# 信息检索与Web搜索

## 第9讲 检索评价&结果摘要

## Evaluation & Snippets

授课人：高曙明

改编自“现代信息检索”网上公开课件（http://ir.ict.ac.cn/~wangbin）

# 关于评价

- 评价很必要，是决策依据，无处不在
  - 大学、教师、学生、城市
- 评价不容易
  - 如何反映本质的好与差
  - 如何保证可比性
- 好的评价体系和评价结果促进发展
- IR方法和系统也需要评价

# IR的评价要素

- **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
  - 最相关的是否返回得最靠前
- **效率 (Efficiency)**
  - 时间开销
  - 响应速度
- **成本**

# 如何评价效果？

- 统一的评价指标
- 标准测试集
  - 一个文档集，一组用于测试的查询
  - 一组相关性判定结果
- 代表性测试集
  - 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)，信息检索的“奥运会”

# 评价任务的例子

- 两个系统，一批查询，对每个查询每个系统分别得到一些结果。目标：哪个系统好?

| 系统&查询 | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| 系统1，查询1 | d3 | d6 | d8 | d10 | |
| 系统1，查询2 | d1 | d4 | d7 | d11 | |
| 系统2，查询1 | d6 | d7 | d8 | d9 | |
| 系统2，查询2 | d1 | d2 | d4 | d13 | |

# 评价指标

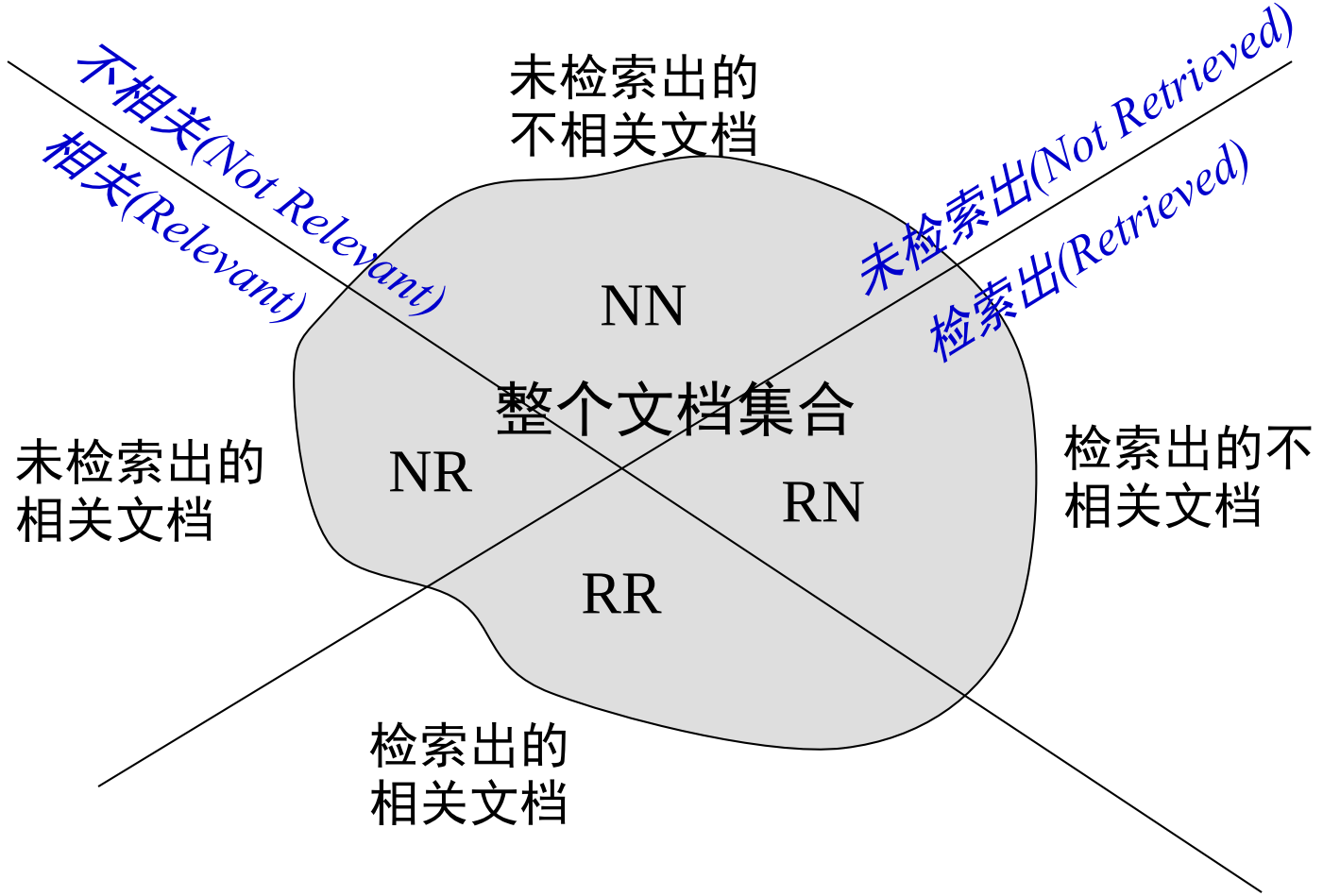
- **评价指标：**某个或某几个可衡量、可比较的值
- **指标分类：**
  - 不考虑序的评价指标
  - 考虑序的评价指标
  - 对单个查询进行评估的指标
  - 对多个查询进行评估的指标

# 评价任务的例子

| 系统&查询 | 1 | 2 | 3 | 4 | … |
|---|---|---|---|---|---|
| 系统1，查询1 | d3 √ | d6 √ | d8 | d10 | |
| 系统1，查询2 | d1 | d4 | d7 | d11 | |
| 系统2，查询1 | d6 √ | d7 | d8 | d9 √ | |
| 系统2，查询2 | d1 | d2 | d4 | d13 | |

对于查询1的标准答案集合 {d3,d4,d6,d9}

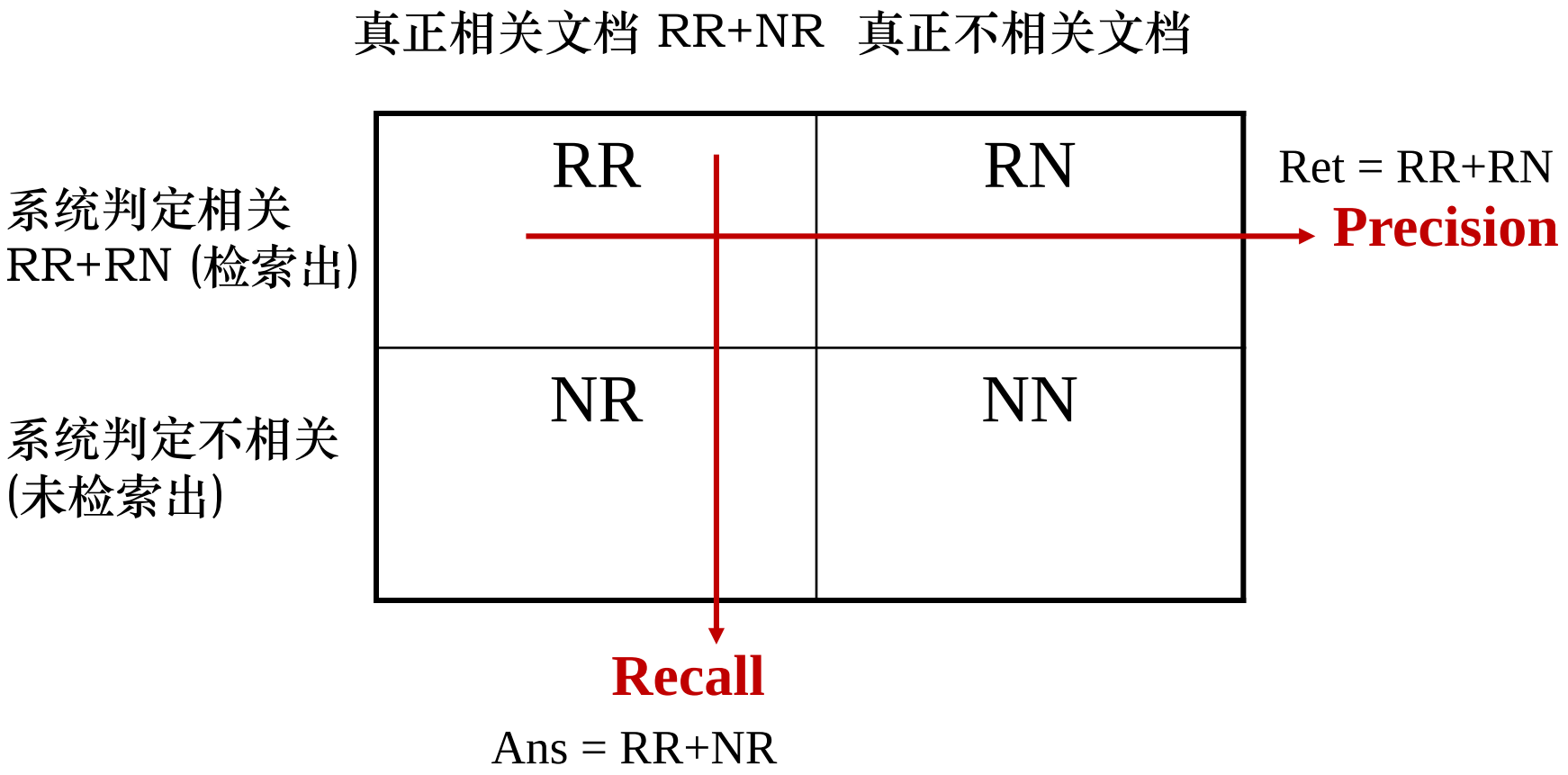
# 整个文档集合的划分


不相关(Not Relevant)
相关(Relevant)
未检索出的
不相关文档
未检索出(Not Retrieved)
检索出(Retrieved)
NN
整个文档集合
未检索出的
相关文档
NR
RN
检索出的不
相关文档
RR
检索出的
相关文档

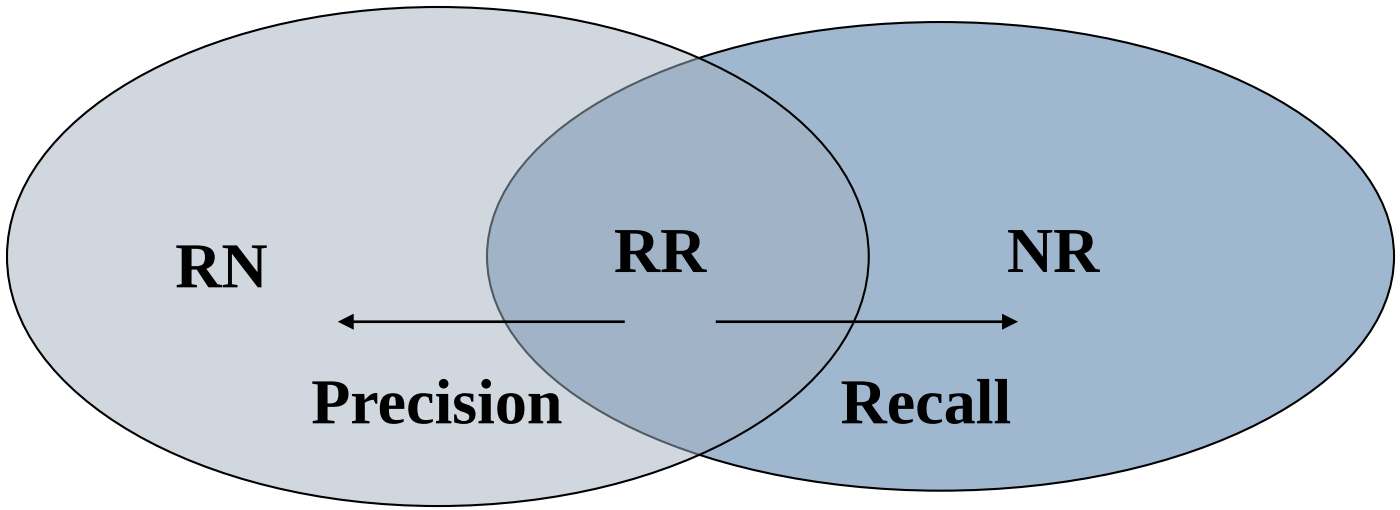
# P/R评价指标

- **正确率(Precision):** RR/(RR + RN)，返回的结果中真正相关结果的比率，也称为**查准率**， P∈ [0,1]
- **召回率(Recall):** RR/(RR + NR)，返回的相关结果数占实际相关结果总数的比率，也称为**查全率**，R∈ [0,1]
- 两个指标分别度量检索效果的某个方面，相对独立，差异可以很大
  - 返回有把握的1篇，P=100%，但R极低
  - 全部文档都返回，R＝100%，但P极低

# P/R与RR/RN/NR/NN之间的关系



# 基于集合的图表示




返回结果Ret　　标准答案Ans

# 评价任务的例子

| 系统&查询 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 系统1，查询1 | d3 √ | d6 √ | d8 | d10 | d11 |
| 系统1，查询2 | d1 | d4 | d7 | d11 | d13 |
| 系统2，查询1 | d6 √ | d7 | d8 | d9 √ | |
| 系统2，查询2 | d1 | d2 | d4 | d13 | d14 |

对于查询1的标准答案集合 {d3,d4,d6,d9}

对于系统1，查询1，正确率2/5，召回率2/4

对于系统2，查询1，正确率2/4，召回率2/4

# 另一个计算例子

- 给定查询Q，本应该有100篇相关文档，某个系统返回200篇文档，其中80篇是真正相关的文档
- Recall=80/100=0.8
- Precision=80/200=0.4
- 结论：召回率较高，但是正确率较低

# P/R评价指标的优点

- 采用P、R两个指标来度量效果，具有灵活性
- 不同的应用、不同的用户可以偏重不同的指标
  - Web搜索：偏重高正确率
  - 情报分析：偏重高召回率
  - 垃圾邮件过滤：偏重高正确率

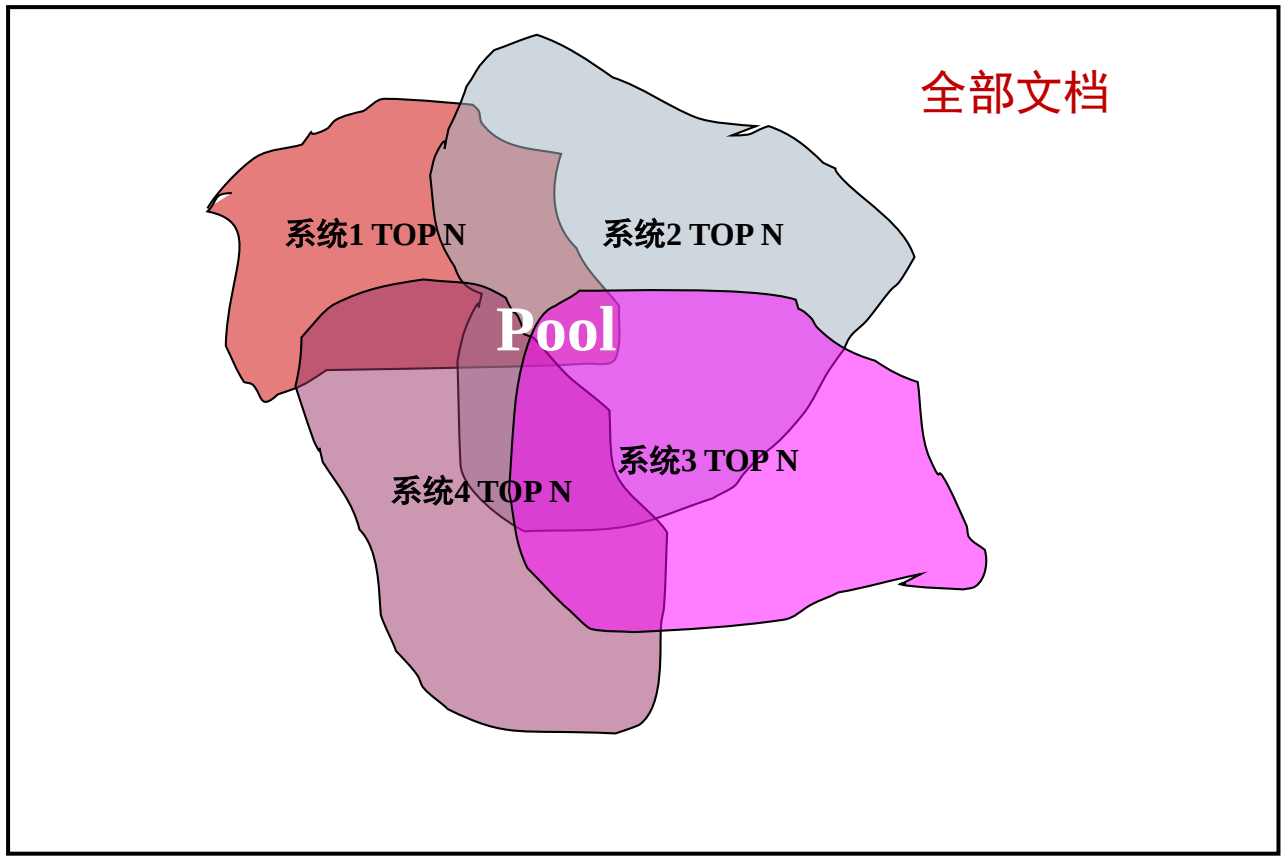
# P/R评价指标的问题

- 召回率难以计算
- 两个指标分别衡量了系统的某个方面，如何综合评价?
- 两个指标都是基于集合进行计算，并没有考虑序的影响
  - 比如，两个系统，对某个查询，返回的相关文档数目一样都是10，但是第一个系统是前10条结果，后一个系统是最后10条结果。显然，第一个系统更优，但是根据基于集合的计算，两者完全一样

# 关于召回率的计算

- **难点：**对于大规模语料集合，列举每个查询的所有相关文档是不可能的事情
- **缓冲池(Pooling)方法：**对多个检索系统的Top N个结果组成的集合进行人工标注，标注出的相关文档集合作为整个相关文档集合。这种做法被验证是可行的(可以比较不同系统的相对效果)，在TREC会议中被广泛采用。

# 4个系统的Pooling



# P和R的融合

- F值(F-measure)：召回率R和正确率P的调和平均值，if P=0 or R=0, then F=0, else 采用下式计算：

$$F=\frac{2}{\frac{1}{P}+\frac{1}{R}}=\frac{2PR}{P+R}\ (P\neq 0, R\neq 0)$$

- $F_\beta$：表示召回率的重要程度是正确率的$\beta$(>=0)倍， $\beta$>1更重视召回率， $\beta$<1更重视正确率

$$F_\beta=\frac{(1+\beta^2)PR}{\beta^2P+R}\ (P\neq 0, R\neq 0)$$

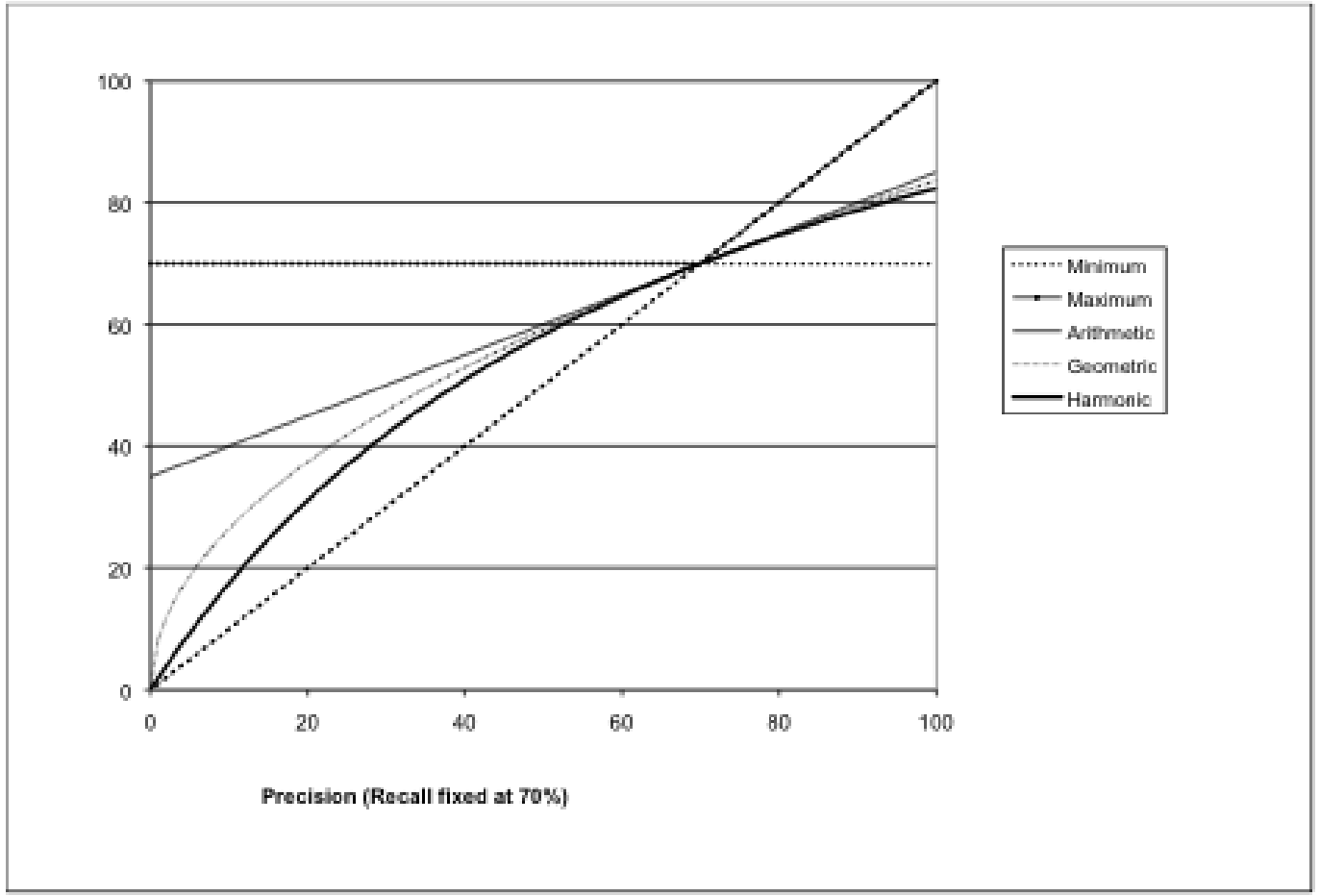
- E(Effectiveness)值：召回率R和正确率P的加权平均值，b>1表示更重视P，E=1-$F_\beta$, ，$b^2=1/\beta^2$

$$E=1-\frac{1+b^2}{\frac{b^2}{P}+\frac{1}{R}}\ (P\neq 0, R\neq 0)$$

# 为什么使用调和平均计算F值

- 如果采用算术平均计算F值，那么一个返回全部文档的搜索引擎的F值就不低于50%，这有些过高。
- 合理做法：不管是P还是R，如果十分低，那么结果应该表现出来，即这样的情形下最终的F值应该有所惩罚
- 采用P和R中的最小值可能达到上述目的
- 但是最小值方法不平滑而且不易加权
- 基于调和平均计算出的F值可以看成是平滑的最小值函数

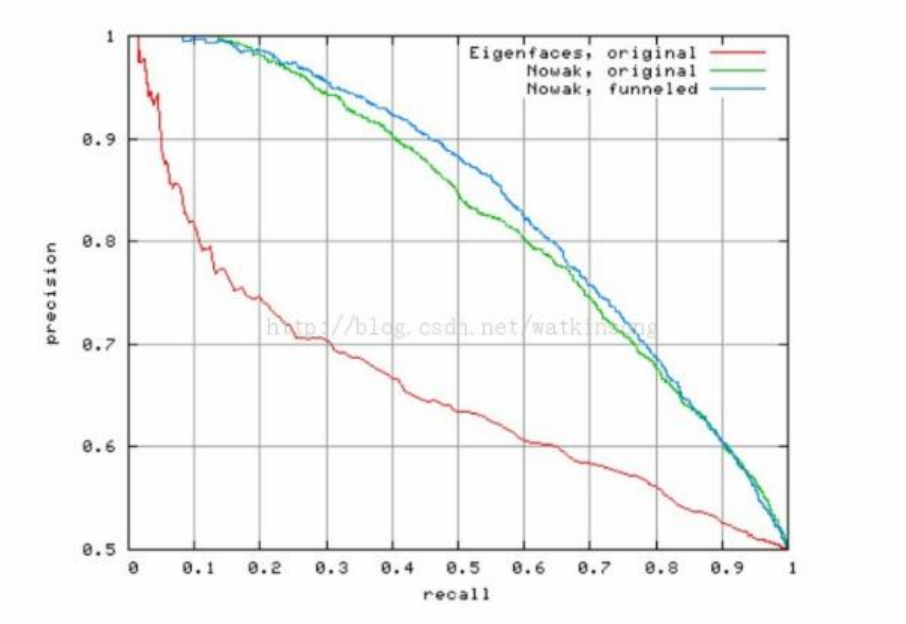
# $F_1$ 及其他平均计算方法



# P-R曲线

- 正确率-召回率曲线(precision vs. recall curve)

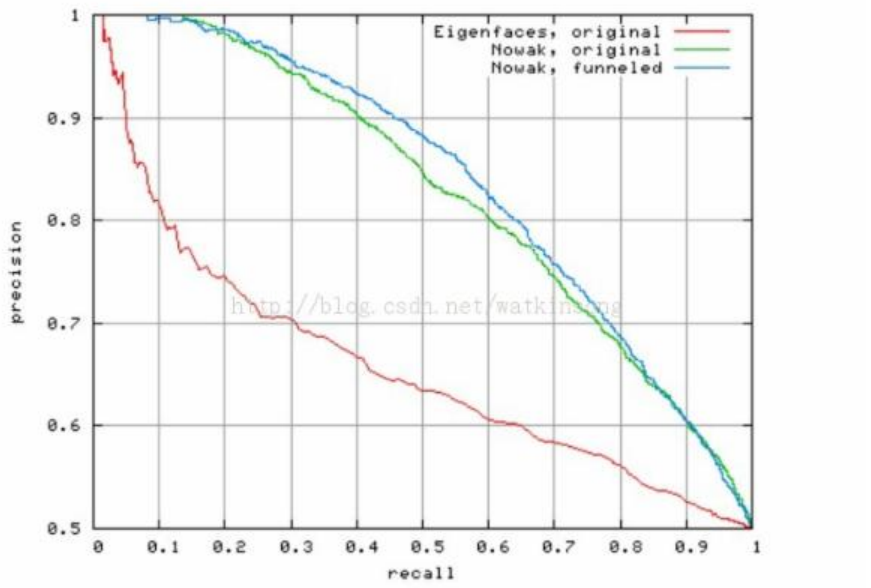
  将返回检索结果看成一个逐步过程，则依序对每一个返回文档子集，都可以得到相应的正确率和召回率，它们的全体形成的一条曲线



# P-R曲线

## P-R曲线特性

- 召回率单调上升，正确率总体下降
- 曲线越向上，系统效果越好

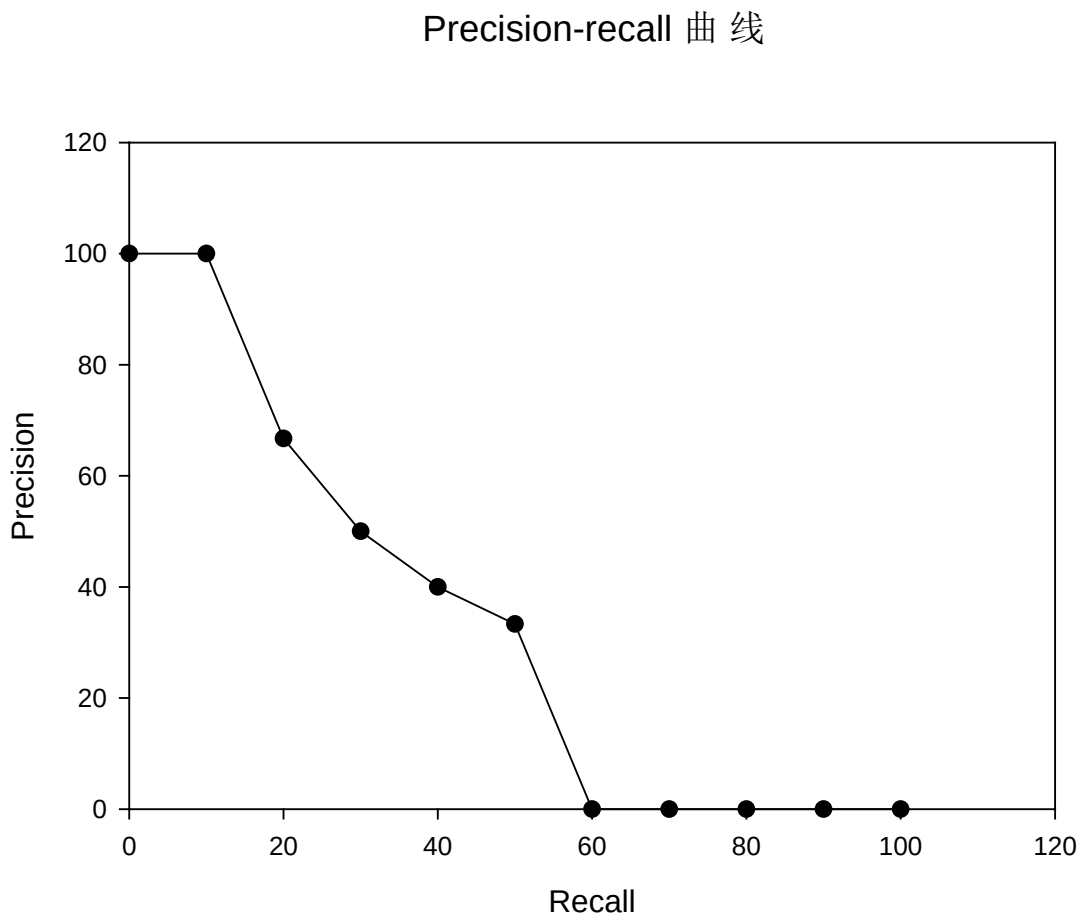


# P-R曲线举例

- 某个查询q的标准答案集合为：
  Rq={d3,d5,d9,d25,d39,d44,d56,d71,d89,d123}
- 某个IR系统对q的检索结果如下：

| | | |
|---|---|---|
| 1. d123 R=0.1,P=1 | 6. d9 R=0.3,P=0.5 | 11. d38 |
| 2. d84 | 7. d511 | 12. d48 |
| 3. d56 R=0.2,P=0.67 | 8. d129 | 13. d250 |
| 4. d6 | 9. d187 | 14. d113 |
| 5. d8 | 10. d25 R=0.4,P=0.4 | 15. d3 R=0.5,P=0.33 |

# P-R曲线举例


Precision-recall 曲 线
120
100
80
60
40
20
0
Precision
0
20
40
60
80
100
120
Recall

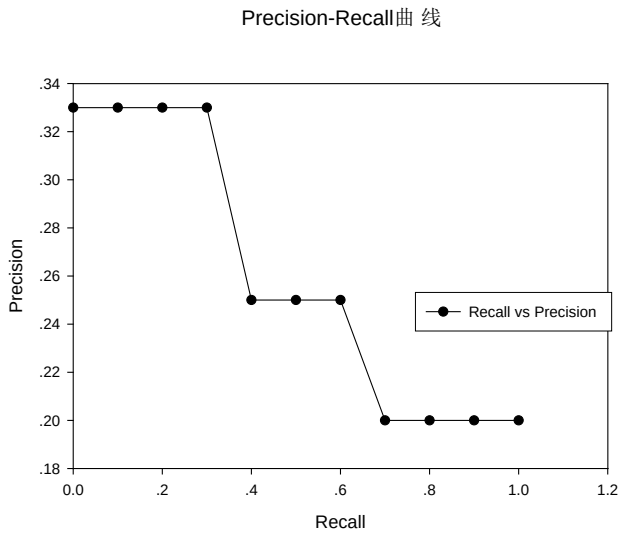
# P-R曲线的插值处理

- **必要性：**召回率不连续，甚至很有限，锯齿状问题
- 对于前面的例子，假设Rq={d3,d56,d129}
  - 3. d56 R=0.33, P=0.33
  - 8. d129 R=0.66, P=0.25
  - 15. d3 R=1, P=0.2
- 不存在10%, 20%,…,90%的召回率点，而只存在 33.3%, 66.7%, 100%三个召回率点
- **处理方法：**对P-R曲线进行插值(interpolate)处理

# 插值生成的P-R曲线图

## □ 插值方法

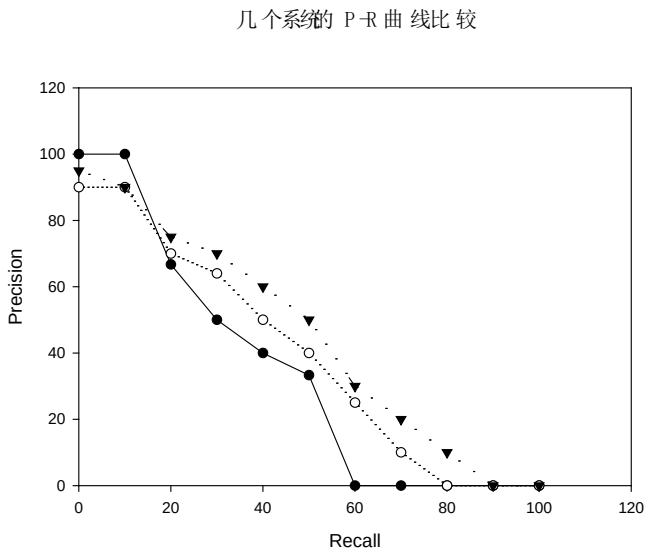
- 对于t%，如果不存在该召回率点，或者该点处具有多个正确率值，则定义t%处的正确率值为其之后的最大的正确率值
- 对于上例，0%,10%,20%,30%上正确率为0.33，40%~60%对应0.25，70%以上对应0.2



# P-R曲线的优缺点

## □ 优点：

- 简单直观
- 考虑了检索结果的覆盖度，也考虑了检索结果的排序

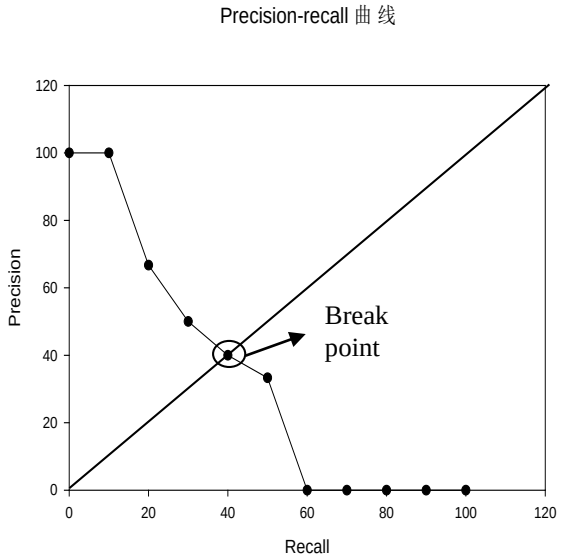



## □ 缺点：

- 单个查询的P-R曲线虽然直观，但是难以明确表示不同检索系统检索结果的优劣
- 处理多个查询的情况更困难

# 基于P-R曲线的单一指标

- 直接进行P-R曲线比较难以量化
- 两种单一指标
  - Break Point：P-R曲线上P=R的点
    - 这样可以直接进行单值比较
  - 平均正确率均值(mean average precision)




对召回率分别为0,0.1,0.2,…,1.0的十一个点上的正确率求平均，再对所有查询求平均值

- 单一指标可以支持基于多个查询的评价

# TREC 概况

- TREC： The Text REtrieval Conference, http://trec.nist.gov
- 由NIST(the National Institute of Standards and Technology)和DARPA(the Defense Advanced Research Projects Agency)联合举办
- 1992年举办第一届会议，每年11月举行，至2006年已有15届，可以看成信息检索的“奥运会”

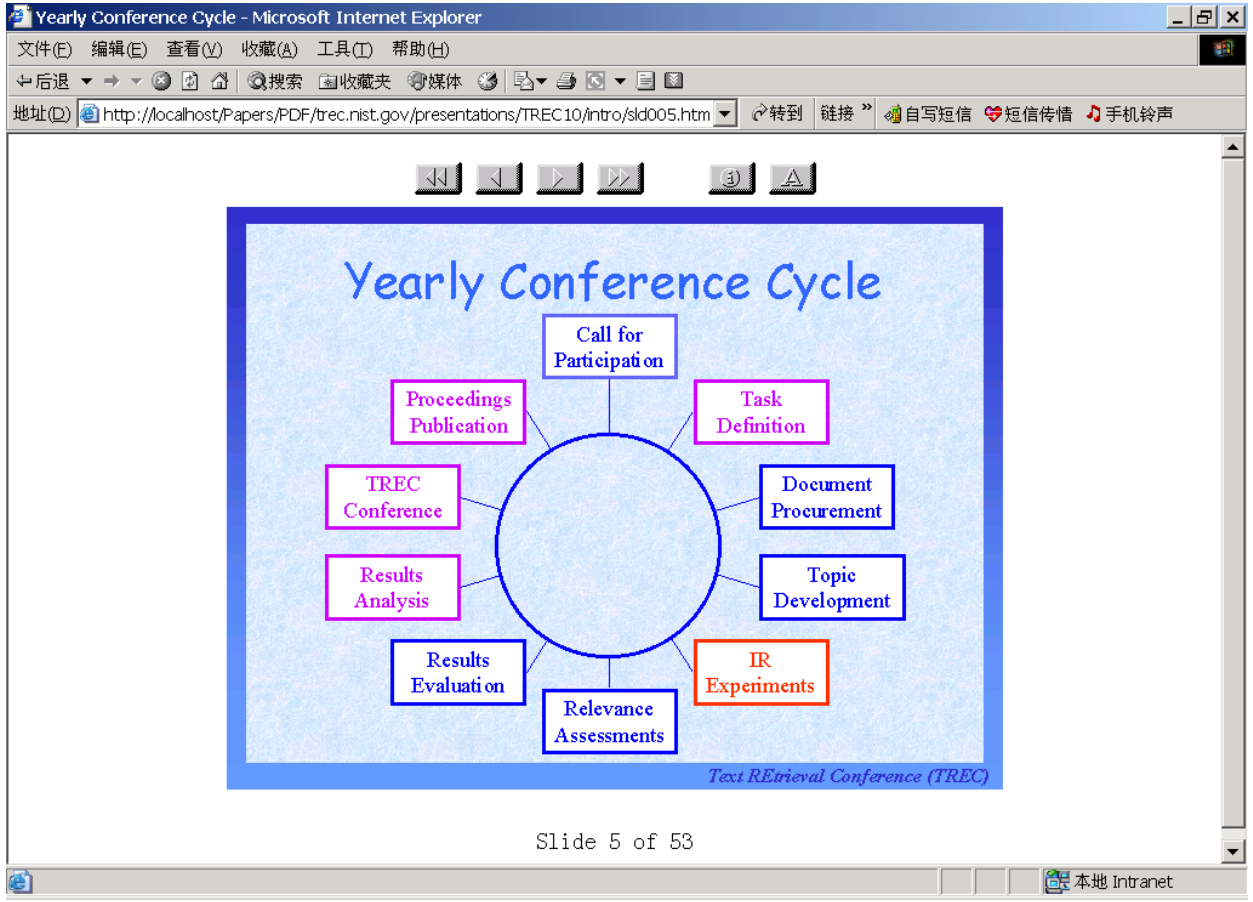
# TREC的目标

1. 提供对大规模文本检索方法和系统进行评估的平台
2. 对评估方法、评估结果以及检索方法进行研讨交流
3. 示范信息检索理论在解决实际问题方面的重大进步，提高信息检索技术从理论走向商业应用的速度
4. 为工业界和学术界提高评估技术的可用性，并开发新的更为适用的评估技术

# TREC的运行方式

- TREC由一个程序委员会管理。这个委员会包括来自政府、工业界和学术界的代表。
- TREC以年度为周期运行。过程为：确定任务→参加者报名→参加者运行任务→返回运行结果→结果评估→大会交流
- 一开始仅仅面向文本，后来逐渐加入语音、图像、视频方面的评测

# TREC的运行方式


Yearly Conference Cycle - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://localhost/Papers/PDF/trec.nist.gov/presentations/TREC10/intro/sld005.htm
转到 链接
自写短信 短信传情 手机铃声
Yearly Conference Cycle
Call for Participation
Task Definition
Document Procurement
Topic Development
IR Experiments
Relevance Assessments
Results Evaluation
Results Analysis
TREC Conference
Proceedings Publication
Text REtrieval Conference (TREC)
Slide 5 of 53
本地 Intranet

# TREC的运行方式

- **确定任务：** NIST提供测试数据和测试问题
- **报名：** 参加者根据自己的兴趣选择任务
- **运行任务：** 参加者用自己的检索系统运行测试问题，给出结果
- **返回结果：** 参加者向NIST返回他们的运行结果
- **结果评估：** NIST使用一套固定的方法和软件对参加者的运行结果给出评测结果
- **大会交流：** 每年的11月召开会议，由当年的参加者们交流彼此的经验

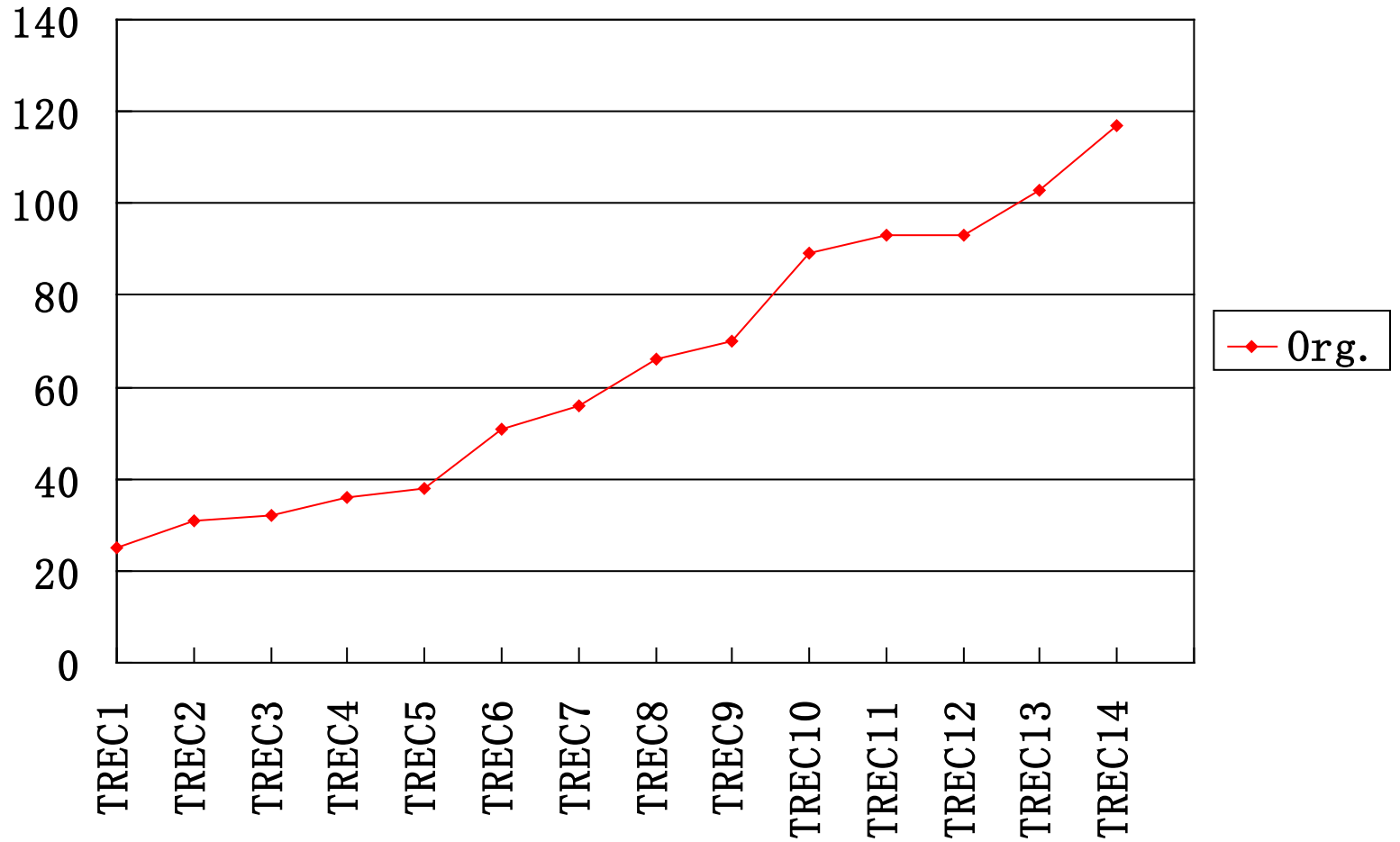
# 测试数据和测试软件

- 由LDC(Linguistic Data Consortium)或者其他单位免费提供，但有些数据需要缴纳费用，一般都必须签订协议
- 每年使用的数据可以是新的，也可以是上一年度已经使用过的
- TREC使用的评估软件是开放的，任何组织和个人都可以用它对自己的系统进行评测

# TREC任务情况

| | | |
|---|---|---|
| TREC1 (92) | 25 | Ad hoc/Routing |
| TREC2 | 31 | Ad hoc/Routing |
| TREC3 | 32 | Ad hoc/Routing |
| TREC4 | 36 | Spanish/Interactive/Database Merging/Confusion/Filtering |
| TREC5 | 38 | Spanish/Interactive/DatabaseMerging/Confusion/Filtering/NLP |
| TREC6 | 51 | Chinese/Interactive/Filtering/NLP/CLIR/Highprecision/SDR/VLC |
| TREC7 | 56 | CLIR/High Precision/Interactive/Query/SDR/VLC |
| TREC8 | 66 | CLIR/Filtering/Interactive/QA/Query/SDR/Web |
| TREC9 | 70 | QA/CLIR(E-C)/Web/Filtering/Interactive/Query/SDR |
| TREC10 | 89 | QA/CLIR/Web/Filtering/Interactive/Video |
| TREC11 (02) | 93 | QA/CLIR/Web/Filtering/Interactive/Video/Novelty/ |
| TREC12 (03) | 93 | QA/Web/Novelty/HARD/Robust/Genomics/ ➔TRECVID单独组织 |
| TREC13 (04) | 103 | QA/Web/Novelty/HARD/Robust/Genomics/Terabyte |
| TREC14 (05) | 117 | QA/HARD/Robust/Enterprise/Genomics/Terabyte/SPAM |
| TREC15 (06) | n/a | QA/Legal/Enterprise/Genomics/Terabyte/SPAM/Blog |
| TREC16 (07) | n/a | QA/Legal/Enterprise/Genomics/Terabyte/SPAM/Blog/Million Query |

# 历届TREC参加单位数示意图



# 参加过TREC的部分单位

| Corp. | University | Asian Organization |
|---|---|---|
| IBM | MIT | Singapore U. (KRDL) |
| AT&T | CMU | KAIST |
| Microsoft | Cambridge U. | Tinghua U. (大陆的清华大学) TREC11 |
| Sun | Cornell U. | Tsinghua U.(中国台湾的清华) TREC7 |
| Apple | Maryland U. | Taiwan U. TREC8&9&10 |
| Fujitsu | Massachusetts U. | Hongkong Chinese U. TREC9 |
| NEC | New Mexico State U. | Microsoft Research China TREC9&10 |
| XEROX | California Berkeley U. | Fudan U. TREC9&10&11(复旦) |
| RICOH | Montreal U. | ICT TREC10&11(中科院计算所) |
| CLRITECH | Johns Hopkins U. | HIT TREC10(哈工大) |
| NTT | Rutgers U. | 北大、软件所、自动化所等 |
| Oracle | Pennsylvania U. | 还有更多的大陆队伍逐渐加入…… |

# 相关性判定及评测

- (Ad hoc任务)Pooling方法：对于每一个topic，NIST从参加者取得的结果中挑选出一部分运行结果，从每个运行结果中取前N个文档，然后用这些文档构成一个文档池，使用人工方式对这些文档进行判断。相关性判断是二值的：相关或不相关。没有进行判断的文档被认为是不相关的。
- NIST使用**trec_eval软件包**对所有参加者的运行结果进行评估，给出大量参数化的评测结果（主要是precision和recall)。根据这些评测数据，参加者可以比较彼此的系统性能。

# 用户判定的有效性

- 只有在用户的评定一致时，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，无法重现实验结果
- 如何度量不同判定人之间的一致性?
- → Kappa 指标

# Kappa 指标

- Kappa是度量判定间一致性的指标

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

- 为类别性判断结果所设计的指标
- 对随机一致性的修正
- $P(A)$ = 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
- k在 [2/3, 1.0]时，判定结果是可以接受的

# Kappa指标算例

| | | Judge 2 Relevance | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Judge 1 Relevance | Yes | 300 | 20 | 320 |
| | No | 10 | 70 | 80 |
| | Total | 310 | 90 | 400 |

Observed proportion of the times the judges agreed

$P(A) = (300 + 70)/400 = 370/400 = 0.925$

$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

$P(E) = P(nonrelevant)^2 + P(\text{relevant})^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic $\kappa = (P(A) - P(E))/(1 - P(E)) =$

$(0.925 - 0.665)/(1 - 0.665) = 0.776$ (still in acceptable range)

# TREC中判定的一致性情况

| 信息需求 | 判断文档数 | 不一致数目 |
|---|---|---|
| 51 | 211 | 6 |
| 62 | 400 | 157 |
| 67 | 400 | 68 |
| 95 | 400 | 110 |
| 127 | 400 | 106 |

# 不一致性带来的影响

- 上述的不一致性很严重。这是否意味着信息检索的实验结果没有意义?
- 不是的
- 不一致性会对指标的绝对数值有很大影响
- 一般不会对系统之间的相对排序产生影响
- 比如，我们想知道A算法是否好于B算法，
  即使判定人员之间的不一致性可能很大，信息检索实验通常也会给出一个比较可靠的答案

# 关于检索评价的研究

- 现有评价体系远没有达到完美程度
  - 对评价的评价研究
  - 指标的相关属性(公正性、敏感性)的研究
  - 新的指标的提出(新特点、新领域)
    - 不考虑召回率的指标 **Precision@N：** 在第N个位置上的正确率，如P@10, P@20
  - 指标的计算(比如Pooling方法中如何降低人工代价？)

# 检索结果片段

- **必要性：** 由于检索正确率难以满足要求，因此用户还需要在返回结果中进一步选择
- **方法：** 提供检索结果的一个短摘要，称结果片段，帮助用户根据该信息来判断结果的相关性
- **摘要类型**
  - 静态摘要：反映文档本质内容，不变化，与查询无关
  - 动态摘要：依赖于查询，它试图解释当前文档返回的原因，个性化生成

# 静态摘要生成

- **简单的启发式方法：**返回文档的前50个左右的单词作为摘要
- **NLP方法：**从文档中返回一些重要句子组成摘要
  - 可以采用简单的NLP启发式方法来对每个句子打分
  - 将得分较高的句子组成摘要
  - 也可以采用机器学习方法，参考第13章
- **高级NLP方法：**通过NLP方法合成或者生成摘要
  - 对大部分IR应用来说，该方法还不够成熟

# 动态摘要生成

- 从查询词项左右两边抽取一些词组成“窗口”
- 出现查询短语的“窗口”优先
- 包含多个查询词项的“窗口”优先
- 最终将所有“窗口”都显示出来作为摘要

# 一个动态摘要的例子

**查询:** “new guinea economic development” Snippets that were extracted from a document: . . . **In recent years, Papua New Guinea has faced severe economic difficulties and** economic growth has slowed, partly as a result of weak governance and civil war, and partly as a result of external factors such as the Bougainville civil war which led to the closure in 1989 of the Panguna mine (at that time the most important foreign exchange earner and contributor to Government finances), the Asian financial crisis, a decline in the prices of gold and copper, and a fall in the production of oil. **PNG’s economic development record over the past few years is evidence that** governance issues underly many of the country’s problems. Good governance, which may be defined as the transparent and accountable management of human, natural, economic and financial resources for the purposes of equitable and sustainable development, flows from proper public sector management, efficient fiscal and accounting mechanisms, and a willingness to make service delivery a priority in practice. . . .

# 参考资料

- 《信息检索导论》第8章
- http://ifnlp.org/ir
  - TREC主页： http://trec.nist.gov
  - *F*-measure: Keith van Rijsbergen
  - 更多有关 A/B 测试的文章
  - Too much A/B testing at Google?
  - Tombros & Sanderson 1998: 动态摘要的最早的几篇文章之一
  - Google VP of Engineering on search quality evaluation at Google

# 课后作业

- 见课程网页：

  **http://10.76.3.31**