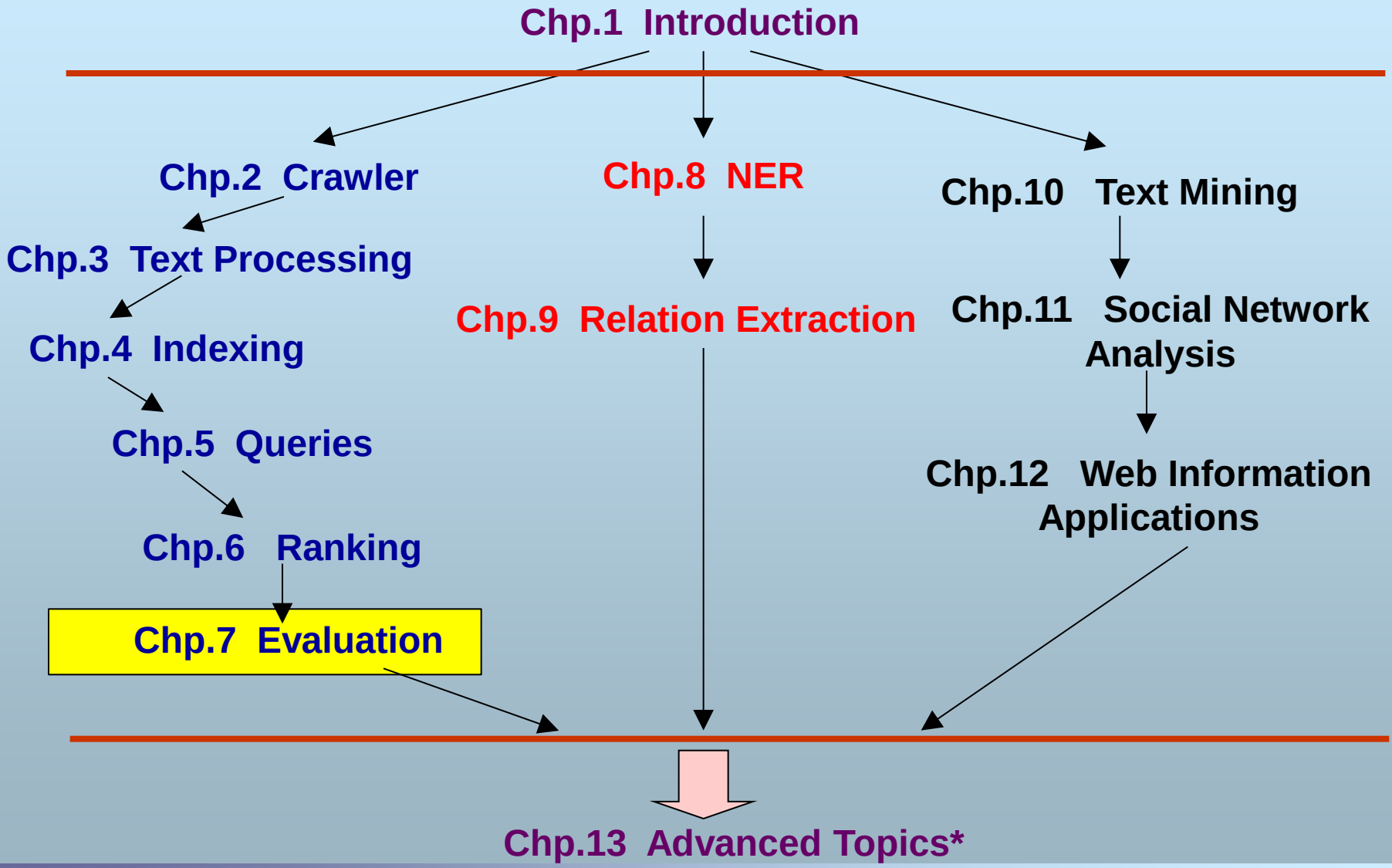
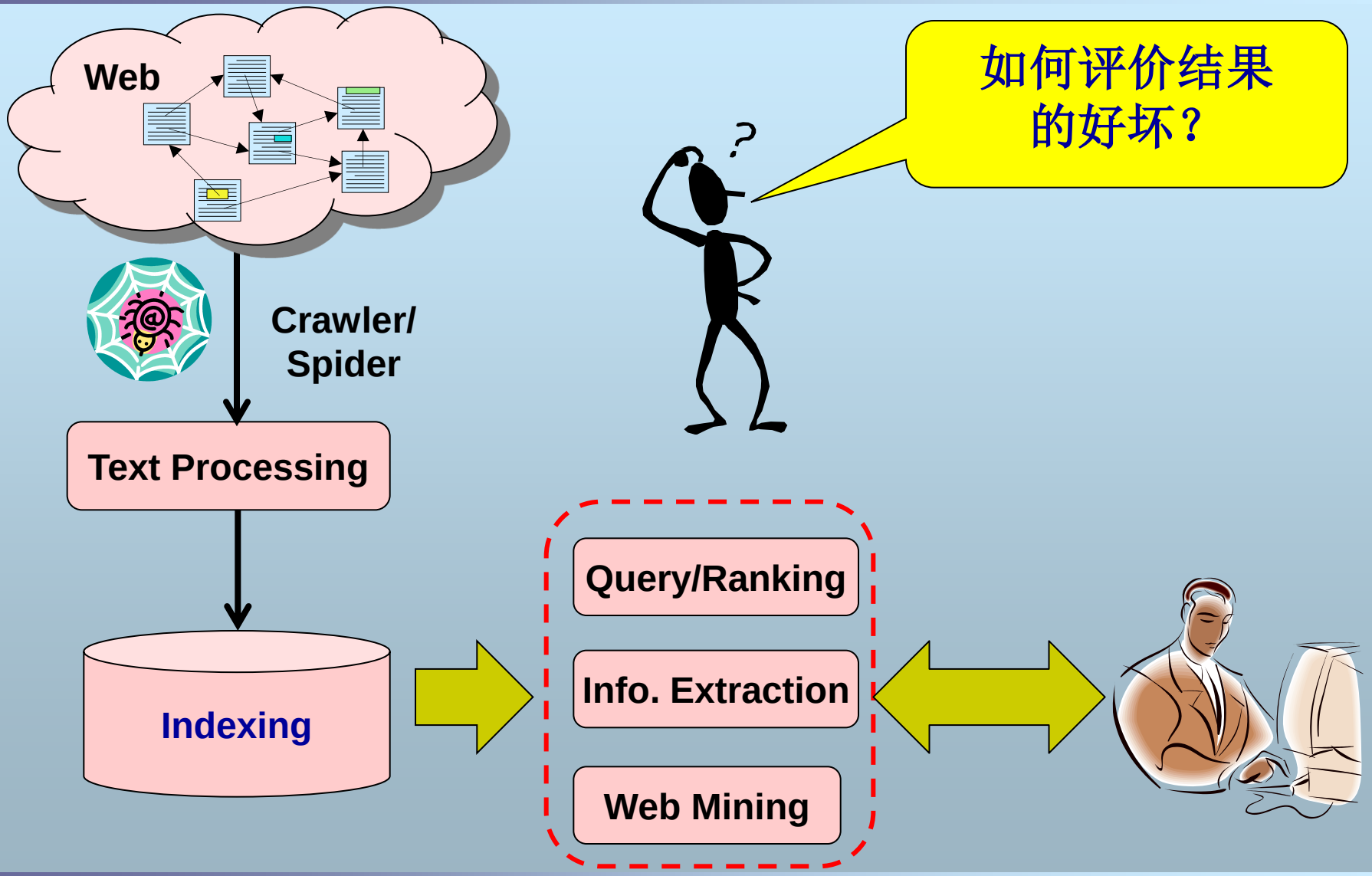
Evaluation

金培权（jpq@ustc.edu.cn）

课程知识结构


Chp.1 Introduction
Chp.2 Crawler
Chp.3 Text Processing
Chp.4 Indexing
Chp.5 Queries
Chp.6 Ranking
Chp.7 Evaluation
Chp.8 NER
Chp.9 Relation Extraction
Chp.10 Text Mining
Chp.11 Social Network Analysis
Chp.12 Web Information Applications
Chp.13 Advanced Topics*

本章讨论的问题


Web
Crawler/
Spider
Text Processing
Indexing
Query/Ranking
Info. Extraction
Web Mining
如何评价结果
的好坏？

本章主要内容

- 信息检索评价概述
- 评价指标

一、IR评价概述

- 评价很难，但是似乎又很容易
  - 主观的，依赖于特定用户的判断
  - 和情景相关的，依赖于用户的需求
  - 认知的，依赖于人的认知和行为能力
  - 时变的，随着时间而变化
- 评价要公平！
  - 例如，在竞技体育中
    - 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

1、IR为什么需要评价？

- 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。

2、IR需要评价什么？

- 最主要的两个方面
  - 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
    - 返回得靠不靠前
- 其他指标
  - 覆盖率(Coverage)
  - 访问量
  - 数据更新速度

3、IR评价的前提

- 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。
- 可以使用多种类型的文档集
  - CACM：ACM通讯标题和摘要（1958-1979）（几千个文档）
  - TREC(Text REtrieval Conference), 美国标准技术研究所, 1992 - (上百万篇文档)，信息检索的“奥运会”
    - AP：Associated Press news corpus （几十万个文档）
    - GOV2：从.gov域名爬取的网页（几百万个文档）

3、IR评价的前提

- 不同的文档集具有完全不同的特征
  - 数据集大小、相关性度量……

Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 Mb	64
AP	242,918	0.7 Gb	474
GOV2	25,205,179	426 Gb	1073

Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

4、IR评价的例子

- 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	

5、IR评价需要考虑的方面

- 评价指标：某个或某几个可衡量、可比较的值
  - 正确率
  - 召回率
  - F-measure
  - MAP
  - MRR
  - NDCG
  - ……
- 评价过程：设计上保证公平、合理

二、IR评价指标

- 效率评价指标（Efficiency）
- 效果评价指标（Effectiveness）

二、IR评价指标

Efficiency metrics

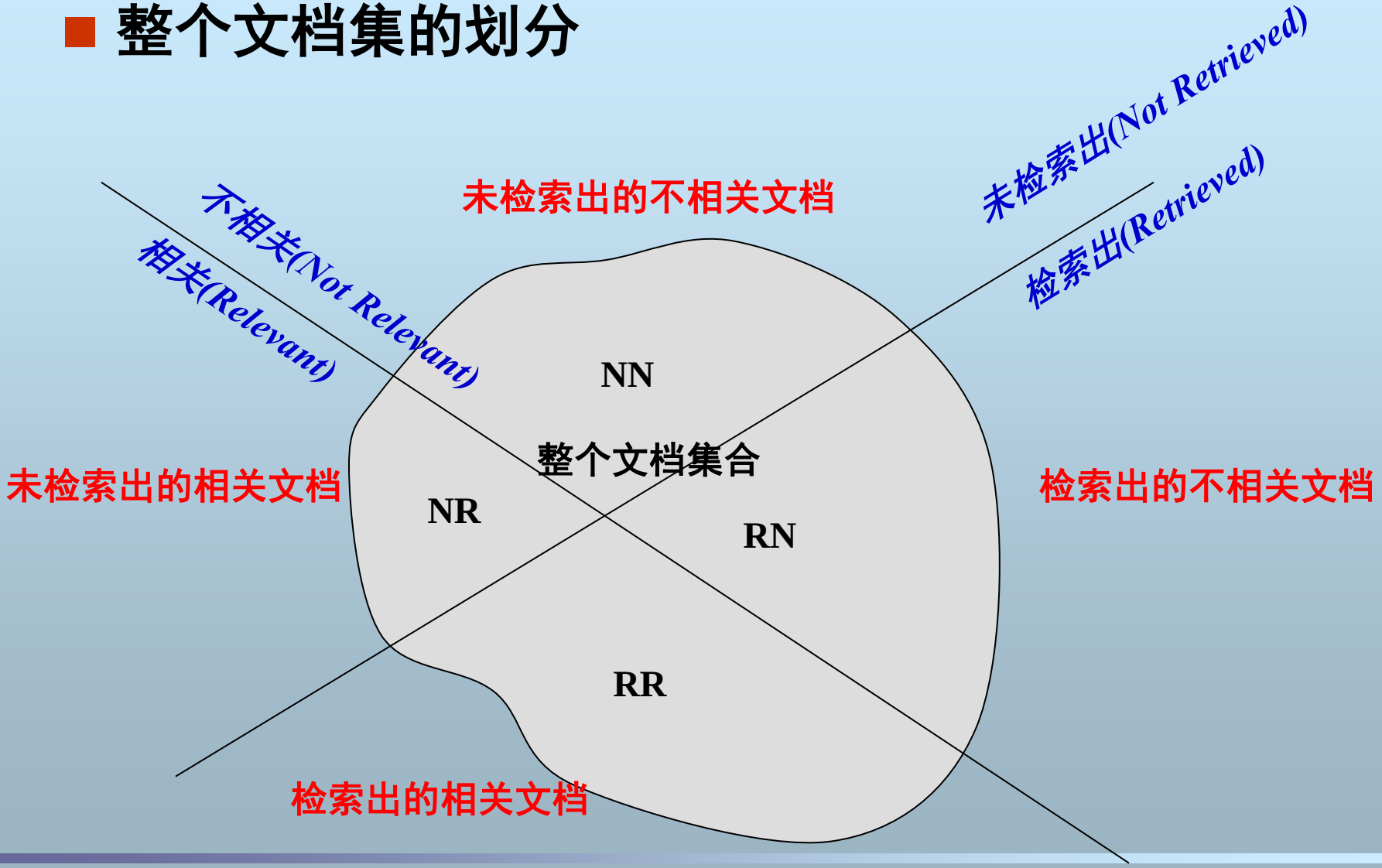
Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

二、IR评价指标

- **Effectiveness metrics**
  - 对单个查询进行评估的指标
    - 基于集合的评价指标
      - 正确率（Precision）、召回率（Recall）、F值
    - 基于序的评价指标
      - P@N、R-Precision、AP等
  - 对多个查询进行评估的指标
    - MAP、MRR
  - 其它的评价指标
    - NDCG

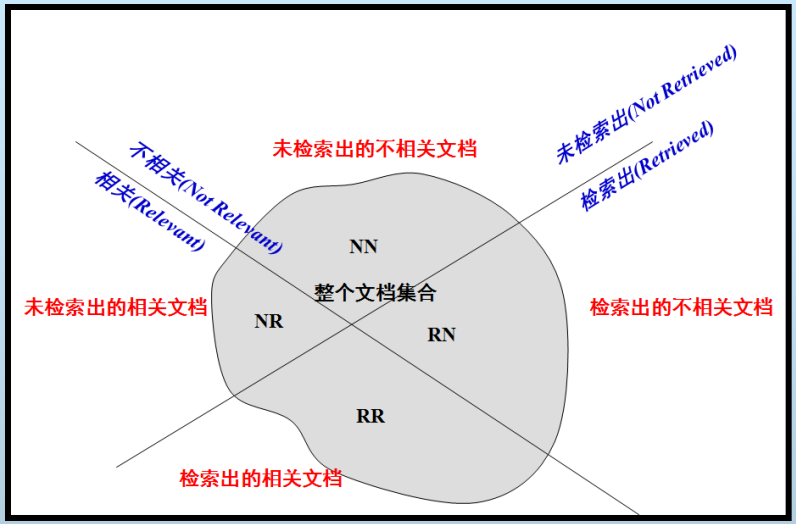
1、正确率和召回率

- 整个文档集的划分



1、正确率和召回率

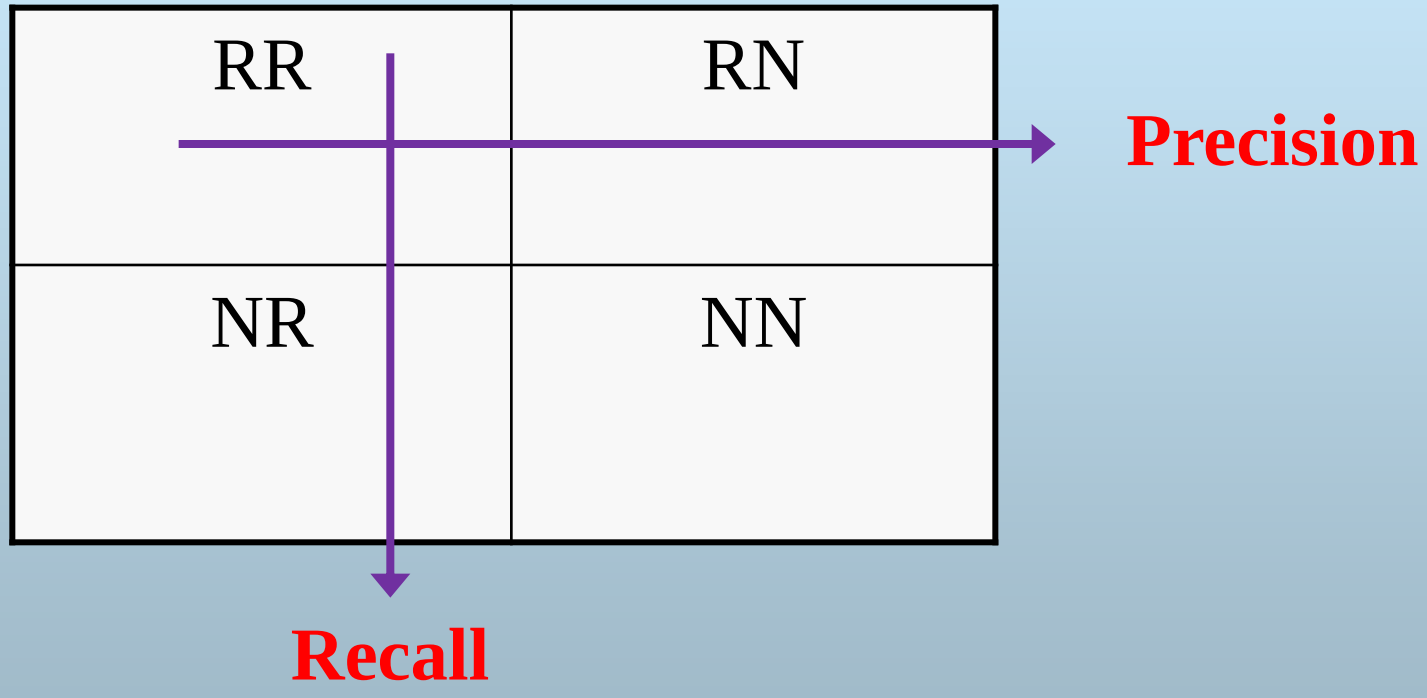
- 正确率(Precision)
  - RR/(RR + RN)
  - 返回的结果中真正相关结果的比率，也称为查准率 P∈ [0,1]
- 召回率(Recall)
  - RR/(RR + NR)
  - 返回的相关结果数占实际相关结果总数的比率，也称为查全率，R∈ [0,1]



- 两个指标分别度量检索效果的某个方面，忽略任何一方面都有失偏颇。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

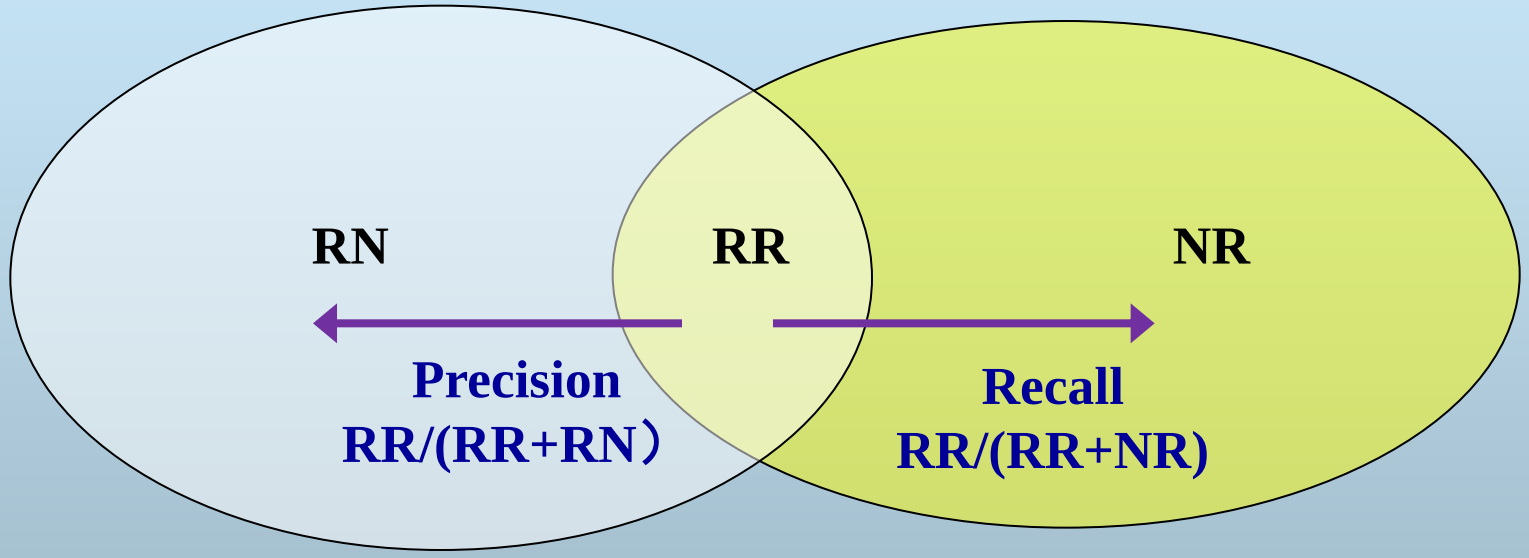
1、正确率和召回率

- 四种关系的矩阵表示



1、正确率和召回率

- 基于集合的表示



1、正确率和召回率

- 举例

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

1、正确率和召回率

- 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.....
- 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——召回率允许低一点
  - 海上目标检测：希望不要漏掉目标——召回率要求高
  - 知识问答：希望返回结果准一点，但不需要结果很全——正确率要求高

1、正确率和召回率

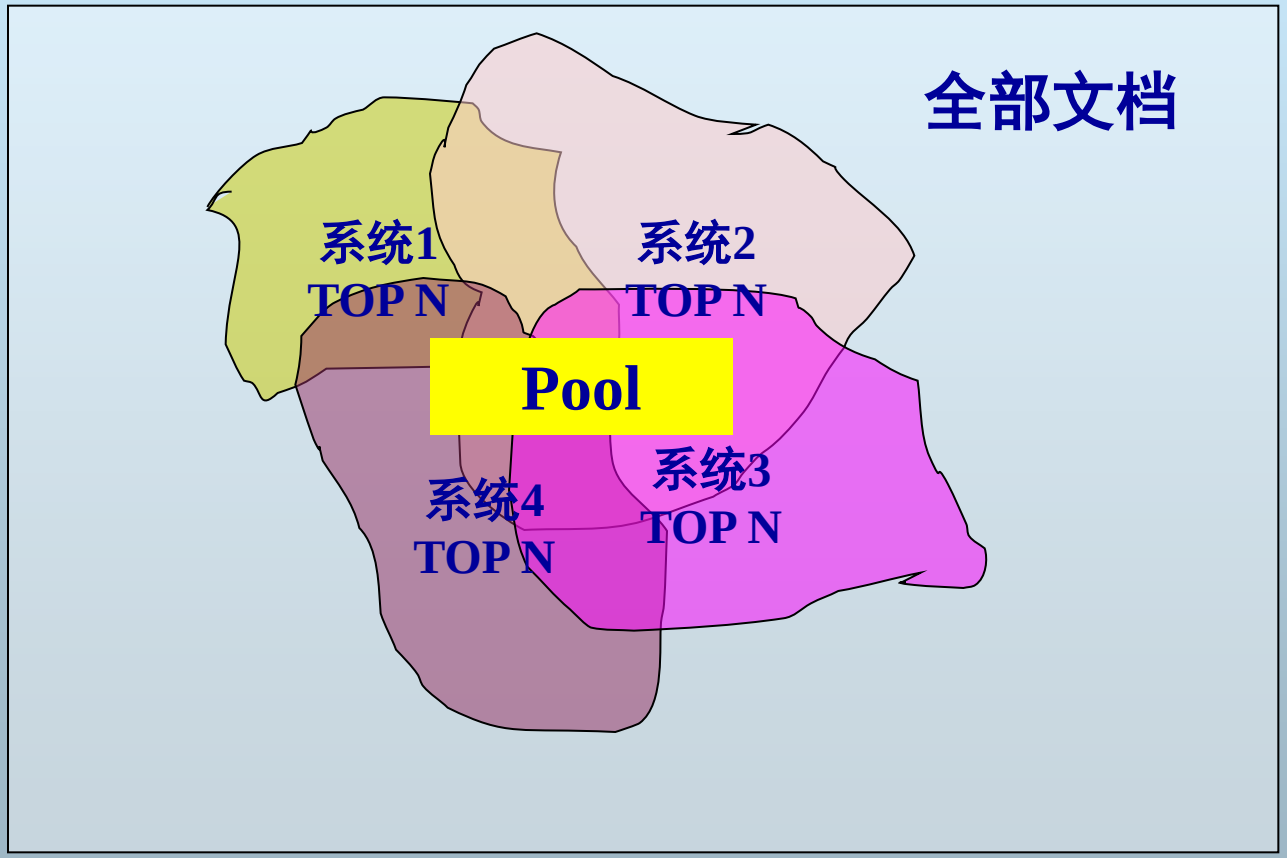
- 召回率的问题：
  - 对于大规模文档集合，列举每个查询的所有相关文档是不可能的事情，因此，不可能准确地计算召回率

1、正确率和召回率

- 解决方法：缓冲池(Pooling)方法
  - Pooling：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Top N个文档（例如N=100），将这些结果文档汇集起来，并进行人工标注，从而得到一个可能相关的文档池“pool”
  - 潜在的假设
    - 绝大多数的相关文档都收录在这个“pool”中
  - 这种做法被验证是可行的(可以比较不同系统的相对效果)——虽然返回不了全部的相关文档，但能够评价各个系统的相对好坏
  - 在TREC评测中被广泛采用（N：50～200）

1、正确率和召回率

- Pooling方法



2、F-measure

- 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：

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

- 更一般的情况——F_β：参数β用于调节召回率和正确率的相对重要程度。$\beta>1$更重视召回率，$\beta<1$更重视正确率。F值即$\beta=1$时的F_1值。F_2值（更重视召回率）和$F_{0.5}$值（更重视正确率）也是常用的指标值

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

2、F-measure

- 为什么不使用其他平均来计算F，比如算术平均

- 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。

正确率P、召回率R、F值是信息检索领域最常用的3个评价指标

3、P@N

- 正确率和召回率的问题
  - 两个指标都是基于(无序)集合进行计算，并没有考虑序的作用
  - 举例：两个系统对某查询都返回20个文档，其中相关文档数都是10，但第一个系统是前10条结果，后一个系统是后10条结果。显然第一个系统优。但两者的P和R一样。
  - 解决方法：引入序的作用
- 考虑序的评价指标
  - P@N、R-precision、AP等

3、P@N

- **P@N**
  - **即Precision@N**
  - **指在第N个位置上的正确率**
  - **对于搜索引擎，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**

3、P@N

举例

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

查询1的标准答案集合为 {d3,d4,d6,d9}
查询2的标准答案集合为 {d1,d2,d13}

系统1查询1: P@2=1, P@5=2/5;　　系统1查询2: P@2=1/2, P@5=2/5;

系统2查询1: P@2=1/2, P@5=2/5;　　系统2查询2: P@2=1, P@5=3/5

4、R-Precision

- **R-Precision**
  - 检索结果中，在所有相关文档总数位置上的正确率
  - 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&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

查询1的标准答案集合为 {d3,d4,d6,d9}　查询2的标准答案集合为 {d1,d2,d13}

系统1查询1: R-Precision=2/4;　系统1查询2: R-Precision=1/3;

系统2查询1: R-Precision=2/4;　系统2查询2: R-Precision=2/3;

5、AP（Average Precision）

- 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: 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
    - 由于每个查询的召回率值不一定就是这11个标准召回率，因此需要对正确率进行插补。
    - 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。
    - 召回率100%点若不存在，正确率可近似为0
  - 简化的AP: 只对返回的相关文档进行计算的AP, AP=(1/1+2/2+3/5+4/10+5/20)/5，倾向那些快速返回结果的系统，没有考虑召回率

5、AP（Average Precision）

Example

1. d123 •(1/1)	6. d9 • (4/6)	11. d38 • (7/11)
2. d84	7. d511	12. d48
3. d56 •(2/3)	8. d129 • (5/8)	13. d250
4. d6 •(3/4)	9. d187	14. d113 • (8/14)
5. d8	10. d25 • (6/10)	15. d3

假设查询的标准答案集合包含10个文档，返回了8个相关文档

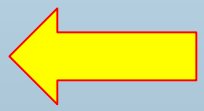
未插值的AP = (1/1+2/3+3/4+4/6+5/8+6/10+7/11+8/14+0+0)/10 【常用】

插值的AP = (1/1+1/1+2/3+3/4+4/6+5/8+6/10+7/11+8/14+0+0)/11

0% 10% 20% 30% 40% 80% 90% 100%

简化的AP = (1/1+2/3+3/4+4/6+5/8+6/10+7/11+8/14)/8

二、IR评价指标

- **Effectiveness metrics**
  - 对单个查询进行评估的指标
    - 基于集合的评价指标
      - 正确率（Precision）、召回率（Recall）、F值
    - 基于序的评价指标
      - P@N、R-Precision、AP等
  - 对多个查询进行评估的指标 
    - MAP、MRR
  - 其它的评价指标
    - NDCG

6、MAP

- MAP(Mean AP)
  - 对所有查询的AP求算术平均
  - 反映在全部查询上的检索效果
- 例如：假设有一个检索系统
  - 对查询1返回4个相关网页，其rank分别为1, 2, 4, 7
  - 对查询2返回3个相关网页，其rank分别为1, 3, 5
  - 查询1共有4个相关文档，查询2共有5个相关文档

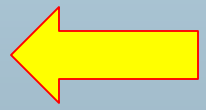
查询1：AP ＝ (1/1+2/2+3/4+4/7)/4＝0.83
查询2：AP ＝ (1/1+2/3+3/5+0+0)/5＝0.45
MAP ＝ (0.83+0.45)/2 ＝ 0.64

7、MRR

- 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
- 例如
  - 两个问题，系统对第一个问题返回的标准答案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＝ (1/2+1/4)/2＝3/8
  - 意味着平均在第8/3个位置处找到相关文档

二、IR评价指标

- **Effectiveness metrics**
  - 对单个查询进行评估的指标
    - 基于集合的评价指标
      - 正确率（Precision）、召回率（Recall）、F值
    - 基于序的评价指标
      - P@N、R-Precision、AP等
  - 对多个查询进行评估的指标
    - MAP、MRR
  - 其它的评价指标
    - NDCG



8、NDCG

- DCG
  - 一种总体观察检索排序效果的方法，利用检索结果序列的相关度加和的思路来衡量。
- 两个假设
  - 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

8、NDCG

- CG(Cumulative Gain)：位于位置1 到p 的检索结果的相关度之和。

$$CG_p = \sum_{i=1}^{p} rel_i$$

  - rel_i 表示第 i 个文档与查询的相关度
    - 可以不仅仅只有相关1和不相关0两种情况，允许有多个相关度级别，比如0, 1, 2, 3
  - 特点
    - CG得分高只能说明检索结果的总体质量比较高
    - 但不能说明结果排序的好坏：CG未考虑相关结果的位置，即前p项中两文档交换不影响计算结果
  - DCG 则希望改变这个特性

8、NDCG

- **DCG(Discounted Cumulative Gain)**
  - **基本思想: 若搜索算法把相关度高的文档排在后面，则应该给予惩罚。一般用log 函数表示这种惩罚。DCG 的计算如下：**

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

Discounted Gain

  - **另一种计算方法：**

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1+i)}$$

更强调排在前面的相关文档的重要性（指数）

8、NDCG

- DCG计算例子
  - 相关度0－3，10个文档的得分如下:
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - discounted gain:
    - ＝ 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
      ＝ 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
  - DCG＝3＋∑ discounted gain:
    - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

8、NDCG

- DCG的值与具体查询有关，和结果列表的长度有关，不利于检索系统之间的对比
  - 不同query的搜索结果有多有少，所以不同query的DCG值就没有办法来做对比
  - 例如，$DCG_5=6.89$，$DCG_{10}=9.61$
- NDCG (Normalized DCG)：对DCG进行规范化
  - 把检索结果按相关度从大到小排序得到一个理想的输出序列
  - 计算此理想序列的DCG, 得到在位置*p* 的ideal DCG（IDCG)
  - 然后以位置*p*的DCG_p与$IDCG_p$比值作为评价指标

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

8、NDCG

- NDCG计算示例
  - 沿用前面例子：相关度0－3，10个文档的得分如下:
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - 理想的输出结果序列: 3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- ideal DCG (IDCG):
  - 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88
- DCG＝3＋∑ discounted gain:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- NDCG: ($NDCG_i = DCG_i / IDCG_i$)

  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

  可以看到任何查询结果位置*p*的NDCG值都规范化为 ≤ 1 的值

本章小结

- 对单个查询进行评估的指标
  - 基于集合的评价指标
    - 正确率（Precision）、召回率（Recall）、F值
  - 基于序的评价指标
    - P@N、R-Precision、AP等
- 对多个查询进行评估的指标
  - MAP、MRR
- 其它的评价指标
  - NDCG