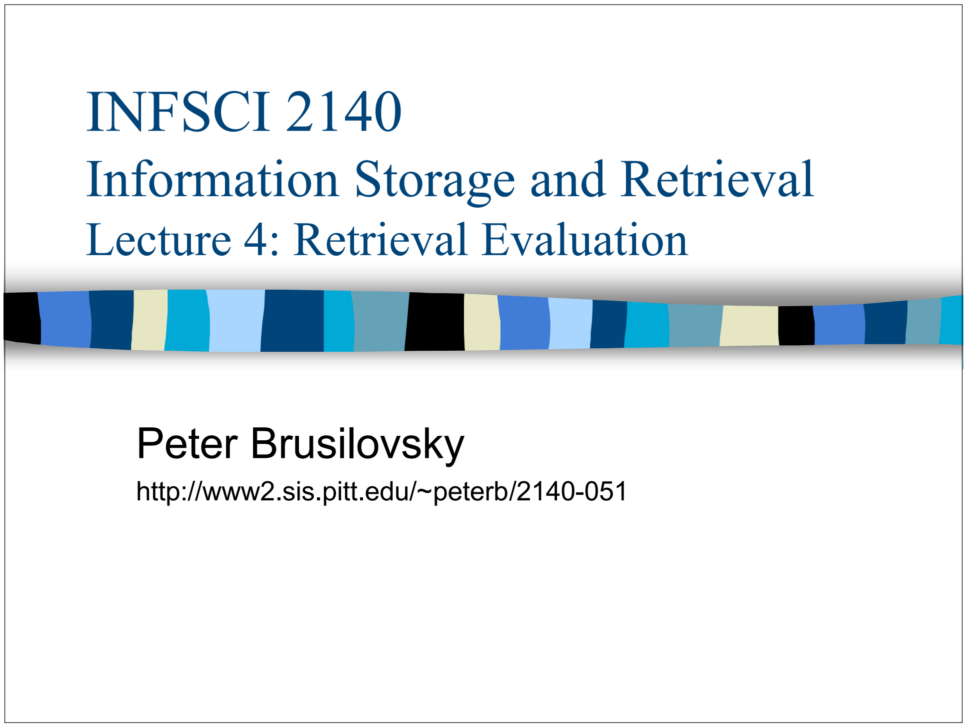
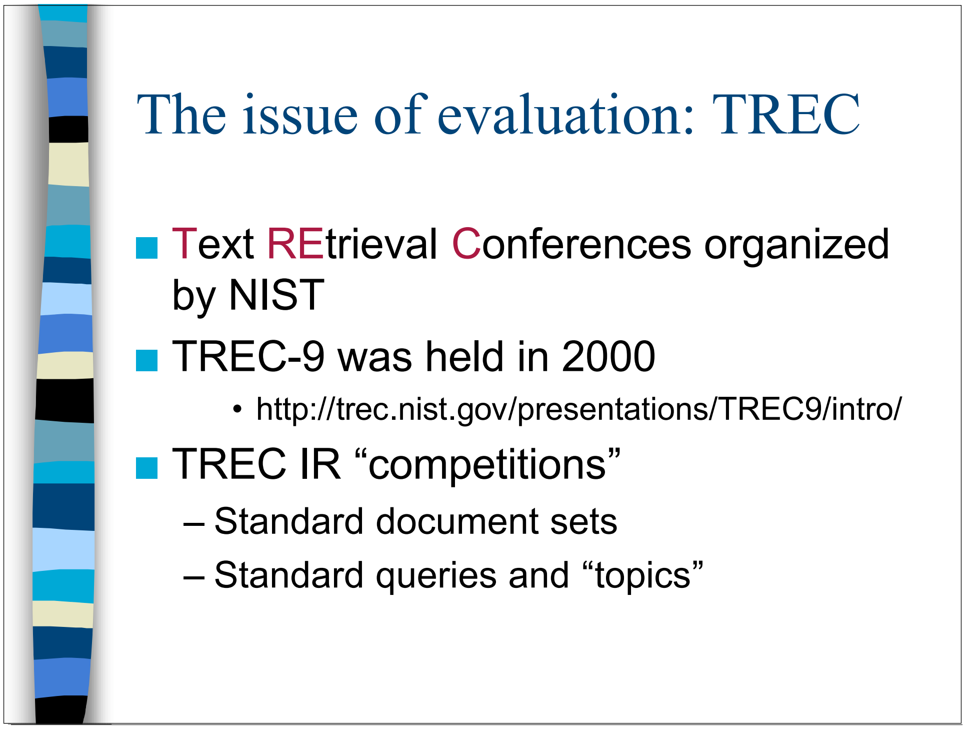


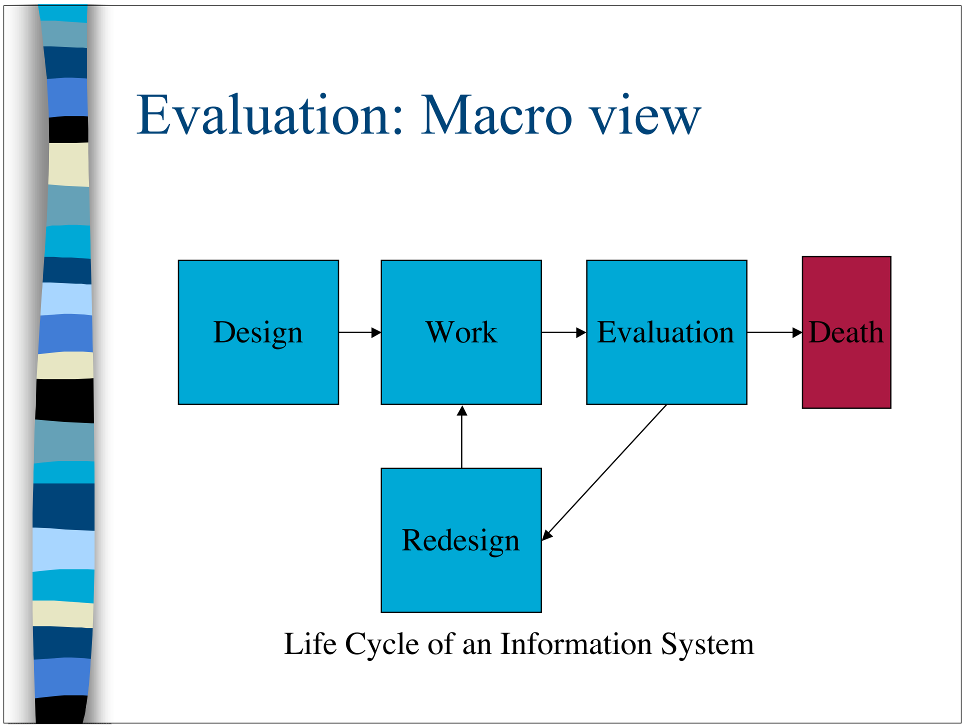
INFSCI 2140
Information Storage and Retrieval
Lecture 4: Retrieval Evaluation

Peter Brusilovsky
http://www2.sis.pitt.edu/~peterb/2140-051



The issue of evaluation: TREC

- Text REtrieval Conferences organized by NIST
- TREC-9 was held in 2000
  - http://trec.nist.gov/presentations/TREC9/intro/
- TREC IR “competitions”
  - Standard document sets
  - Standard queries and “topics”



Evaluation: Macro view

Design

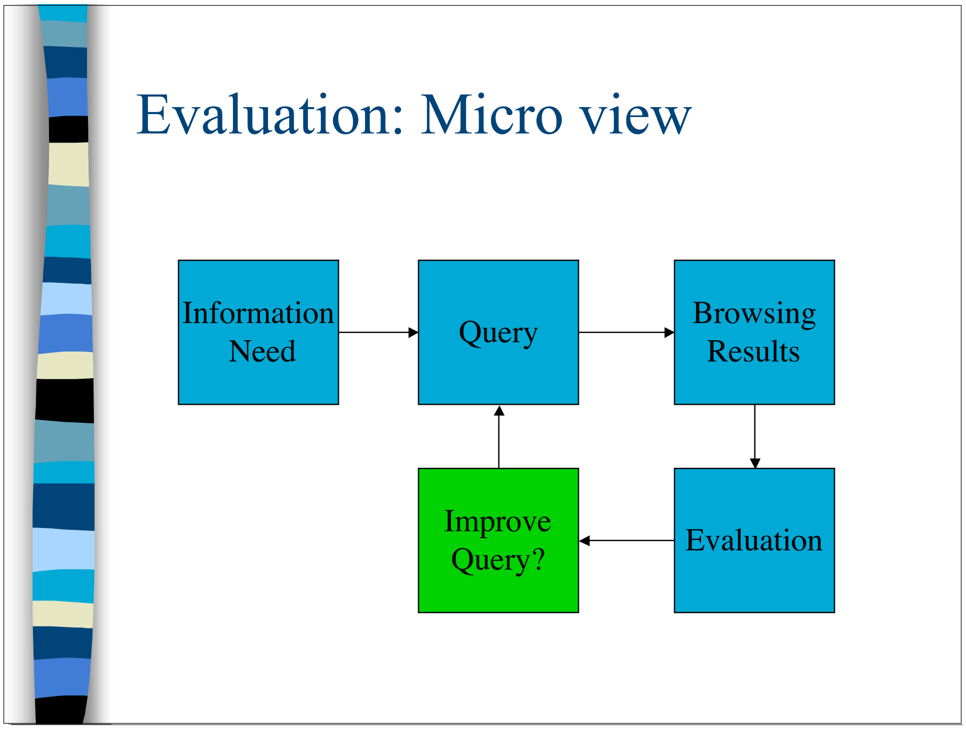
Work

Evaluation

Death

Redesign

Life Cycle of an Information System



Evaluation: Micro view

Information Need

Query

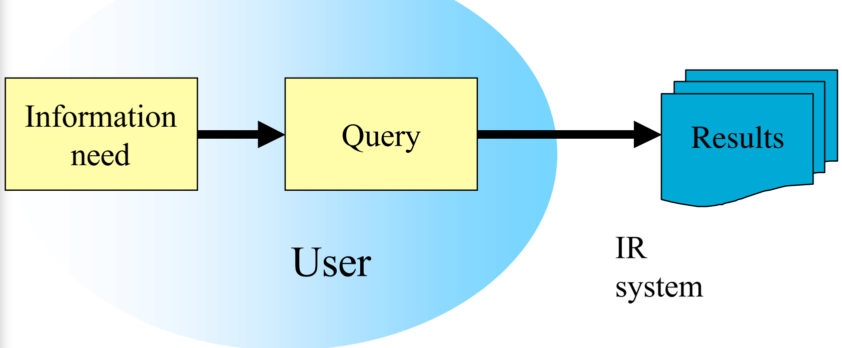
Browsing Results

Evaluation

Improve Query?

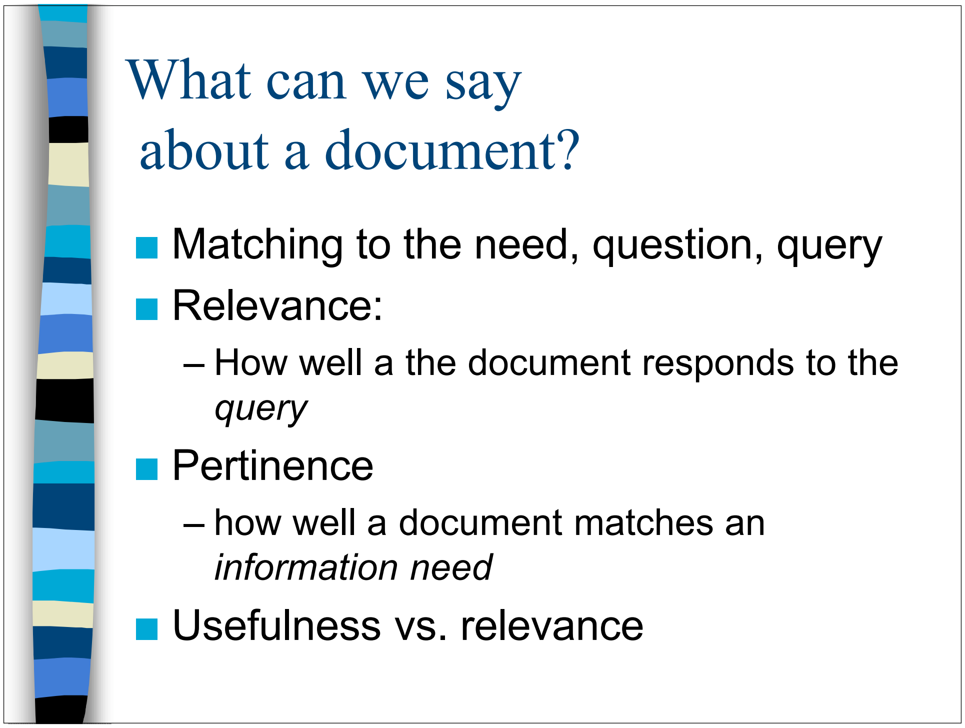
Effectiveness

- The effectiveness of a retrieval system is related to the user satisfaction
  - i.e. is related to the ectosystem



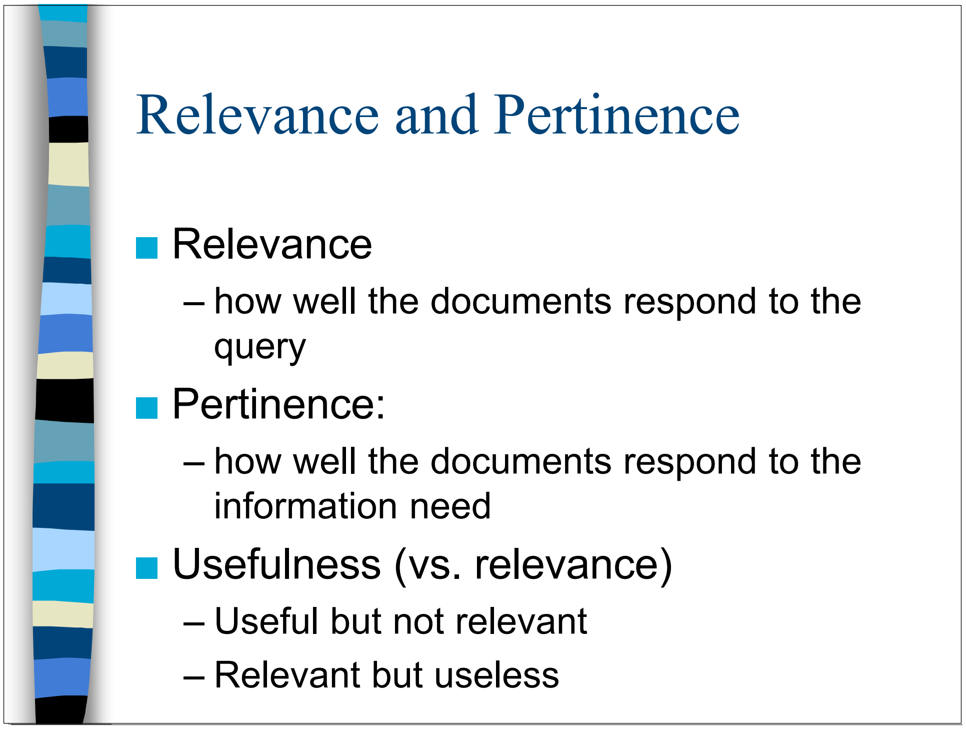
How Good is the Model?

- Was the query language powerful enough to represent the need?
- Were we able to use query syntax to express what we need
  - Operators
  - Weights
- Were the words from the limited vocabulary expressive enough?



What can we say about a document?

- Matching to the need, question, query
- Relevance:
  - How well a the document responds to the *query*
- Pertinence
  - how well a document matches an *information need*
- Usefulness vs. relevance



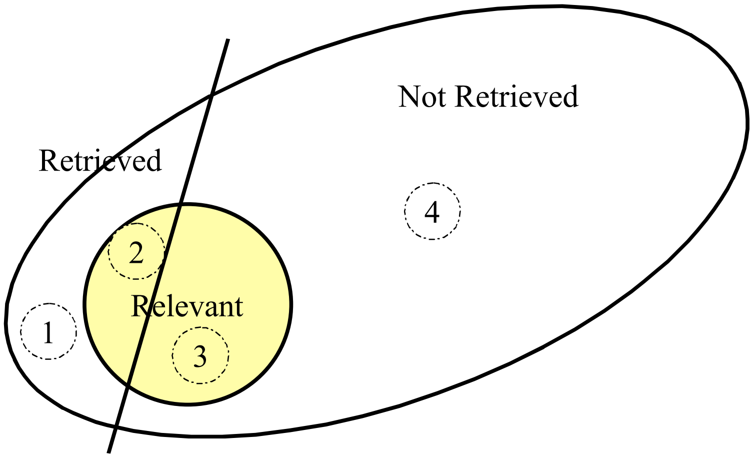
Relevance and Pertinence

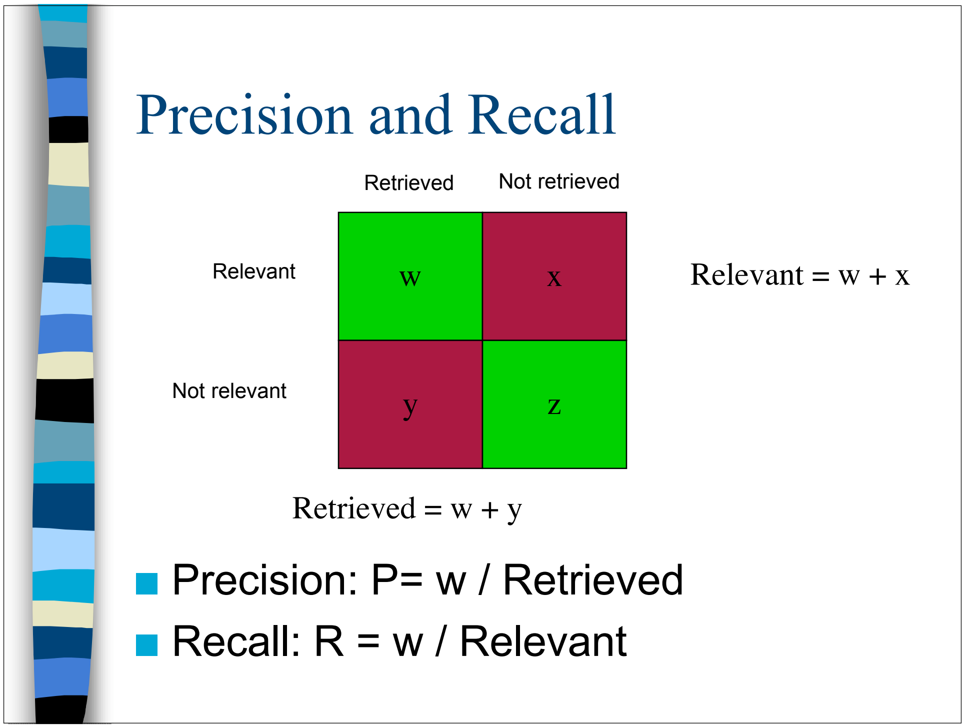
- Relevance
  - how well the documents respond to the query
- Pertinence:
  - how well the documents respond to the information need
- Usefulness (vs. relevance)
  - Useful but not relevant
  - Relevant but useless

How can we measure it?

- Binary measure (yes/no)
- N-ary measure:
  - 3 very relevant
  - 2 relevant
  - 1 barely relevant
  - 0 not relevant
- N=?: consistency vs. expressiveness

Precision and Recall





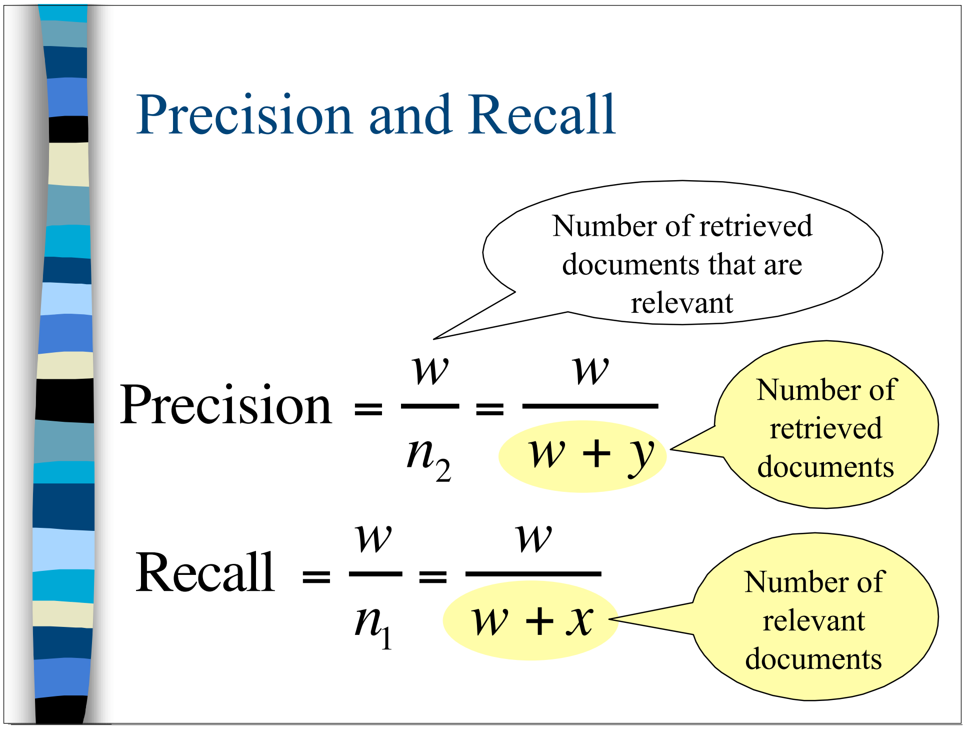
Precision and Recall

	Retrieved	Not retrieved
Relevant	w	x
Not relevant	y	z

Relevant = w + x

Retrieved = w + y

- Precision: P= w / Retrieved
- Recall: R = w / Relevant



Precision and Recall

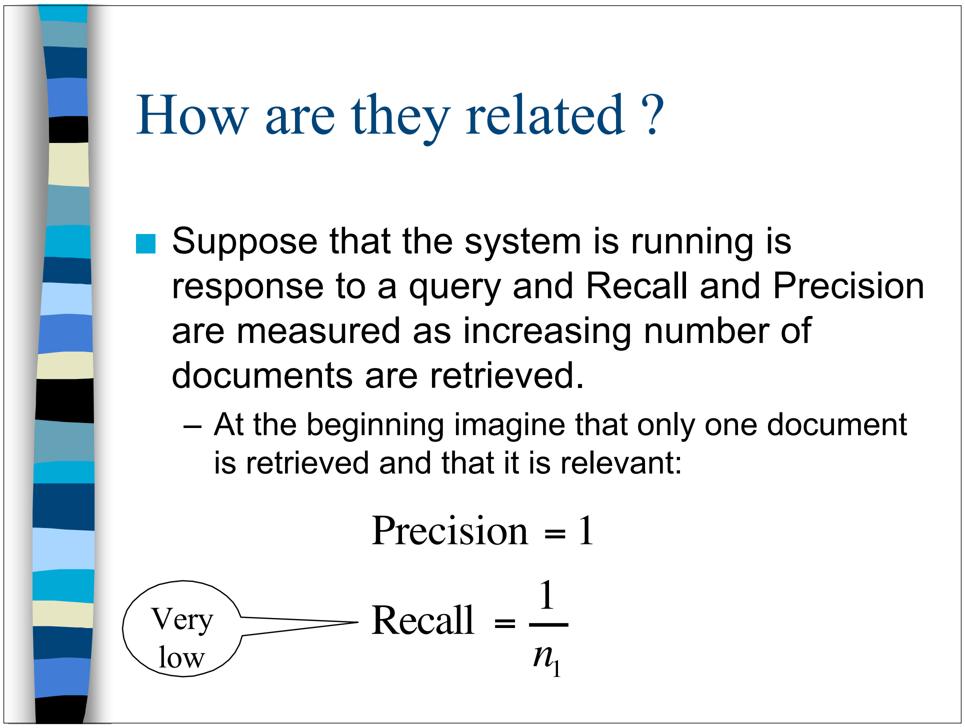
Number of retrieved documents that are relevant

$$\text{Precision} = \frac{w}{n_2} = \frac{w}{w+y}$$

Number of retrieved documents

$$\text{Recall} = \frac{w}{n_1} = \frac{w}{w+x}$$

Number of relevant documents



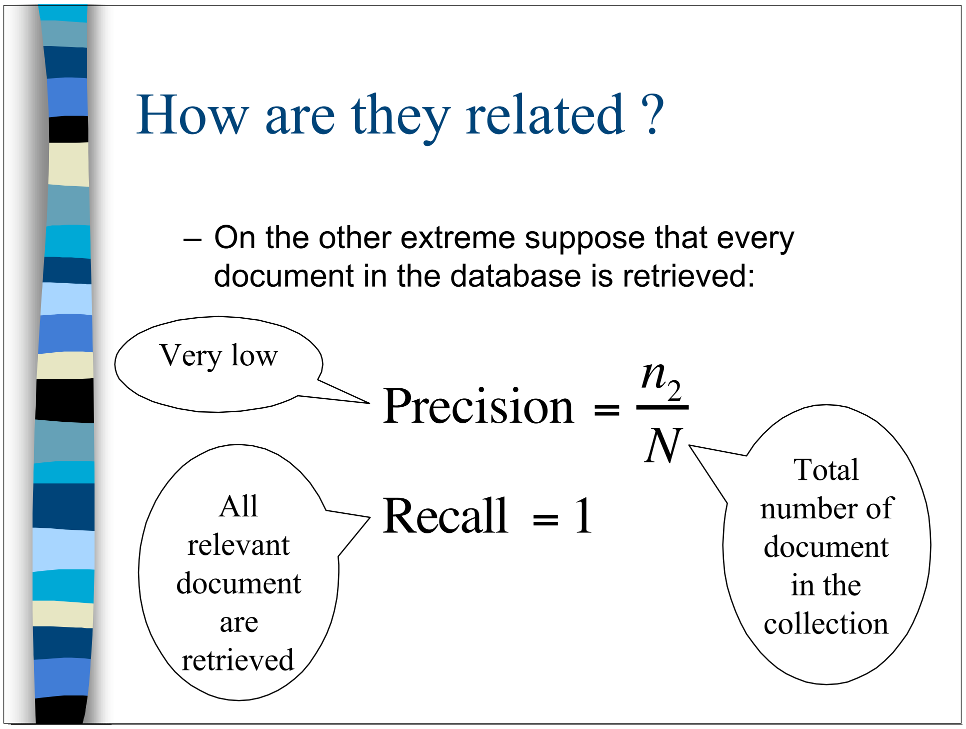
How are they related ?

- Suppose that the system is running is response to a query and Recall and Precision are measured as increasing number of documents are retrieved.
  - At the beginning imagine that only one document is retrieved and that it is relevant:

$$\text{Precision} = 1$$

Very low

$$\text{Recall} = \frac{1}{n_1}$$



How are they related ?

- On the other extreme suppose that every document in the database is retrieved:

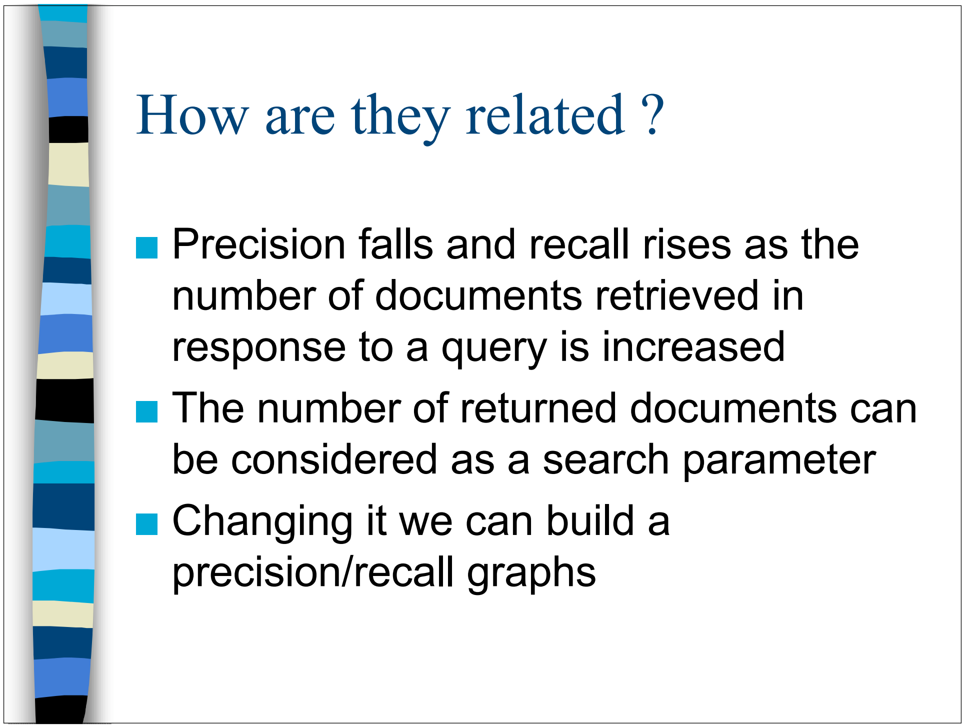
Very low

$$\text{Precision} = \frac{n_2}{N}$$

Total number of document in the collection

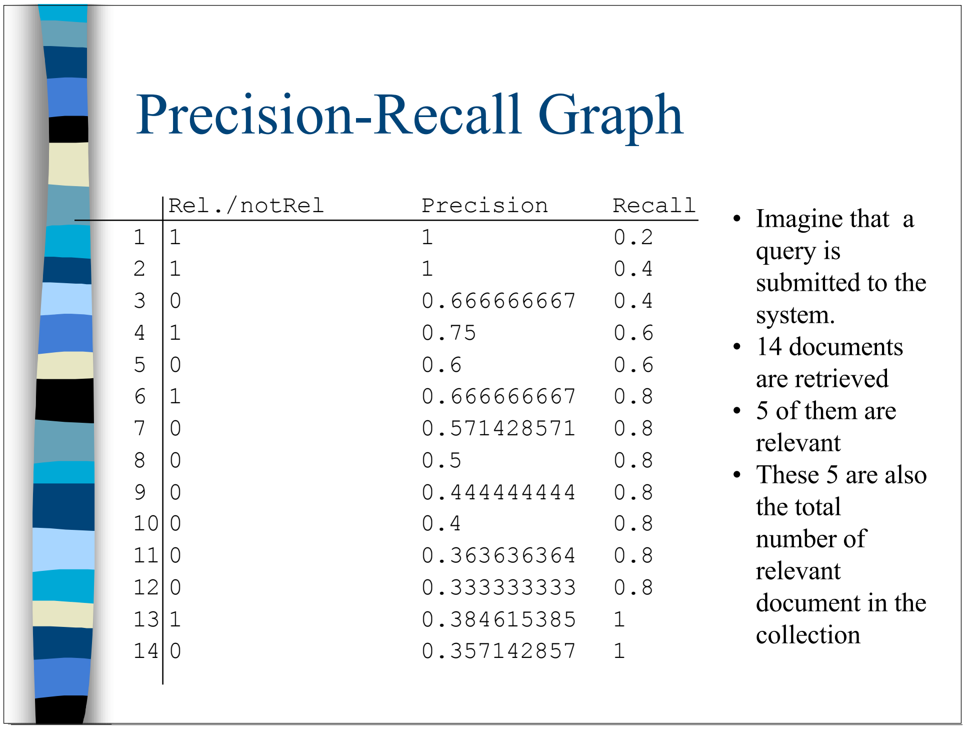
All relevant document are retrieved

$$\text{Recall} = 1$$



How are they related ?

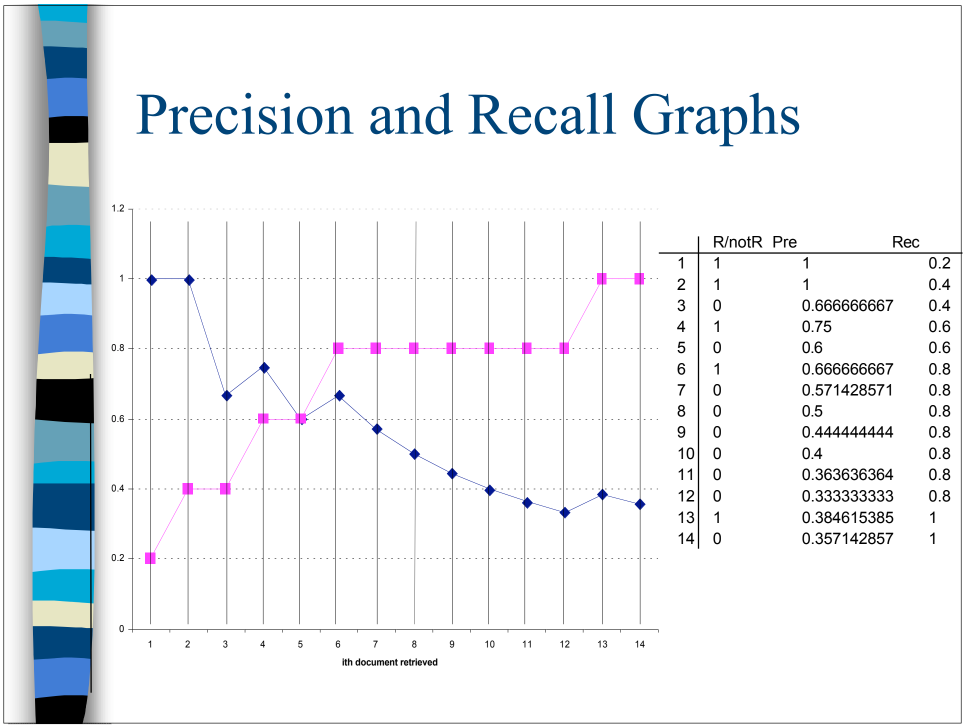
- Precision falls and recall rises as the number of documents retrieved in response to a query is increased
- The number of returned documents can be considered as a search parameter
- Changing it we can build a precision/recall graphs



Precision-Recall Graph

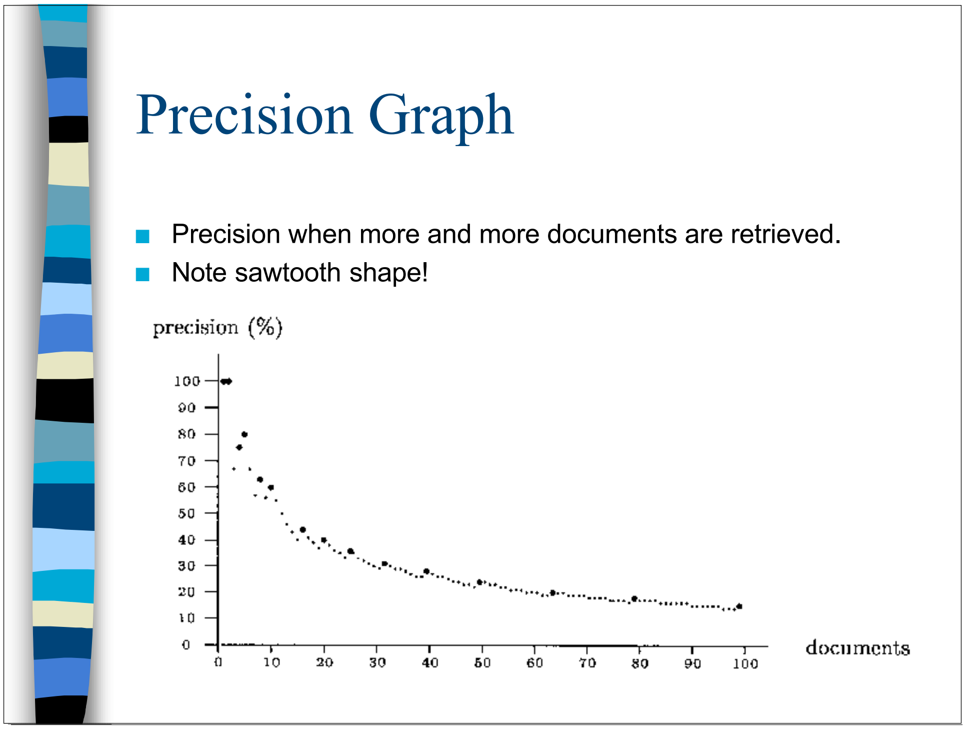
	Rel./notRel	Precision	Recall
1	1	1	0.2
2	1	1	0.4
3	0	0.666666667	0.4
4	1	0.75	0.6
5	0	0.6	0.6
6	1	0.666666667	0.8
7	0	0.571428571	0.8
8	0	0.5	0.8
9	0	0.444444444	0.8
10	0	0.4	0.8
11	0	0.363636364	0.8
12	0	0.333333333	0.8
13	1	0.384615385	1
14	0	0.357142857	1

- Imagine that a query is submitted to the system.
- 14 documents are retrieved
- 5 of them are relevant
- These 5 are also the total number of relevant document in the collection



Precision and Recall Graphs

	R/notR	Pre	Rec
1	1	1	0.2
2	1	1	0.4
3	0	0.666666667	0.4
4	1	0.75	0.6
5	0	0.6	0.6
6	1	0.666666667	0.8
7	0	0.571428571	0.8
8	0	0.5	0.8
9	0	0.444444444	0.8
10	0	0.4	0.8
11	0	0.363636364	0.8
12	0	0.333333333	0.8
13	1	0.384615385	1
14	0	0.357142857	1

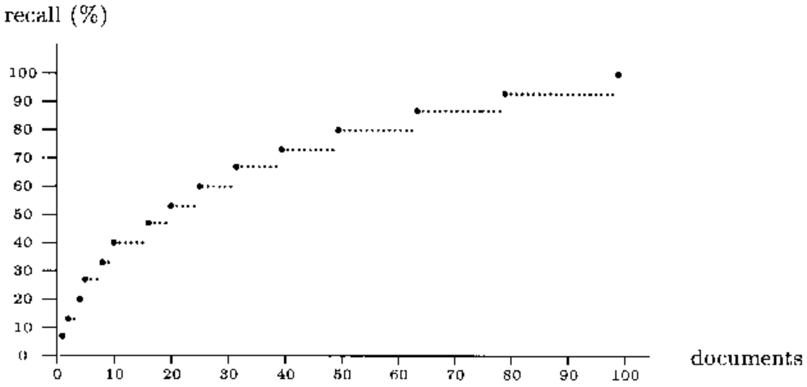


Precision Graph

- Precision when more and more documents are retrieved.
- Note sawtooth shape!

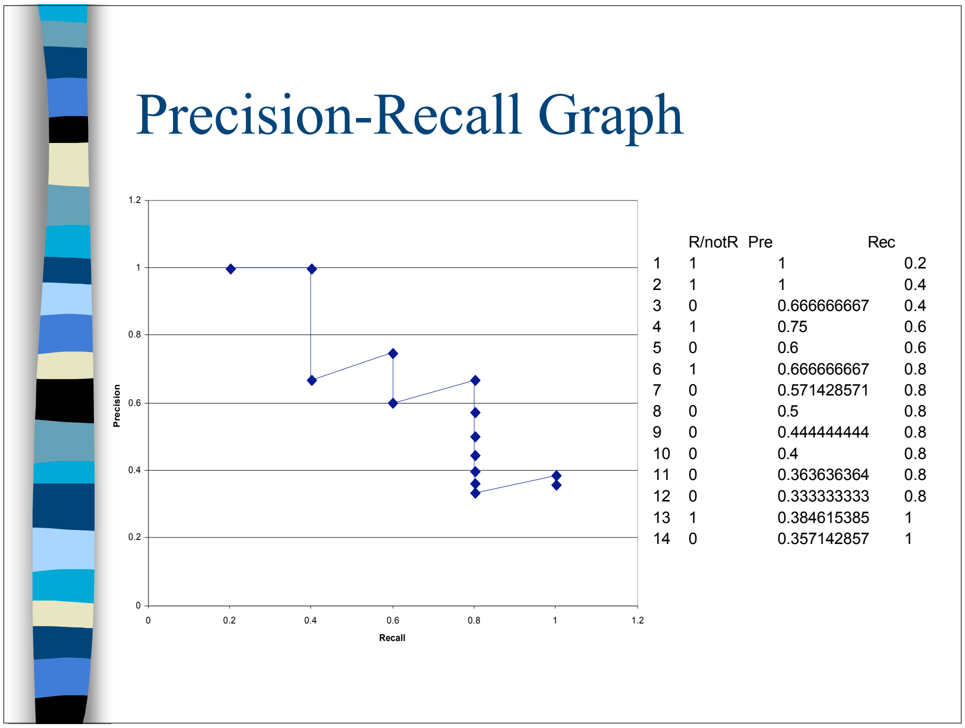
Recall Graph

- Recall when more and more documents are retrieved.
- Note terraced shape!



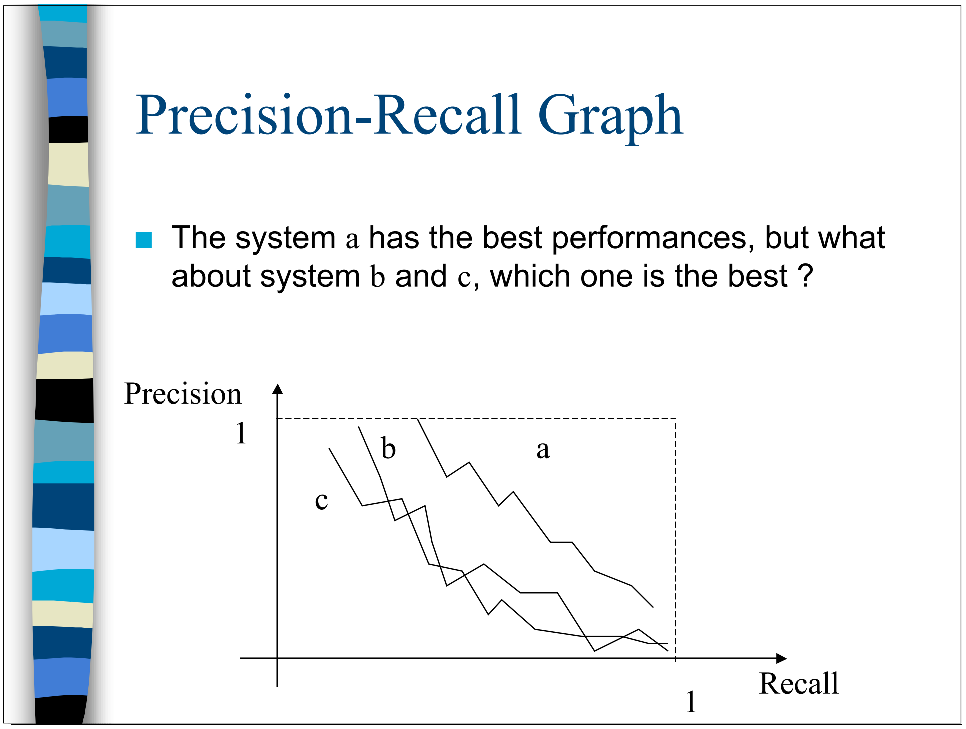
Precision-Recall Graph

- Sequences of points (p, r)
- Similar to y = 1 / x:
  - – Inversely proportional!
- Sawtooth shape
- Use smoothed graphs
- How we can compare different IR systems using precision-recall graphs?



Precision-Recall Graph

	R/notR	Pre	Rec
1	1	1	0.2
2	1	1	0.4
3	0	0.666666667	0.4
4	1	0.75	0.6
5	0	0.6	0.6
6	1	0.666666667	0.8
7	0	0.571428571	0.8
8	0	0.5	0.8
9	0	0.444444444	0.8
10	0	0.4	0.8
11	0	0.363636364	0.8
12	0	0.333333333	0.8
13	1	0.384615385	1
14	0	0.357142857	1



Precision-Recall Graph

- The system a has the best performances, but what about system b and c, which one is the best ?

Fallout

- The proportion of not relevant document that are retrieved (it should be low for a good IR system)
- Fallout measures how well the system filters out not-relevant documents

$$F = \frac{y}{N - n_1}$$

Number of not relevant documents that are retrieved

Total number of not relevant documents

Generality

- Proportion of relevant documents in the collection. It is more related to the query rather than to the retrieval process

$$G = \frac{n_1}{N}$$

Number of relevant documents in the collection

Total number documents

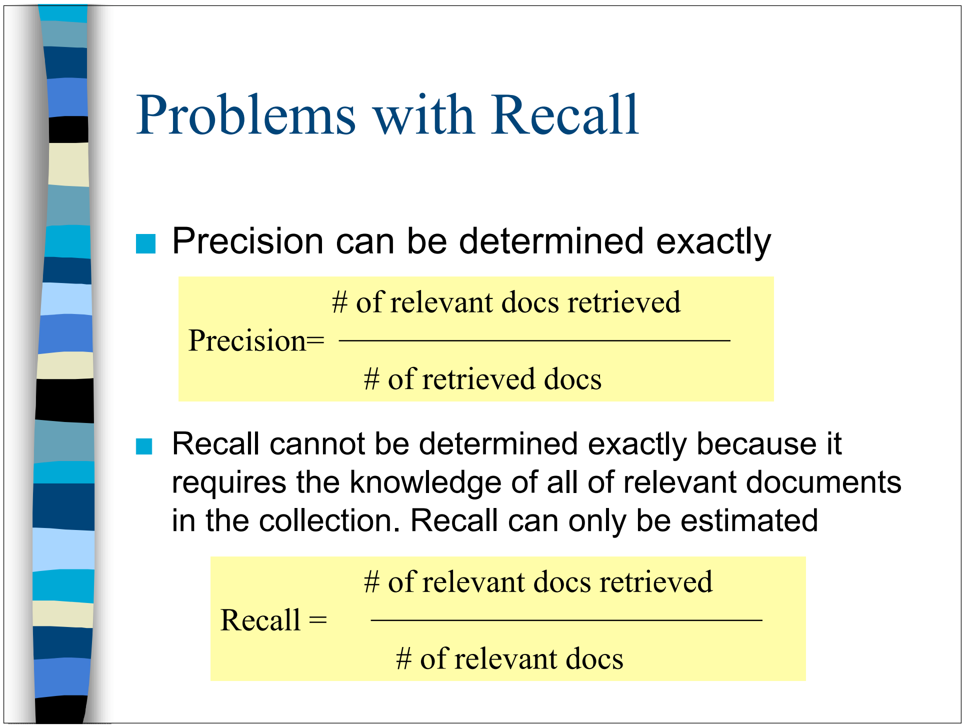
Exercise 1

Imagine that an IR system retrieved 10 document in answer to a query, but only the document number 1, 3, 5, 7 are relevant.

Calculate Precision, Recall and Fallout considering that there are other 6 relevant documents that were not retrieved and that the total number of documents in the collection is 100 (included the 10 retrieved).

Problems of recall & precision

- Hard to find recall
- Neither shows effectiveness
  - Comparing the graphs
  - F-measure
  - Relative performance as another single measure
- Recall & precision may not be important for the user



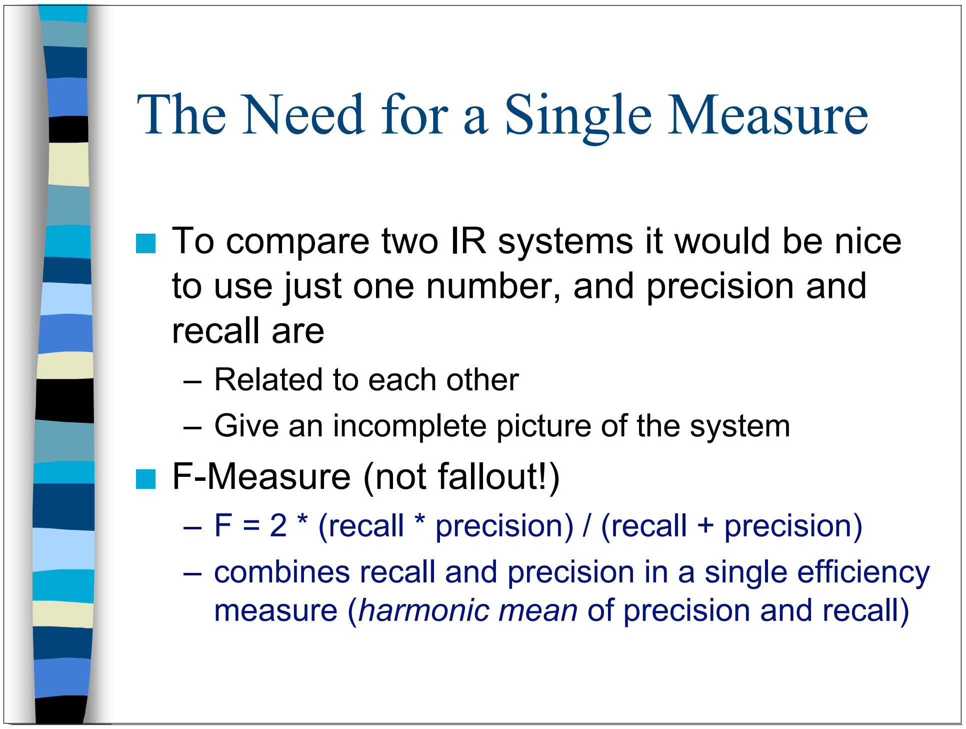
Problems with Recall

- Precision can be determined exactly

$$\text{Precision} = \frac{\text{\# of relevant docs retrieved}}{\text{\# of retrieved docs}}$$

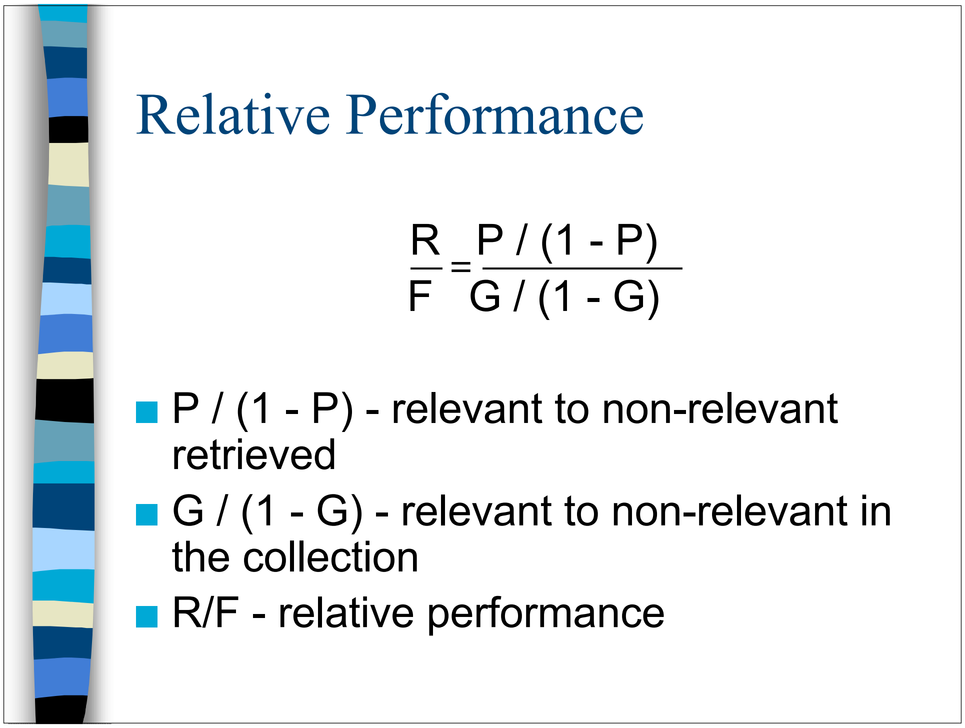
- Recall cannot be determined exactly because it requires the knowledge of all of relevant documents in the collection. Recall can only be estimated

$$\text{Recall} = \frac{\text{\# of relevant docs retrieved}}{\text{\# of relevant docs}}$$



The Need for a Single Measure

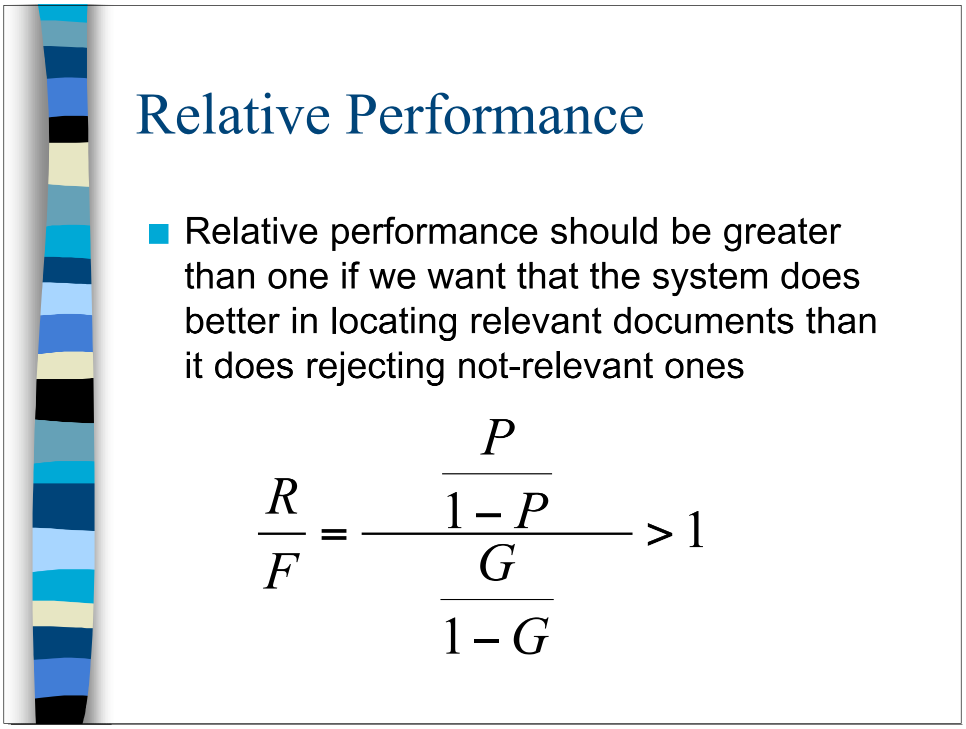
- To compare two IR systems it would be nice to use just one number, and precision and recall are
  - Related to each other
  - Give an incomplete picture of the system
- F-Measure (not fallout!)
  - F = 2 * (recall * precision) / (recall + precision)
  - combines recall and precision in a single efficiency measure (*harmonic mean* of precision and recall)



Relative Performance

$$\frac{R}{F} = \frac{P / (1 - P)}{G / (1 - G)}$$

- P / (1 - P) - relevant to non-relevant retrieved
- G / (1 - G) - relevant to non-relevant in the collection
- R/F - relative performance



Relative Performance

- Relative performance should be greater than one if we want that the system does better in locating relevant documents than it does rejecting not-relevant ones

$$\frac{R}{F} = \frac{\dfrac{P}{1-P}}{\dfrac{G}{1-G}} > 1$$

Precision and Recall: User View

- It is not clear how important they are for the users:
  - – Precision in usually more important that recall, because users appreciate outputs that do not contain not relevant documents
  - – This, of course, depends on the kind of user: high recall is important for an attorney that needs to determine all the legal precedents to a case.

What does the user want? Restaurant case

- The user wants to find a restaurant serving Sashimi. She can use 2 IR systems. How we can say which one is better?

User - oriented measures

- Coverage ratio:

  known_relevant_retrieved / known_ relevant
- Novelty ratio:
  – new_relevant / Relevant
- Relative recall
  – relevant_retrieved /wants_to_examine
- Recall Effort:
  – wants_to_examine / had_to_examine

Coverage and Novelty

- **Coverage Ratio**: proportion of relevant documents known to the user that are actually retrieved
  – A high coverage ratio would give to the user some confidence that the system is locating all he relevant documents
- **Novelty Ratio**: proportion of relevant retrieved documents that were unknown to the user
  – A high novelty ratio suggests that the system is effective in locating documents previously unknown to the user

Coverage and Novelty

- For example if the user knows that there are 16 relevant documents (but they are not all the relevant documents) and the system **retrieve 10 relevant documents** included 4 of those that the user knows we have:

$$\text{Coverage ratio} = \frac{4}{16} \qquad \text{Novelty ratio} = \frac{6}{10}$$

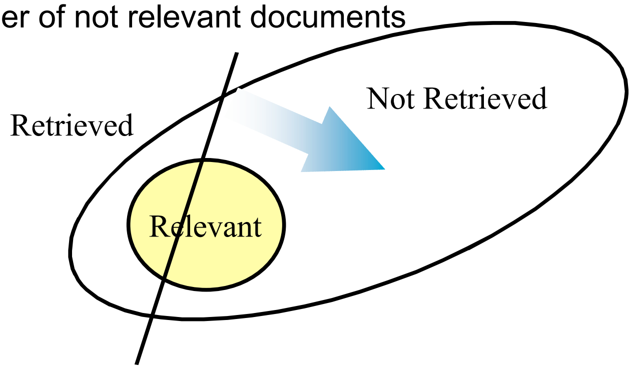
- User may expect 40 relevant documents in total

Relative Recall and Effort

- **Relative Recall**: The ratio of relevant retrieved documents examined by the user to the number of documents the user would have liked to examine
  - If the system has retrieved 5 relevant documents among 20 - how large is the relative recall?
- **Relative Effort**: The ratio of number of relevant documents desired to the number of documents examined by the user to find the number of relevant documents desired
  - this ratio go to 1 if the relevant docs are the first examined, to early 0 if the user would need to examine hundreds of documents to find the desired few.

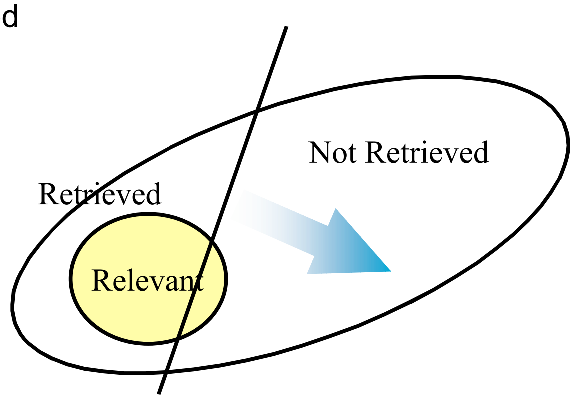
What happen when we increase the number of documents retrieved?

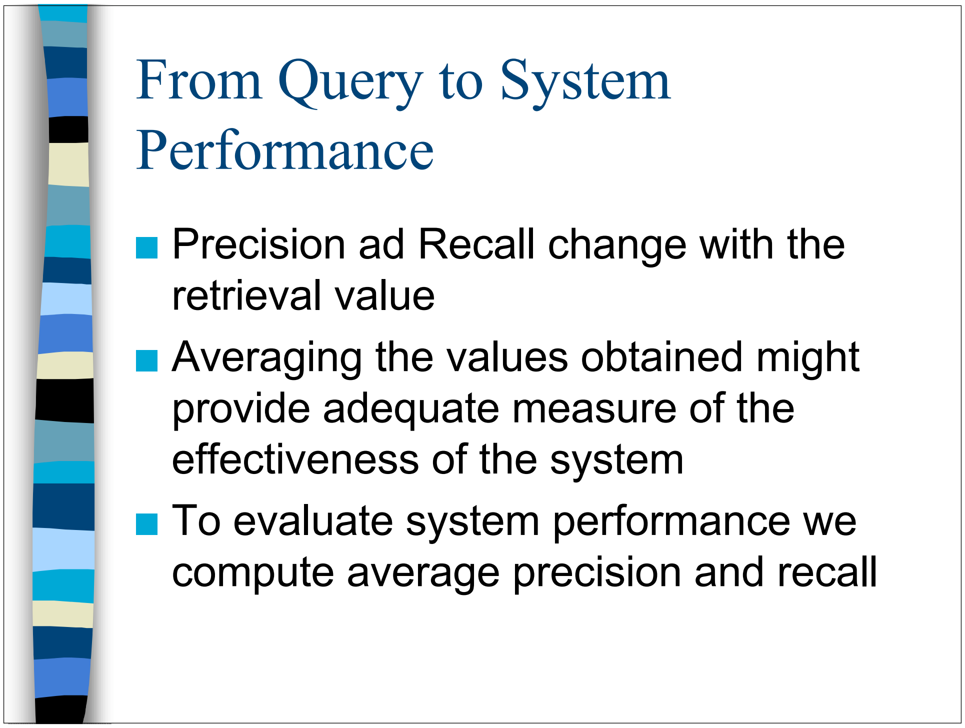
- At **low retrieval volumes** when we increase the number of documents retrieved , the number of relevant documents increase more rapidly than the number of not relevant documents



What happen when we increase the number of documents retrieved?

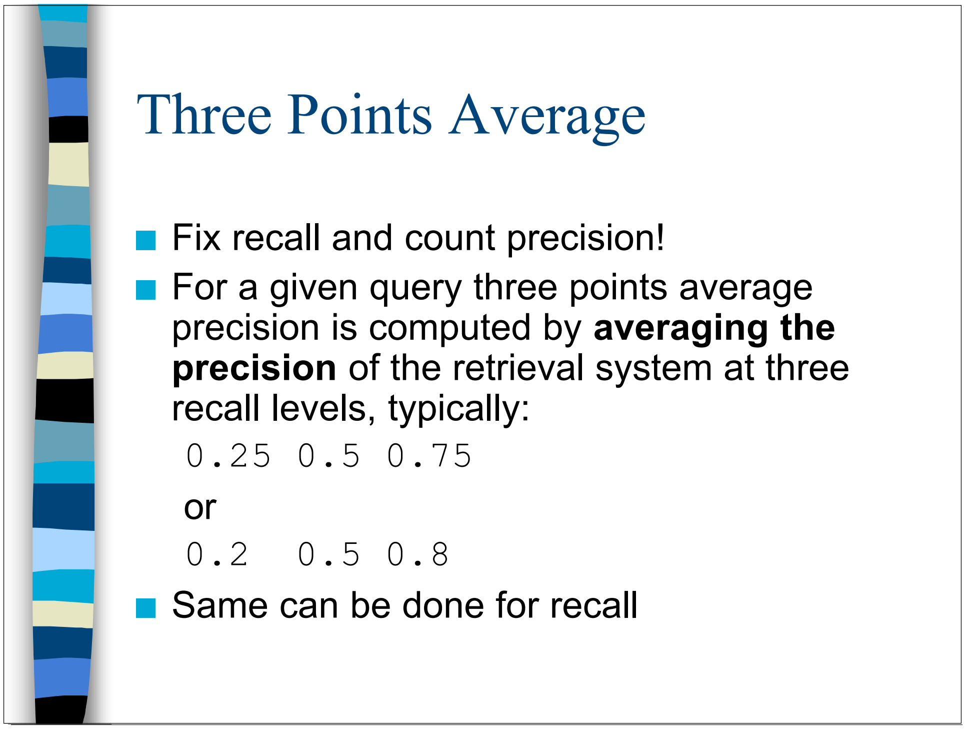
- At **high retrieval volumes** when we increase the number of document retrieved the situation is reversed





From Query to System Performance

- Precision ad Recall change with the retrieval value
- Averaging the values obtained might provide adequate measure of the effectiveness of the system
- To evaluate system performance we compute average precision and recall



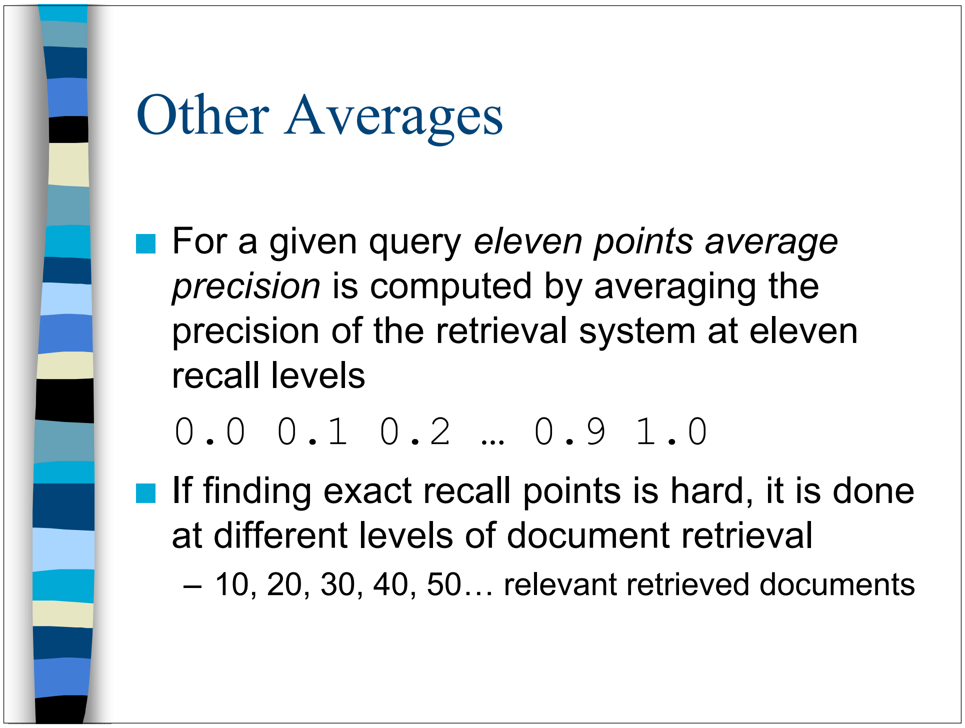
Three Points Average

- Fix recall and count precision!
- For a given query three points average precision is computed by **averaging the precision** of the retrieval system at three recall levels, typically:

  `0.25 0.5 0.75`

  or

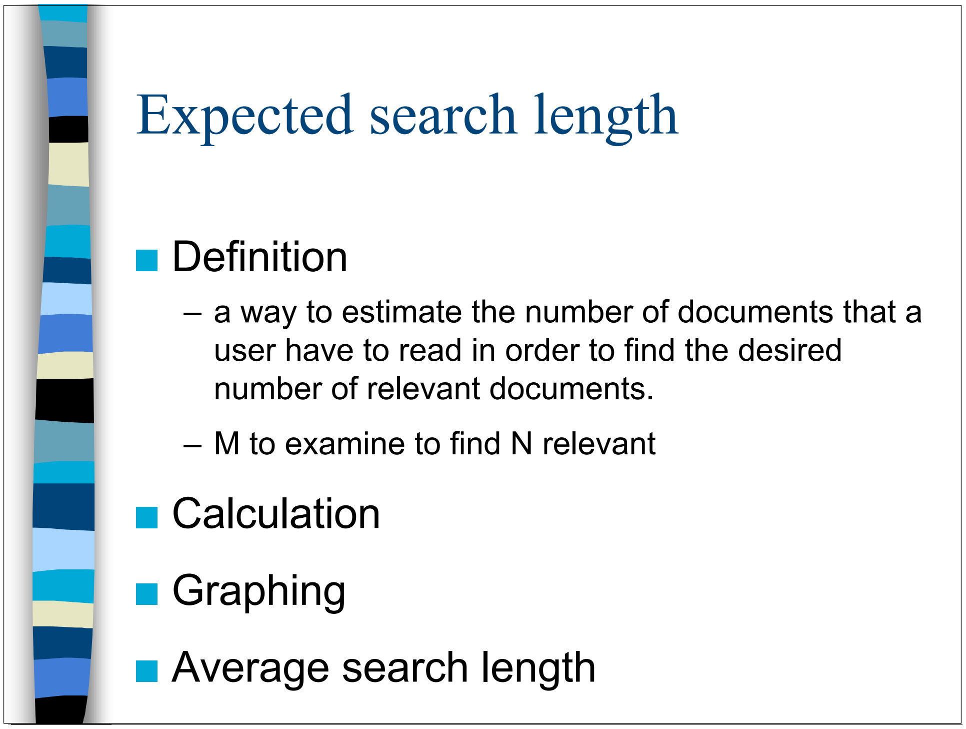
  `0.2  0.5 0.8`
- Same can be done for recall



Other Averages

- For a given query *eleven points average precision* is computed by averaging the precision of the retrieval system at eleven recall levels

  `0.0 0.1 0.2 … 0.9 1.0`
- If finding exact recall points is hard, it is done at different levels of document retrieval
  - 10, 20, 30, 40, 50… relevant retrieved documents



Expected search length

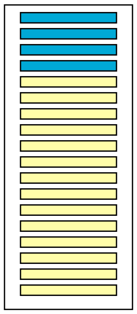
- Definition
  - a way to estimate the number of documents that a user have to read in order to find the desired number of relevant documents.
  - M to examine to find N relevant
- Calculation
- Graphing
- Average search length

Taking the Order into Account

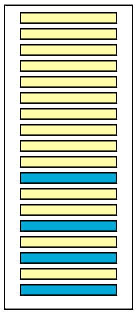
- Results of search is not a set, but a sequence
- Recall and Precision fail to take into account the sequentiality effect in presenting the retrieval results
- Two documents that contains the same information can be judged by the system in a different way
  - the first in the list is considered relevant
  - the second one, maybe separated from the first by many other documents, is considered much less relevant

Frustration

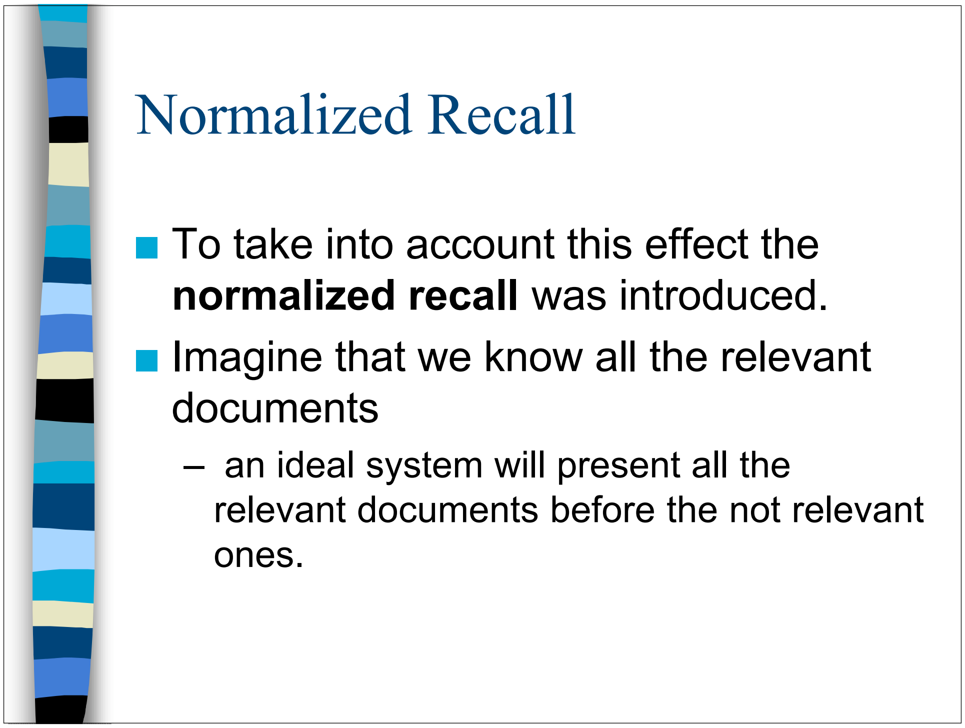
- Two systems can give a very different perception if they just organize the same documents in a different way:



All the relevant documents in the first positions

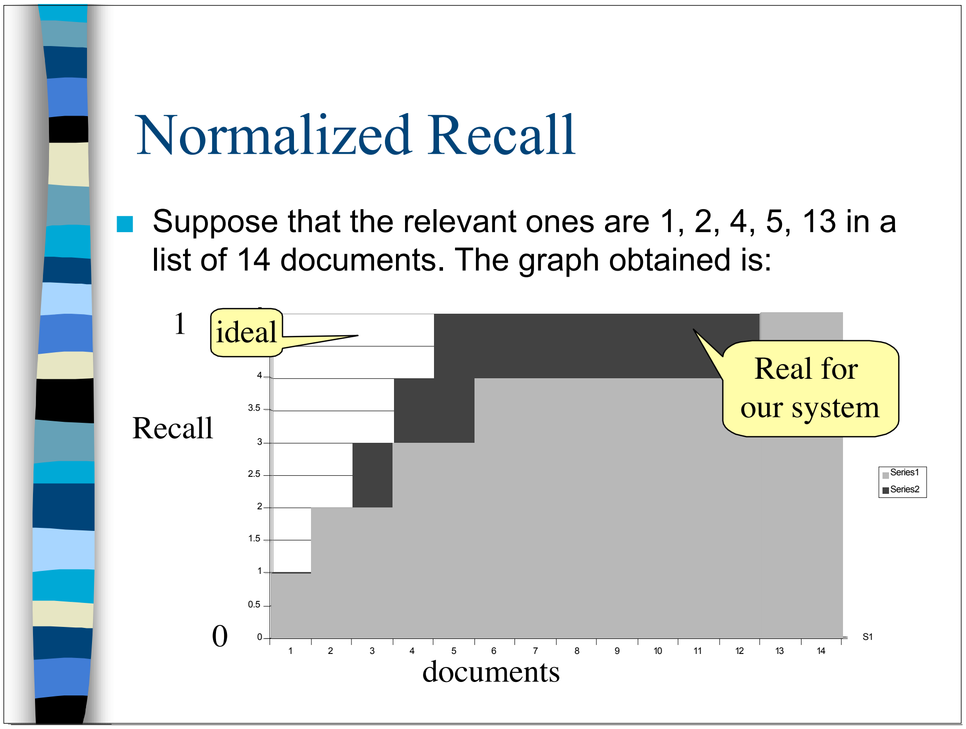


Relevant documents scattered in the list at the end of the list



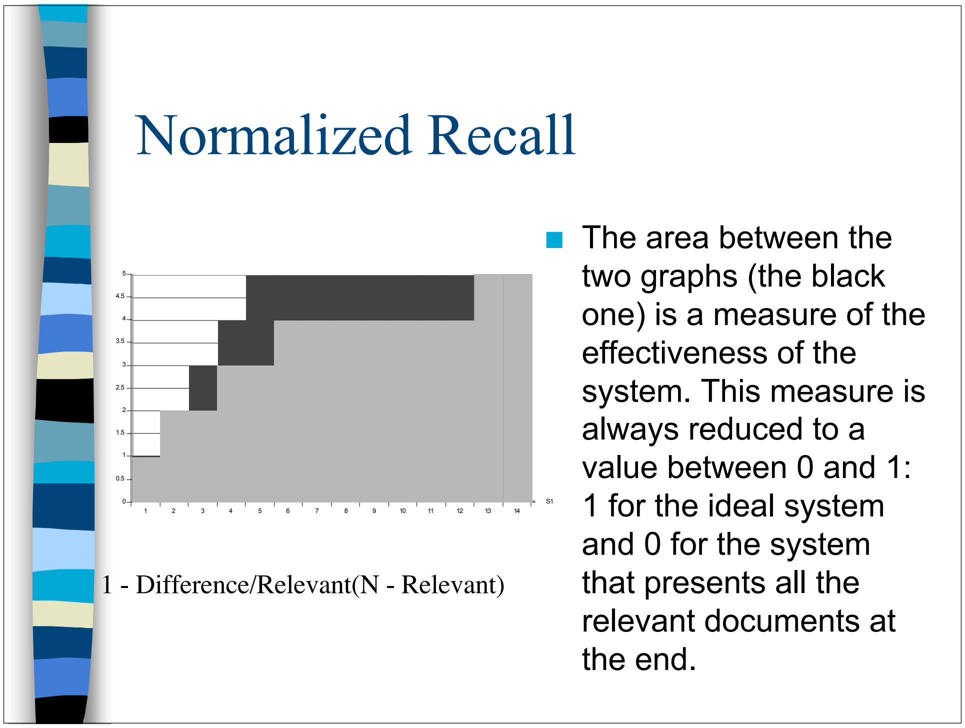
Normalized Recall

- To take into account this effect the **normalized recall** was introduced.
- Imagine that we know all the relevant documents
  – an ideal system will present all the relevant documents before the not relevant ones.



Normalized Recall

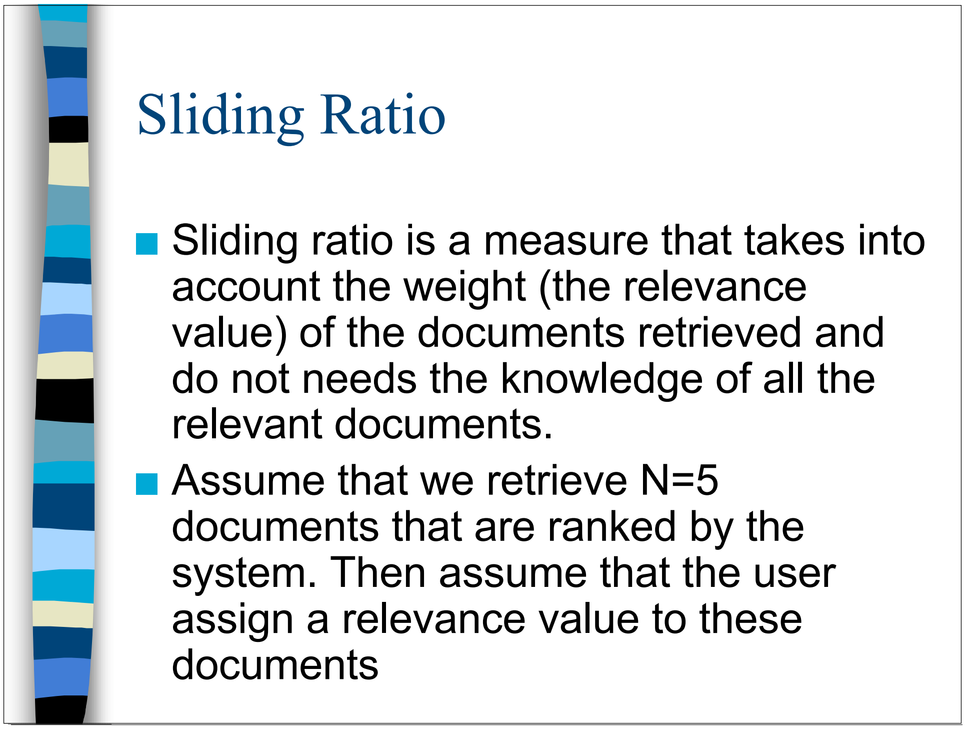
- Suppose that the relevant ones are 1, 2, 4, 5, 13 in a list of 14 documents. The graph obtained is:



Normalized Recall

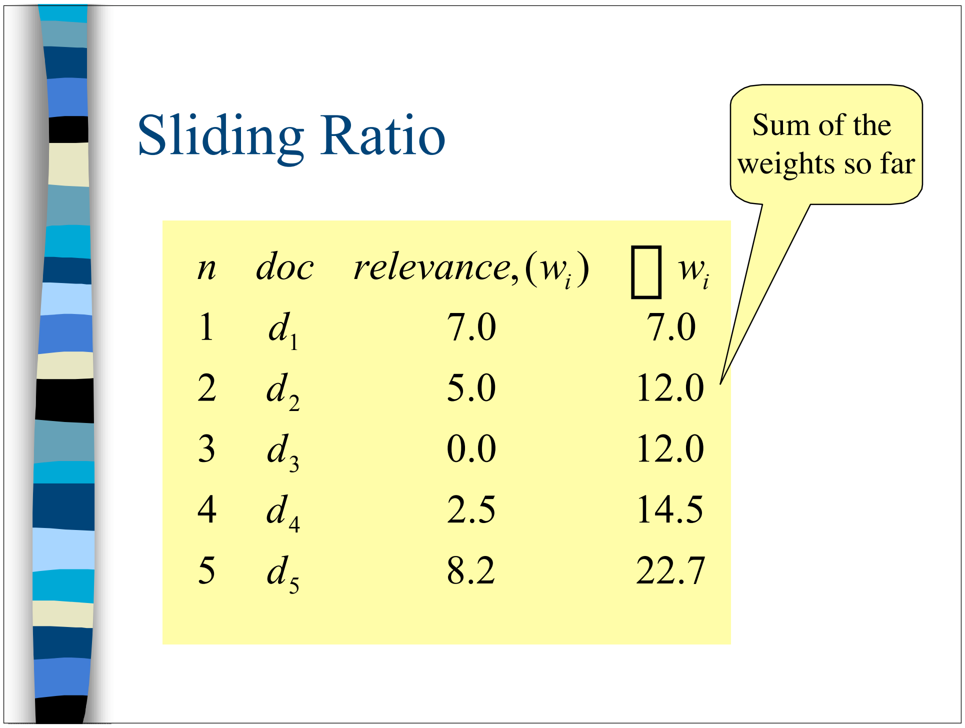
1 - Difference/Relevant(N - Relevant)

- The area between the two graphs (the black one) is a measure of the effectiveness of the system. This measure is always reduced to a value between 0 and 1: 1 for the ideal system and 0 for the system that presents all the relevant documents at the end.



Sliding Ratio

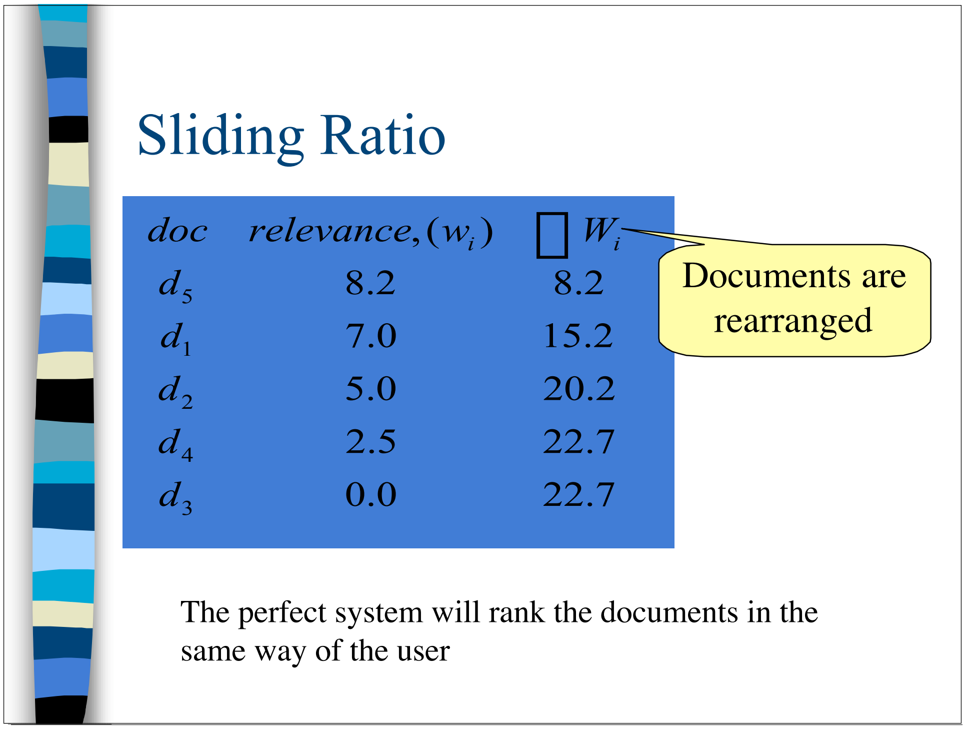
- Sliding ratio is a measure that takes into account the weight (the relevance value) of the documents retrieved and do not needs the knowledge of all the relevant documents.
- Assume that we retrieve N=5 documents that are ranked by the system. Then assume that the user assign a relevance value to these documents



Sliding Ratio

Sum of the weights so far

n	doc	$relevance, (w_i)$	□ w_i
1	d_1	7.0	7.0
2	d_2	5.0	12.0
3	d_3	0.0	12.0
4	d_4	2.5	14.5
5	d_5	8.2	22.7

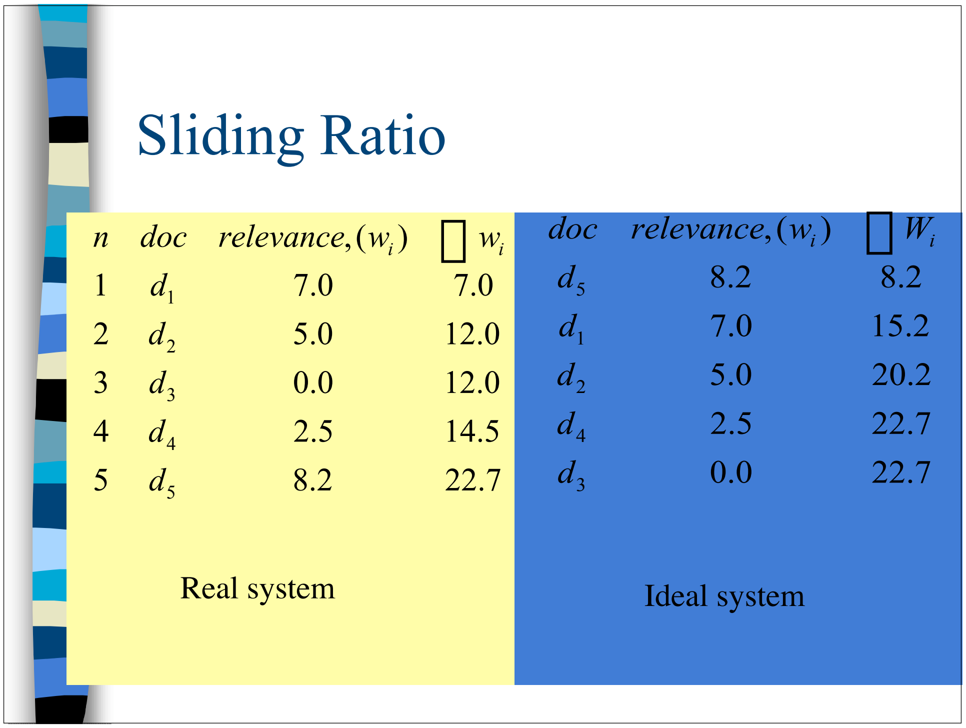


Sliding Ratio

doc	$relevance, (w_i)$	□ W_i
d_5	8.2	8.2
d_1	7.0	15.2
d_2	5.0	20.2
d_4	2.5	22.7
d_3	0.0	22.7

Documents are rearranged

The perfect system will rank the documents in the same way of the user

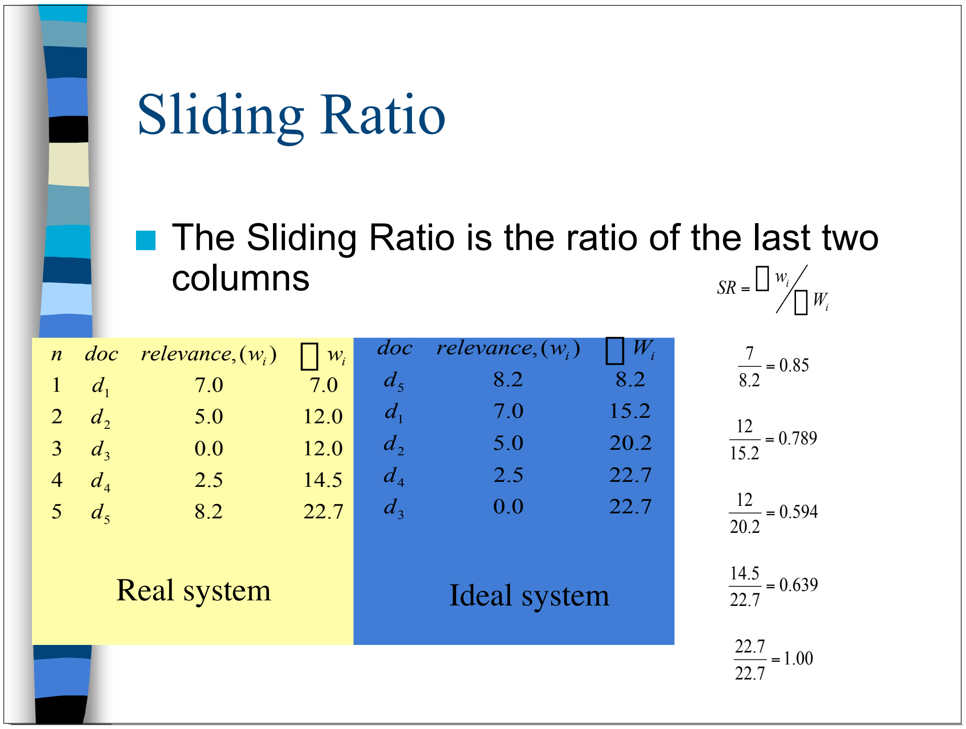


Sliding Ratio

n	doc	$relevance, (w_i)$	$\sum w_i$	doc	$relevance, (w_i)$	$\sum W_i$
1	d_1	7.0	7.0	d_5	8.2	8.2
2	d_2	5.0	12.0	d_1	7.0	15.2
3	d_3	0.0	12.0	d_2	5.0	20.2
4	d_4	2.5	14.5	d_4	2.5	22.7
5	d_5	8.2	22.7	d_3	0.0	22.7

Real system

Ideal system



Sliding Ratio

- The Sliding Ratio is the ratio of the last two columns

$$SR = \sum w_i \Big/ \sum W_i$$

n	doc	$relevance, (w_i)$	$\sum w_i$	doc	$relevance, (w_i)$	$\sum W_i$
1	d_1	7.0	7.0	d_5	8.2	8.2
2	d_2	5.0	12.0	d_1	7.0	15.2
3	d_3	0.0	12.0	d_2	5.0	20.2
4	d_4	2.5	14.5	d_4	2.5	22.7
5	d_5	8.2	22.7	d_3	0.0	22.7

Real system

Ideal system

$$\frac{7}{8.2} = 0.85$$

$$\frac{12}{15.2} = 0.789$$

$$\frac{12}{20.2} = 0.594$$

$$\frac{14.5}{22.7} = 0.639$$

$$\frac{22.7}{22.7} = 1.00$$

Sliding Ratio

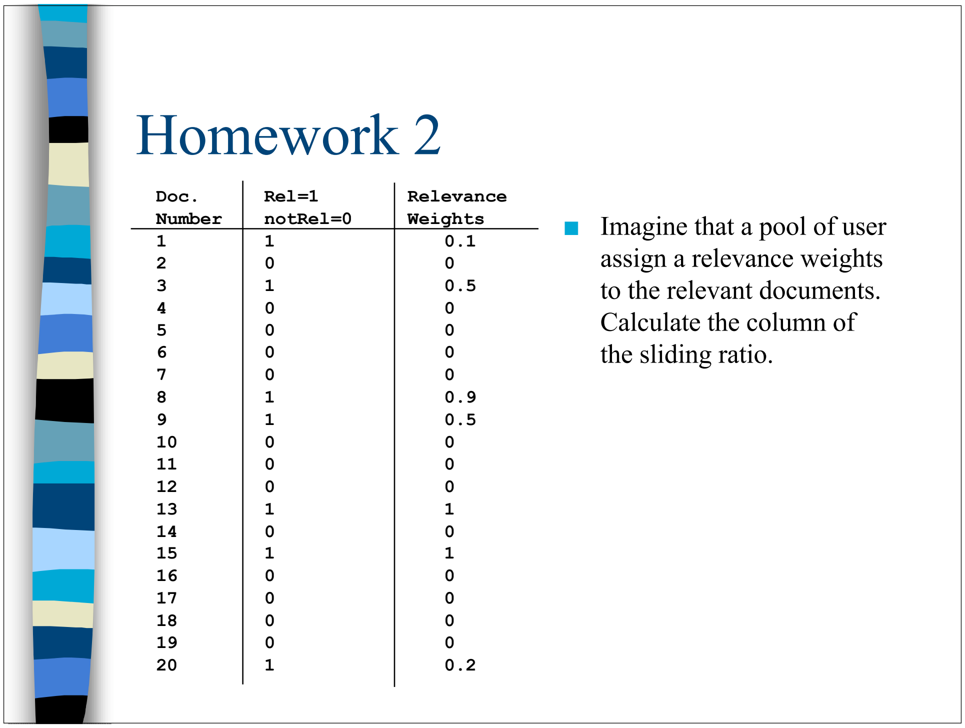
- If the number of retrieved documents N is large enough then SR is a reasonably accurate picture of the retrieval system performances

Homework 1

Imagine that an IR system retrieved 20 document in answer to a query, but only documents number 1, 3, 8, 9, 13, 15, and 20 are relevant.

Calculate Precision, Recall, Fallout and the ratio Recall/Fallout considering that there are other 5 relevant documents that were not retrieved and that the total number of documents in the collection is 100 (included the 20 retrieved).

Explore this problem using graphing applet



Homework 2

Doc.	Rel=1	Relevance
Number	notRel=0	Weights
1	1	0.1
2	0	0
3	1	0.5
4	0	0
5	0	0
6	0	0
7	0	0
8	1	0.9
9	1	0.5
10	0	0
11	0	0
12	0	0
13	1	1
14	0	0
15	1	1
16	0	0
17	0	0
18	0	0
19	0	0
20	1	0.2

- Imagine that a pool of user assign a relevance weights to the relevant documents. Calculate the column of the sliding ratio.