计算机信息检索

第2章 信息检索的评价(IR evaluation)

张金柱

课前思考题

- 为什么要评价？
- 评价什么？
- 如何评价？
- 怎么基于检索结果，计算各项评价指标？

提纲

- 信息检索的评价
  - 基本指标：召回率、正确率
  - 其他指标：F值、AP、MAP
- TREC会议概况

从竞技体育谈起

- **评价时刻在你我身边**
  - 110米栏世界纪录：2012年9月8日，美国选手梅里特，12秒80，创造了新的世界纪录。
  - 男子马拉松世界最好成绩：基梅托，肯尼亚人，于2014年9月28日的柏林马拉松，人类首破2小时03分大关。
- **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：马拉松中的犯规
  - 指标要一样：速度、耐力

为什么要评估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。

IR中评价什么？

- **效率(Efficiency)—可以采用通常的评价方法**
  - 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
  - 返回得靠不靠前
- **其他指标**
  - 覆盖率(Coverage)
  - 访问量
  - 数据更新速度

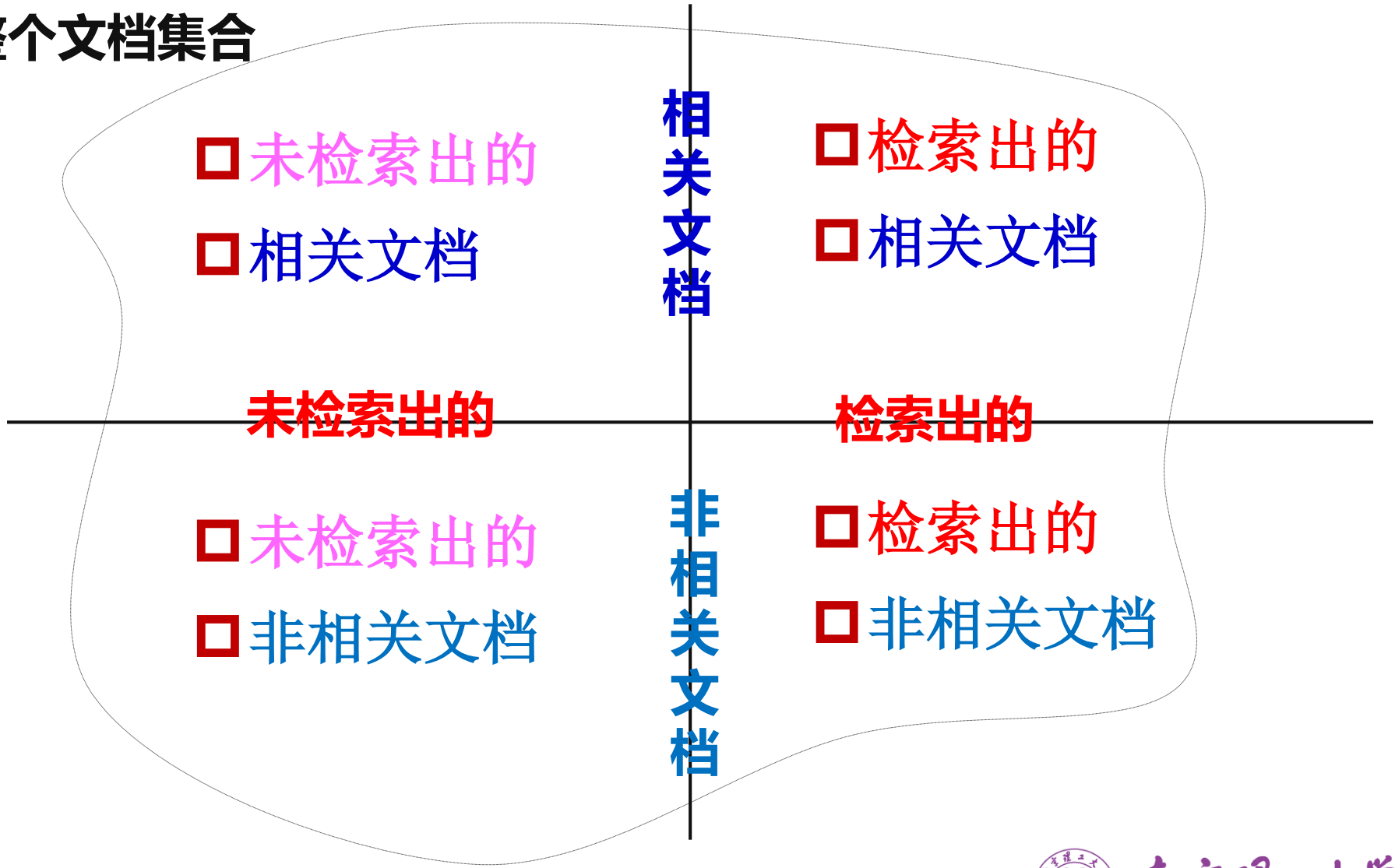
如何评价效果？

- **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(TextREtrievalConference), 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”

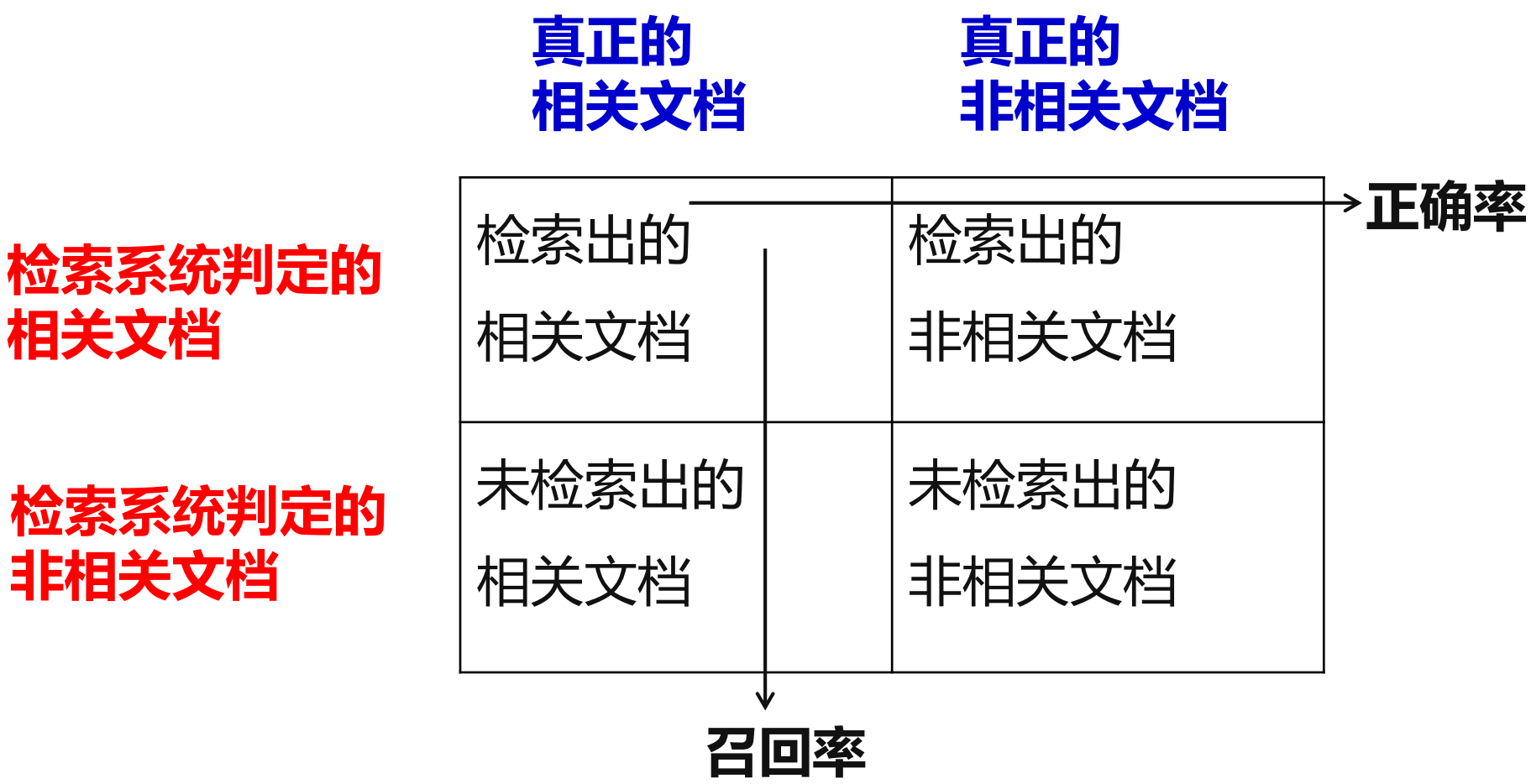
评价指标分类

- 对单个查询进行评估的指标
  - 对单个查询得到结果
- 对多个查询进行评估的指标(通常用于对系统的评价)
  - 求平均

评价指标



评价指标



评价指标

- **正确率(Precision)，返回的检索结果中真正相关结果数占结果总数的比率，也称为查准率，P∈[0,1]**

  =检索出的相关文档数/检索出的所有文档数

- **召回率(Recall)，返回的检索结果中真正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]**

  =检索出的相关文档数/真正的相关文档总数

- **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1篇，P=100%，但R极低；
  - 全部返回，R＝1，但P极低

召回率和正确率一个计算例子

- 一个例子：某一查询Q，真正的相关文档数为100篇，某个系统返回200篇文档，其中80篇是真正相关的文档
  - Precision=80/200=0.4
  - Recall=80/100=0.8
- 结论：召回率较高，但是正确率较低

关于正确率和召回率的讨论(1)

- 准确率和召回率互相影响，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高，当然如果两者都低，那是什么地方出问题了。
- “宁可错杀一千，不可放过一人”
  - 偏重召回率，忽视正确率。冤杀太多。
- 判断是否有罪：
  - 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

关于正确率和召回率的讨论(2)

- **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。

关于召回率的计算

- 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，很难准确地计算召回率
- 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，在TREC会议中被广泛采用。

正确率和召回率的问题

- **两个指标分别衡量了系统的某个方面，但是为比较带来了难度，究竟哪个系统好？**
  - 解决方法：单一指标，将两个指标融成一个指标
- **两个指标都是基于集合进行计算，并没有考虑序的作用**
  - 举例：两个系统，对某个查询，系统返回的真正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

评价指标(3)—P和R融合

- F值：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：

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

- E值：召回率R和正确率P的加权平均值，b>1表示更重视P

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

评价指标(4) - 引入序的作用

- **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。**

评价指标(5)—引入序的作用

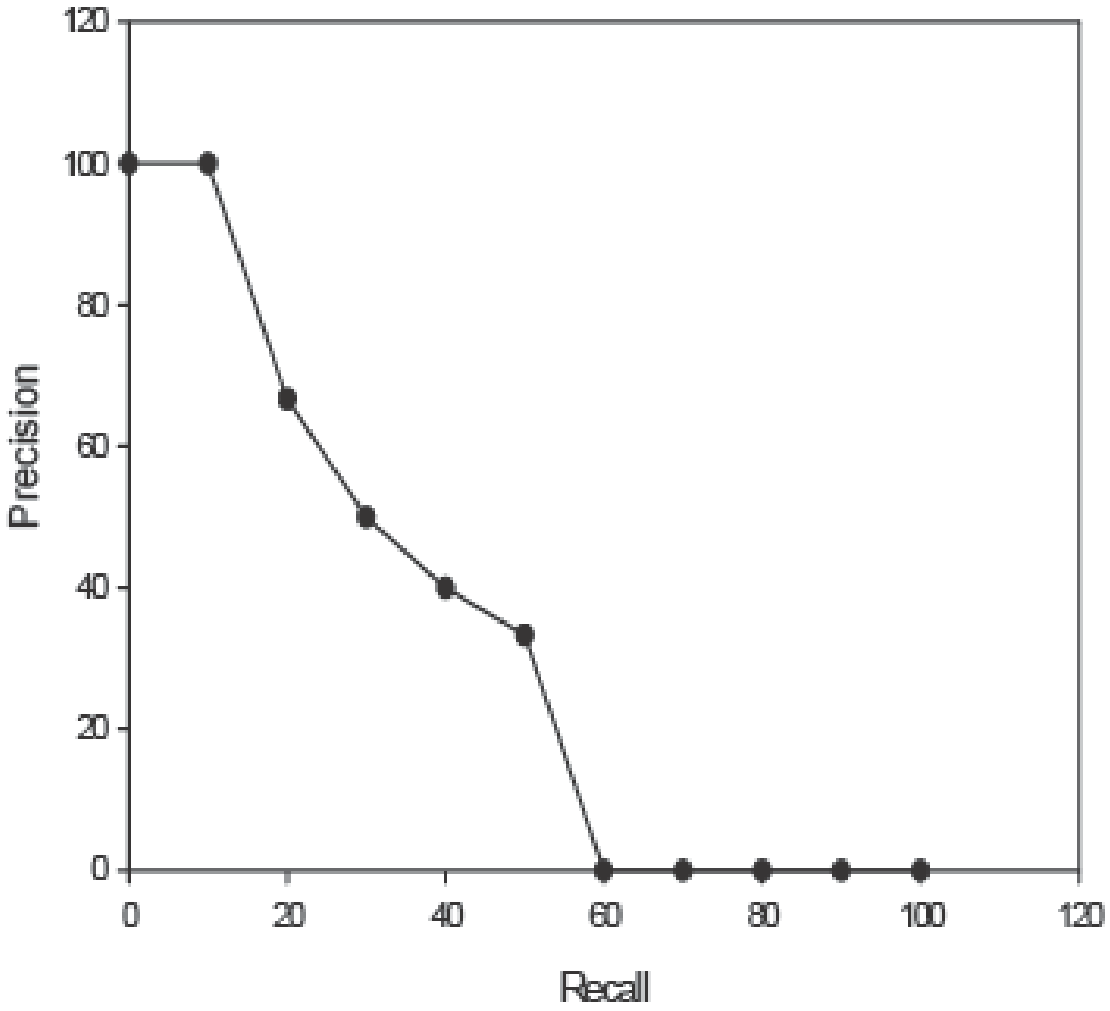
- **正确率-召回率曲线(precision versus recall curve)**
  - 检索结果以排序方式排列，用户不可能马上看到全部文档，因此，在用户观察的过程中，正确率和召回率在不断变化(vary)。
  - 可以求出在召回率分别为0%,10%,20%,30%,…,90%,100%上对应的正确率，然后描出图像

P-R曲线的例子

- 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



P-R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

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(通常用于对系统的评价)
  - 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**

评价指标

- 宏平均和微平均的例子
- 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，则：
  - 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

评价指标测试

- **两个查询q1、q2的标准答案数目分别为100个和200个，某系统对q1检索出150个结果，其中正确数目为60，前100条结果中有20条正确结果；系统对q2检索出400个结果，其中正确数目为100，前200条结果中有80条正确结果。则：**
  - Q1：P1=:60/150=0.4, R1=60/100=0.6
  - Q2：P2=100/400=0.25, R2=100/200=0.5
  - MacroP=(P1+P2)/2=0.325 MacroR=(R1+R2)/2=0.55
  - MicroP=(60+100)/(150+400)=0.29
  - MicroR=(60+100)/(100+200)=0.53

面向用户的评价指标

- 前面的指标都没有考虑用户因素。而相关不相关由用户判定。
- 假定用户已知的相关文档集合为U，检索结果和U的交集为Ru，则可以定义覆盖率(Coverage) C=|Ru|/|U|，表示系统找到的用户已知的相关文档比例。
- 假定检索结果中返回一些用户以前未知的相关文档Rk，则可以定义出新率(Novelty Ratio) N=|Rk|/(|Ru|+|Rk|)，表示系统返回的新相关文档的比例。

其他评价指标

本章小结

- 为什么要评价？
- 如何评价？
- 各种评价指标(正确率、召回率、平均正确率)的定义及计算方法
- 基本指标：正确率、召回率

课后练习题

- **两个系统A,B，两个查询q1,q2, 它们的标准答案分别是：**
  - Rq1={d1,d4,d28,d39,d56},
  - Rq2={d3,d7,d16,d45,d86,d97}
- **A、B 检索的结果分别如下表所示。试计算出每个系统对每个查询的P、R、F、P-R曲线指标。请写出计算过程和结果。**

系统-查询	返回结果数	正确答案位置
A-q1	20	2-d4;6-d28;10-d56
A-q2	20	1-d7;5-d3;14-d86; 20-d97
B-q1	20	1-d1;8-d28;12-d39;20-d56
B-q2	20	2-d3; 4-d7; 15-d45, 17-d97

问题？