Evaluation

Evaluation

- The major goal of IR is to search document relevant to a user query.
- The evaluation of the performance of IR systems relies on the notion of relevance.

What constitute relevance ?

Relevance

- Relevance is subjective in nature i.e. it depends upon a specific user's judgment.
- Given a query, the same document may be judged as relevant by one user and non-relevant by another user. Only the user can tell the true relevance.
- however not possible to measure this "true relevance"
- Most of the evaluation of IR systems so far has been done on document test collections with known relevance judgments.

- Another issue with relevance is the degree of relevance.
- Traditionally, relevance has been visualized as a binary concept i.e. a document is judged either as relevant or not relevant whereas relevance is a continuous function (a document may exactly what the user want or it may be closely related)

Why System Evaluation?

- There are many retrieval models/ algorithms/ systems, which one is the best?
- What is the best component for:
  - Ranking function (dot-product, cosine, …)
  - Term selection (stop word removal, stemming…)
  - Term weighting (TF, TF-IDF,…)
- How far down the ranked list will a user need to look to find some/all relevant documents?

Evaluation of IR Systems

- The IR evaluation models can be broadly classified as **system driven** model and **user-centered** model.
- System driven model focus on measuring how well the system can rank documents
- user–centered evaluation model attempt to measure the user's satisfaction with the system.

Evaluation of IR Systems

- The evaluation of IR system is the process of assessing how well a system meets the information needs of its users (Voorhees, 2001).
- Criteria's for evaluation

    -Coverage of the collection

    -Time lag

    - Presentation format

    - User effort

    - Precision

    - Recall

Evaluation Criteria : Precision and Recall

- Of these criteria, recall and precision have most frequently been applied in measuring information retrieval.
- Both these criteria are related with the effectiveness aspect of IR system i.e. its ability to retrieve relevant documents in response to user query.
- These measures are based on relevance judgments.

Effectiveness : User Centred

- Effectiveness is purely a measure of the ability of the system to satisfy user in terms of the relevance of documents retrieved
- Aspects of effectiveness include:

  - whether the documents being returned are relevant to the user

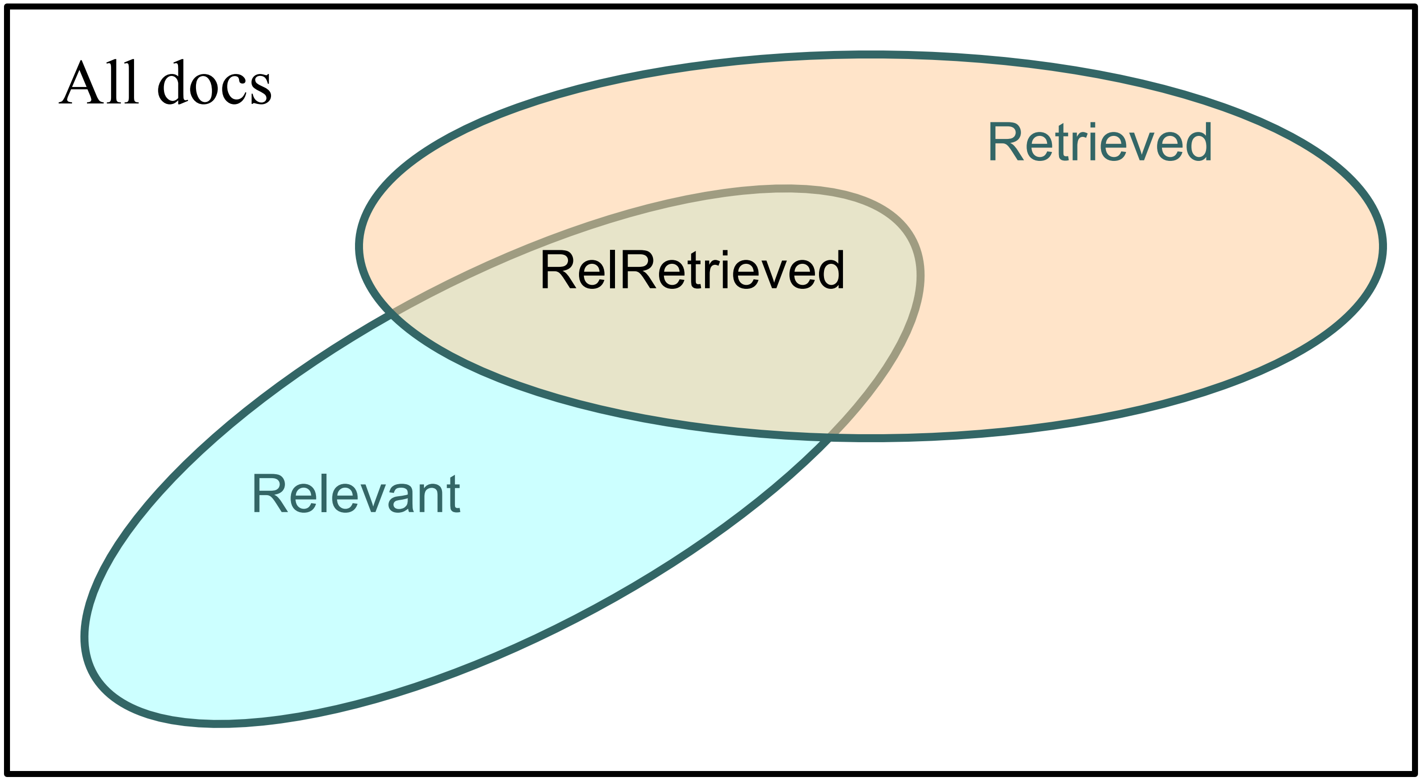
  -whether they are presented in the order of relevance

  - whether a significant number of relevant documents in the collection are being returned to the user etc

Evaluation of IR Systems

- Traditional goal of IR is to retrieve *all* and *only* the relevant documents in response to a query
- ‘All’ is measured by *recall*: the proportion of relevant documents in the collection which are retrieved i.e. P(retrieved|relevant)
- ‘Only’ is measured by *precision*: the proportion of retrieved documents which are relevant

Precision vs. Recall



$$\text{Precision} = \frac{|\text{RelRetrieved}|}{|\text{Retrieved}|} \qquad \text{Recall} = \frac{|\text{RelRetrieved}|}{|\text{Relevant}|}$$

Precision and Recall

- These definitions of precision and recall are based on binary relevance judgment, which means that every retrievable item is recognizably “relevant”, or recognizably “not relevant”.
- Hence, for every search result all retrievable documents will be either

  (i) relevant or non-relevant and

  (ii) retrieved or not retrieved.

Precision and Recall

	Relevant	Non Relevant	
Retrieved	$A \cap B$	$\bar{A} \cap B$	B
Not Retrieved	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	

$$\text{Precision} = \frac{|A \cap B|}{|B|} \qquad \text{Recall} = \frac{|A \cap B|}{|A|}$$

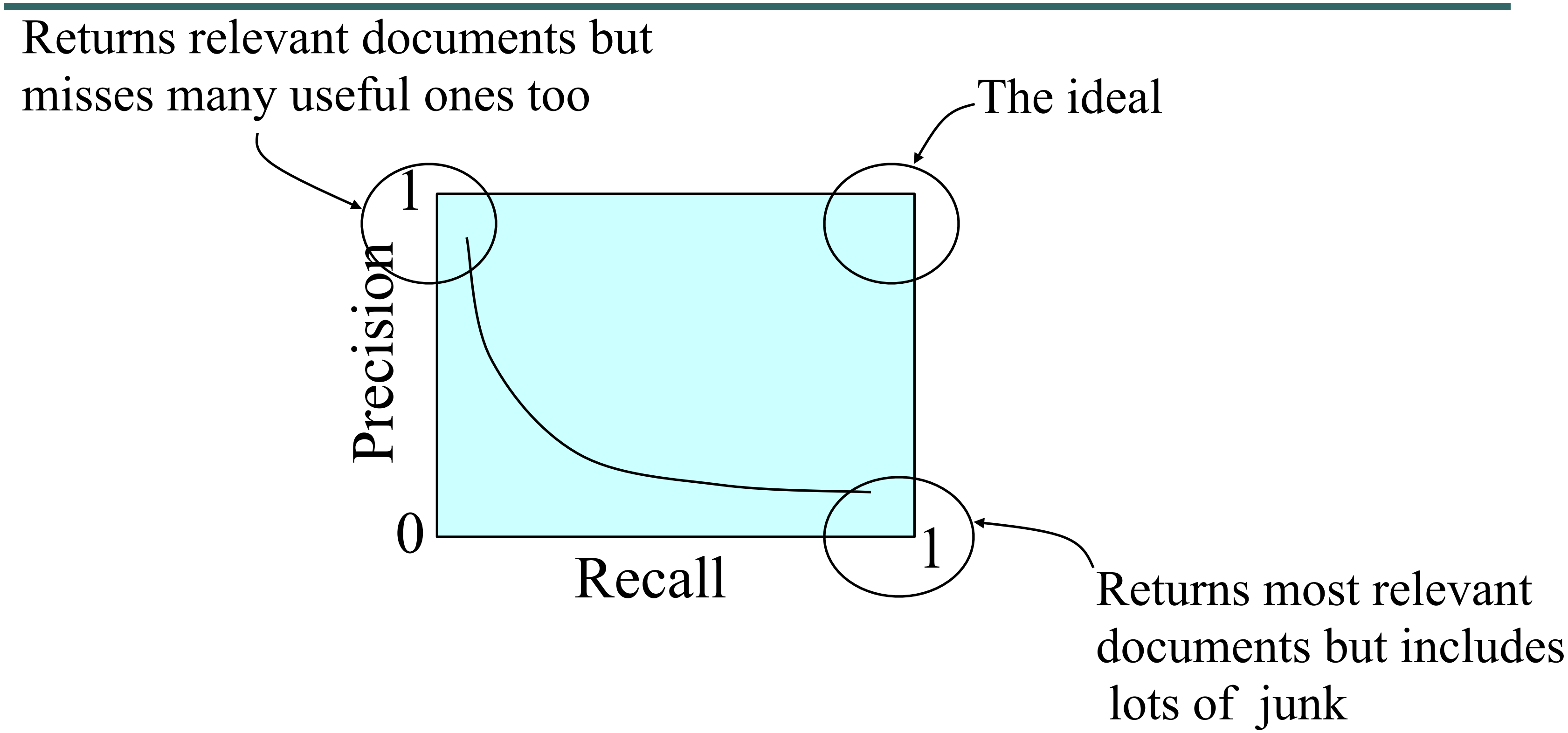
where, A is set of relevant docments,

$|A|$ = No. of relevant documents in the collection(NR_{rel})

B is set of retrieved documents

and $|B|$ = No. of retrieved documents(NR_{ret})

Trade-off between Recall and Precision



Computing Precision and Recall Test collection approach

- The total number of relevant documents in a collection must be known in order for recall to be calculated.
- To provide a framework of evaluation of IR systems, a number of test collections have been developed (Cranfield, TREC etc.).
- These document collections are accompanied by a set of queries and relevance judgments.

IR test collections

Collection	Number of documents	Number of queries
Cranfield	1400	225
CACM	3204	64
CISI	1460	112
LISA	6004	35
TIME	423	83
ADI	82	35
MEDLINE	1033	30
TREC-1	742,611	100

Fixed Recall Levels

- One way to evaluate is to look at average precision at fixed recall levels
  - Provides the information needed for precision/recall graphs

Document Cutoff Levels

- Another way to evaluate:
  - Fix the number of documents retrieved at several levels:
    - top 5
    - top 10
    - top 20
    - top 50
    - top 100
  - Measure precision at each of these levels
  - Take (weighted) average over results
- focuses on how well the system ranks the first k documents.

Computing Recall/Precision Points

- For a given query, produce the ranked list of retrievals.
- Mark each document in the ranked list that is relevant according to the gold standard.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.

Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document. Never reach 100% recall

Interpolating a Recall/Precision Curve

- Interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0$, $r_1 = 0.1$, …, $r_{10}=1.0$
- The interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level greater than or equal to *j*-th level.

Example: Interpolated Precision

Precision at observed recall points:

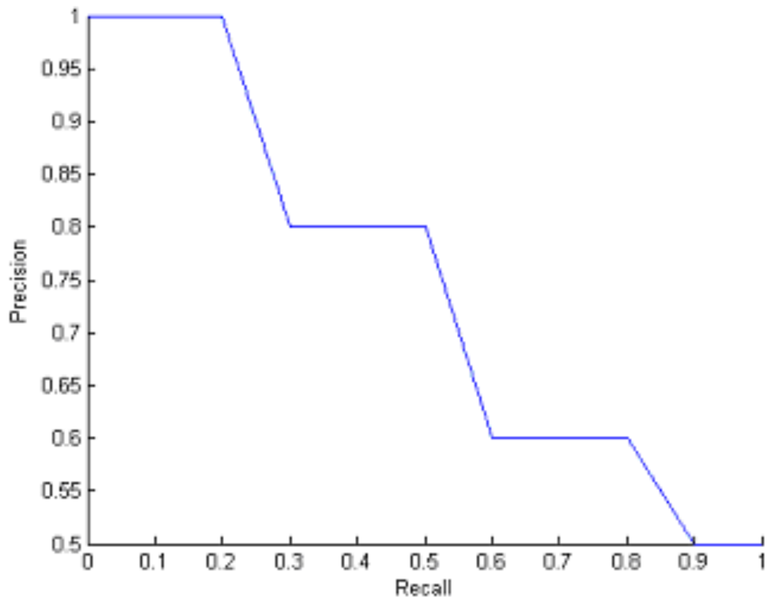
Recall	Precision
0.25	1.0
0.4	0.67
0.55	0.8
0.8	0.6
1.0	0.5

The interpolated precision :

0.0	1.0
0.1	1.0
0.2	1.0
0.3	0.8
0.4	0.8
0.5	0.8
0.6	0.6
0.7	0.6
0.8	0.6
0.9	0.5
1.0	0.5

Interpolated average precision = 0.745

Recall-Precision graph



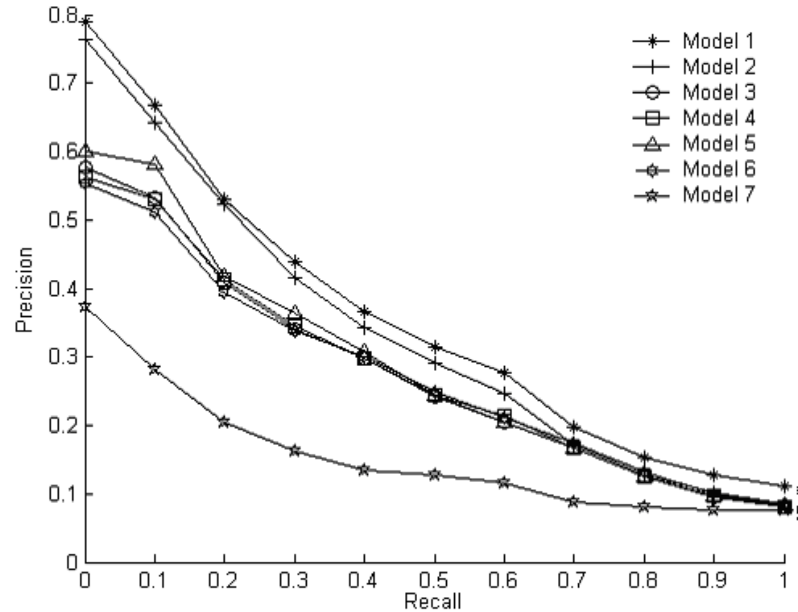
Average Recall/Precision Curve

- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

Average Recall/Precision Curve

Model

1. doc = “atn”, query = “ntc”
2. doc = “atn”, query = “atc”
3. doc = “atc”, query = “atc”
4. doc = “atc”, query = “ntc”
5. doc = “ntc”, query = “ntc”
6. doc = “ltc”, query = “ltc”
7. doc = “nnn”, query= “nnn”



- 1.000000 1.000000 1.000000 0.250000 0.250000 0.150000 0.150000 0.060606 0.016611 0.016611 0.016611
- 1.000000 1.000000 0.333333 0.333333 0.022727 0.022727 0.030075 0.026042 0.026042 0.026042 0.026042
- 1.000000 0.068966 0.081081 0.031250 0.024155 0.020115 0.022663 0.023377 0.021692 0.010138 0.010138
- 0.111111 0.111111 0.166667 0.085714 0.078431 0.078431 0.087719 0.086957 0.063636 0.034335 0.034335
- 1.000000 1.000000 1.000000 1.000000 0.200000 0.200000 0.200000 0.029703 0.029703 0.029703 0.029703
- 1.000000 1.000000 0.636364 0.142857 0.142857 0.135922 0.100000 0.055866 0.024974 0.014123 0.014123

Problems with Precision/Recall

- Can’t know true recall value
  - except in small collections
- Precision/Recall are related
  - A combined measure sometimes more appropriate
- Assumes batch mode
  - Interactive IR is important and has different criteria for successful searches
- Assumes a strict rank ordering matters.

Other measures: R-Precision

- R-Precision is the precision after *R* documents have been retrieved, where *R* is the number of relevant documents for a topic.
  - It de-emphasizes exact ranking of the retrieved relevant documents.
  - The average is simply the mean R-Precision for individual topics in the run.

Other measures: F-measure

- F-measure takes into account both precision and recall. It is defined as harmonic mean of recall and precision.

$$F = \frac{2PR}{P + R}$$

- Compared to arithmetic mean both need to be high for harmonic mean to be high.

E-measure

- E-measure is a variant of F-measure that allows weighting emphasis on precision over recall. It is defined as:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- Value of β controls the trade-off between precision and recall.

  Setting β to 1, gives equal weight to precision and recall ($E = F$)

  $\beta>1$ weight precision more whereas $\beta<1$ gives more weight to recall.

Normalized recall (NR)

- *Normalized recall* measures how close is the set of the retrieved document to an ideal retrieval in which the most relevant NR_{rel} documents appear in first NR_{rel} positions.
- Let's introduce two terms Ideal Rank (IR) and Average Rank (AR)

Ideal Rank (IR)

- If relevant documents are ranked 1,2,3, … then Ideal rank (IR) is given by

$$IR = \frac{\sum_{r=1}^{NR_{rel}} r}{NR_{rel}}$$

Average Rank (AR)

- Let the average rank (AR) over the set of relevant documents retrieved by the system be:

$$AR = \frac{\sum_{r=1}^{NR_{re}} Rank_r}{NR_{rel}}$$

- $Rank_r$ represents the rank of the r^{th} relevant document

Difference between AR and IR

- The difference between AR and IR, given by AR-IR, represents a measure of the effectiveness of the system.
- This difference ranges from 0 (for the perfect retrieval) to $(N-NR_{rel})$ for worst case retrieval

Normalized Recall

- The expression AR-IR can be normalized by dividing it by ($N-NR_{rel}$) and then by subtracting the result from 1, we get the normalized recall (NR) given by:

$$\mathrm{N}R = 1 - \frac{AR - IR}{(N - NR_{rel})}$$

- This measure ranges from 1 for the best case to 0 for the worst case.

- The expression AR-IR can be normalized by dividing it by (N-NR_{rel}) and then by subtracting the result from 1, we get the normalized recall (NR) given by:

$$\mathrm{N}R = 1 - \frac{AR - IR}{(N - NR_{rel})}$$

- This measure ranges from 1 for the best case to 0 for the worst case.

Evaluation Problems

- Realistic IR is interactive; traditional IR methods and measures are based on non-interactive situations
- Evaluating interactive IR requires human subjects (no gold standard or benchmarks)

[Ref.: See Borlund, 2000 & 2003; Borlund & Ingwersen, 1997 for IIR evaluation]