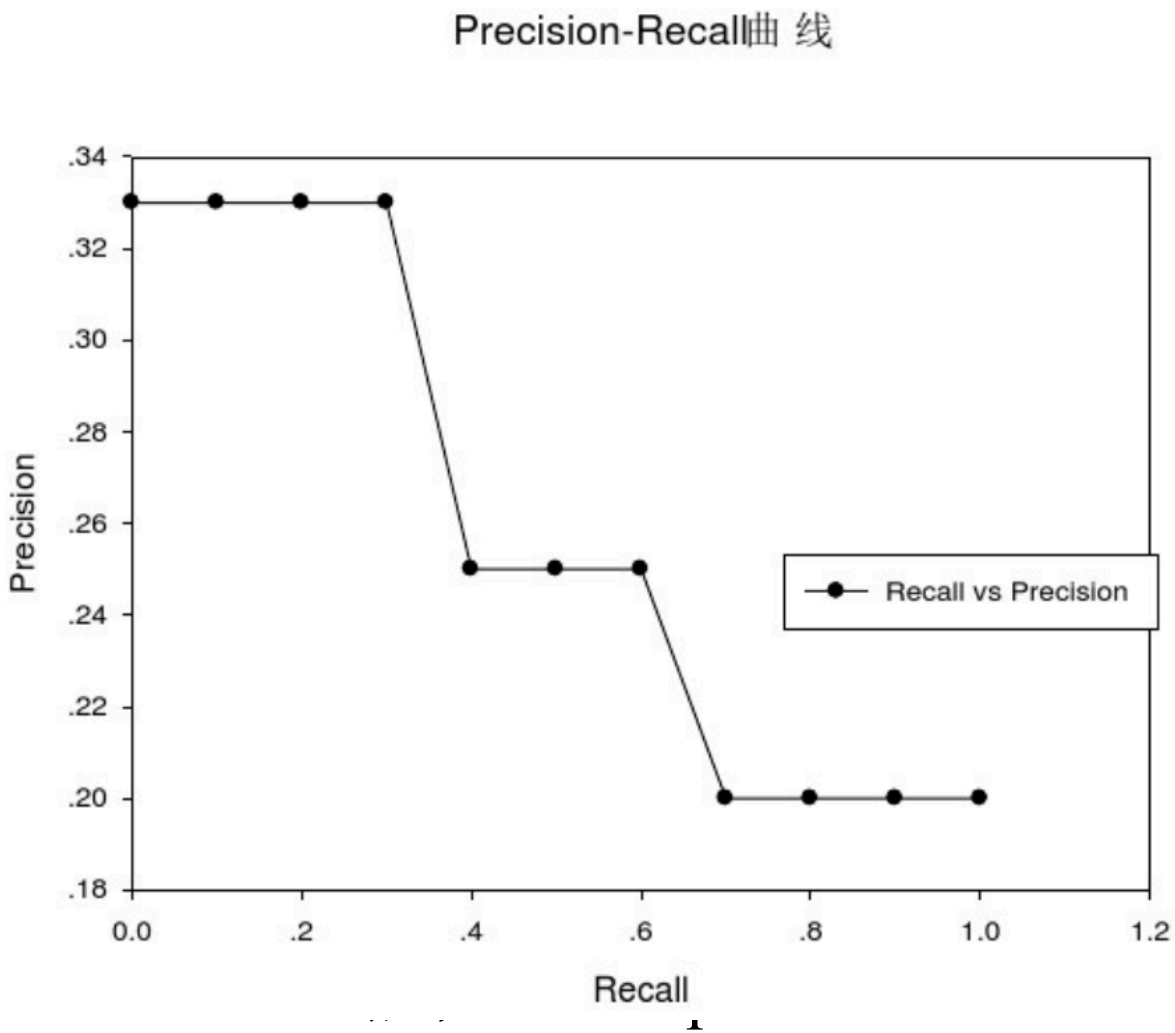
- P/R曲线
  - 检索结果以排序方式排列，用户在观察过程中，正确率和召回率在不断变化
  - 可以求出召回率分别在0%，10%，20%……100%上对应的准确率，然后绘出图像

准确率、召回率、F值

- P/R曲线—例子
  - 某个查询q的标准答案集合为
    - $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
  - 某个IR系统对q的检索结果如下

1. d_{123} R=0.1, P=1	9. d_{187}
2. d_{84}	10. d_{25} R=0.4, P=0.4
3. d_{56} R=0.2, P=0.67	11. d_{38}
4. d_6	12. d_{48}
5. d_8	13. d_{250}
6. d_9 R=0.3, P=0.5	14. d_{113}
7. d_{511}	15. d_3 R=0.5, P=0.33
8. d_{129}	





1. d_{123}	9. d_{187}
2. d_{84}	10.d_{25}
3. d_{56} R=0.33, P=0.33	11.d_{38}
4. d_6	12.d_{48}
5. d_8	13.d_{250}
6. d_9	14.d_{113}
7. d_{511}	15.d_3 R=1, P=0.2
8. d_{129} R=0.66, P=0.25	

的召回率点，只存

个召回率点

下存在的召回率点进行

- 对于t%，如果不存在该召回率点，则定义t%为t%到（t+10）%中最大正确率值
- 对于上例，0%，10%，20%，30%上正确率为0.33，40%~60%正确率为0.25，70%以上对应0.2

准确率、召回率、F值

- P/R曲线
  - 优点
    - 简单直观
    - 既考虑了检索结果的覆盖度，有考虑了检索结果的排序情况
  - 缺点
    - 单个查询的P/R曲线虽然直观，但是难以明确表示两个查询的检索结果的优劣

准确率、召回率、F值

- F值（F-measure）
  - 准确率和召回率的调和平均值
    - 如果P=0 or R=0，那么F=0
    - 否则，F值如下计算

$$F = \frac{2}{1/P + 1/R} = \frac{2PR}{P+R}$$

1. d_{123}	9. d_{187}
2. d_{84}	10.d_{25}
3. d_{56} R=0.33, P=0.33, F=0.33	11.d_{38}
4. d_6	12.d_{48}
5. d_8	13.d_{250}
6. d_9	14.d_{113}
7. d_{511}	15.d_3 R=1, P=0.2, F=0.33
8. d_{129} R=0.66, P=0.25, F=0.36	

准确率、召回率、F值

- F值（F-measure）
  - E值（E-measure）
    - F值的变种

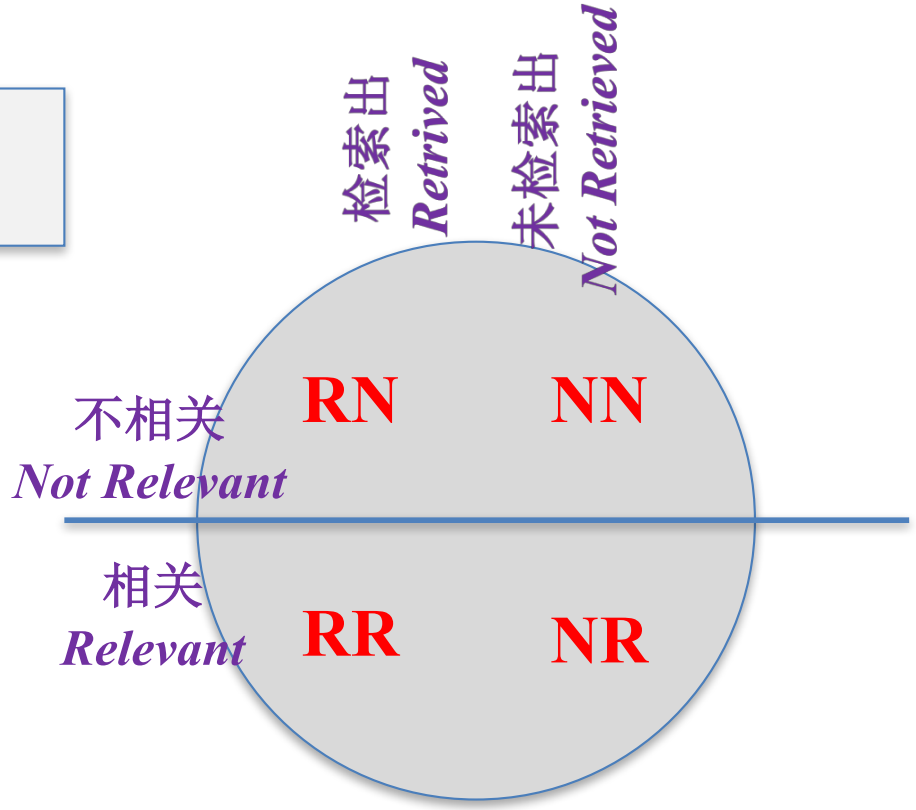
$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\beta^2/_R + 1/_P} \qquad (\beta \geq 0)$$

    - β值
      - $\beta = 1$：准确率、召回率同等重要（E=F）
      - $\beta > 1$：召回率更重要
      - $\beta < 1$：准确率更重要

准确率、召回率、F值

- 精确率（Accuracy）
  - 精确率是所有判定中正确的比率

$$accuracy = \frac{RR + NN}{RN + RR + NR + NN}$$



准确率、召回率、F值

- 精确率（Accuracy）
  - 精确率不适合IR的原因
    - 由于和查询相关的文档占文档集的极少数，所以即使什么都不返回也会得到很高的精确率
    - 信息检索用户希望找到某些文档并且能够容忍结果中有一定的不相关性
    - 返回一些即使不好的文档也比不反回任何文档强

准确率、召回率、F值

- 准确率、召回率、F值的应用领域
  - 拼写校对
  - 中文分词
  - 文本分类
  - 人脸识别
  - ……

准确率、召回率、F值

- 准确率、召回率、F值的讨论（1）
  - “宁可错杀一千，不可放过一人”
    - 偏重召回率，忽视正确率
  - 判断一个人是否有罪
    - 如果没有证据证明你无罪，那么判定你有罪
      - 召回率高，有些人受冤枉
    - 如果没有证据证明你有罪，那么判定你无罪
      - 召回率低，有些人逍遥法外

准确率、召回率、F值

- 准确率、召回率、F值的讨论（2）
  - 不同的用户，对准确率、召回率的要求不同
    - 垃圾邮件过滤
      - 宁愿漏掉一些垃圾邮件，但是尽量少将正常邮件判定成垃圾邮件（侧重召回率）

有些用户希望返回的结果全一点，他有时间挑选

有些用户希望返回的结果准一点，他不需要结果很全就能完成任务

主要内容

- 准确率、召回率、F值
- 单值概括（不考虑召回率）
  - MAP，R-Precision，准确率直方图，Precision@N，RR&MRR，Bpref，NDCG
- 相关评测
- 一致性检验

单值概括（不考虑召回率）

- MAP（Mean Average Precision ）
  - 平均准确率（AP）
    - 对不同召回率点上的准确率进行平均
    - 例1
      - (1 + 1 + 0.75 + 0.667 + 0.38)/5 = 0.7594
    - 例2
      - (1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429)/6 = 0.625
  - MAP
    - 查询集合中，每个查询的平均准确率的平均值

单值概括（不考虑召回率）

- R-Precision
  - 给定一个查询q，排序结果列表中第R个位置的准确率
    - 针对查询q，相关文档的总数为R

R = 相关文档总数 = 6

R-Precision = 4/6 = 0.67

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

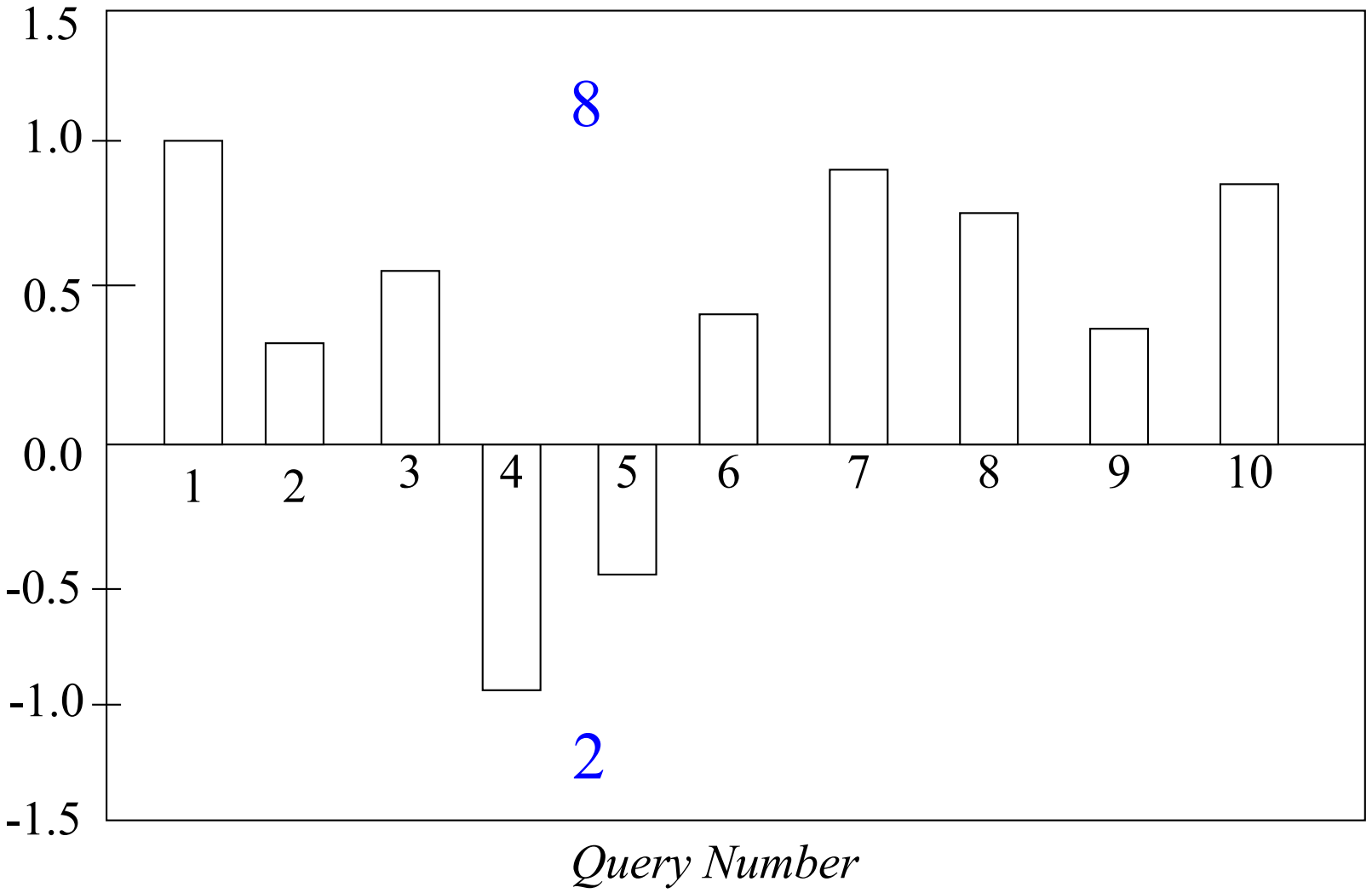
单值概括（不考虑召回率）

- 准确率直方图
  - 多个查询的 *R-Precision* 测度
  - 用来比较两个算法的检索纪录

$$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$

  - RP_A和RP_B是针对第i个查询检索算法A和B的*R-Precision*值
    - $RP_{A/B}=0$: 对于第 i 个查询，两个算法有相同的性能
    - $RP_{A/B}>0$: 对于第 i 个查询，算法 A 有较好的性能
    - $RP_{A/B}<0$: 对于第 i 个查询，算法 B 有较好的性能

单值概括（不考虑召回率）



单值概括（不考虑召回率）

- Precision@N
  - 第N个位置上的准确率
    - P@5，P@10，P@20.....
  - Precision@N例子
    - P@1=1
    - P@2=1
    - P@5=3/5=0.6
    - P@10=4/10=0.4
    - ……

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

单值概括（不考虑召回率）

- RR&MRR

*RR*评价是基于*2*元相关判断基础上的，因此*RR*与*MRR*都不能区分一个高相关性的文档与低

  - RR（*Reciprocal Ranking*）排序倒数
    - 第一个相关文档出现位置的倒数
    - RR=1/r，其中r为第一个相关文档在结果中出现的位置
  - MRR（*Mean Reciprocal Ranking*）平均排序倒数
    - 是在*RR*的基础上对多个查询的*RR*结果取平均值
    - *MRR=0.25* 就意味着检索系统平均在返回结果的第四个位置找到相关文档

$$MRR = \frac{\sum_{q=1}^{n} \frac{1}{rank_q}}{n}$$

单值概括（不考虑召回率）

- Bpref（Binary preference）
  - 基本思想
    - 相关性判断不完全的情况下
      - 计算在进行了相关性判断的文档集合中，在判断到相关文当前，需要判断的不相关文档的篇数
    - 相关性判断完整的情况下
      - 利用Bpref和MAP进行评价的结果很一致

Buckley, C. & Voorhees, E.M. Retrieval Evaluation with Incomplete Information, Proceedings of SIGIR 2004.

单值概括（不考虑召回率）

- Bpref（Binary preference）
  - 对每个查询，已判定结果中有R个相关结果

$$Bpref = \frac{1}{R}\sum_{r}(1 - \frac{|n\text{排在}r\text{前面}|}{R})$$

  - r是相关文档，n是Top R篇不相关文档集合的子集
  - 例子
    - $S = \{D_1, D_2^*, D_3, D_4, D_5^*, D_6, D_7^*, D_8, D_9, D_{10}\}$
    - 其中D_2、D_5 和 D_7 是相关文档，D_3 和 D_4 为未经判断的文档
    - $R=3;\quad bpref= 1/3\ [(1 -1/3) + (1 -1/3) + (1 -2/3)]$

单值概括（不考虑召回率）

- NDCG（Normalized Discounted Cumulative Gain）
  - 针对一个查询，很少有文档完全相关或者完全不相关
  - 相关度级别
    - 如0，1，2，3，4
    - 相关度级别越高的结果越多越好
    - 相关度级别越高的结果越靠前越好

单值概括（不考虑召回率）

- NDCG（Normalized Discounted Cumulative Gain）
  - CG（Cumulative Gain）

$$CG_n = \begin{cases} G[1], & if\ i = 1 \\ CG[i-1] + G[i], & otherwise \end{cases}$$

n	doc #	relevance (gain)	CG_n
1	588	1.0	1.0
2	589	0.6	1.6
3	576	0.0	1.6
4	590	0.8	2.4
5	986	0.0	2.4
6	592	1.0	3.4
7	984	0.0	3.4
8	988	0.0	3.4
9	578	0.0	3.4
10	985	0.0	3.4
11	103	0.0	3.4
12	591	0.0	3.4
13	772	0.2	3.6
14	990	0.0	3.6

单值概括（不考虑召回率）

- NDCG（Normalized Discounted Cumulative Gain）
  - Discounted CG Vector
    - 用户更关心的是排序靠

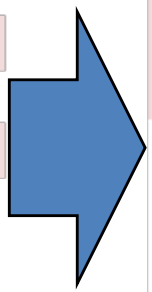
$$DCG_n = \begin{cases} CG[i], & i < b \\ DCG[i-1] + \dfrac{G[i]}{\log_b i}, & i \geq b \end{cases}$$

n	doc #	rel (gain)	CG_n	$\log_n$	DCG_n
1	588	1.0	1.0	-	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

单值概括（不考虑召回率）

- NDCG（Normalized Discounted Cumulative Gain）
  - 理想情况下的排序（IDCG）

n	doc #	rel (gain)	CG_n	$\log_n$	DCG_n
1	588	1.0	1.0	0.00	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44



n	doc #	rel (gain)	CG_n	$\log_n$	$IDCG_n$
1	588	1.0	1.0	0.00	1.00
2	592	1.0	2.0	1.00	2.00
3	590	0.8	2.8	1.58	2.50
4	589	0.6	3.4	2.00	2.80
5	772	0.2	3.6	2.32	2.89
6	576	0.0	3.6	2.58	2.89
7	986	0.0	3.6	2.81	2.89
8	984	0.0	3.6	3.00	2.89
9	988	0.0	3.6	3.17	2.89
10	578	0.0	3.6	3.32	2.89
11	985	0.0	3.6	3.46	2.89
12	103	0.0	3.6	3.58	2.89
13	591	0.0	3.6	3.70	2.89
14	990	0.0	3.6	3.81	2.89

单值概括（不考虑召回率）

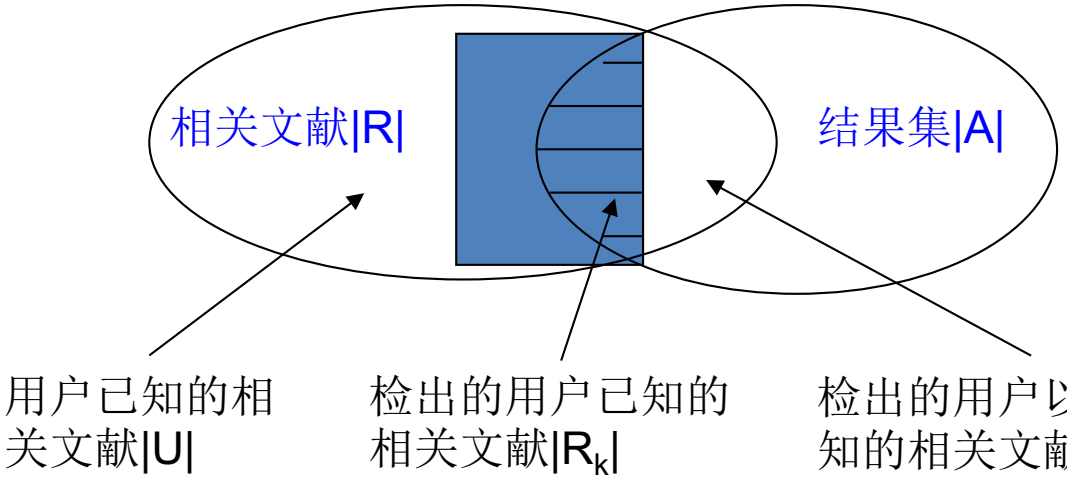
- NDCG（Normalized Discounted Cumulative Gain）
  - 对所得结果进行归一化

$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

n	doc #	rel (gain)	DCG_n	$IDCG_n$	$NDCG_n$
1	588	1.0	1.00	1.00	**1.00**
2	589	0.6	1.60	2.00	**0.80**
3	576	0.0	1.60	2.50	**0.64**
4	590	0.8	2.00	2.80	**0.71**
5	986	0.0	2.00	2.89	**0.69**
6	592	1.0	2.39	2.89	**0.83**
7	984	0.0	2.39	2.89	**0.83**
8	988	0.0	2.39	2.89	**0.83**
9	578	0.0	2.39	2.89	**0.83**
10	985	0.0	2.39	2.89	**0.83**
11	103	0.0	2.39	2.89	**0.83**
12	591	0.0	2.39	2.89	**0.83**
13	772	0.2	2.44	2.89	**0.84**
14	990	0.0	2.44	2.89	**0.84**

单值概括（不考虑召回率）

- 面向用户的测度方法
  - 覆盖率
    - 实际检出的相关文献中用户已知的相关文献所占比例
  - 新颖率
    - 实际检出的相关文献中用户未知的相关文献所占比例



$$coverage = \frac{|R_k|}{|U|}$$

$$novelty = \frac{|R_u|}{|R_u| + |R_k|}$$

主要内容

- 准确率、召回率、F值
- 单值概括（不考虑召回率）
  - MAP，R-Precision，准确率直方图，Precision@N，RR&MRR，Bpref，NDCG
- 相关评测
- 一致性检验

相关评测

- TREC评测（Text Retrieval Conference）
  - 由NIST和DARPA联合举办， http://trec.nist.gov
    - NIST： the National Institute of Standards and Technology，美国国家标准技术协会
    - DARPA： the Defense Advanced Research Projects Agency，美国国防部
  - 1992年第一届（文本、语音、图像、视频等）
  - 支持在信息检索领域的基础研究，提供对大规模文本检索方法的评估办法

相关评测

- TREC评测（Text Retrieval Conference）
  - TREC以年度为周期运行
    - **确定任务**：NIST提供测试数据和测试问题
    - **参赛者报名**：参赛者根据自己的兴趣选择任务
    - **参赛者运行任务**：参赛者用自己的检索系统运行测试问题，给出结果
    - **返回运行结果**：向NIST返回运行结果，以便评估
    - **结果评估**：NIST用一套固定的方法和软件对参赛者的运行结果给出评测结果
    - **大会交流**：每年11月召开会议，由当年的参赛者交流彼此的经验

相关评测

- TREC评测（Text Retrieval Conference）
  - TREC中名词定义
    - Track
      - TREC的每个子任务，如QA、Filtering、Web、Blog等
    - Topic
      - 预先确定的问题，用来向检索系统提问
      - Topic->query（自动或手工）
      - Question（QA）
    - Document （TIPSTER&TREC CDs、WT2G、WT10G、GOV2）
      - 包括训练集和测试集合
    - Relevance Judgments
      - 相关性评估，人工或者半自动

相关评测

- TREC评测（
  - Topic的一般
    - Title
      - 标题，通
    - Description
      - 描述，一
    - Narrative
      - 详述，更

<num>Number 351

<title>Falkland Petroleum Exploration

<desc>Description:

What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

<narr>Narrative:

Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.

相关评测

- TREC评测（Text Retrieval Conference）
  - Topic的使用方式
    - 可以利用Topic文本中的部分或者全部，构造适当的查询条件
    - 可以使用任何方式构造查询条件（手工、自动）
      - 提交查询结果时要注明查询条件的产生方式

相关评测

- 其他评测
  - TRECVID
    - Video检索方面的评测
    - 2003年从TREC中的Video Track任务中分离出来
  - MUC（Message Understanding Conference）
    - DARPA组织的有关信息抽取的评测会议
    - 1991年开始，1997年为最后一届（后来演变为ACE评测）
    - 最后两届加入了命名实体识别和共指消解任务

相关评测

- 其他评测
  - ACE评测（Automatic Content Extraction）
    - 美国NIST组织，主要面向新闻领域的文本，抽取其中的实体、关系和事件
    - 2000年开始，每年1届
  - DUC评测（Document Understanding Conference）
    - 2001年开始，由NIST组织
    - 面向文档摘要的评测会议（单文档摘要、多文档摘要....）

相关评测

- 其他评测
  - NTCIR（NII Test Collection for IR systems）
    - 日本国立情报学研究所组织的关于亚洲语言相关的IR评测
  - CLEF
    - 有关欧洲语言相关的IR评测（跨语言）
  - TAC
    - 2008年，将DUC任务和TREC中QA任务合并
  - 863评测、中文信息学会的倾向性分析评测等

主要内容

- 准确率、召回率、F值
- 单值概括（不考虑召回率）
  - MAP，R-Precision，准确率直方图，Precision@N，RR&MRR，Bpref，NDCG
- 相关评测
- 一致性检验

一致性检验

- 用户判定的有效性
  - 只有在用户的判定一致时，相关性判定的结果才可用
  - 如果结果不一致，那么不存在标准答案
  - 如何度量不同判定人之间的一致性？
    - Kappa指标

一致性检验

- Kappa
  - 度量判定间一致性的指标
  - 为类别型判断结果所设计的指标

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

  - P(A)：观察到的一致性判断比例
  - P(E)：随机情况下所期望的一致性判断比例

κ值在$\left[\frac{2}{3}, 1.0\right]$时，判定结果是可以接受的

一致性检验

- Kappa
  - 观察到的两个人一致性判断比率

$$P(A) = \frac{300 + 70}{400} = 0.925$$

第1个人相关性判定结果	第2个人相关性判定结果		
	yes	no	total
yes	300	20	320
no	10	70	80
total	310	90	400

  - 边缘统计量

$$P(nonrelevant) = \frac{80 + 90}{400 + 400} = 0.2125$$

$$P(relevant) = \frac{320 + 310}{400 + 400} = 0.7878$$

  - 两个人的随机一致性比率

$$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$$

  - Kappa统计量

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776$$

本章小结

- 为什么要进行评价
- 掌握基本的评价指标
  - 准确率、召回率、F值
  - 单值概括
    - MAP、R-Precision、P@N.....
- 了解相关的评测会议
- 掌握一致性检验的判定方法