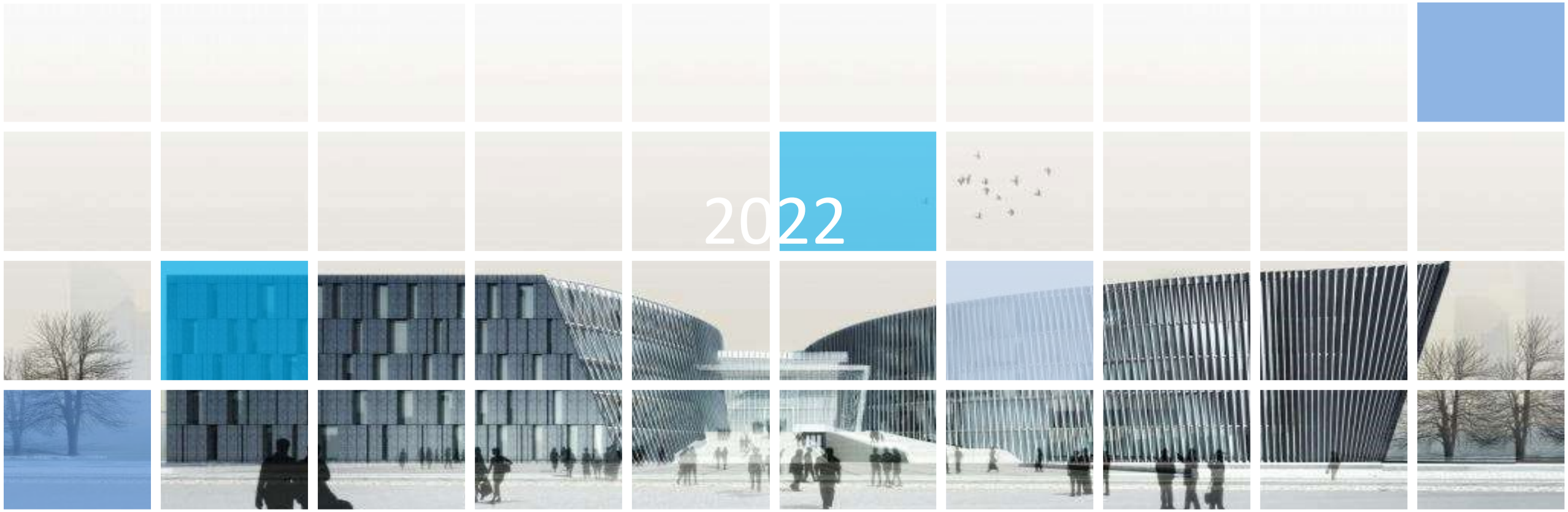


南京大学信息管理学院

信息检索

邓三鸿

njuir@sina.com

10

PART Ten

信息检索的评价

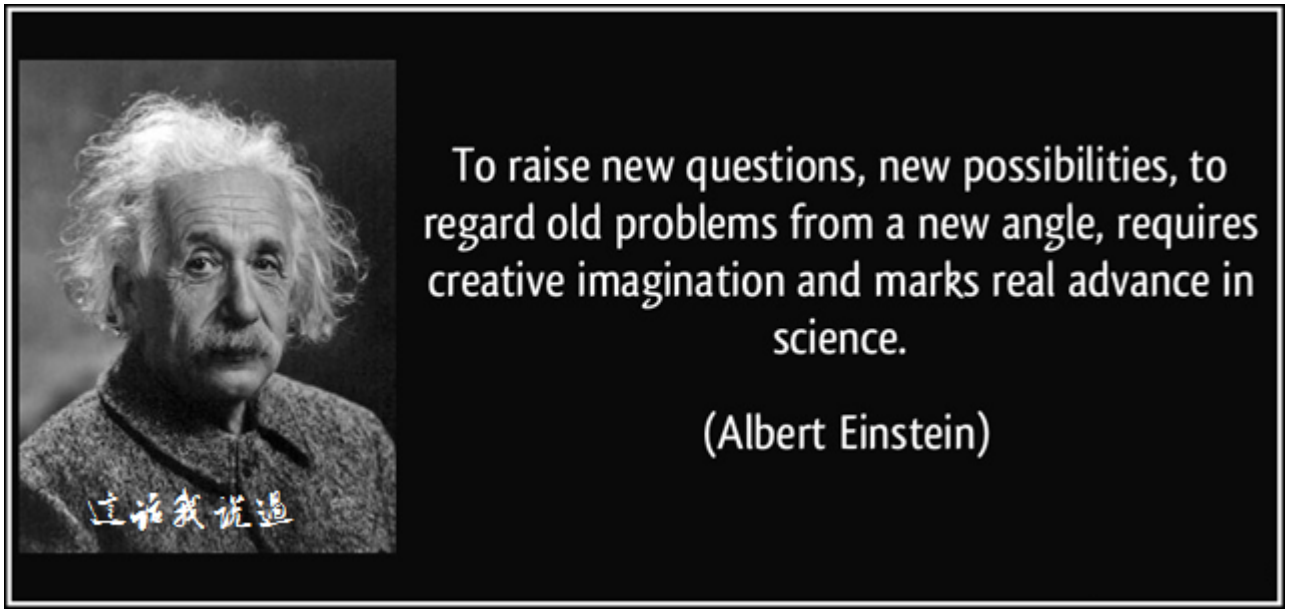
Evaluation of IR

关于评价

评价（Evaluation）：**发现和收集关于某种活动的数据，从中判断该项活动的质量及达到预期目标程度的行为。简单地说，评价就是对系统的价值和效率进行测评。**

信息检索系统评价：**根据给定的指标体系，采用一定的方法和程序，对信息检索系统的功能、特性和运营状况进行评测，或对有关假设、预期效益、性能值进行验证，以确定系统达到了何种水平、投入成本是否值得、是否可以改进和如何改进，乃至系统是否应生存下去。**

IR评价的意义



了解已有检索系统的功能，找出缺陷并改进；

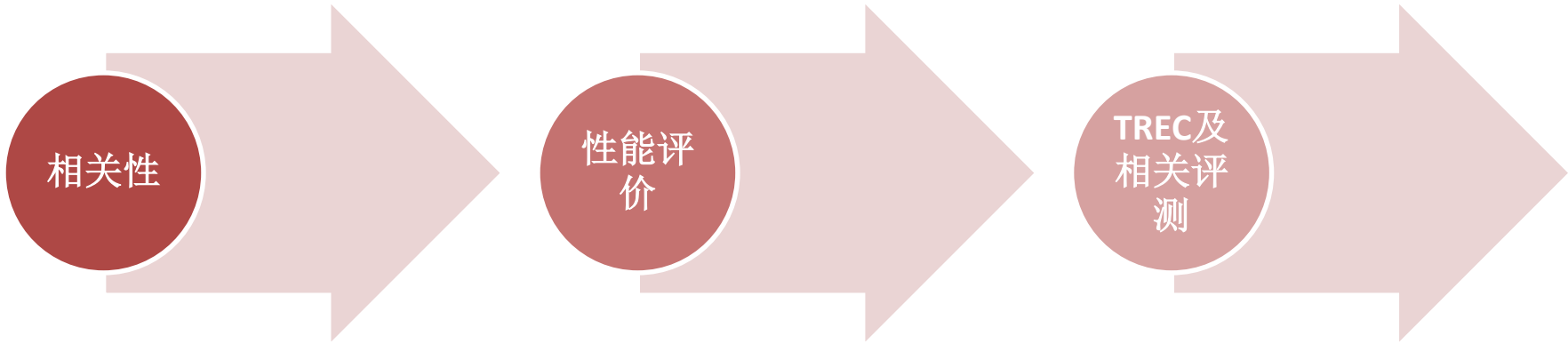
比较各种检索系统的优劣；

提高效率和效益；

有助于新的检索系统的设计；

丰富信息检索的理论。

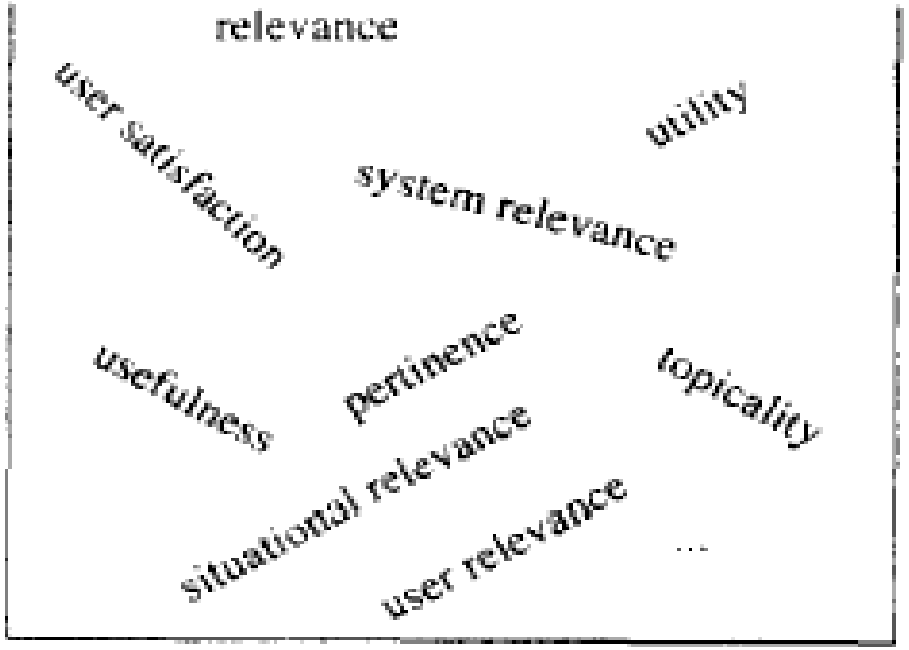
IR评价的相关内容



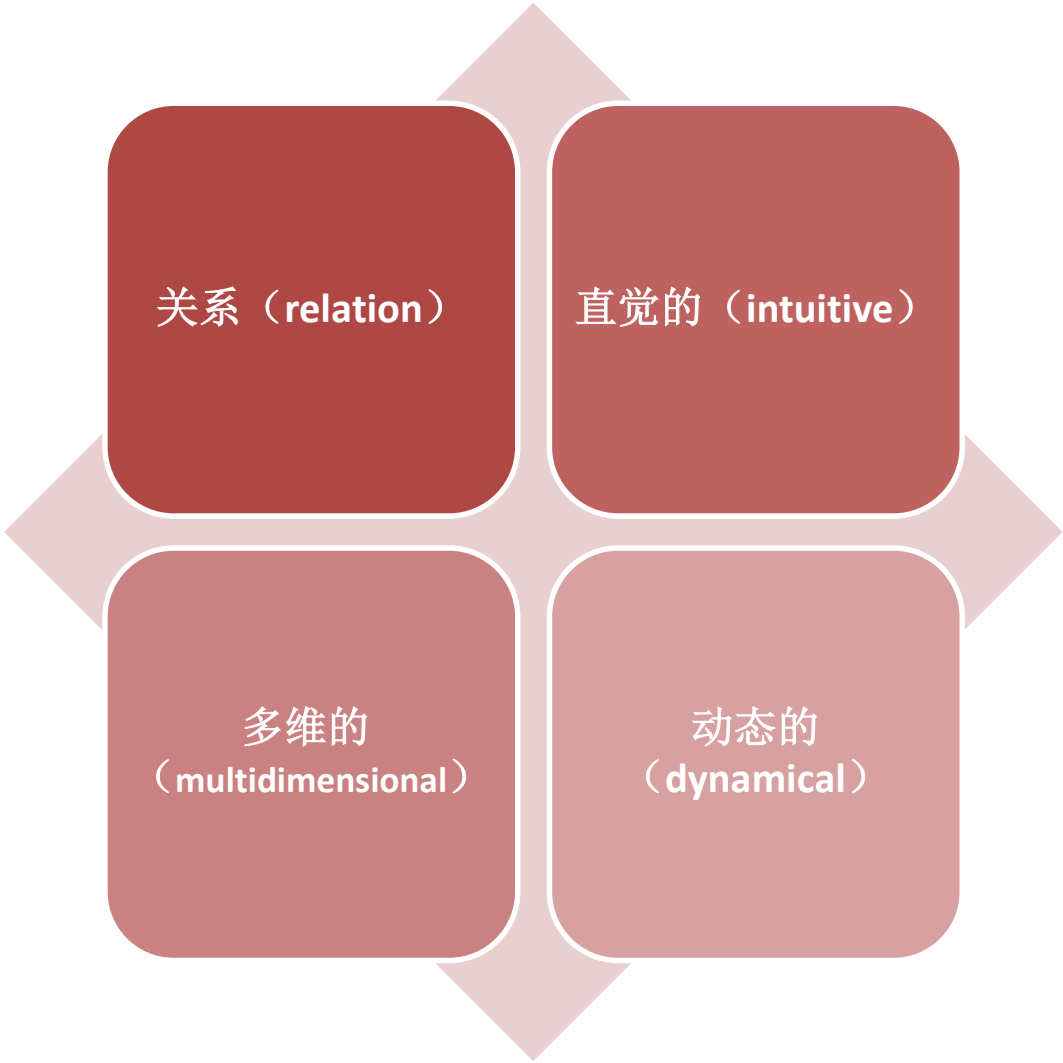
确定性与相关性

数据检索是“确定性”的
信息检索是“相关性”的

信息检索中的“相关性”主要是指检索系统针对用户的信息需求从文档集合中检出的**文档与用户需求之间的一种匹配关系**。



相关性的本质特征


关系（relation）
直觉的（intuitive）
多维的
（multidimensional）
动态的
（dynamical）

米扎罗的相关性问题模型

- **信息源**

  Surrogate < Document < Information

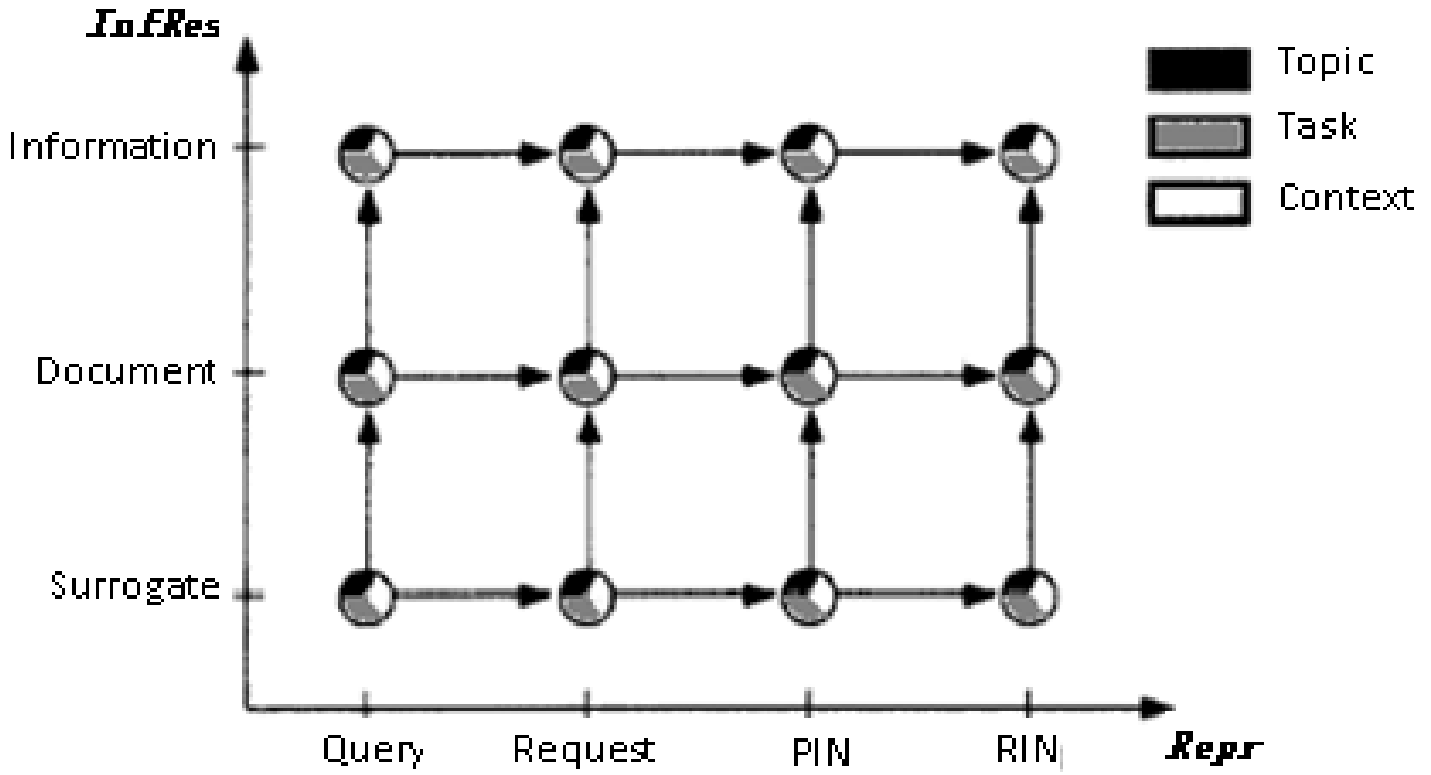
- **用户信息需求**

  RIN < PIN < Request < Query

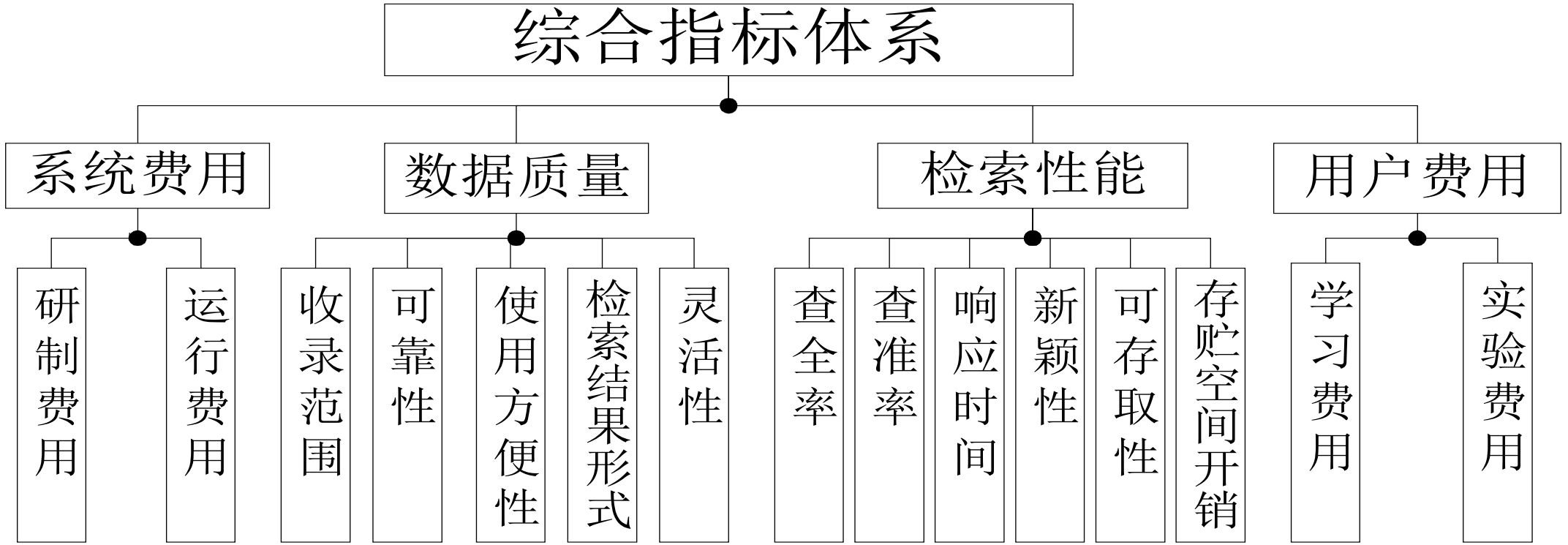
- **时间**

- **组件**

  主题、任务、情境或语境



综合评价体系（参考）


综合指标体系
系统费用
研制费用
运行费用
数据质量
收录范围
可靠性
使用方便性
检索结果形式
灵活性
检索性能
查全率
查准率
响应时间
新颖性
可存取性
存贮空间开销
用户费用
学习费用
实验费用

信息检索性能评价及评价指标

- *系统性能指标*
  - 时间效率、空间开销、响应速度……
- **系统角度的相关性判断及评价指标**
  - P、R、F、E……
- 用户角度的相关性判断及评价指标
  - 涵盖率、新颖率……

基本条件

- **一个文档集合*C*。**
  - 系统将从该集合中按照用户查询检出相关文档
- **一组用户查询{q_1, q_2, …, q_n}。**
  - 每个用户查询q_i描述了用户的信息需求
- **对应每个用户查询的标准相关文档集{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

评价任务示例

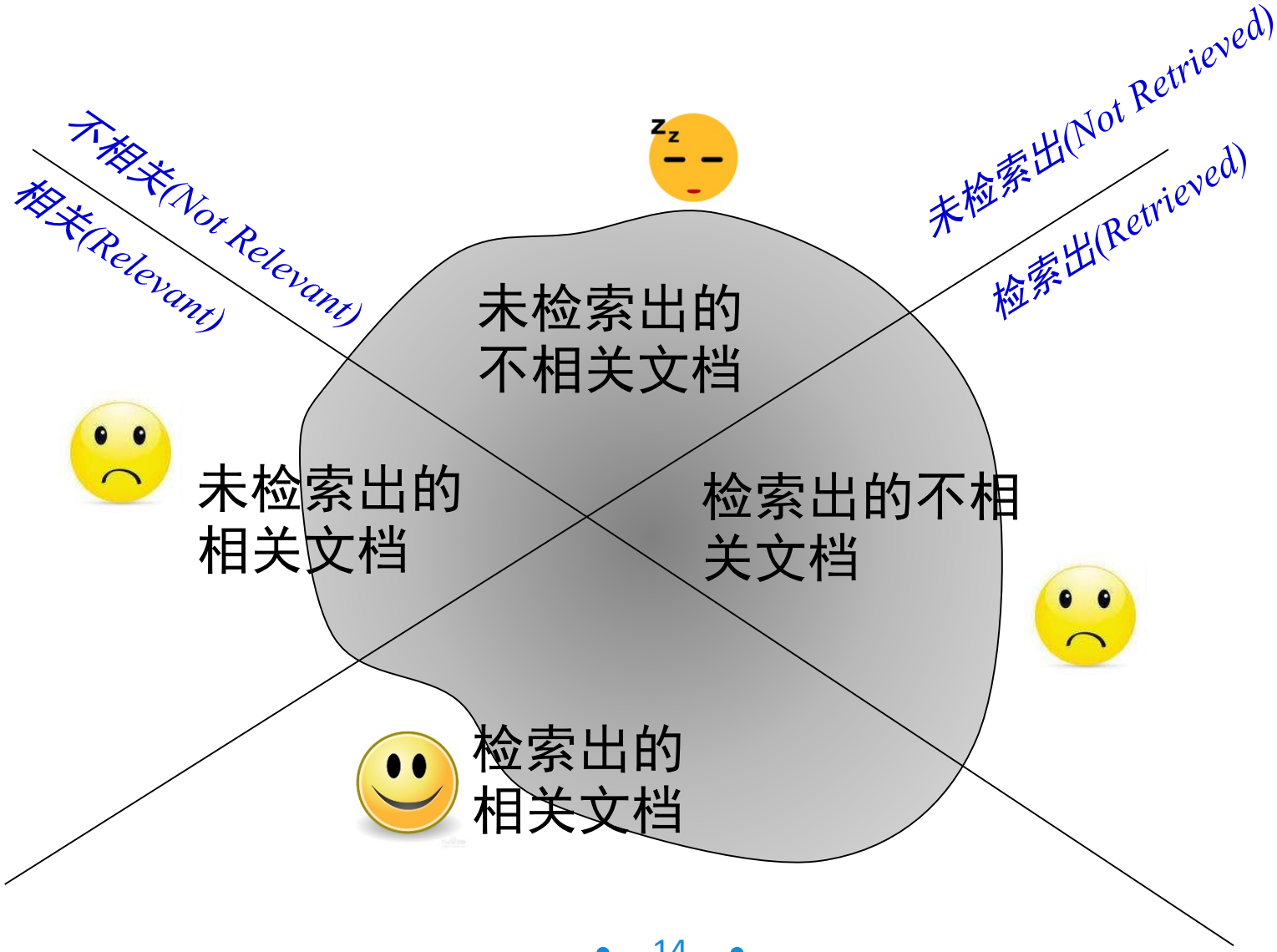
系统&查询	1	2	3	4	…
系统1，查询1	d_3	d_6	d_8	d_{10}	
系统1，查询2	d_1	d_4	d_7	d_{11}	
系统2，查询1	d_6	d_7	d_3	d_9	
系统2，查询2	d_1	d_2	d_4	d_{13}	

缓冲池（Pooling）

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，**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 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会议等多种测试中被广泛采用。

整个文档集合的划分


不相关(Not Relevant)
相关(Relevant)
未检索出(Not Retrieved)
检索出(Retrieved)
未检索出的
不相关文档
未检索出的
相关文档
检索出的不相
关文档
检索出的
相关文档

检索性能评价2*2表

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

基本评价指标

- 准确率（Precison）
- 召回率（Recall）
- 调和指标F、E
- 平均准确率（AP）

查准率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

查准率/准确率（Precision ratio）

——检出的相关信息数量与检出的信息总量的比率

$$P=\frac{\text{检出的相关信息数量}}{\text{检出的信息总量}}=\frac{a}{a+b}$$

查全率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

查全率/召回率（Recall ratio）

——检出的信息数量与检索系统中相关信息总量之间的比率

$$R=\frac{\text{检出的相关信息数量}}{\text{系统中的相关信息数量}}=\frac{a}{a+c}$$

Ps:漏检率与误检率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

$$\text{漏检率（M）} = \frac{\text{未检出的相关文献}}{\text{文档中相关文献总量}} \times 100\% = \frac{c}{a+c} \cdot 100\%$$

$$\text{误检率（N）} = \frac{\text{检出的不相关文献量}}{\text{检出的文献总量}} \times 100\% = \frac{b}{a+b} \cdot 100\%$$

R+M=1， P+N=1

囊括值

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

- 囊括值（Generality）-系统中相关信息数量与系统的信息总量的比率

$$\text{Generality} = \frac{\text{系统中相关信息数量}}{\text{系统的信息总量}} = \frac{a+c}{a+b+c+d}$$

非相关检出率

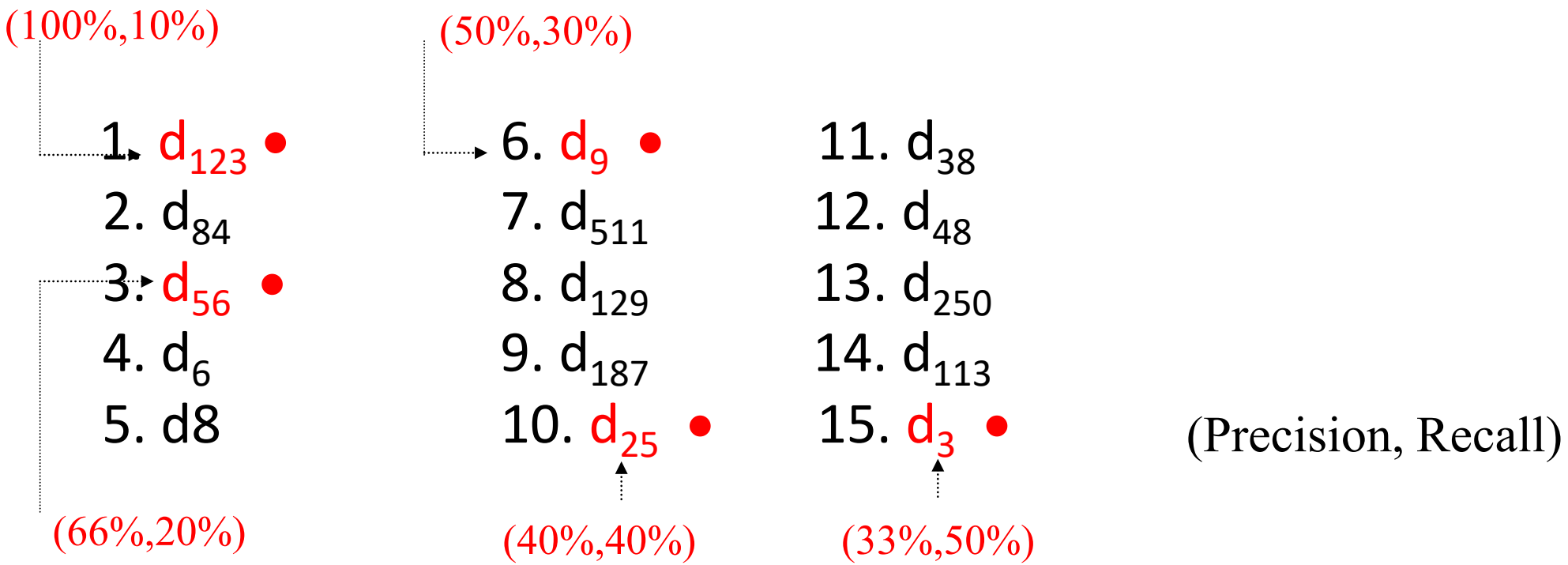
	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

- 非相关检出率（Fallout ratio）-检出的非相关信息数量与系统中的非相关信息总量的比率

$$\text{Fallout}=\frac{\text{检出的非相关信息数量}}{\text{系统中的非相关信息数量}}=\frac{b}{b+d}$$

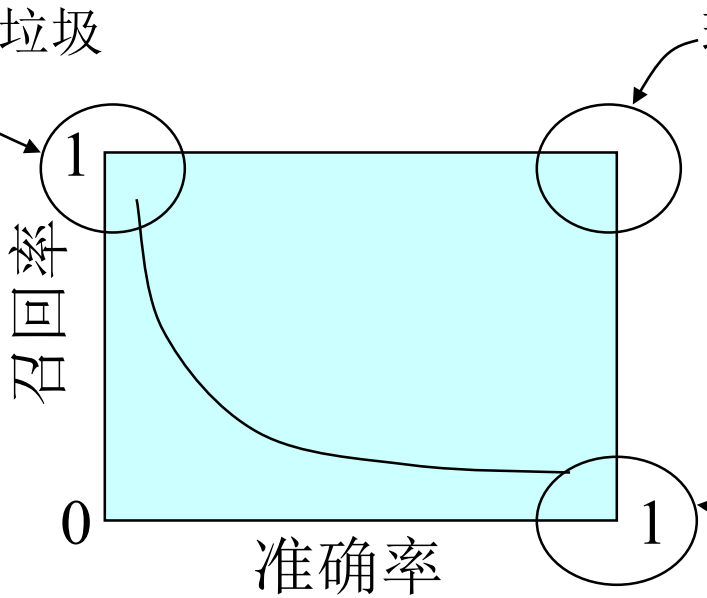
P-R例子

- $R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$
- 通过某一个检索算法得到的**排序结果**：

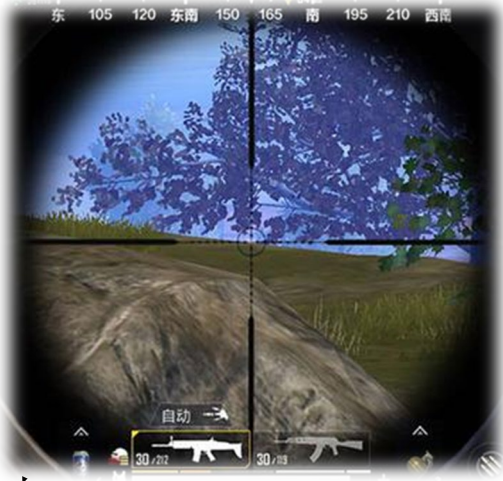


P-R的关系曲线

返回了大多数相关文档
但是包含很多垃圾



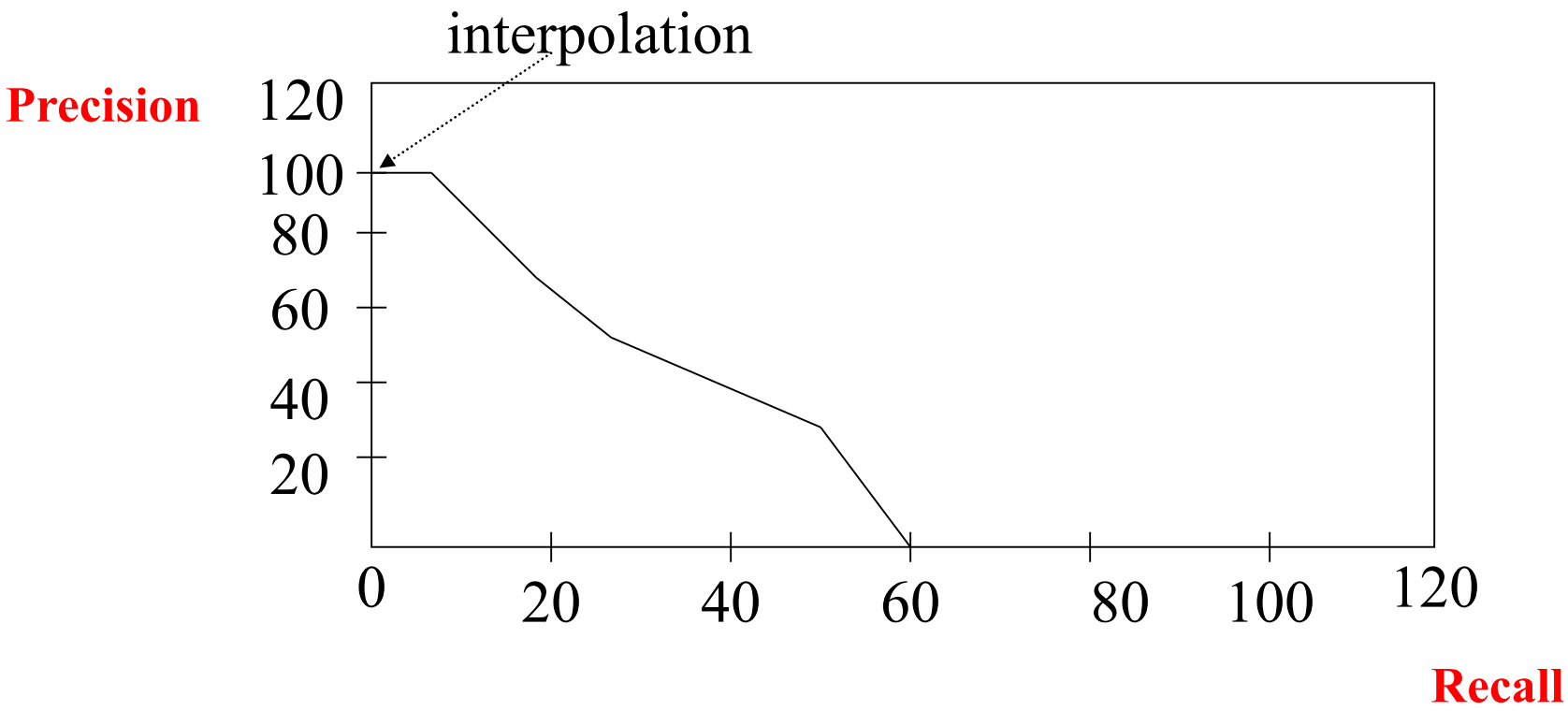
理想情况



返回最相关的文本
但是漏掉了很多
相关文本

11点标准P-R曲线

- 11个标准查全率水平所对应的查准率: 0%, 10%, 20%, …, 100%



平均准确率

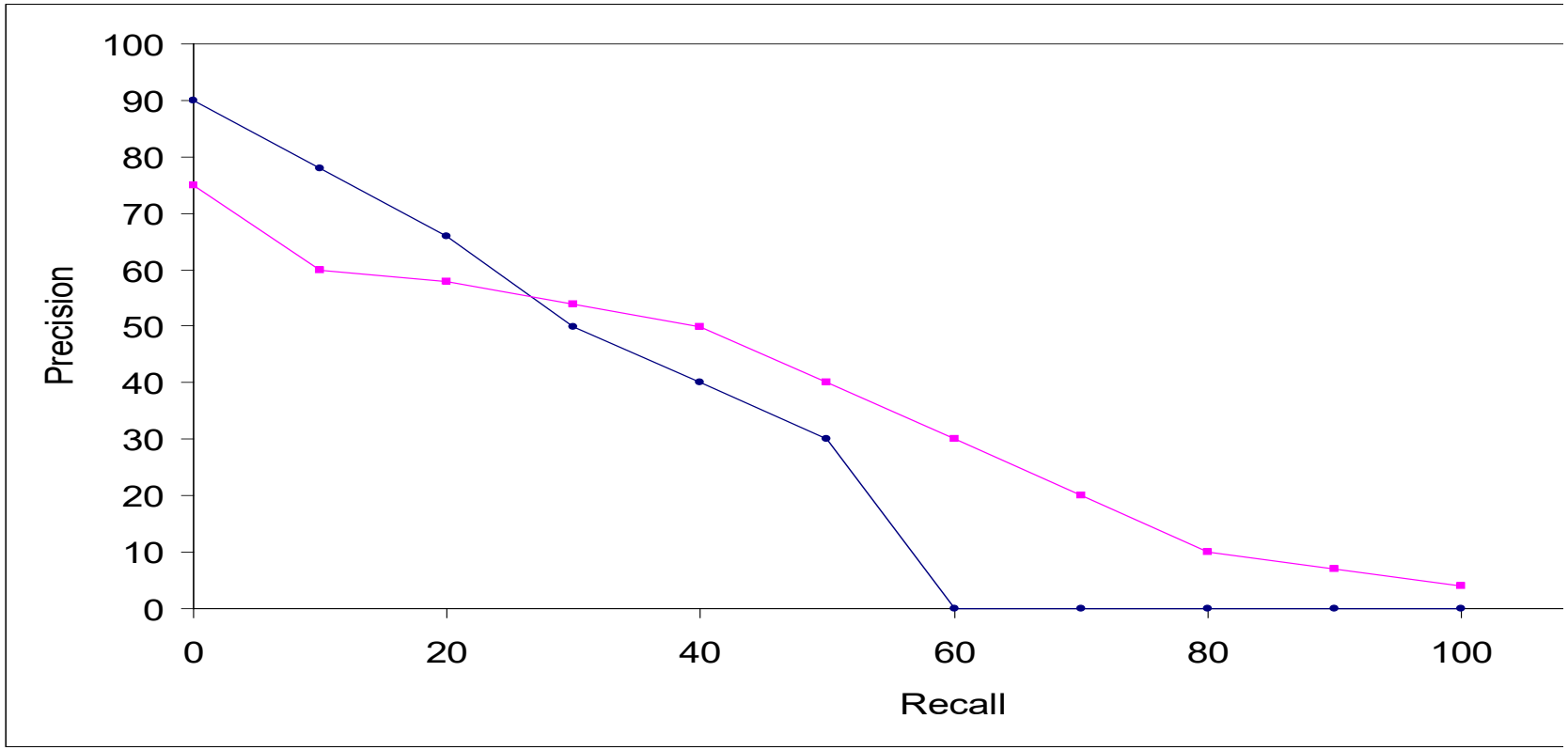
为了评价某一算法对于**所有测试查询**的检索性能，对**每个召回率水平下的准确率**进行平均化处理，公式如下：

$$\overline{P}(r)=\frac{\sum_{i=1}^{N_q}P_i(r)}{N_q}$$

- N_q：——使用的查询总数
- $P_i(r)$ ——在召回率为r时的第i个查询的准确率

比较示例

- 对多个查询，进行平均，有时该曲线也称为：准确率/召回率的值。
- 如下为两个检索算法在多个查询下的准确率/召回率的值。
  - 第一个检索算法在低召回率率下，其准确率较高。
  - 另一个检索算法在高召回率下，其准确率较高



P-R评价的问题

- 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？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，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

- 召回率难以计算

  解决方法：Pooling方法，或者不考虑召回率？

P-R的综合-F值

调和平均值（ Harmonic Mean ）是将准确率和召回率加权平均的评价方法

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

F∈[0,1]

P	R	F
1	1	1.00
0.01	0.01	0.01
0.5	0.5	0.50
0.01	1	0.02
1	0.01	0.02

例子

1. d123
2. d84
3. d56 ●
4. d6
5. d8

6. d9
7. d511
8. d129 ●
9. d187
10. d25

11. d38
12. d48
13. d250
14. d113
15. d3 ●

(33.3%,33.3%)　(25%,66.6%)　(20%,100%)

$$F(3)=\frac{2}{\frac{1}{0.33}+\frac{1}{0.33}}=0.33 \qquad F(8)=\frac{2}{\frac{1}{0.25}+\frac{1}{0.67}}=0.36 \qquad F(15)=\frac{2}{\frac{1}{0.20}+\frac{1}{1}}=0.33$$

P-R的综合-E指数

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

E∈[0,1]

b为用户指定的参数，可以允许用户调整P和R的相对重要程度

- b=1时，E=1-F。这表示E指数和F指数互补
- b>1时，表示准确率P的重要性大于召回率R
- b<1时，表示召回率R的重要性大于准确率P

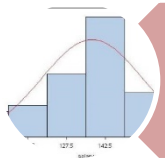
单值评价指标

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MAP

P@10

R准确率



准确率直方图

MAP

- **M**ean **A**verage **P**recision, 平均准确率均值
- 单个查询的平均准确率是逐个考察排序中每个新的相关文档，然后对其准确率值进行平均后的平均值;
- 查询集合的平均准确率是每个查询的平均准确率MAP的平均值，MAP的计算公式如下:

$$MAP = \frac{1}{r}\sum_{i=1}^{r}\frac{i}{\text{第}i\text{个相关文档的位置}}$$

•r 为相关文档数

- MAP是反映系统在**全部查询**上性能的单值指标
- 系统检索出来的相关文档位置越靠前， MAP就可能越高.
- 如果系统没有返回相关文档，则MAP默认为0.

计算MAP举例

$$MAP = \frac{1}{r}\sum_{i=1}^{r}\frac{i}{\text{第}i\text{个相关文档的位置}}$$

- $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

AP=(1+0.66+0.5+0.4+0.33)/5=0.578

p@10

p@10——系统对于查询返回的**前10个结果的准确率**.

- 对于搜索引擎系统来讲，由于没有一个搜索引擎系统能够保证搜集到所有的网页，所以召回率很难计算，因而**准确率成为目前的搜索引擎系统主要关心的指标**
- 考虑到用户在查看搜索引擎结果时，往往希望在第一个页面（如10个结果）就找到自己所需的信息，因此P@10能比较真实有效地反映在真实应用环境下所表现的性能

p@10计算

- $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

0.4

R-Precision

- 单个查询的*R*准确率是指检索出**R篇文档时的准确率**.
- R是当前检索中**相关的文档**总数
- 查询集合中所有查询的R准确率是每个查询的R准确率的平均值.

$$R-\Pr ecision=\frac{\text{前}R\text{篇文档中实际相关文档数}}{R}$$

R-Precision计算

$$R-\mathrm{Precision}=\frac{\text{前}R\text{篇文档中实际相关文档数}}{R}$$

- $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

10-precision=4/10=0.4

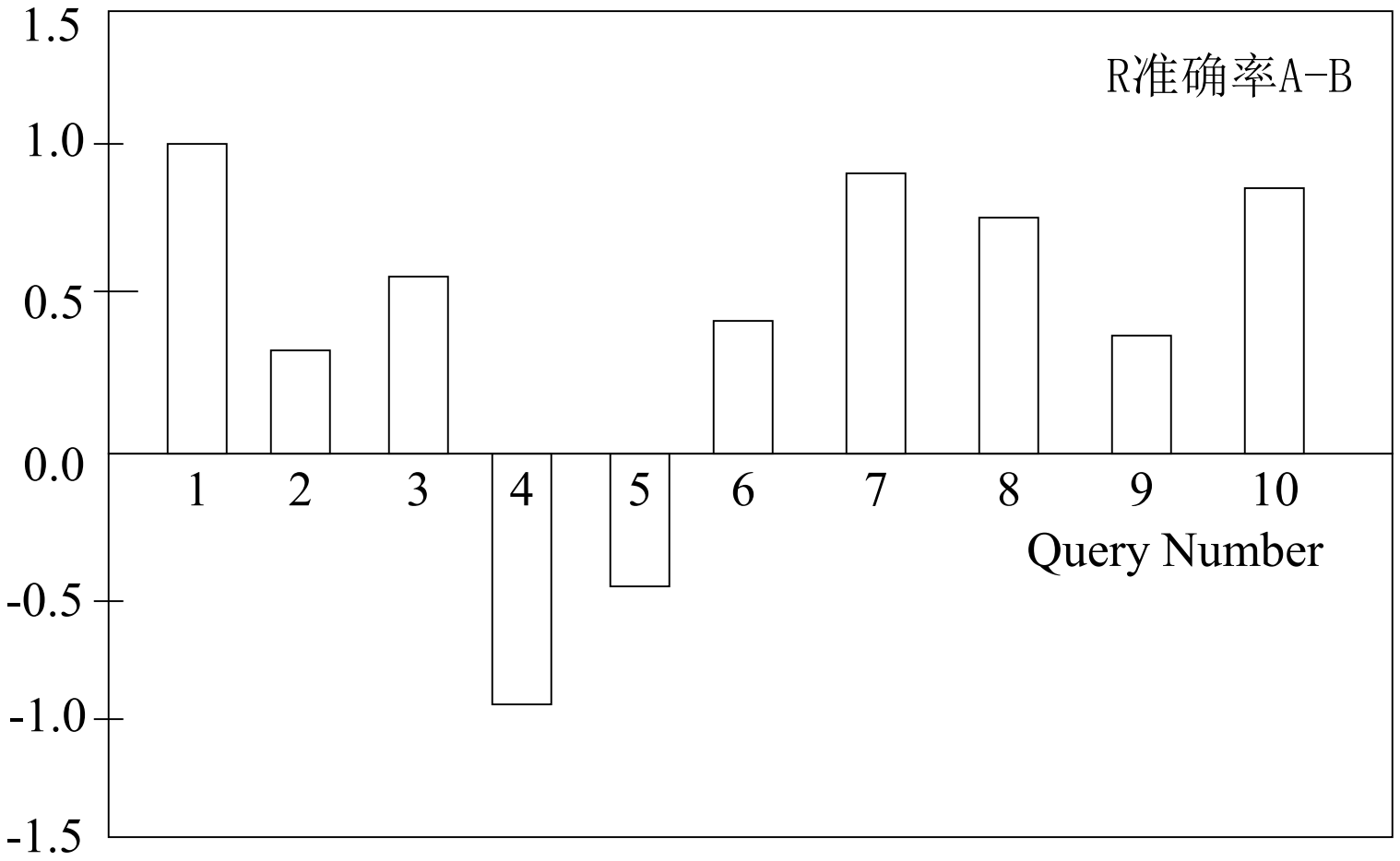
准确率直方图

- 多个查询的**R-Precision**测度
- **用来比较两个算法的检索纪录**
- 用$RP_A(i)$和$RP_B(i)$分别表示使用检索算法A和检索算法B检索第i个查询时得到的R准确率,它们之间的差值

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

- $RP_{A-B}=0$: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

准确率直方图：例



单指标评价小结

- 随着信息技术以及互联网的发展，信息检索研究所采用的数据集越来越大，因此构建完整的相关判断越来越难；
- 在相关判断不完整的情况下，采用现有评价方法得出的测试结果会有失公正；
- 对于搜索引擎这样的对高相关性文档进行检索的任务来讲，传统的评价方法也无法很好地对任务评测；
- 特殊指标： ~~*Bpref*、*N(D)CG*、*单一相关文档检索的评价*~~

面向用户的评价

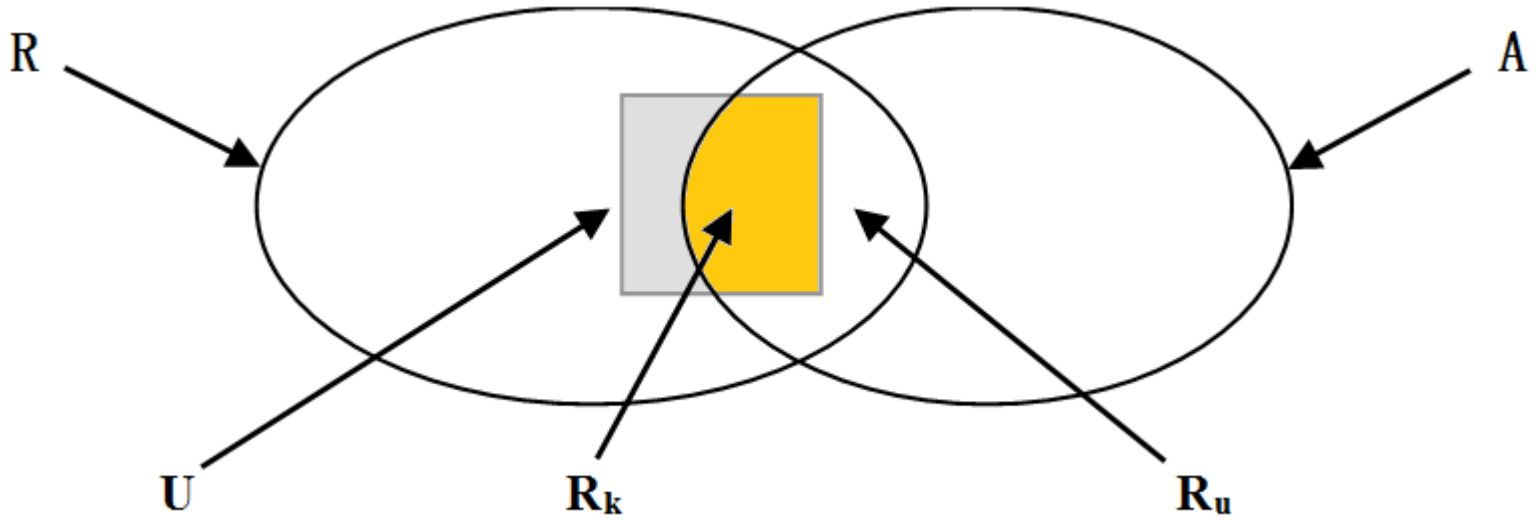
- 面向用户的测度方法/User-Oriented Measures
  - 覆盖率：实际检出的相关文献中用户一致的相关文献所占比例

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

  - 新颖率：检出的相关文献中用户未知的相关文献所占的比例

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

覆盖率和新颖率



R ——相关文档集

A ——返回文档集

U ——用户的相关文档集

R_k ——返回的、用户已知的文档集

R_u ——返回的，用户未知的文档集

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

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

统一评测

- 同一个算法在不同的数据条件下得到的结果差异很大；
- 没有统一的测试方法和共同的数据集合，几乎不可能比较不同算法；
- 数据采集需花费很大的人力物力，而由政府学术机构或者学术团体组织的开放技术评测，可以为科研提供一种统一的、普遍认可的评价基准和大型测试集，节省了各个研究者重复采集数据而造成的重复付出，对整个领域的科学研究和技术进步起到很大的推动作用；
- 通过技术评测可以提出新的研究问题

国外的评测1

- **The Cranfield Experiments**, by *Cyril W. Cleverdon*

  1957 –1968 （上百篇文档集合）

  http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

- **SMART System**, by *Gerald Salton*

  1964-1988 （数千篇文档集合）

Gerald Salton, 1927-1995

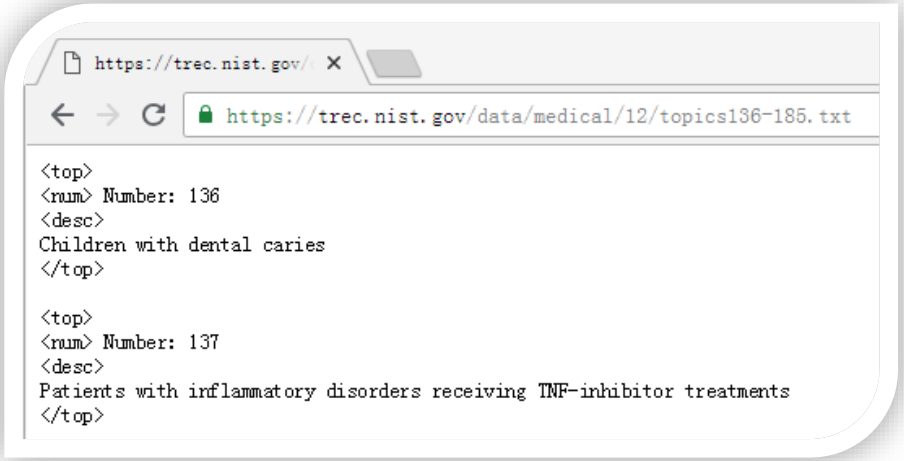
国外的评测II

- **TREC评测**
  - 文本检索会议（Text Retrieval Conference，TREC）是信息检索(IR) 界为进行检索系统和用户评价而举行的活动，它由美国国家标准技术协会(NIST) 和美国高级研究计划局（DARPA） 共同资助，始于1992年。
  - 检索评测中的奥运会！！
- **NTCIR评测**
  - NTCIR(NACSIS Test Collection for IR Systems)始于1998年，是由日本国立信息学研究所（National Institute of Informatics，简称NII）主办的搜索引擎评价型国际会议
- **CLEF评测**
  - CLEF于2000年开始筹办，是欧洲各国共同合作进行的一项长期研究计划，主要想通过评测信息科技技术，促进欧洲语言中的各种单一语言以及多语言信息技术的发展，
  - CLEF的目标只在于跨语言信息检索以及多语言信息检索方面

TREC评测

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  - 1992年开始，每年一次
  - 由美国国防部Defense Advanced Research Projects Agency （DARPA）和美国国家标准技术研究所National Institute of Standards and Technology （NIST）联合发起
  - 参加者免费获得标准训练和开发数据
  - 参加者在参加比赛时收到最新的测试数据，并在限定时间内作出答案，返给组织者
  - 组织者对各参赛者的结果进行评价
  - 包括检索、过滤、问答等多个主题

TREC目的



- 促进基于大规模测试文档集的检索研究。
- 为了反映现实系统的主题多样性，必须保证有足够的实验语料集，TREC的文献集合一般在G级左右，包括50~100万篇文献（近几年更大，可达数千万记录，T级存储）；
- 建立一个开放的论坛来交流研究思想，使与会者能交流研究的成果与心得，促进企业学术机构和政府部门之间的交流沟通。
- 通过展示检索方法在解决实际问题中的有效性，来加速实验室技术的商业化产品转换。
- 通过提供大型的语料库、统一的测试程序，有系统地整理评测结果，达到改善文本检索评价和检验方法的目标。

TREC早期任务

- Ad hoc检索任务（传统的批处理检索）
- 类似图书馆里的书籍检索，即书籍库（数据库、文档集合）相对稳定不变，而用户的查询要求是千变万化的
- 主要研究任务包括对大数据库的索引查询、查询的扩展等
- 固定主题检索任务（Information Routing）
- 用户的查询要求相对稳定，而文档集常常发生变化
- 研究的主要任务不是索引，而是对**用户兴趣的建模**，即如何为用户兴趣建立合适的数学模型

TREC评测的评价方法

- 概括表统计
- 准确率-召回率平均值
- 文档级别平均值
  - 平均准确率

发布Track → 报名 → 用户测试与提交 → 评估 → 交流

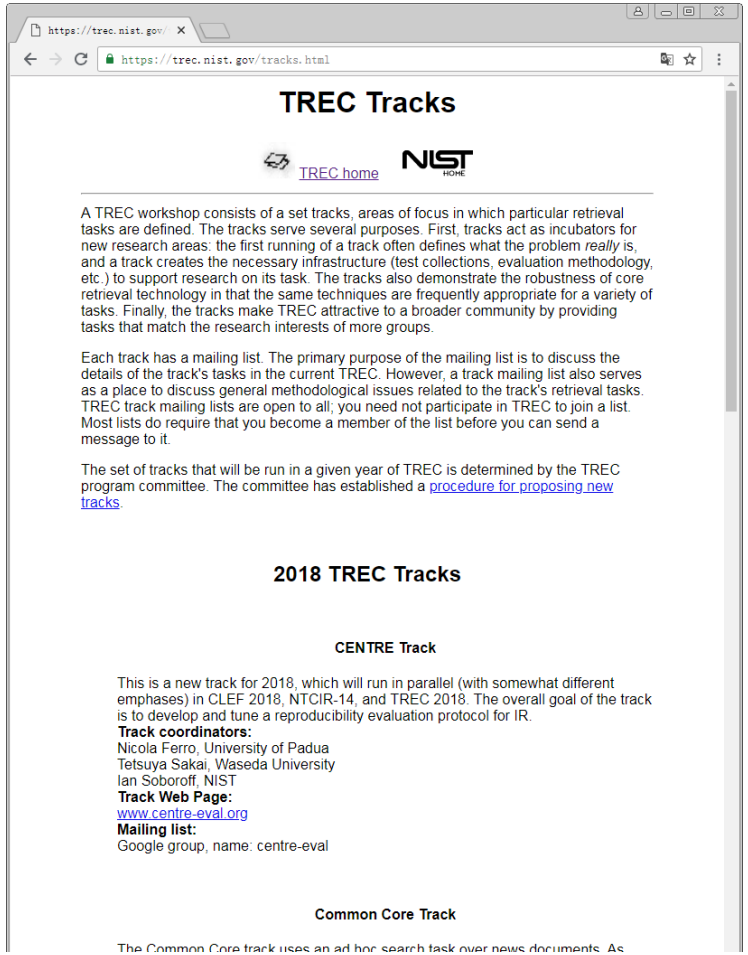
参加过TREC的部分单位

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评测的任务（Tracks）

https://trec.nist.gov/tracks.html

- 2018 TREC Tracks
  - CENTRE Track
  - Common Core Track
  - Complex Answer Retrieval Track
  - Incident Streams Track
  - News Track
  - Precision Medicine Track
  - Real-Time Summarization Track



https://trec.nist.gov/tracks.html

TREC Tracks

TREC home NIST HOME

A TREC workshop consists of a set tracks, areas of focus in which particular retrieval tasks are defined. The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups.

Each track has a mailing list. The primary purpose of the mailing list is to discuss the details of the track's tasks in the current TREC. However, a track mailing list also serves as a place to discuss general methodological issues related to the track's retrieval tasks. TREC track mailing lists are open to all; you need not participate in TREC to join a list. Most lists do require that you become a member of the list before you can send a message to it.

The set of tracks that will be run in a given year of TREC is determined by the TREC program committee. The committee has established a procedure for proposing new tracks.

2018 TREC Tracks

CENTRE Track

This is a new track for 2018, which will run in parallel (with somewhat different emphases) in CLEF 2018, NTCIR-14, and TREC 2018. The overall goal of the track is to develop and tune a reproducibility evaluation protocol for IR.
Track coordinators:
Nicola Ferro, University of Padua
Tetsuya Sakai, Waseda University
Ian Soboroff, NIST
Track Web Page:
www.centre-eval.org
Mailing list:
Google group, name: centre-eval

Common Core Track

The Common Core track uses an ad hoc search task over news documents. As

51

NTCIR

http://research.nii.ac.jp/ntcir/index-en.html

- NTCIR评测

  NTCIR（NACSIS Test Collection for IR Systems）始于1998年，是由日本国立信息学研究所（National Institute of Informatics，简称NII）主办的搜索引擎评价型国际会议

- 主要评测任务
  - ✓ 传统的日文、中文、韩文、英文的单语ad hoc任务.
  - ✓ 最重要的任务是跨语言信息检索。若以C、J、K、E分别代表中文、日文、韩文、英文，则有C→CJKE、J→CJKE、K→CJKE、E→CJKE等极为复杂的检索任务。
  - ✓ 另外一个比较重要的任务是中枢语言信息检索，这个任务是模拟在语言资源不足的情况下进行跨语言信息检索。

    如要进行C→K的跨语言信息检索，但是没有中韩双语词典，只好借用中英词典以及英韩词典，此时，英语就被视为中枢语言。

CLEF

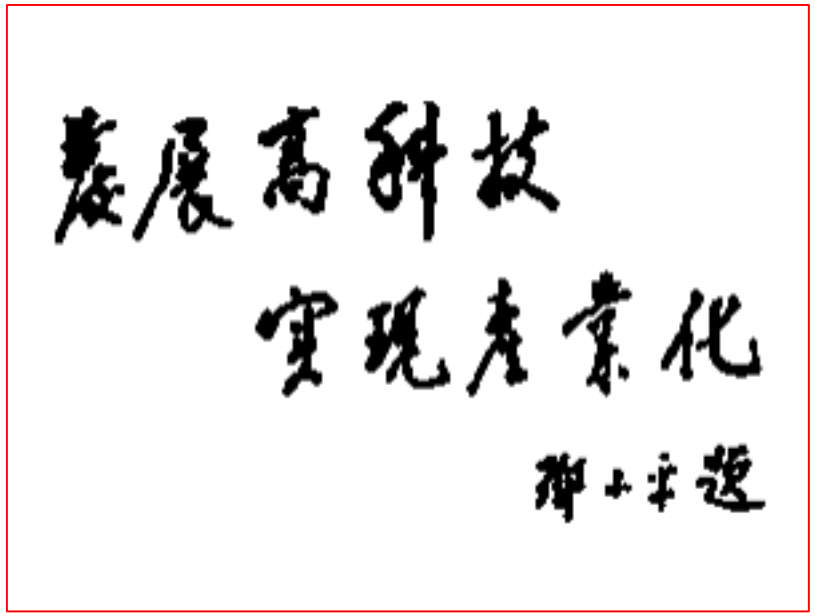
http://clef.isti.cnr.it

http://www.clef-initiative.eu/

- CLEF（Cross-Language Evaluation Forum）评测
  - CLEF（2000-2009）是欧洲各国共同合作进行的一项长期研究计划，主要想通过评测信息科技技术，促进欧洲语言中的各种单一语言以及多语言信息技术的发展。
  - CLEF的目标只在于跨语言信息检索以及多语言信息检索方面
- CLEF的评测任务
  - 跨语言文本检索：包括三个子任务，即单语检索、双语检索以及多语检索。
  - 跨语言专利数据检索：主要是使用专业领域上下文的信息进行单语言以及跨语言的信息检索。
  - 交互式跨语言检索（Interactive Cross-Language Retrieval (iCLEF) ）:尝试模拟实际检索环境下使用者与检索系统的互动情形，以改善信息检索系统的性能。
  - 多语问答:是一种跨语言QA检索评测
  - 图像跨语言检索/跨语言语间检索

国内863评测介绍

- 全名
  - 863计划中文信息处理与智能人机接口技术评测（1991-2005）
- 组织者
  - 国家高技术研究发展计划（863计划）
- 方式
  - 通过网络进行
  - 各单位在自己的环境中运行参评系统
  - 2005年11月召开研讨会
- 2005年度评测内容
  - 机器翻译
  - **信息检索**
  - 语音识别



863评测介绍—信息检索评测

- 项目：相关网页检索
- 任务定义：给定主题，返回数据中与该主题相关的网页。
- 数据：CWT100g (中文Web测试集100g)
  - 根据天网搜索引擎截止**2004**年**2**月**1**日发现的中国范围内提供**Web**服务的**1,000,614**个主机，从中采样**17,683**个站点，在**2004**年**6**月搜集获得**5,712,710**个网页（有效网页：**5,594,521**）
  - 包括网页内容和**Web**服务器返回的信息
  - 真实容量为**90GB**。

主题

- **主题**（Topic）模拟了用户需求，由若干字段组成，描述了用户所希望检索的信息。主题和查询的区别在于：主题是对信息需求的陈述，查询则是信息检索系统的实际输入。
- 主题由4个字段组成：
  - **编号**（num）
  - **标题**（title）
  - **描述**（desc）
  - **叙述**（narr）

主题实例

- <top>
- <num>编号：020
- <title> 下载"香奈儿"
- <desc> 描述：mp3格式歌曲“香奈儿”的下载地址
- <narr> 叙述：仅检索具有歌曲“香奈儿”下载地址的网页。有关“香奈儿”的介绍不在检索范围内。提供非mp3格式下载地址的页面不在检索之列。
- </top>

查询的构造

- **自动方式和人工方式**
- 自动方式是指在没有任何人为因素的影响下根据主题构造查询的方式
- 除此之外的方式均为人工方式
  - 只允许以人工方式构造查询，不允许在检索过程中加入任何人为因素。
  - 最多返回1000条排序结果

	MAP	R-Precision	P@10
第一名	自动化所0.3175	哈工大0.3672	清华0.6280
第二名	哈工大 0.3107	自动化所0.3607	哈工大0.6240
第三名	清华大学0.2858	清华0.3293	自动化所0.5540

重要会议/小组

SIGIR Special Interest Group on Information Retrieval

http://sigir.org/

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

https://trec.nist.gov/

https://www.ccf.org.cn/

http://www.cipsc.org.cn/

小结