现代信息检索

Modern Information Retrieval

第5讲 检索评价

IR Evaluation

*改编自"An introduction to Information retrieval"网上公开的课件，地址 http://nlp.stanford.edu/IR-book/

提纲

1. 有关检索评价
2. 评价指标
3. 相关评测
4. 实验设计

提纲

1. 有关检索评价
2. 评价指标
3. 相关评测
4. 实验设计

关于评价

- 评价无处不在，也很必要
  - 工作、生活、娱乐
- 评价很难，但是似乎又很容易
  - 人的因素、标准、场景
- 评价是检验学术进步的重要标准

从竞技体育谈起

- 世界记录 vs. 世界最好成绩
  - 男子一百米短跑世界纪录：标准赛道标准气候条件的历史最快纪录
  - 男子一百米短跑今年最好成绩：可能比世界纪录更快，但是不符合标准条件
- 评价要公平！
  - 环境要基本一致：天气、风速、跑道等等
  - 比赛过程要一样：竞走中的犯规
  - 指标要一样：速度、耐力

为什么要评价IR？

- 通过评估可以评价不同技术的优劣，不同因素对系统的影响，从而促进本领域研究水平的不断提高
  - 类比：110米栏各项技术---起跑、途中跑、跨栏、步频、冲刺等等
- 信息检索系统的目标是较少消耗情况下尽快、全面返回准确的结果。
- 计算机应用学科偏重于研究“更好的”方法/算法/模型，即所谓Betterness
  - 需要一种公平可靠的方法和指标体系进行评价

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)：本讲主要内容
  - 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
  - 是否排名靠前
- 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，所有文档-查询对完全标注)
  - **SMART System,** Gerald Salton, 1964-1988 (数千篇文档集合，所有文档-查询对完全标注)
  - **TREC(Text REtrieval Conference)**, Donna Harman, 美国标准技术研究所, 1992 - (上百万篇文档，仅标注“缓冲池”)，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	d3	d9	
系统2，查询2	d1	d2	d4	d13	

评价的几个问题

- 评价指标：某个或某几个可衡量、可比较的值
- 评价过程：设计上保证公平、合理
- IR中评价的难点：相关性（Relevance）是一个主观概念，文档相关性依赖于查询
  - 数据标注工作量庞大

提纲

1. 有关检索评价
2. 评价指标
3. 相关评测
4. 实验设计

评价指标分类

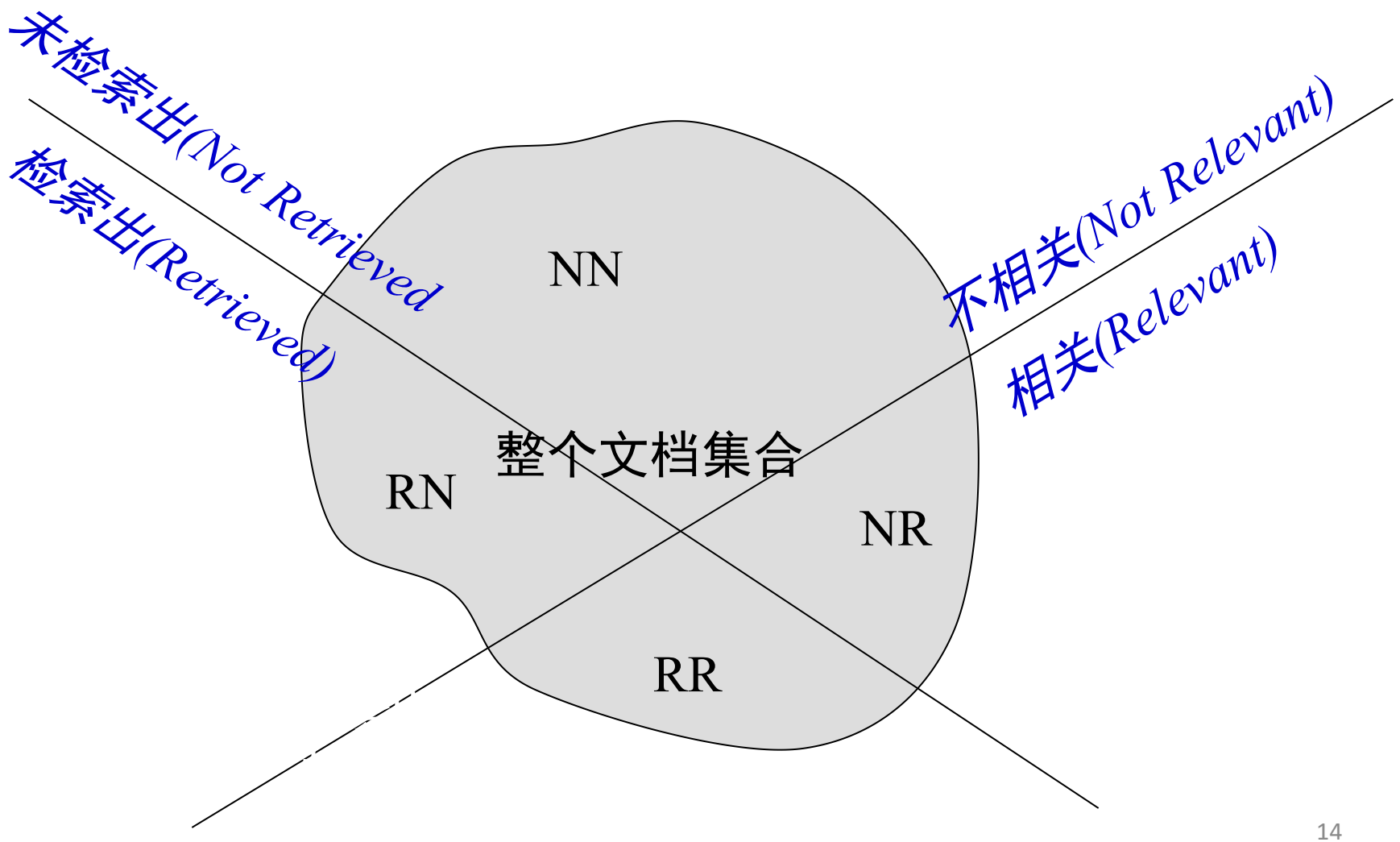
- 对单个查询进行评估的指标
  - 在单个查询上检索系统的得分
- 对多个查询进行评估的指标
  - 在多个查询上检索系统的得分

回到例子

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	d2	d9 √	
系统2，查询2	d1	d2	d4	d13	

查询1的标准答案集合 {d3,d4,d6,d9}

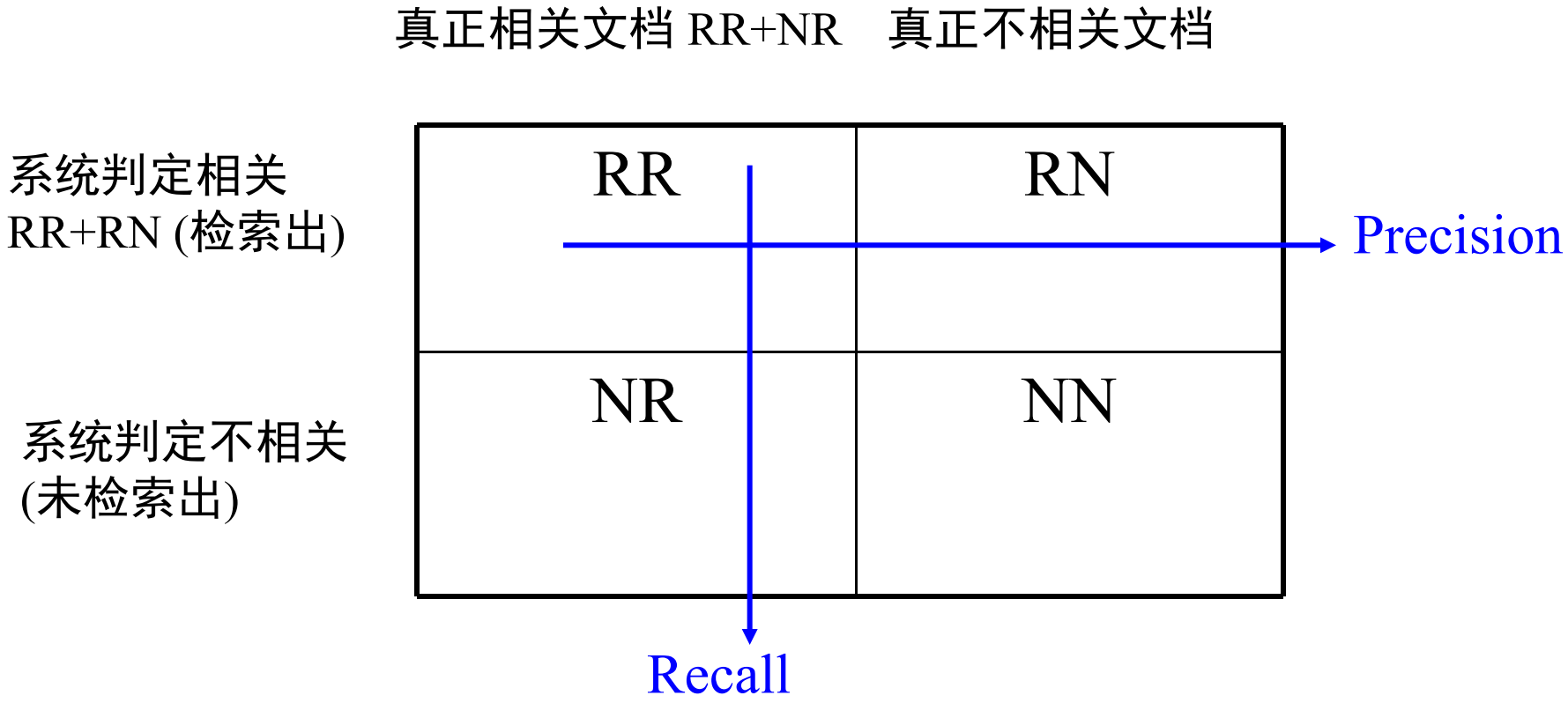
整个文档集合的划分


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

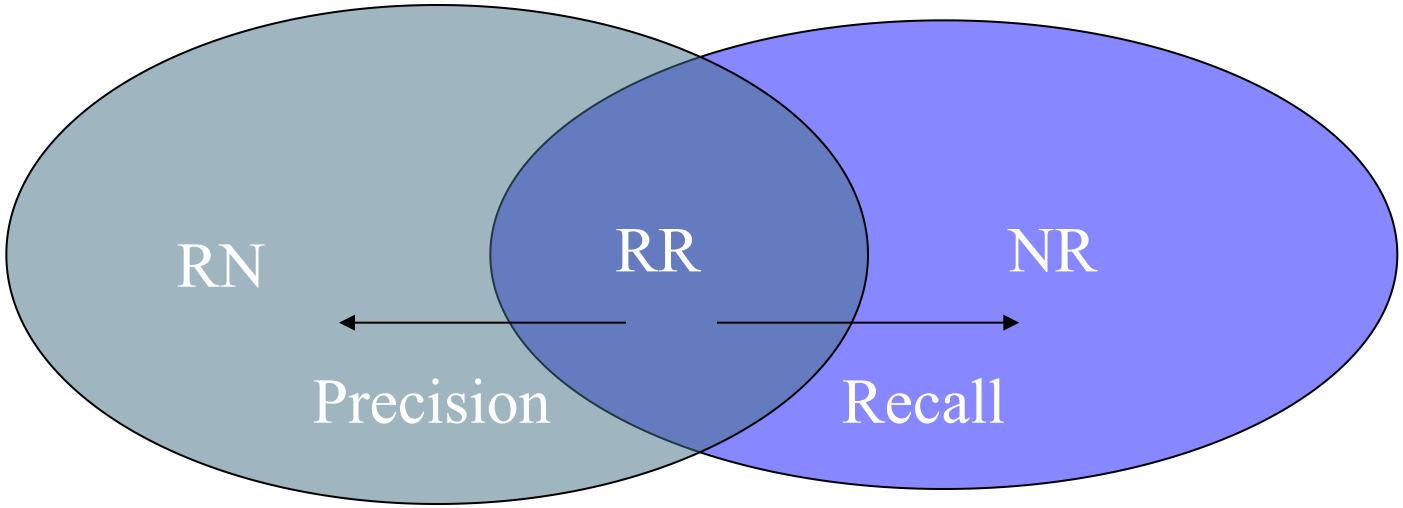
评价指标

- **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]
- **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]
- 两个指标分别度量检索效果的某个方面，忽略任何一个方面都有失偏颇。两个极端情况：返回有把握的1篇，P=100%，但R极低；全部文档都返回，R＝1，但P极低

四种关系的矩阵表示


真正相关文档 RR+NR
真正不相关文档
系统判定相关
RR+RN (检索出)
系统判定不相关
(未检索出)
RR
RN
NR
NN
Precision
Recall

基于集合的维恩图(venn diargram)



回到例子

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	d2	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

课堂提问：另一个计算例子

- 一个例子：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

正确率和召回率的应用领域

- 拼写校对
- 中文分词
- 文本分类
- 人脸识别
- ……

分类任务的常用指标

关于正确率和召回率的讨论

- 虽然Precision和Recall都很重要，但是不同的应用、不用的用户可能会对两者的要求不一样。因此，实际应用中应该考虑这点。
  - 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。
  - 有些用户希望返回的结果全一点，他有时间挑选；有些用户希望返回结果准一点，他不需要结果很全就能完成任务。

课堂提问：

- 正确率和召回率的定义或者计算有什么问题或不足？

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	d2	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

正确率和召回率的问题

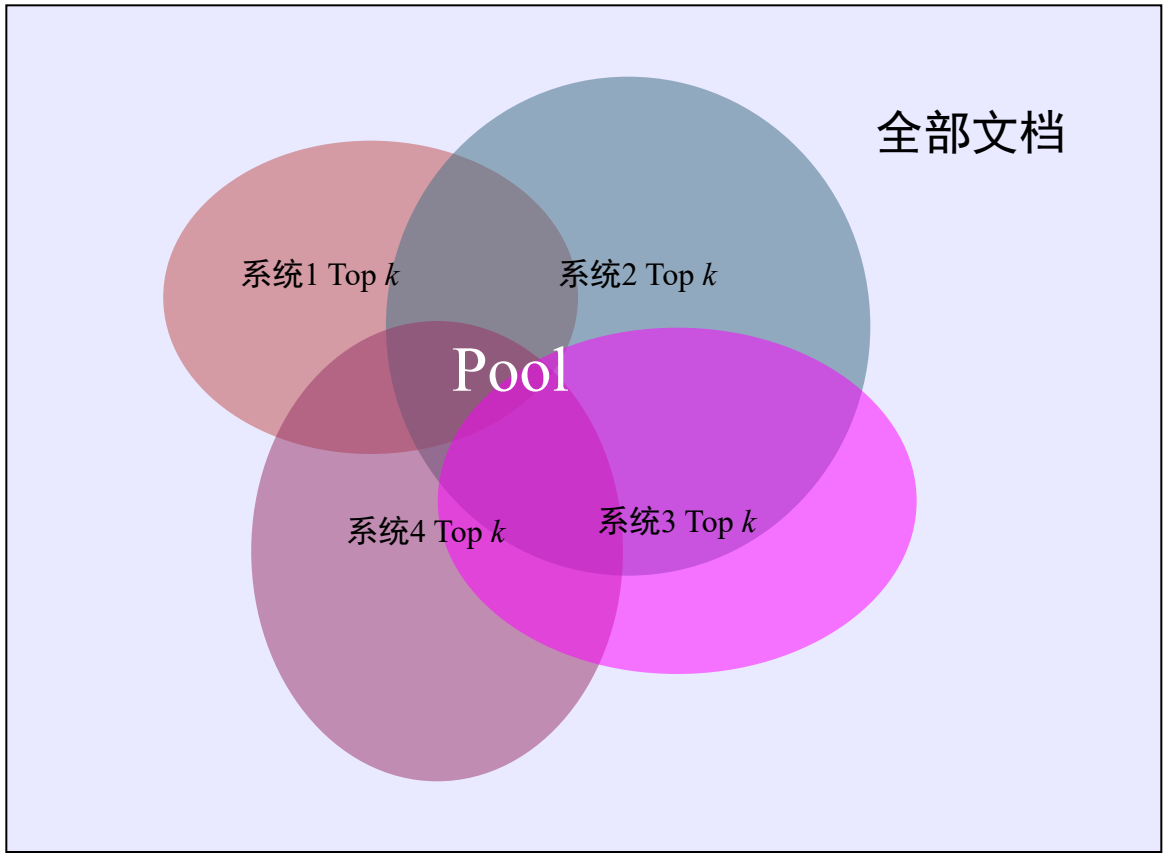
- 召回率难以计算
  - 解决方法：Pooling（缓冲池）方法，或者不考虑召回率
- 两个指标分别衡量了系统的某个方面，但是也为比较带来了难度，究竟哪个系统好？大学最终排名也只有一个指标。
  - 解决方法：单一指标，将两个指标融成一个指标
- 两个指标都是基于(无序)集合进行计算，并没有考虑（排）序的作用
  - 举例：两个系统，对某个查询，返回的相关文档数目一样都是10，但是第一个系统是前10条结果，后一个系统是最后10条结果。显然，第一个系统优。但是根据上面基于集合的计算，显然两者指标一样。
  - 解决方法：引入序的作用

问题一：召回率难以计算

- 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，因此，这种情况几乎不可能准确地计算召回率

- 缓冲池(Pooling)方法：对多个检索系统的Top k 个结果组成的集合(并集)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



课堂提问

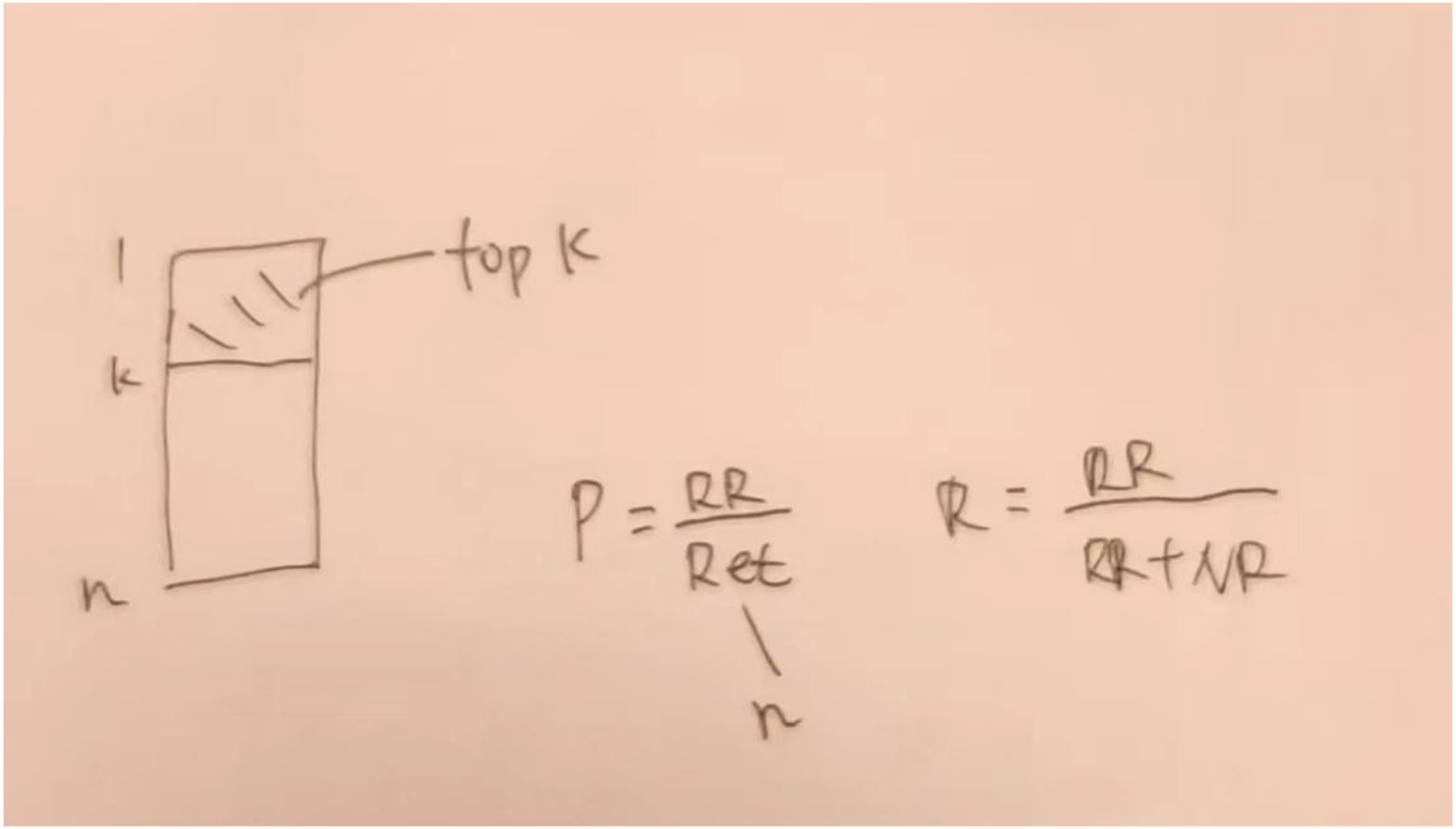
- (某个系统的某个查询)通过Pooling计算出的召回率、正确率和真正的召回率、正确率的大小之间有什么关系？
- 情况1(常见情况)：如果只有部分结果进行了Pooling操作，那么显然在计算正确率时有 RR_P<=RR，分母都是Ret不变，此时计算出来的正确率会小于真实的正确率。而对于召回率，计算中的分子分母都变小，所以结果不一定。
- 情况2：如果所有的返回文档（即Ret）都进行了Pooling，那么正确率计算时的分子分母都不变，此时计算出的正确率等于真实的正确率。此时，由于分子不变，而分母显然小于真实的相关文档总数，所以计算出来的召回率大于真实的召回率。

RR_P: pooling计算得到的返回相关文档数量

RR: 真实返回相关文档数量

Ret：返回文档数量

课堂提问



问题二：P和R需要融合

- 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：

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

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

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

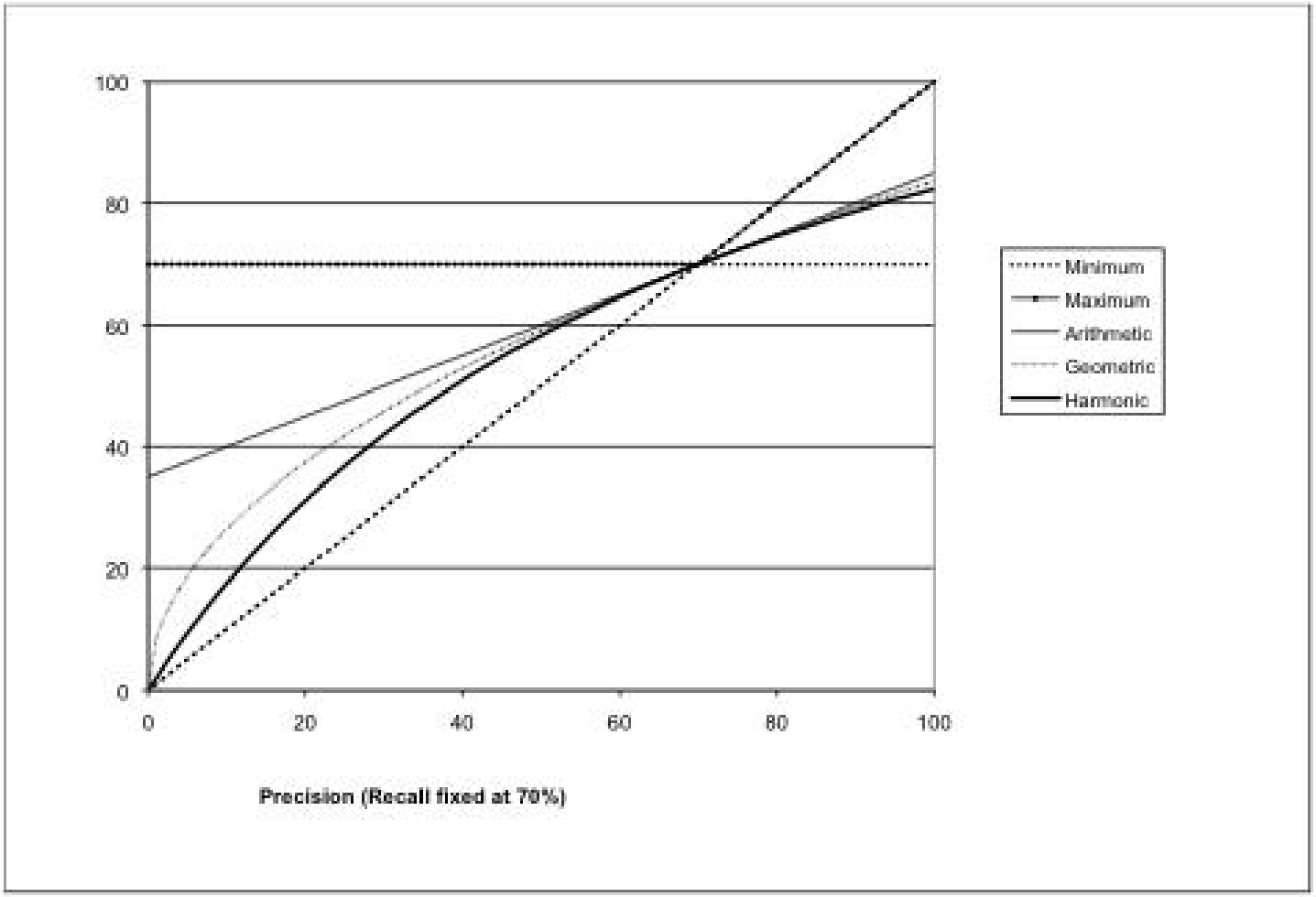
- E(Effectiveness)值：召回率R和正确率P的加权平均值，b>1表示更重视P，E=1- $F_{\beta,}$, $b^2=1/\beta^2$

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

为什么使用调和平均计算F值

- 为什么不使用其他平均来计算F，比如算术平均
  - 调和平均、几何平均、算术平均
- 调和平均能**平衡精确率和召回率**的影响，特别是当两者相差较大时
  - 调和平均倾向于接近较小的数值，因此只有当精确率和召回率都较高时，F值才会高，确保模型在两个指标上都表现良好，防止一个指标表现很好时掩盖另一个指标表现较差的情况
- 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。
  - 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

F_1 及其他平均计算方法



图例自上到下：最小值，最大值，算术平均，几何平均，调和平均

精确率(Accuracy)

- 精确率是所有判定中正确的比率
  - accuracy = (RR+NN)/(RN + RR + NR + NN)
- 为什么通常使用P、R、F而不使用精确率？
- Web信息检索当中精确率为什么不可用？

课堂练习

- 计算P、R、F1

	相关	不相关
返回	18	2
未返回	82	1,000,000,000

- 下面的一个搜索引擎无论对于什么查询都返回0结果，为什么该引擎例子表明使用精确率是不合适的？

精确率不适合IR的原因

- 由于和查询相关毕竟占文档集的极少数，所以即使什么都不返回也会得到很高的精确率
- 什么都不返回可能对大部分查询来说可以得到 99.99%以上的精确率(类别不平衡问题！)

- 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返回任何文档强

- 因此，实际中常常使用P、R和F1，而不使用精确率

问题三：P、R没有考虑结果的序

- 一种引入序的方法：R-Precision
  - 检索结果中，在**所有相关文档总数**位置上的准确率，如某个查询的相关文档总数为80，则计算检索结果中在前80篇文档的正确率。

系统1，查询1　d3√　d6 √　d8　d10　d11
系统2，查询1　d6 √　d7　d2　d9 √

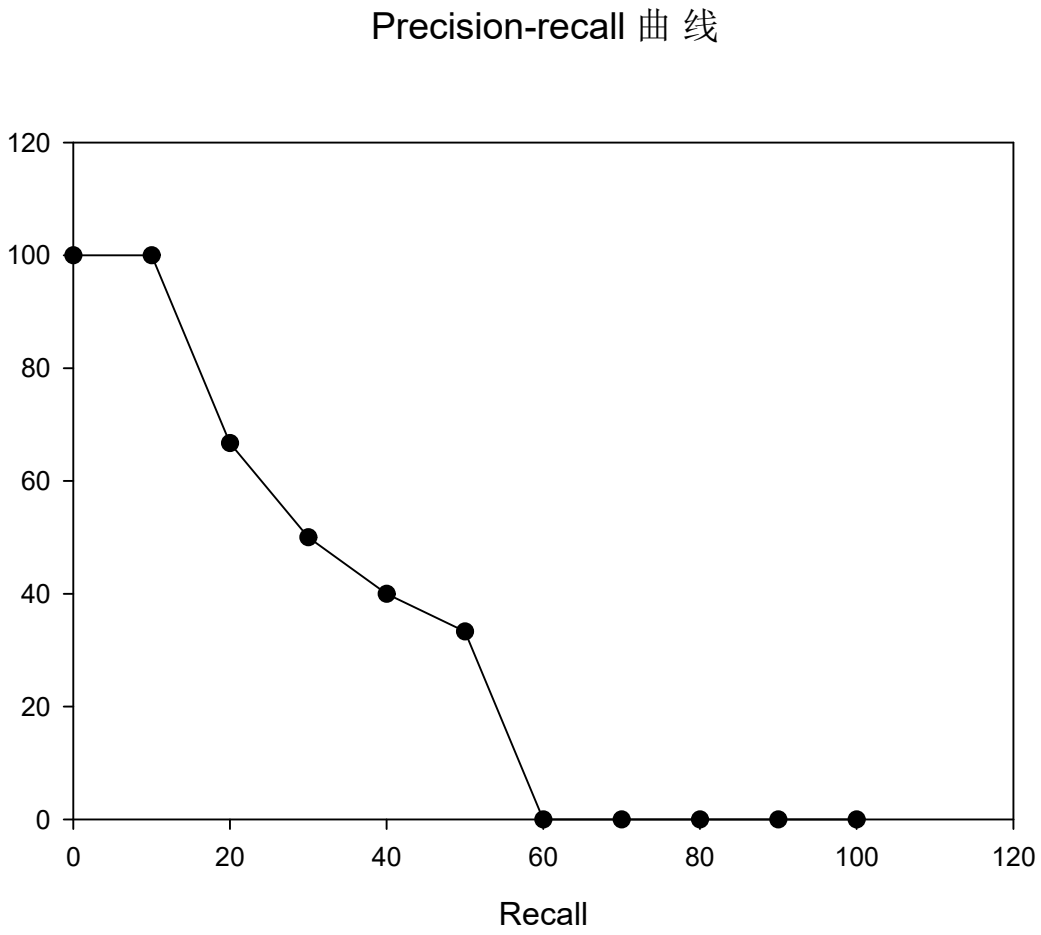
对于查询1的标准答案集合 {d3,d4,d6,d9}

系统1 R-Precision=2/4　系统2 R-Precision=2/4

正确率-召回率 (P-R) 曲线

- 正确率-召回率 曲线(precision versus recall curve)
  - 检索结果以排序方式排列，用户不可能马上看到全部文档，因此，在用户观察的过程中，正确率和召回率在不断变化(vary)。
  - 可以求出在召回率分别为0%,10%,20%,30%,…,90%,100%上对应的正确率，然后描出图像
  - 在上面的曲线对应的系统结果更好
    - 即线下面积（Area Under the Curve, 缩写AUC）

P-R曲线


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

P-R曲线的例子

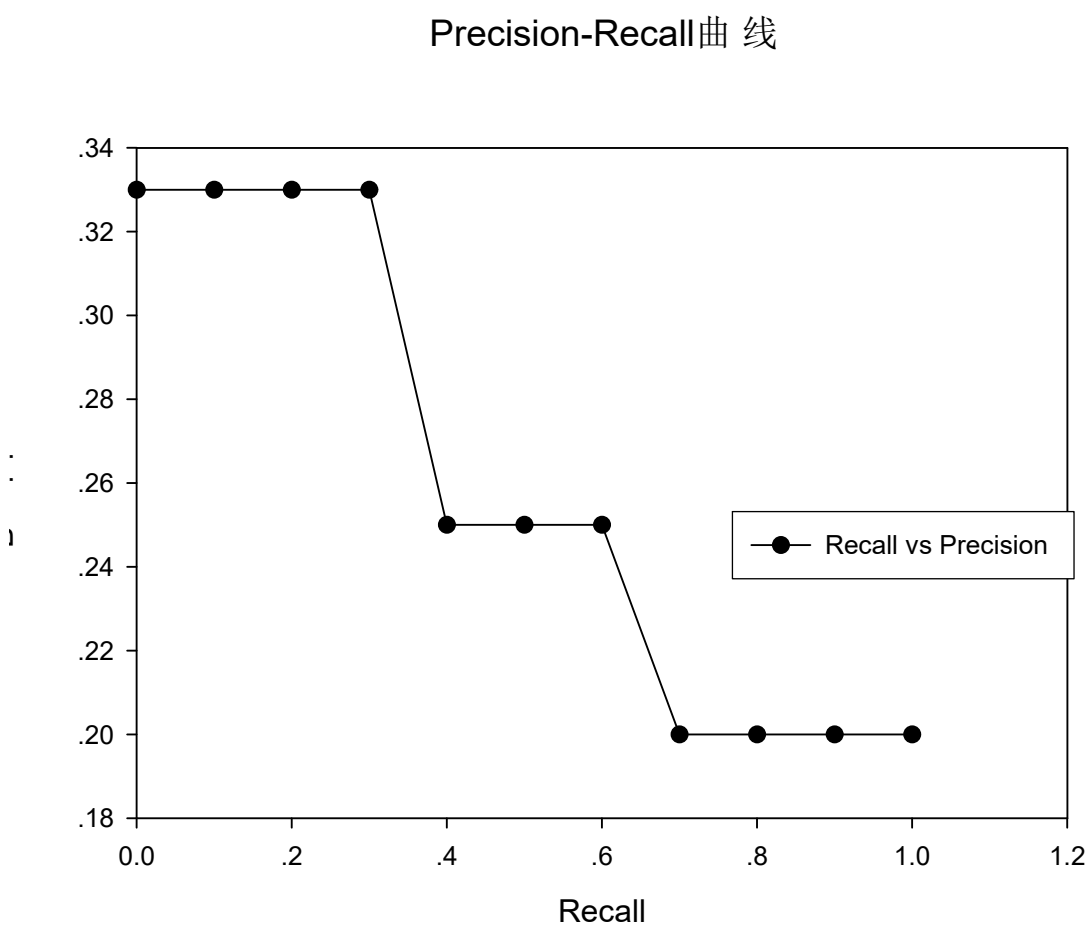
- 某个查询q的标准答案(即相关文档)集合为：R_q={d3,d5,d9,d25,d39,d44,d56,d71,d89,d123}
- 某个IR系统对q的检索结果如下：

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 曲线的插值问题

- 对于前面的例子，若假设R_q={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
- 在这种情况下，需要利用存在的召回率点对不存在的召回率点进行插值(interpolate)
- 对于t%，如果不存在该召回率点，则定义t%为从t%到(t+10)%中最大的正确率值。
- 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曲线图


Precision-Recall曲 线
.34
.32
.30
.28
.26
.24
.22
.20
.18
0.0
.2
.4
.6
.8
1.0
1.2
Recall
Recall vs Precision

P-R曲线的优缺点

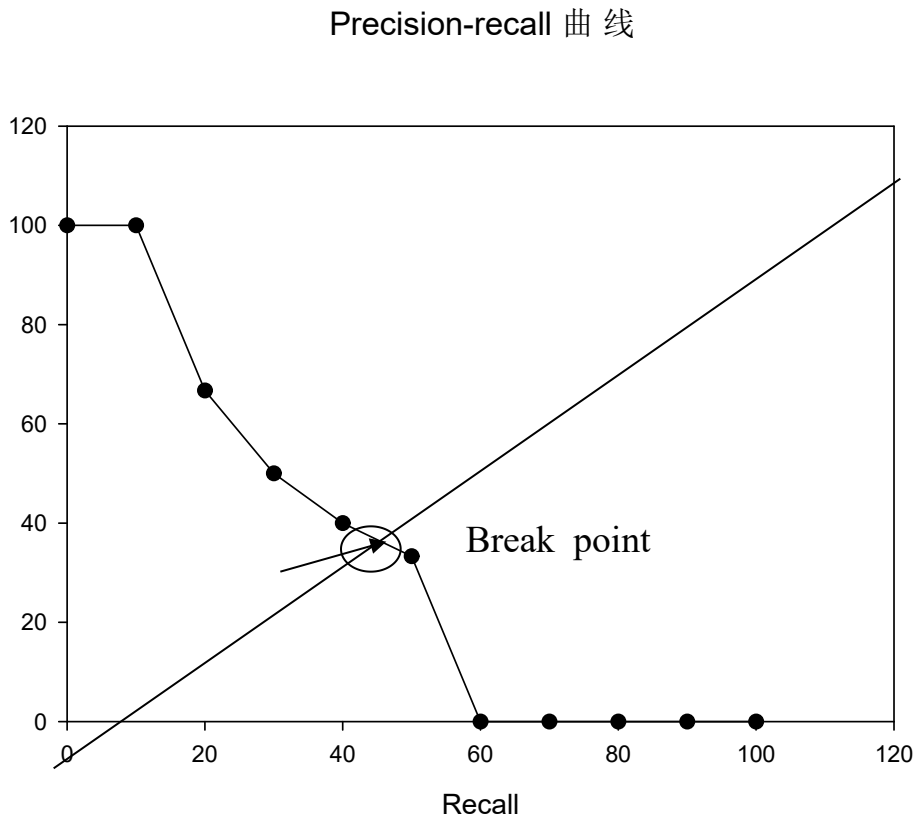
- 优点：
  - 简单直观
  - 既考虑了检索结果的覆盖度，又考虑了检索结果的排序情况
- 缺点：
  - 单个查询的P-R曲线虽然直观，但是难以明确表示两个查询的检索结果的优劣

基于P-R曲线的单一指标

- Break Point：P-R曲线上 P=R的那个点
  - 这样可以直接进行单值比较

- 11点平均正确率(11 point average precision)：在召回率分别为0,0.1,0.2,…,1.0的十一个点上的正确率求平均，也叫插值AP（interpolated average precision）

P-R曲线中的break point



Break point:数值越大，系统效果越好。理想状态下，P/R曲线相交于(100%, 100%)，线下面积为1，系统效果最优。

未插值的平均正确率
Uninterpolated Average Precision

信息检索领域的最常见评价指标之一

- 有别于前述（11点）插值平均准确率
- 计算每个已知相关文档所在排名的正确率：
  - 对每个**返回相关**文档计算该返回**位置**的正确率
  - 对于**未返回相关文档**，正确率记为 **0**
  - 对所有已知相关文档对应的正确率求**算术平均**
- 特点：
  - 计算更简便，不进行插值平滑
  - 综合考虑正确率和召回率
  - 强调排名靠前文档正确率

引入序的作用

- 平均正确率(Average Precision, AP)：对不同召回率点上的正确率进行平均
  - **未插值的AP**(常用): 某个查询Q共有6个相关结果，某系统排序返回了5篇相关文档，其位置分别是第1，第2，第5，第10，第20位，则
    AP=(1/1+2/2+3/5+4/10+5/20+0)/6
  - **插值的AP**:在召回率分别为0, 0.1, 0.2, … , 1.0的十一个点上的正确率求平均，等价于11点平均

不考虑召回率的指标

- AP通过已知相关文档的总数考虑召回率。也有一些完全不考虑召回率的指标
- Precision@N：在第N个位置上的正确率。大量统计数据表明，大部分搜索引擎用户只关注前一、两页的结果，因此，P@10, P@20对大规模搜索引擎来说是很好的评价指标
- bpref、NDCG：后面详细介绍。

评价指标分类

- 对单个查询进行评估的指标
  - 对单个查询得到一个结果

- 对多个查询进行评估的指标←
  - 在多个查询上检索系统的得分求平均

评价指标

- 平均的求法：
  - 宏平均(Macro Average)：对每个查询求出某个指标，然后对这些指标进行算术平均
  - 微平均(Micro Average)：将所有查询视为一个查询，将各种情况的文档总数求和，然后进行指标的计算
    - 如：Micro Precision=(对所有查询检出的相关文档总数)/(对所有查询检出的文档总数)
  - 宏平均对所有查询一视同仁，微平均受返回相关文档数目比较大的查询影响
- 例如 MAP(Mean AP)：对所有查询的AP求宏平均
  - 一个常见指标

回到例子

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	d2	d9 √	
系统2，查询2	d1 √	d2 √	d4	d13 √	d14

假设查询1已知有4个相关文档，查询2已知有3个相关文档，

系统1查询1: P=2/5, R=2/4, F=4/9,AP=1/2;系统1查询2: P=2/5, R=2/3, F=1/2,AP=7/15;

系统2查询1: P=2/4, R=2/4, F=1/2,AP=3/8;系统2查询2: P=3/5, R=3/3. F=3/4,AP=11/12;

系统1的MacroP=2/5, MacroR=7/12, MacroF=17/36, MAP=29/60, MicroP=4/10, MicroR=4/7, MicroF=8/17

系统2的MacroP=11/20, MacroR=3/4, MacroF=5/8, MAP=31/48, MicroP=5/9, MicroR=5/7, MicroF=5/8

课堂提问

- 两个查询q1、q2的标准答案数目分别为100个和50个，某系统对q1检索出80个结果，其中正确数目为40，系统对q2检索出30个结果，其中正确数目为24，求MacroP/MacroR/MicroP/MicroR：

P1=40/80=0.5, R1=40/100=0.4
P2=24/30=0.8, R2=24/50=0.48
MacroP=(P1+P2)/2=0.65,
MacroR=(R1+R2)/2=0.44
MicroP=(40+24)/(80+30)=0.58
MicroR=(40+24)/(100+50)=0.43

关于相关性（Relevance）

- 相关性试图度量用户的满意程度，但是用户是否满意取决于很多因素。
  - 用户工作量：用户是否容易构建查询、进行搜索以及浏览返回结果
  - 响应时间：输入到输入之间的等待时间
  - 结果呈现方式：用户方便浏览获得答案
  - 文档集覆盖度：文档集对相关文档的覆盖程度

- 搜索引擎往往还会考虑多样性(diversity)：结果的多样性，比如输入“苹果”，可以是公司、产品、操作系统、水果等等。

其他评价指标

- 不同的信息检索应用或者任务还会采用不同的评价指标
- MRR(Mean Reciprocal Rank)：对于某些IR系统(如问答系统或主页发现系统)，只关心第一个标准答案返回的位置(Rank)，越靠前越好，这个位置的倒数称为RR，对问题集合求平均，则得到MRR
  - 例子：两个问题，系统对第一个问题返回的标准答案的Rank是2，对第二个问题返回的标准答案的Rank是4，则系统的MRR为 (1/2+1/4)/2=3/8
  - 常用于问答等相关内容单一的任务

一些较新的面向排序任务的评价指标

- Bpref
- GMAP
- NDCG

Bpref

- Bpref：Binary preference，2005年首次引入到TREC的Terabyte任务中
- 在大语料上，人工标注非常稀疏，代价也很大
  - 很多任务中缓冲池深度=10
- 基本的思想：在相关性判断(Relevance Judgment) 不完全的情况下，仅考虑已经过人工判断的文档，即**标注相关文档和标注不相关文档的相对排序**
  - **即所谓二元优先级（binary preference, 简称bpref）**
- 实验结果表明，相关性判断完全的情况下，利用Bpref和MAP进行评价的结果很一致，但是相关性判断不完全的情况下，Bpref更鲁棒。

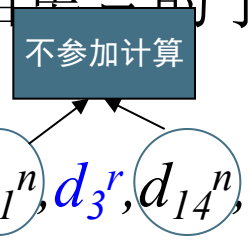
*Buckley, C. & Voorhees, E.M. Retrieval Evaluation with Incomplete Information, Proceedings of SIGIR 2004

原始定义

- 对n（标注不相关文档）排在r（标注相关文档）前面的情况进行惩罚
- 对每个Topic，已判定结果中有R个相关结果

$$bpref = \frac{1}{R}\sum_r \left(1 - \frac{\min(|\, n\text{排在}r\text{前面}\,|, R)}{\mathrm{R}}\right)$$

  - r是相关文档，n是Top R篇不相关文档集合的子集
  - 例子：R=4



$d_{15}^{r}, d_{13}^{n}, d_{10}^{u}, d_{12}^{n}, d_{9}^{r}, d_{7}^{u}, d_{4}^{n}, d_{6}^{n}, d_{5}^{u}, d_{2}^{r}, d_{1}^{n}, d_{3}^{r}, d_{14}^{n}, \ldots$

$bpref$=1/4*(1-0+1-2/4+1-4/4+1-4/4)=3/8

特定情况

- 当R很小(1 or 2)时，原公式不合适

$$bpref\,10 = \frac{1}{R}\sum_{r}\left(1 - \frac{|\,n\text{排在}r\text{前面}\,|}{10+\mathrm{R}}\right)$$

  - r是相关文档，n是Top 10+R篇不相关文档集合的子集

更新的定义

- 对每个Topic，已判定结果集合中有R个相关文档，N个不相关文档，则

$$bpref = \frac{1}{R}\sum_{r}\left(1 - \frac{\min(|\, n\text{排在}r\text{前面}\,|, R)}{\min(\mathrm{R},\mathrm{N})}\right)$$

Bpref can be thought of as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones. Bpref and mean average precision are very highly correlated when used with complete judgments. But when judgments are incomplete, rankings of systems by bpref still correlate highly to the original ranking, whereas rankings of systems by MAP do not.

*参看trec_eval工具8.0修正说明(bpref_bug文件)

思考：怎样对评价指标进行评价？

- bPref比MAP“好”：如何判断？
- [Buckley & Voorhees, SIGIR 2004] 的做法
  - 假定缓冲池的相关性数据是“完整”的
  - 假定使用完整缓冲池得到的MAP是可靠的
  - 模拟产生不同完整程度的缓冲池：从完整缓冲池中分别随机选取1%, 2%, 3%, 5%, 10%, 20%, 30%, … , 90% 的不相关和相关文档
  - 使用不同完整程度的缓冲池对参赛系统重新排序，计算新排序和原始排序（即100%完整度缓冲池产生的排序）的一致性

具体步骤

- 评测：对若干系统(S_1--S_n)按照评价指标排序
- 创建“完整”缓冲池：TREC pooling
  - 将系统S_1--S_n的排名前k文档合并，人工标注
  - k：缓冲池深度
- 基于“完整”缓冲池计算MAP，得到系统排序
  - $R_{MAP,100\%}$=S_1, S_2, S_3, …, S_n
  - 认为该排序是可靠的
- 创建不完整缓冲池
  - 从“完整”缓冲池随机抽取1%, 2%, 3%, 5%, 10%, 20%, 30%, … , 90% 的不相关和相关文档

具体步骤

- 基于不同完整度的缓冲池计算评价指标
- 得到不同的系统排序，例如：
  - $R_{Bpref,30\%}$:基于30%完整度的缓冲池计算得到的BPref所获得的系统排序
- 计算不同排序列表之间的排名相关性（Spearman's Rank Correlation）
- 主要结论：
  - 当缓冲池的完整度较高的情况下，$R_{MAP,x\%}$和$R_{MAP,100\%}$(即上一页的可靠排序)高度相关
  - 当缓冲池的完整度低的情况下，$R_{MAP,x\%}$和$R_{MAP,100\%}$(即上一页的可靠排序)相关性弱
    - MAP不适用
    - 而$R_{Bpref,x\%}$和$R_{MAP,100\%}$以及$R_{Bpref,100\%}$依然高度相关
  - 因此，BPref比MAP更适合于相关性标注稀疏数据集上的系统评测

上述实验方法是否有问题？

- 随机抽取的方式不能反映缓冲池的创建方式
- 缓冲池：不同系统的top-k返回文档的集合
  - k：缓冲池“深度”
- 现在常用做法是通过改变缓冲池深度来模拟不同的相关性判断完整度

GMAP

- GMAP(Geometric MAP): TREC2004 Robust 任务引进
- 先看一个例子

系统	Topic	AP	Increase	MAP
系统A	Topic 1	0.02	-	0.113
	Topic 2	0.03	-	
	Topic 3	0.29	-	
系统B	Topic 1	0.08	+300%	0.107
	Topic 2	0.04	+33.3%	
	Topic 3	0.20	-31%	

- 从MAP来看，系统A好于系统B，但是从每个查询来看，3个查询中有2个Topic B比A有提高，其中一个提高的幅度达到300%

GMAP

- 几何平均值：考虑系统在多个不同查询上性能的稳定性

$$GMAP = \sqrt[n]{\prod_{i=1}^{n} AP_i} = \exp(\frac{1}{n}\sum_{i=1}^{n} \ln AP_i)$$

- 上面那个例子 $GMAP_a$=0.056, $GMAP_b$=0.086
- $GMAP_a$<$GMAP_b$
- GMAP和MAP各有利弊，可以配合使用，如果存在难Topic时，GMAP更能体现细微差别

NDCG

- 每个文档不仅仅只有相关和不相关两种情况，而是有相关度级别，比如0,1,2,3。我们可以假设，对于返回结果：
  - 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

*Jarvelin, K. & Kekalainen, J. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems,* **2002**, *20*, 422-446

（线性）NDCG

- Direct Gain

$$G' = \langle 3, 2, 3, 0, 0, 1, 2, 2, 3, 0, \ldots \rangle.$$

- Cumulated Gain(CG) vector

$$\mathrm{CG}[i] = \begin{cases} \mathrm{G}[1], & \text{if } i = 1 \\ \mathrm{CG}[i-1] + \mathrm{G}[i], & \text{otherwise.} \end{cases} \quad (1)$$

$$\mathrm{CG}' = \langle 3, 5, 8, 8, 8, 9, 11, 13, 16, 16, \ldots \rangle.$$

- Discounted CG vector(${}^b\log i$表示以b为底对i取对数)

$$\mathrm{DCG}[i] = \begin{cases} \mathrm{CG}[i], & \text{if } i < b \\ \mathrm{DCG}[i-1] + \mathrm{G}[i]/{}^b\log i, & \text{if } i \geq b. \end{cases} \quad (2)$$

$$\mathrm{DCG}' = \langle 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61, \ldots \rangle.$$

NDCG

- BV(Best Vector)：假定m个3，l个2，k个1，其他都是0（以上假设4级相关性）

$$\mathrm{BV}[i] = \begin{cases} 3, & \text{if } i \le m, \\ 2, & \text{if } m < i \le m+l, \\ 1, & \text{if } m+l < i \le m+l+k, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\mathrm{I}' = \langle 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, \ldots \rangle.$$

$$\mathrm{CG}'_{\mathrm{I}} = \langle 3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, \ldots \rangle$$
$$\mathrm{DCG}'_{\mathrm{I}} = \langle 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 11.21, 11.53, 11.83, 11.83, 11.83, \ldots \rangle.$$

NDCG

- Normalized (D)CG

$$norm\text{-}vect(\mathrm{V},\mathrm{I}) = \langle \mathrm{v}_1/\mathrm{i}_1, \mathrm{v}_2/\mathrm{i}_2, \ldots, \mathrm{v}_k/\mathrm{i}_k \rangle. \quad (5)$$

$$\begin{aligned} \mathrm{nCG}' &= norm\text{-}vect(\mathrm{CG}', \mathrm{CG}'_\mathrm{I}) \\ &= \langle 1, 0.83, 0.89, 0.73, 0.62, 0.6, 0.69, 0.76, 0.89, 0.84, \ldots \rangle. \end{aligned}$$

- nDCG[i]=DCG'[i]/DCG'$_I$[i]
- nDCG={1, 0.83, 0.87, 0.78, 0.71, 0.69...}
- N (D)CG@k：表示第k个位置上的N(D)CG值

NDCG

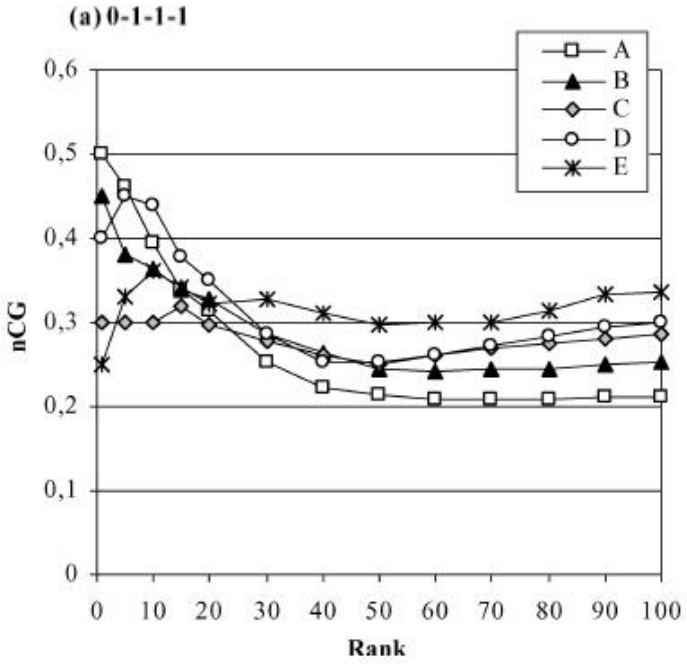


Fig. 3(a). Normalized cumulated gain (nCG) curves, binary weighting.

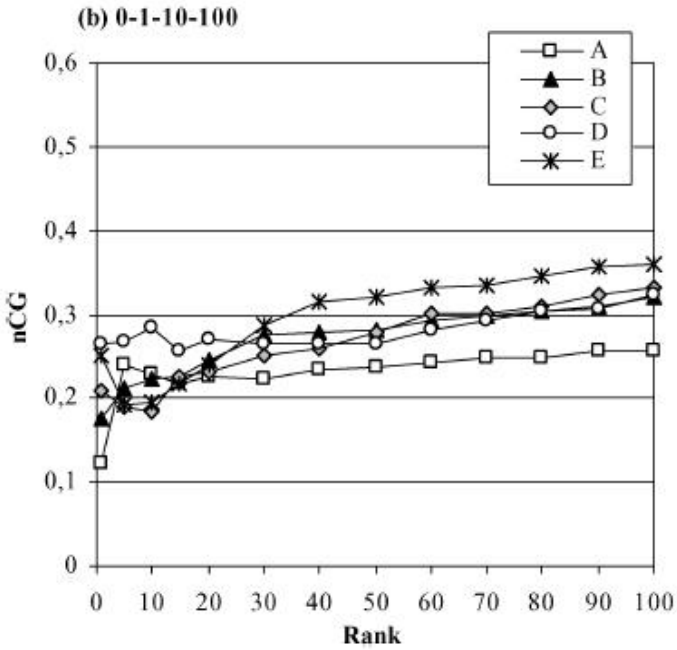


Fig. 3(b). Normalized cumulated gain (nCG) curves, nonbinary weighting.

NDCG

- 优点：
  - 图形直观，易解释
  - 支持非二值的相关度定义，比P-R曲线更精确
  - 能够反映用户的行为特征(如：用户的持续性persistence)
- 缺点：
  - 相关度的定义难以一致
  - 需要参数设定

指数NDCG

- 加大指标对相关度的敏感性，由线性变化改为指数变化，相关度3、2、1 在计算时用2^3、2^2、2^1

$$\mathrm{NDCG}(Q,k)=\frac{1}{|Q|}\sum_{j=1}^{|Q|}Z_{j,k}\sum_{m=1}^{k}\frac{2^{R(j,m)}-1}{\log(1+m)} \qquad (8\text{-}9)$$

- 据说搜索引擎公司常用这个公式

关于评价方面的研究

- 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)
  - 指标的计算(比如Pooling方法中如何降低人工代价？查询集或文档集合发生变化怎么办？)

推荐阅读：Ben Carterette, James Allan and Ramesh Sitaraman, Minimal Test Collections for Retrieval Evaluation, SIGIR06, best paper & best student paper.

提纲

1. 有关检索评价
4. 评价指标
3. 相关评测
4. 实验设计

TREC 概况

- The Text REtrieval Conference,TREC, 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2019年已有28届，可以看成信息检索的“奥运会”

TREC的目标(1)

- 总目标：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
- 1.鼓励对基于大测试集合的信息检索方法的研究
- 2.提供一个可以用来交流研究思想的论坛，增进工业界、学术界和政府部门之间的互相了解；

TREC的目标(2)

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(1)

- 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

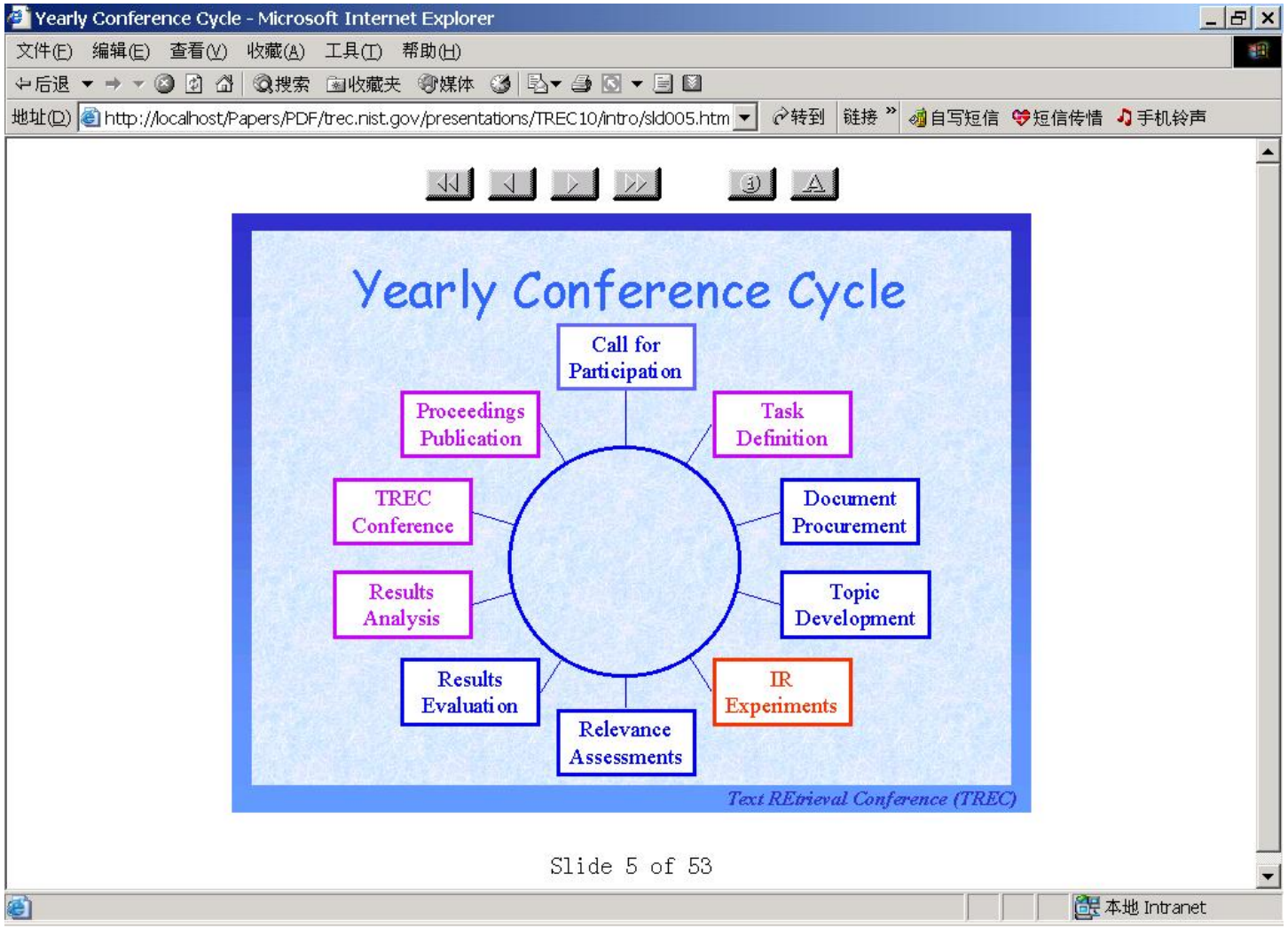
TREC的运行方式(2)

- 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，以便评估

TREC的运行方式(3)

- 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

TREC的运行方式(4)


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

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

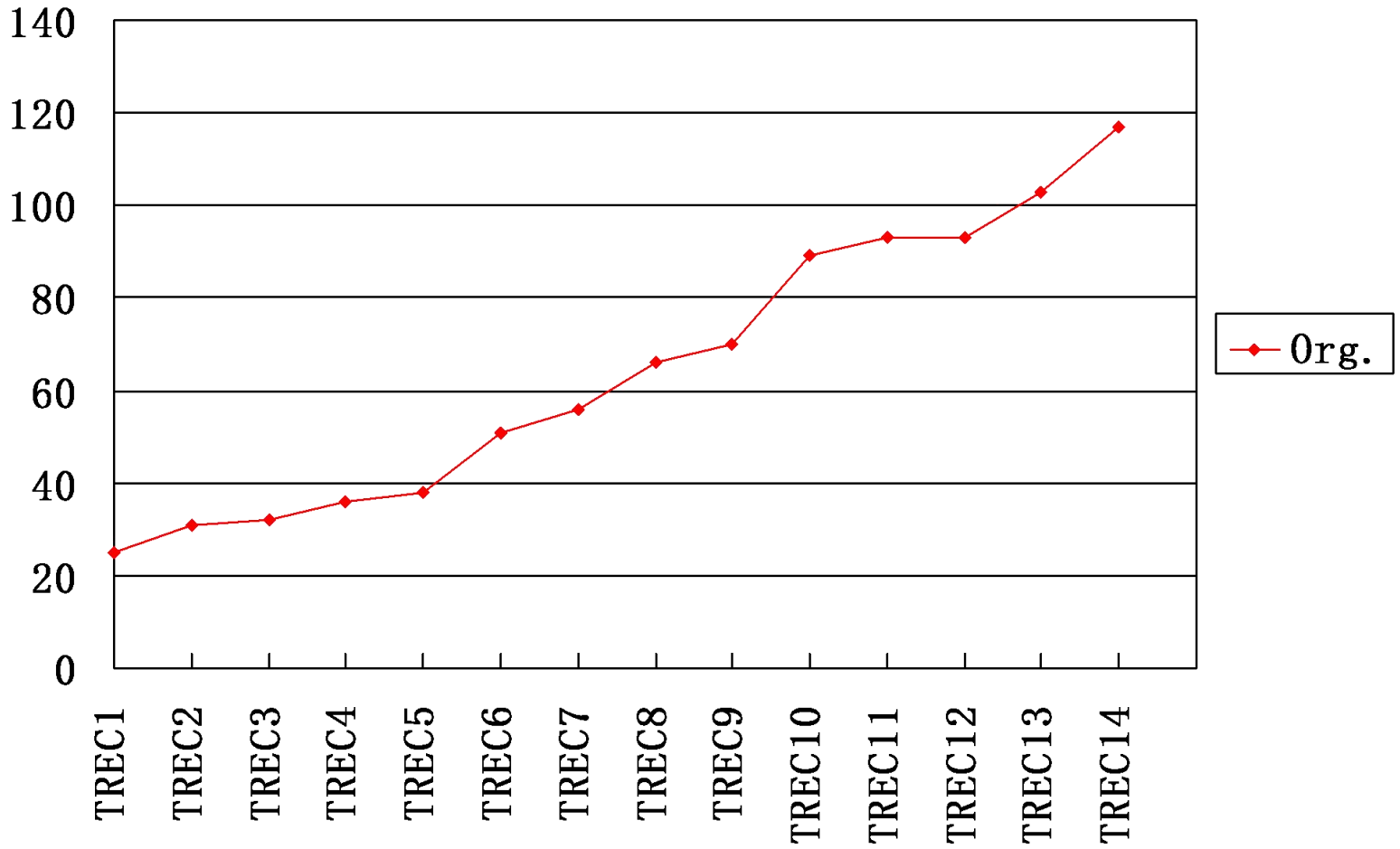
TREC任务

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

TREC任务

- 近年来顺势时代变化增加了多个项目
- Blog/Microblog
- Legal
- Covid
- Deep Learning
- 等等...

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) TREC11
Sun	Cornell U.	Tsinghua U.(Taiwan) 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……

TREC中名词定义

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

Topic的一般结构

- Title：标题，通常由几个单词构成，非常简短
- Description：描述，一句话，比Title详细，包含了Title的所有单词
- Narrative：详述，更详细地描述了哪些文档是相关的

Topic示例

<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.

使用Topic的方式

- 按照会议要求，可以利用Topic文本中的部分或者全部字段，构造适当的查询条件
- 可以使用任何方式构造查询条件，这包括手工的和自动的两大类。但提交查询结果时要注明产生方式。

评测方法

- 基于无序集合的评测：返回结果无顺序
  - Set Precision/Set Recall
- 基于有序集合的评测：
  - P@n/Average Precision/Reciprocal Rank
- 其他评测方法
  - Filtering Utility

相关性评估过程(1)

- (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中一部分运行结果，从每个运行结果中取头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。

相关性评估过程(2)

- 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。
- 其他track也有相应的公开评测工具

人工标注的有效性

- 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 (1)

- Kappa是度量判定间一致性的指标
- 为类别性判断结果(判定的结果是类别型)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

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

Kappa (2)

- *k在* [2/3, 1.0]时，判定结果是可以接受的
- 如果k值比较小，那么需要对判定方法进行重新设计

计算kappa统计量

		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

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

Pooled marginals (汇总边际)

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

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

Probability that the two judges agreed by chance $P(E) =$

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

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

- 事实上对系统之间的相对排序没有明显影响
  - 比如，我们想知道A算法是否好于B算法
  - 信息检索实验会给出一个可靠的答案，即使判定人员之间的不一致性可能很大

大型搜索引擎的评价

- Web下召回率难以计算
- 搜索引擎常使用top k的正确率来度量, 比如, $k = 10$ . . .
- . . . 或者使用一个考虑返回结果所在位置的指标，比如正确答案在第一个返回会比第十个返回的系统给予更大的指标
- 搜索引擎也往往使用非相关度指标
  - 比如：第一个结果的点击率
  - 仅仅基于单个点击使得该指标不太可靠 (比如你可能被检索结果的摘要所误导，等点进去一看，实际上是不相关的) . . .
  - 当然，如果考虑点击历史的整体情况会相当可靠
  - 比如：一些基于用户行为的指标
  - 比如: A/B 测试

A/B 测试

- 目标: 测试某个独立的创新点
- 先决条件：大型的搜索引擎已经在上线运行
- 很多用户使用老系统
- 将一小部分(如 1%)流量导向包含了创新点的新系统
- 对新旧系统进行自动评价，并得到某个评价指标，比如第一个结果的点击率
- 于是，可以通过新旧系统的指标对比来判断创新点的效果
- 这也可能是大型搜索引擎最信赖的方法

上述相关性定义的补充

- 上述相关性定义是针对独立的查询-文档对
- 另一种定义：边缘相关性(marginal relevance)
- 边缘相关性指的是结果列表中位置k上的文档相对于其前面的文档 $d_1 \ldots d_{k-1}$ 中包含的信息之外所带来的额外信息。同一篇文档，后面再次出现不能带来更多的信息。

- 课堂练习
  - 为什么边缘相关性更能表示用户的真实满意度？
  - 给出一个P、R类相关性指标不能反映用户满意度而边缘相关性却能否反映用户满意度的例子。
  - 在实际系统中，使用边缘相关性的难点在哪？

关于评价的可信度

- 在多组数据集上评价，得到相似的结果
- 进行统计显著性检测(statistical significance test)
  - 配对t检验
  - 符号检验
  - Wilcoxon符号秩检验

查询	系统一的 AP (X)	系统二的 AP (Y)	X-Y
1	0.0273	0.0323	-0.005
2	0.5725	0.5796	-0.0071
3	0.1388	0.1772	-0.0384
4	0.1196	0.1066	0.013
5	0.0015	0.0033	-0.0018
6	0.1069	0.1093	-0.0024
7	0.2127	0.2311	-0.0184
8	0.3617	0.4414	-0.0797
9	0.0005	0.001	-0.0005
10	0.1636	0.1426	0.021
11	0.3618	0.4206	-0.0588
12	0.7412	0.7412	0
13	0.6814	0.7866	-0.1052
14	0.0019	0.0023	-0.0004
15	0.0362	0.0113	0.0249
均值	0.2352	0.2524	-0.0172

提纲

1. 有关检索评价
4. 评价指标
3. 相关评测
4. 实验设计

IR系统评价实验设计

- “In this paper, we propose a ...”
  - Is the proposed method valid?
  - Is the proposed method effective?
  - Can the proposed method be generalized to other datasets?

- 一个正确的实验方案是对系统效果评判的基础
- Donald Metzler: 在我审过并且拒掉的论文中，85%被拒的原因都是实验方法不正确

实验方案设计中要考虑到的问题

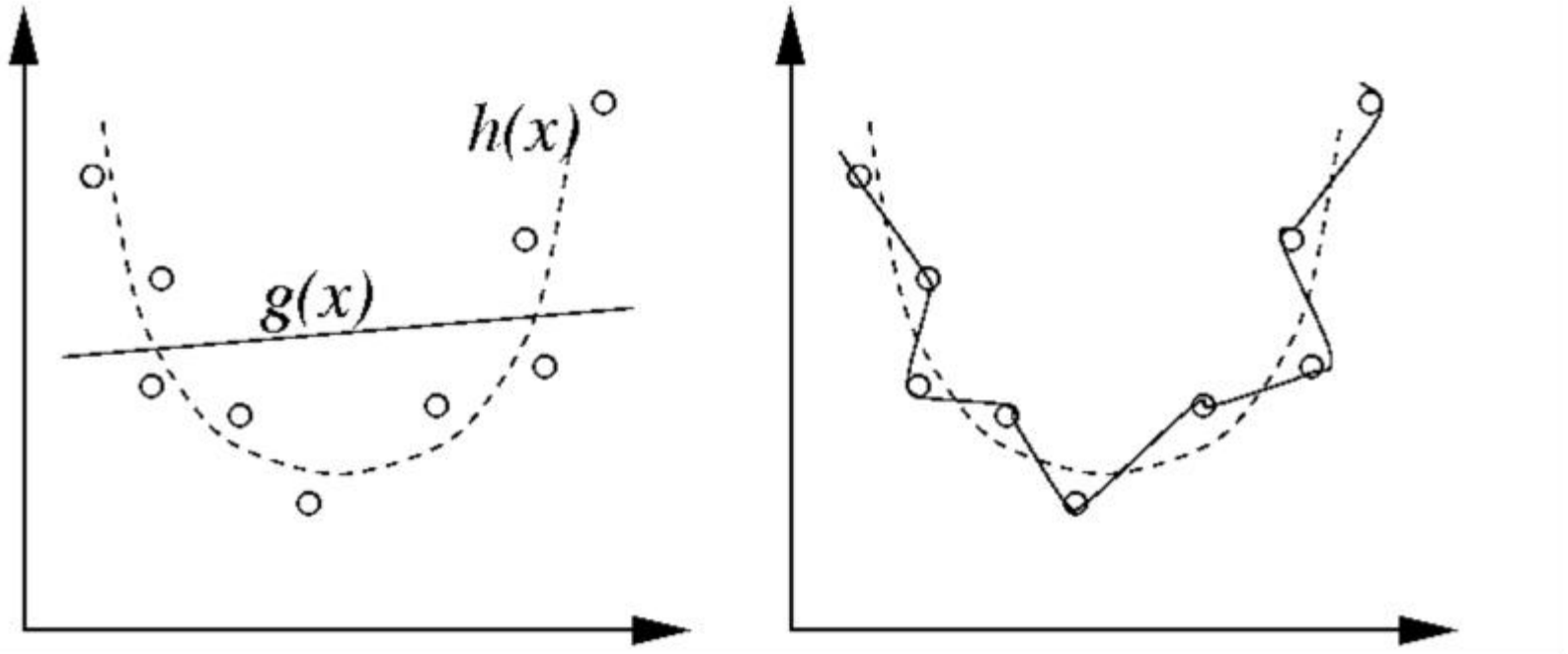
- Baseline –实验比较的基准
  - 计算机科学研究往往是关于“更好”的新模型、方法、算法，需要通过实验与基准比较，证明确实“更好”
  - baseline是否适用？
  - baseline是否够好？
  - 是否是一个公平的比较？
- Dataset — 数据集
  - 使用的数据集是否有足够的代表性
- 模型往往有待调参数（例如回转长度归一因子），因此需要一种 训练—测试方案
  - 在训练集上学习得到优化模型
  - 在测试集上评价

过拟合 (Over-fitting)

- Over-fitting occurs when a statistical model describes **random error or noise** instead of the underlying relationship

- Over-fitting generally occurs when a model is **excessively complex**, such as having **too many degrees of freedom**, in relation to the **amount of data** available

示例：过度描述随机误差



过拟合

- Over-fitting occurs when a statistical model describes random error or noise instead of the underlying relationship
- Over-fitting generally occurs when a model is excessively complex, such as having too many degrees of freedom, in relation to the amount of data available
  - Lack of training data
- A model which has been over-fit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data

评价方案

- Hold-out evaluation (单次检验)
  - 在一个数据集上训练，在另一个无交叉的测试集上评价
- Leave-one-out (留一交叉检验)
  - 将每个数据单元作为一个数据子集(例如将50个查询划分为50个子集)
  - 对于每个子集i，用其它(N-1)个数据子集训练，在i上测试
  - 最后将N个测试子集上得到的结果取平均值
  - 适用于仅有极少量数据的场合
- k-fold Cross-validation (K折交叉检验)
  - 非常流行的IR评价方案
  - 将数据集划分为n个子集，每个子集可包含多个数据单元
  - 例如将50个查询划分为5个子集，每个包含10个查询
  - 对于每个子集i，用其它(n-1)个数据子集训练，在i上测试
  - 最后将n个测试子集上得到的结果取平均值

测试数据

- 常碰到的论文评审意见：数据集代表性不够，因此实验结论尚需经过其它数据集上实验检验
  - 用更多、更大、更具有代表性的数据集

- Hold-out / k-fold cross-validation
  - Can the training – test data represent the nature of the data?
  - 换句话说，为什么这样划分数据子集？
  - Monte Carlo method
    - Randomly Re-samples the dataset

测试语料与查询

- 理论上，需要同时划分文档语料与查询
- 但在IR实验中，通常只划分查询
- 可能的问题：在同样的语料上进行训练和测试可能导致不公平的评价
- 但是IR研究通常是针对特定的数据类型，例如网页、博客、微博等，应用领域相对固定
  - 因此只划分查询也是可行的，也是最常用的实验方案

统计测试 (Statistical Test)

- “Evaluation shows improvement over baseline”
  - Is the improvement significant? 提高是否显著？
  - Is the improvement obtained by chance? 是否只是偶然得到的结果?

- 训练－测试评价方案和统计测试是IR研究的基本组成部分

比较下列两组“提高”

Q1	0.1	0.2
Q2	0.15	0.17
Q3	0.25	0.28
Q4	0.05	0.12
Q5	0.34	0.45
Q6	0.66	0.82
Q7	0.36	0.40
Q8	0.68	0.78
Q9	0.12	0.14
Q10	0.15	0.18
Av.	0.286	0.354

Imp=23.78%, p-value=0.01562

Significant at 0.05 level

Q1	0.1	0.8
Q2	0.15	0.03
Q3	0.25	0.58
Q4	0.05	0.05
Q5	0.34	0.55
Q6	0.66	0.52
Q7	0.36	0.25
Q8	0.68	0.60
Q9	0.12	0.84
Q10	0.15	0.10
Av.	0.286	0.432

Imp=51.05, p-value=0.4258

Insignificant improvement

统计测试

- 统计测试检验在评价实验中不同系统间表现出的性能差异是否是偶然得到的 (obtained by chance)
- 如何选择正确的统计测试方法是IR实验中的一个重要问题
- 参考文献：Selecting Statistical Tests
  http://www.cios.org/readbook/rmcs/ch19.pdf

本讲小结

- 信息检索的评价方法
  - 不考虑序检索评价指标(即基于集合)：P、R、F
  - 考虑序的评价指标：P/R曲线、MAP、NDCG
- 信息检索评测语料及会议
- 检索实验的设计

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?
  - Google VP of Engineering on search quality evaluation at Google