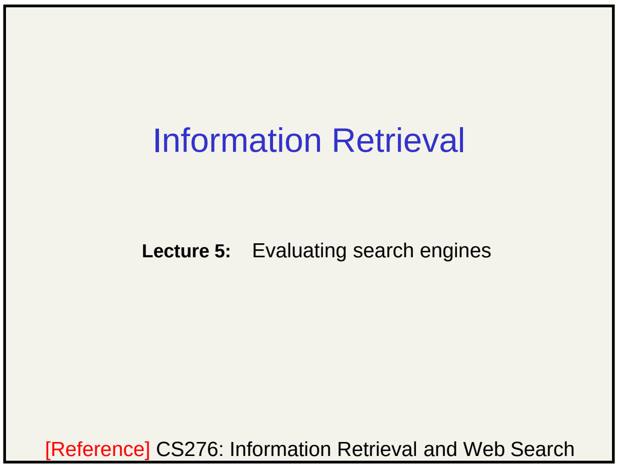
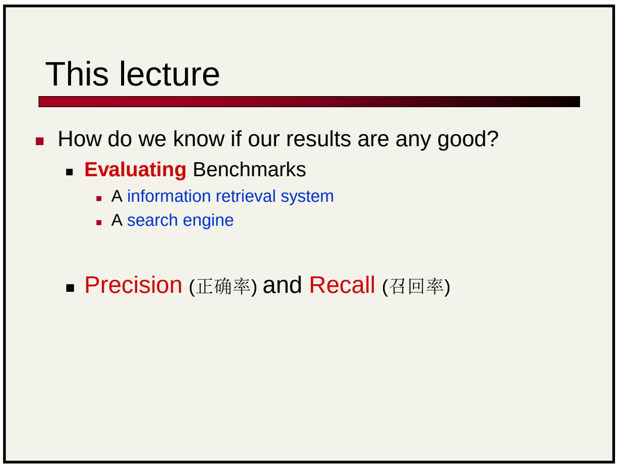


Information Retrieval

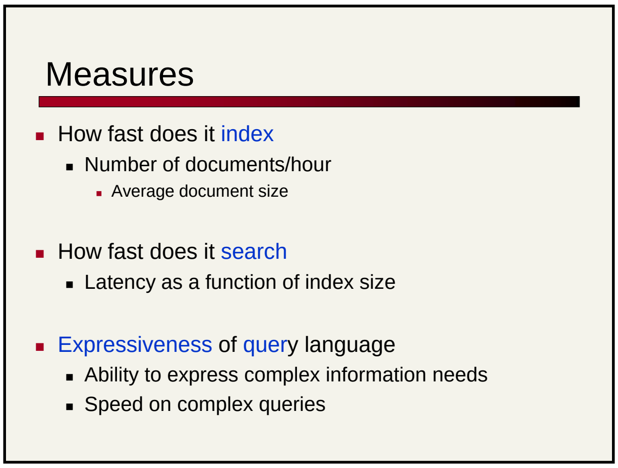
Lecture 5: Evaluating search engines

[Reference] CS276: Information Retrieval and Web Search



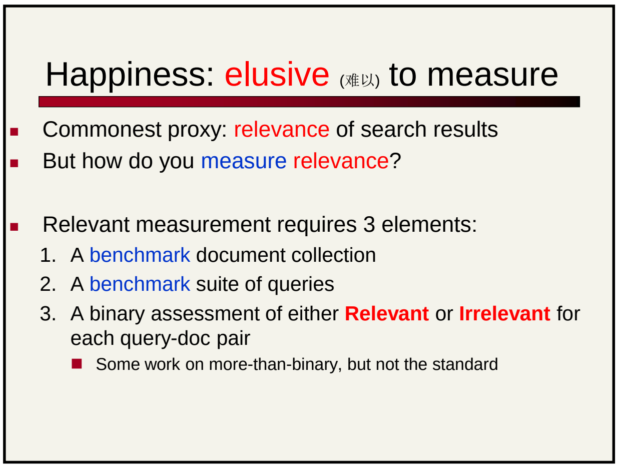
This lecture

- How do we know if our results are any good?
  - **Evaluating** Benchmarks
    - A information retrieval system
    - A search engine
  - Precision (正确率) and Recall (召回率)



Measures

- How fast does it index
  - Number of documents/hour
    - Average document size
- How fast does it search
  - Latency as a function of index size
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries



Happiness: elusive (难以) to measure

- Commonest proxy: relevance of search results
- But how do you measure relevance?
- Relevant measurement requires 3 elements:

1. A benchmark document collection
2. A benchmark suite of queries
3. A binary assessment of either **Relevant** or **Irrelevant** for each query-doc pair
   - Some work on more-than-binary, but not the standard

Standard relevance benchmarks

- **TREC**
  - Text Retrieval Conference
  - National Institute of Standards and Testing (NIST) has run a large IR test bed for many years
    - Since 1992
  - TREC Ad Hoc
    - The first 8 TREC evaluations between 1992 to 1999
    - 6 CDs
    - 1.89 million documents (189万篇文档)
    - 450 information needs
      - Topics and specified in detailed text passages

Evaluating an IR system

- Note: the **information need** is translated into a **query**
- Relevance is assessed relative to the **information need** **not** the **query**
- E.g.,
  - **Information need** : *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
  - **Query** : ***wine red white heart attack effective***
- Evaluate : whether the doc addresses the information need, not whether it has those words

Accuracy

- Given a query an engine classifies each doc
  - "Relevant" or "Irrelevant"
- Accuracy of an engine :
  - the fraction of these classifications that is correct
- Why is this not a very useful evaluation measure in IR?

Unranked retrieval evaluation: Precision and Recall

- **Precision** : fraction of retrieved docs
  - relevant = P(relevant|retrieved)
- **Recall** : fraction of relevant docs
  - retrieved = P(retrieved|relevant)

	Relevant	Not Relevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Precision/Recall

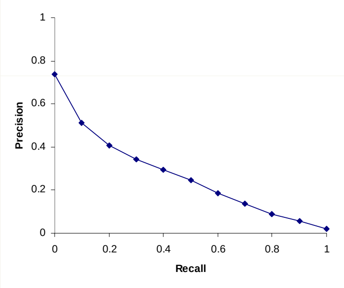
- You can get high recall (but low precision) by retrieving all docs for all queries
- Recall is a non-decreasing function of the number of docs retrieved
- **In a good system**
  - precision decreases as either number of docs retrieved or recall increases
  - A fact with strong empirical confirmation

Evaluation ranked results

- Graphs are good, but people want summary measures
  - Precision at fixed retrieval level
    - Perhaps most appropriate for web search: all people want are good matches on the first one or two results pages
    - But has an arbitrary parameter of ***k***
  - **11-point** interpolated average precision
    - The standard measure in the TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (the value for 0 is always interpolated!), and average them
    - Evaluates performance at all recall levels

Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)

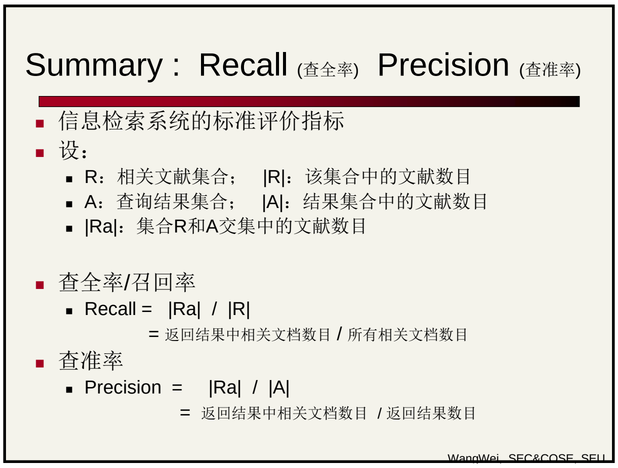


Example : a query *q*

- Assume *Rq* = $\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$
- Ranking the documents in the answer set A
  - $\mathbf{d_{123}}$ d_{84} $\mathbf{d_{56}}$ d_6 d_8 $\mathbf{d_9}$ d_{511} d_{129} d_{187} $\mathbf{d_{25}}$ d_{38} d_{48} d_{250} d_{113} $\mathbf{d_3}$

Document	Precision	Recall
d_{123} :	P =1/1 =100%,	R = 1/10 =10%
d_{56} :	P = 2/3 =66%,	R = 2/10 =20%
d_9 :	P = 3/6 =50%,	R = 3/10 =30%
d_{25} :	…	…
d_3 :	…	…

…………………

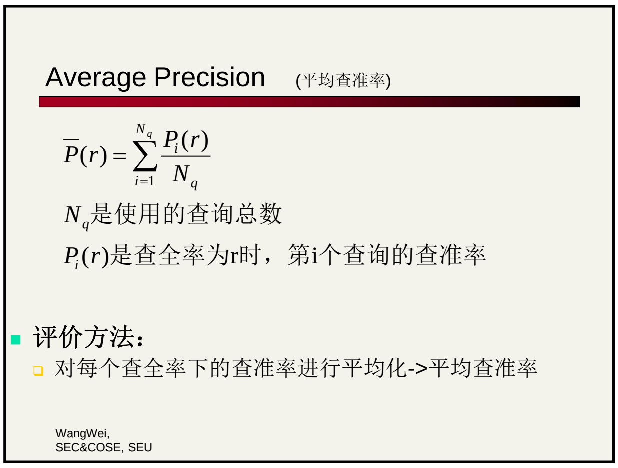


Summary : Recall (查全率) Precision (查准率)

- 信息检索系统的标准评价指标
- 设：
  - R：相关文献集合； |R|：该集合中的文献数目
  - A：查询结果集合； |A|：结果集合中的文献数目
  - |Ra|：集合R和A交集中的文献数目
- 查全率/召回率
  - Recall = |Ra| / |R|

    = 返回结果中相关文档数目 / 所有相关文档数目
- 查准率
  - Precision = |Ra| / |A|

    = 返回结果中相关文档数目 / 返回结果数目



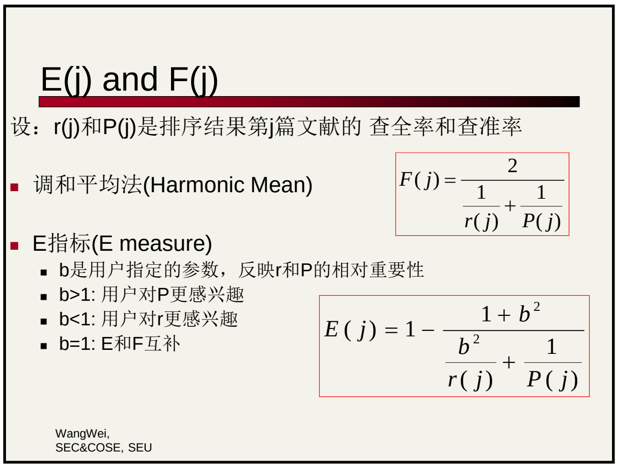
Average Precision (平均查准率)

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

N_q是使用的查询总数

$P_i(r)$是查全率为r时，第i个查询的查准率

- 评价方法：
  - 对每个查全率下的查准率进行平均化->平均查准率



E(j) and F(j)

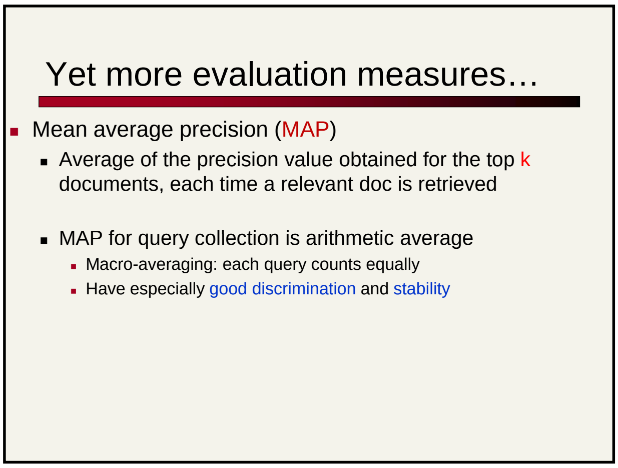
设：r(j)和P(j)是排序结果第j篇文献的 查全率和查准率

- 调和平均法(Harmonic Mean)

$$F(j)=\frac{2}{\frac{1}{r(j)}+\frac{1}{P(j)}}$$

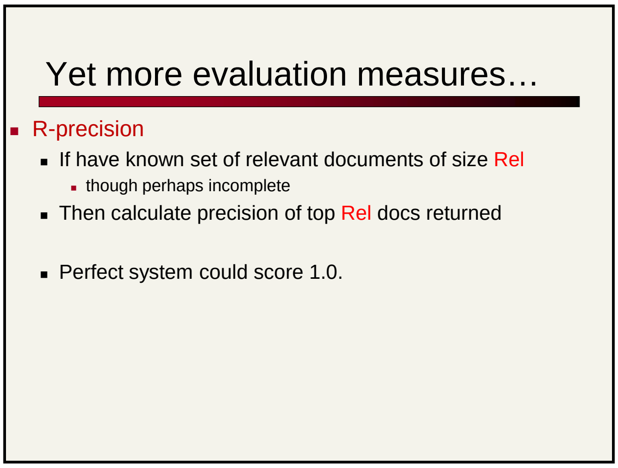
- E指标(E measure)
  - b是用户指定的参数，反映r和P的相对重要性
  - b>1: 用户对P更感兴趣
  - b<1: 用户对r更感兴趣
  - b=1: E和F互补

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



Yet more evaluation measures…

- Mean average precision (MAP)
  - Average of the precision value obtained for the top k documents, each time a relevant doc is retrieved
  - MAP for query collection is arithmetic average
    - Macro-averaging: each query counts equally
    - Have especially good discrimination and stability



Yet more evaluation measures…

- R-precision
  - If have known set of relevant documents of size Rel
    - though perhaps incomplete
  - Then calculate precision of top Rel docs returned
  - Perfect system could score 1.0.