Jul – Nov 2022
IT458

Evaluating IR Systems

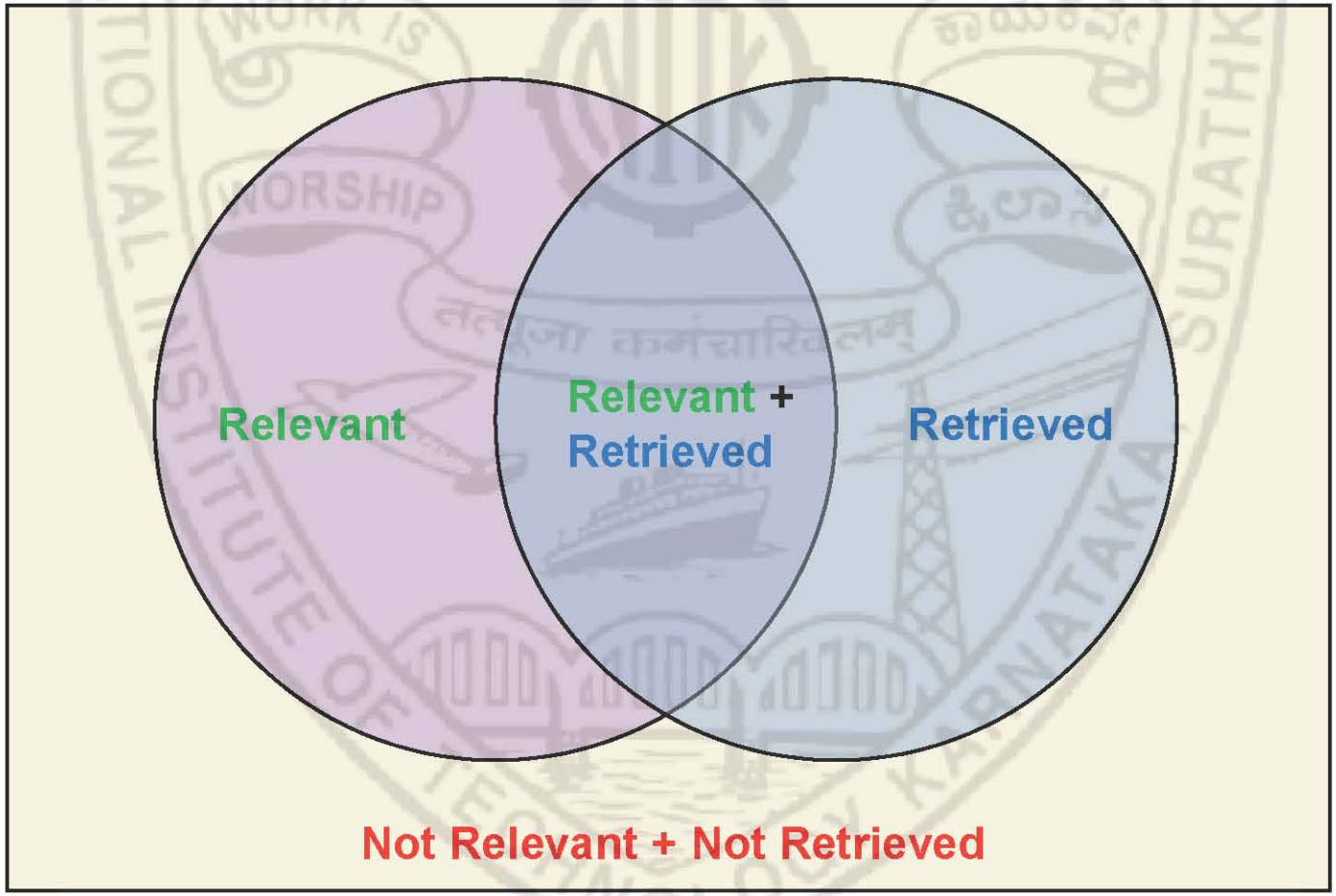
Standard Metrics and Evaluation Techniques

What is relevance?

Relevance is the	measure degree estimate appraisal relation	**of a**	correspondence utility connection satisfaction fit bearing Matching
existing between a	document article textual form reference information provided fact	**and a**	query request information used point of view information requirement
as determined by	person judge user requester Information specialist		

Tefko Saracevic. (1975) Relevance: A Review of and a Framework for Thinking on the Notion in Information Science. Journal of the American Society for Information Science, 26(6), 321-343;

A View of Relevancy...



IR System Evaluation

- A way to measure how well the IR system meets the information needs of its users.

- Process of associating quantitative evaluation metrics to results produced by an IR system.
  - Measuring ***System Relevance*** !!

* Measuring User relevance -- quite challenging as a same result set might be interpreted differently by distinct users.

IR System Evaluation - Multidimensional aspects

- For understanding -
  - How well are the documents **represented**?
  - How **suitable** is the IR model for the given corpus?
  - Was user **information need** captured effectively?
  - Which **ranking** scheme is better?
  - How to **improve** overall system further?
  - How does this IR system compare to other systems, **objectively**?
  - ....

Evaluation Criteria

- Effectiveness
  - System-only, human+system
- Efficiency
  - Retrieval time, indexing time, index size
- Usability
  - Learnability, novice use, expert use

IR Effectiveness Evaluation

- **System-centered strategy**
  - Given **documents, queries** and **relevance judgments**
    - **Try several variations** on the retrieval system
    - **Measure** which model performs better.
      - i.e. **ranks more** good docs near the top

IR Effectiveness Evaluation

- **User-centered strategy**
  - **Given several users**, and **at least 2 retrieval systems**
    - Have each user try the **same task** on both systems
    - **Measure** which system works the "**best**"

Evaluation Techniques

- Types:
  - By inspection of examples
  - by demonstration/improvised demonstration
  - on test data (*benchmarking*)
  - on common test data (*comparative benchmarking*)
  - on common, unseen test data (*learnability*)

Good Measures of Effectiveness of IR Systems

- Capture some aspect of what the user wants
- Have predictive value for other situations
  - Verifiable performance for different queries/document collections.
- Easily compared
  - Optimally, expressed as a single number
- Achieve a meaningful improvement
- Achieve reliable improvement in unseen cases
  - Can be verified using statistical tests

IR Test/Reference Collections

- Representative document collection
  - Size, sources, genre, topics, …
- "Random" sample of representative queries
  - Built somehow from "formalized" topic statements
- Known binary relevance
  - For each topic-document pair (topic, not query!)
  - Assessed by humans, used only for evaluation
- Measure of effectiveness
  - Used to compare alternate systems

IR Test/Reference Collections

- Most used evaluation method in IR.

- A reference collection is composed of:
  - A set D of pre-selected documents
  - A set I of information need descriptions used for testing
  - A set of relevance judgements associated with each pair $[i_m, d_j]$, $i_m \in I$ and $d_j \in D$
    - The relevance judgement has a value of 0 if document d_j is non-relevant to i_m, and 1 otherwise.

IR Test/Reference Collections

- Most used evaluation method in IR.

- A reference collection is composed of:
  - A set ***D*** of pre-selected documents
  - A set ***I*** of information need descriptions used for testing
  - A set of relevance judgements associated with each pair $[i_m, d_j]$, $i_m \in I$ *and* $d_j \in D$
    - The relevance judgement has a value of 0 if document d_j is non-relevant to i_m, and 1 otherwise.

* Relevance judgements are produced by human specialists.

IR Test/Reference Collections

- Some sample **test reference collection** composed **of** documents, queries, and relevance judgements –
  - First testset collection – Cranfield-2 collection
  - Modern open test collections
    - TREC document collection
    - 20 NewsGroups
    - Reuters-21578

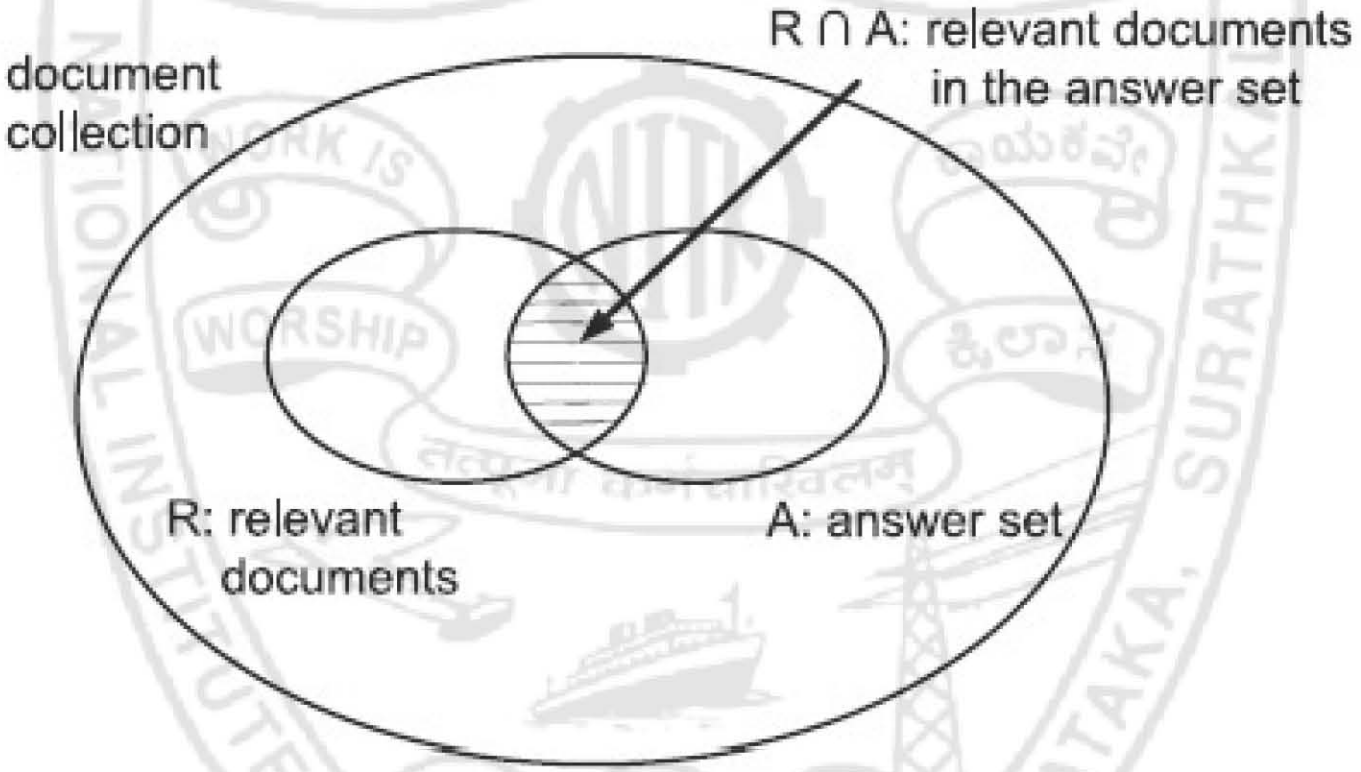
ACM Special Interest Group on Information Retrieval http://sigir.org/resources/

Jul – Nov 2022
IT458

Evaluating IR Systems

Set based Metrics

Set Based Evaluation Metrics



I: an information need/query
R: the set of relevant documents for I
A: answer set for I, generated by an IR system
R ∩ A: the intersection of the sets R and A

Set-Based Metrics

- **Precision**
  - How much of what was found is relevant?
    - *Often of interest, particularly for interactive searching*
- **Recall**
  - How much of what is relevant was found?
    - *Particularly important for law, patents, medical search…*
- **Fallout**
  - How much of what was irrelevant was rejected?
    - *Useful when different size collections are compared*

Set based Metrics - Precision

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Miss
Not relevant	False Alarm	Irrelevant Rejected

- **Precision**:
  - The fraction of documents retrieved that are **relevant** to the user's information need.

$$\text{Precision} = \frac{\text{Relevant} \quad \text{Retrieved}}{\text{Retrieved}}$$

Set based Metrics - Recall

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Miss
Not relevant	False Alarm	Irrelevant Rejected

- **Recall:**
  - fraction of the documents that are **relevant** to the query that are **successfully retrieved**.

$$\text{Recall} = \frac{\text{Relevant} \quad \text{Retrieved}}{\text{Relevant}} = 1 - Miss$$

Set based Metrics - Fallout

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Miss
Not relevant	False Alarm	Irrelevant Rejected

- **Fallout:**
  - Fraction of **non-relevant documents** that are retrieved, out of all non-relevant documents available

$$\text{Fallout} = \frac{\text{Irrelevant Rejected}}{\text{Not Relevant}} = 1 - FA$$

Set based Metrics

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Miss
Not relevant	False Alarm	Irrelevant Rejected

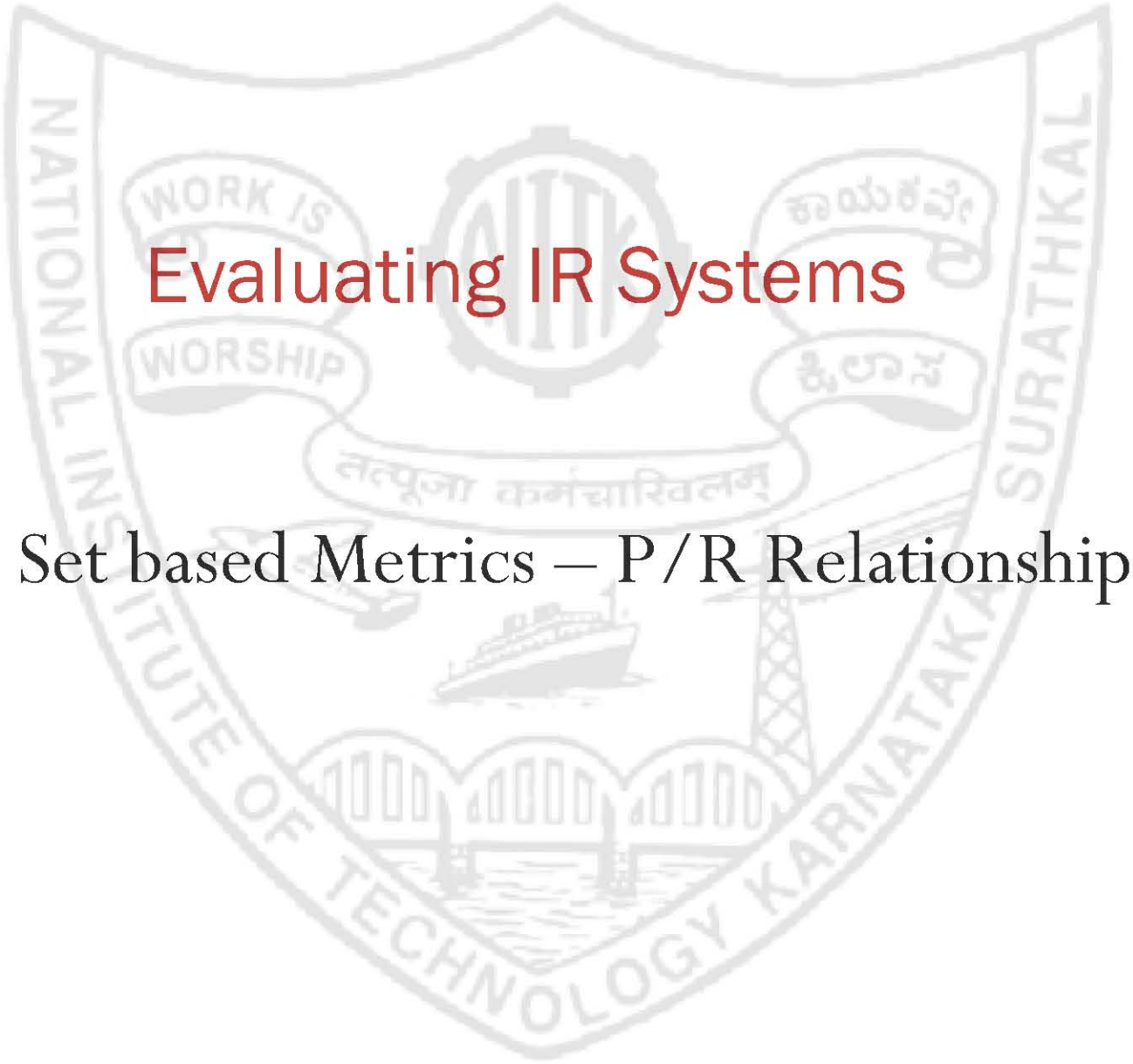
User-Oriented

$$\text{Precision} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Relevant}} = 1 - Miss$$

$$\text{Fallout} = \frac{\text{Irrelevant} \cap \text{Rejected}}{\text{Not Relevant}} = 1 - FA$$

System-Oriented



Evaluating IR Systems

Set based Metrics – P/R Relationship

Precision and Recall - Observations

- P & R - assumption is that all docs in the *retrieved set A* have been examined

- P & R vary as the user proceeds with their examination of the set A.

Solution: represent and analyze P-R relationship as a plot/curve of precision versus recall

P-R relationship – An example

- Consider a reference collection and a set of test queries

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

Set of relevant docs for a query q1

- A new IR algorithm yields the following ranking for q_1

01. d_{123} ●	06. d_9 ●	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} ●	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} ●	15. d_3 ●

(relevant docs are marked with a green bullet)

P-R relationship – An example

$$R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

01. d_{123} ●	06. d_9 ●	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} ●	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} ●	15. d_3 ●

- Observations:
  - document d_{123}, ranked as number 1, is relevant.
    - corresponds to 10% of all relevant documents

    → **precision is 100% at 10% recall**
  - document d_{56}, ranked as number 3, is the next relevant.
    - 2 documents out of 3 are relevant,
    - 2 out of 10 relevant documents have been identified.

    → **precision is 66.6% at 20% recall**

    .

    .
  - Compute precision upto 100% recall

P-R relationship – An example

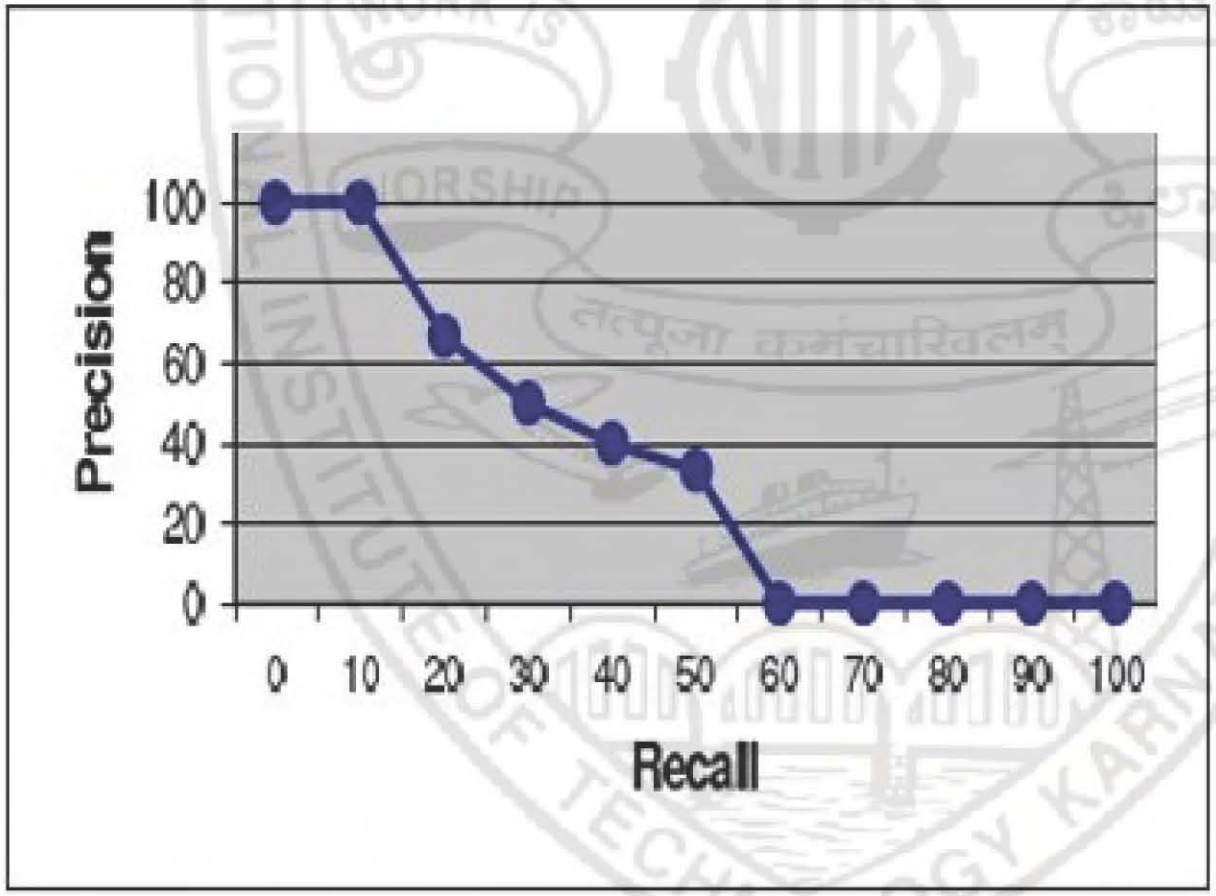
- Finally, a plot of all such values is plotted, i.e. the P-R curve
  - Called the **11-point interpolated average precision**.
    - P& R @ 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

      +

      P&R @ 0

P-R relationship – An example

- **The P-R curve as a 11-point interpolated plot**



Recall	Precision
0	100
10	100
20	66.6
30	50
40	40
50	33.3
60	0
70	0
80	0
90	0
100	0

Exercise

- Consider a reference collection and a set of test queries
  - set of relevant docs for a query q_2:

$$R_{q_2} = \{d_3, d_{56}, d_{129}\}$$

  - The previous IR algorithm processes query q2 and returns a ranking, as above.

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} ●	08. d_{129} ●	13. d_{810}
04. d_{32}	09. d_4	14. d_5
05. d_{124}	10. d_{130}	15. d_3 ●

(relevant docs are marked with a green bullet)

  - Plot the PR graph for this algorithm.

Precision and Recall - Summary

- Extensively used to evaluate retrieval performance of IR algorithms.
  - Also adopted for other domains – ML, Data mining…
- Some problems with these two measures:
  - proper estimation of maximum recall for a query requires detailed knowledge of all the documents in the collection.
    - Solution: Use *relative recall*.

Jul – Nov 2022
IT458

Evaluating IR Systems

Range based Metrics

Range based metrics

- Web searches
  - very high recall – *often unnecessary for users*
  - More relevant docs at top of the ranking the better
- *Solution:* Precision @ K
- *Commonly used* –
  - Precision at 5 *(P@5)*
  - Precision at *10 (P@10)*

P@5 and P@10 – An example

- Consider again the ranking for a sample query q1 (used earlier)

01. d_{123} ●	06. d_9 ●	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} ●	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} ●	15. d_3 ●

- Here, P@5 = 40%

  P@10 = 40%

* P@5 and P@10 are usually computed and averaged over a sample of 100 queries.

Precision@k and Recall@k

- Together used to evaluate quality of ranking.
- P@K
  - fraction of retrieved top-K documents that are **relevant**
- R@K
  - fraction of **relevant** documents that are retrieved in the top-K

* Assumption: the user will only examine the top-K results

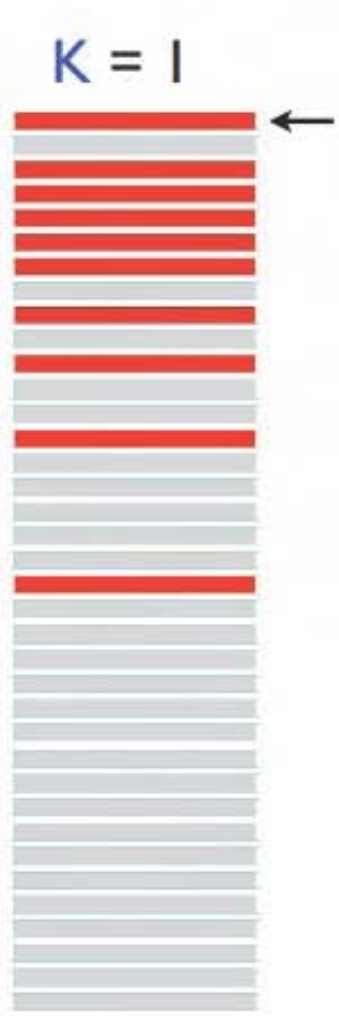
P/R @ *K* - Insights

- Advantages:
  - easy to compute
  - easy to interpret

- Issues:
  - the value of *K* has a huge impact on the metric.
    - the ranking or performance of the system beyond chosen *K* is inconsequential .
  - how do we pick *K*?

P/R @K - Observations

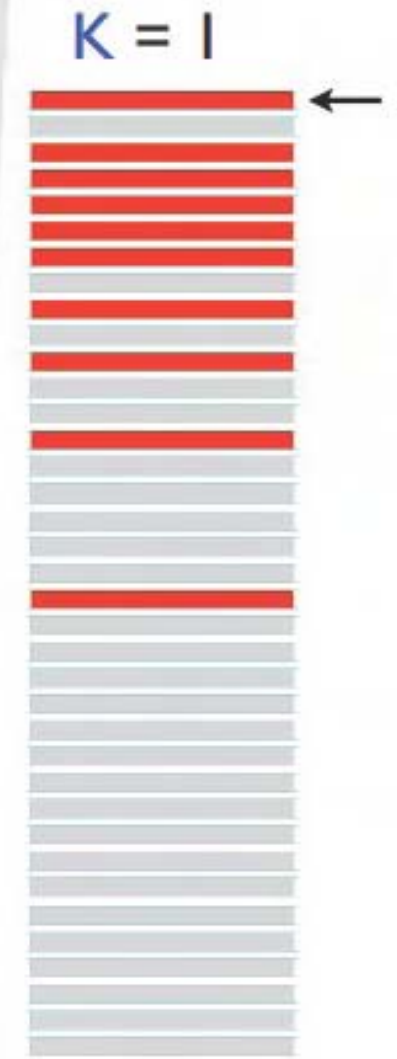
- Assume that there are 20 relevant documents for a query. We want to assess the range based performance of this system.
- The IR system generated ranked list is shown :
  - Relevant docs at positions – 1, 3, 4, 5, 6, 7, 9, 11, 13, 19.



P/R @K - Observations

Relevant docs at positions – 1, 3, 4, 5, 6, 7, 9, 11, 13, 19.

Kvalue	Precision@K	Recall@K
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

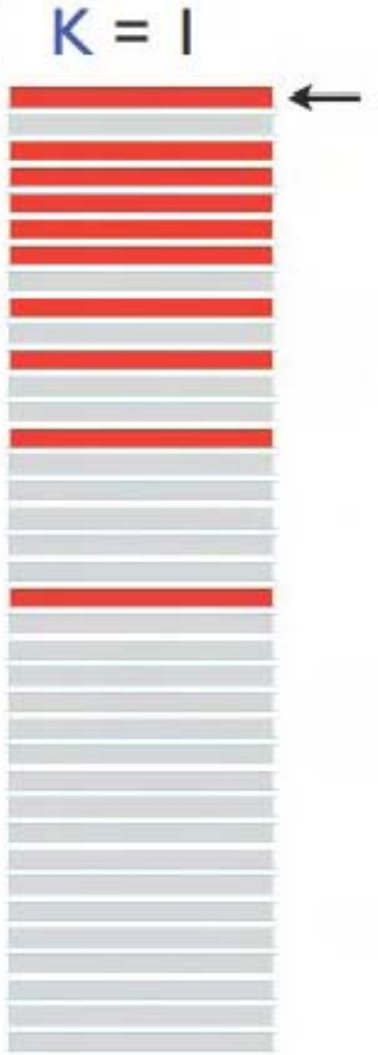


P/R @K - Observations

- At K = 1

Relevant docs at positions – 1, 3, 4, 5, 6, 7, 9, 11, 13, 19.

K value	Precision@K	Recall@K
1	1/1 = 1.0	1/20 = 0.05
2		
3		
4		
5		
6		
7		
8		
9		
10		

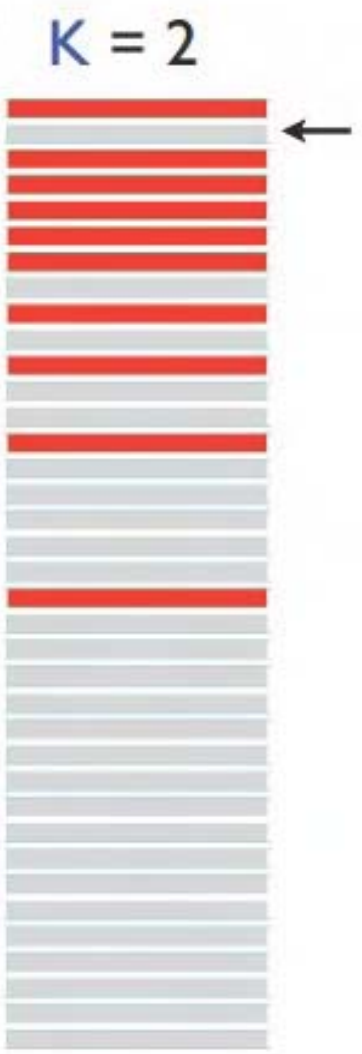


P/R @K - Observations

- At K = 2

Relevant docs at positions – 1, 3, 4, 5, 6, 7, 9, 11, 13, 19.

K value	Precision@K	Recall@K
1	1/1 = 1.0	1/20 = 0.05
2	½ = 0.5	1/20 = 0.05
3		
4		
5		
6		
7		
8		
9		
10		



P/R @K - Observations

- At K = 3

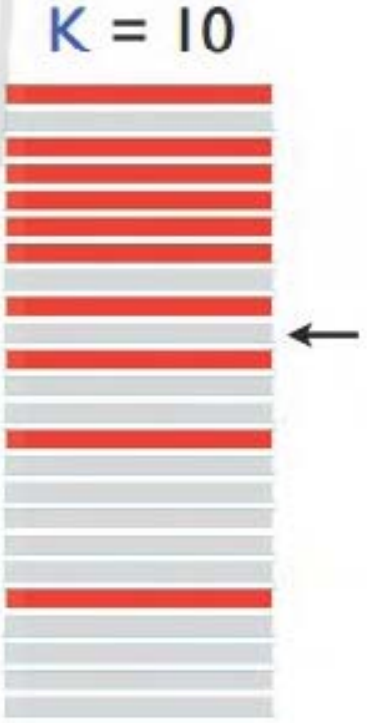
Relevant docs at positions – 1, 3, 4, 5, 6, 7, 9, 11, 13, 19.

Kvalue	Precision@K	Recall@K
1	1/1 = 1.0	1/20 = 0.05
2	½ = 0.5	1/20 = 0.05
3	2/3=0.67	2/20 = 0.01
4		
5		
6		
7		
8		
9		
10		

P/R @K - Observations

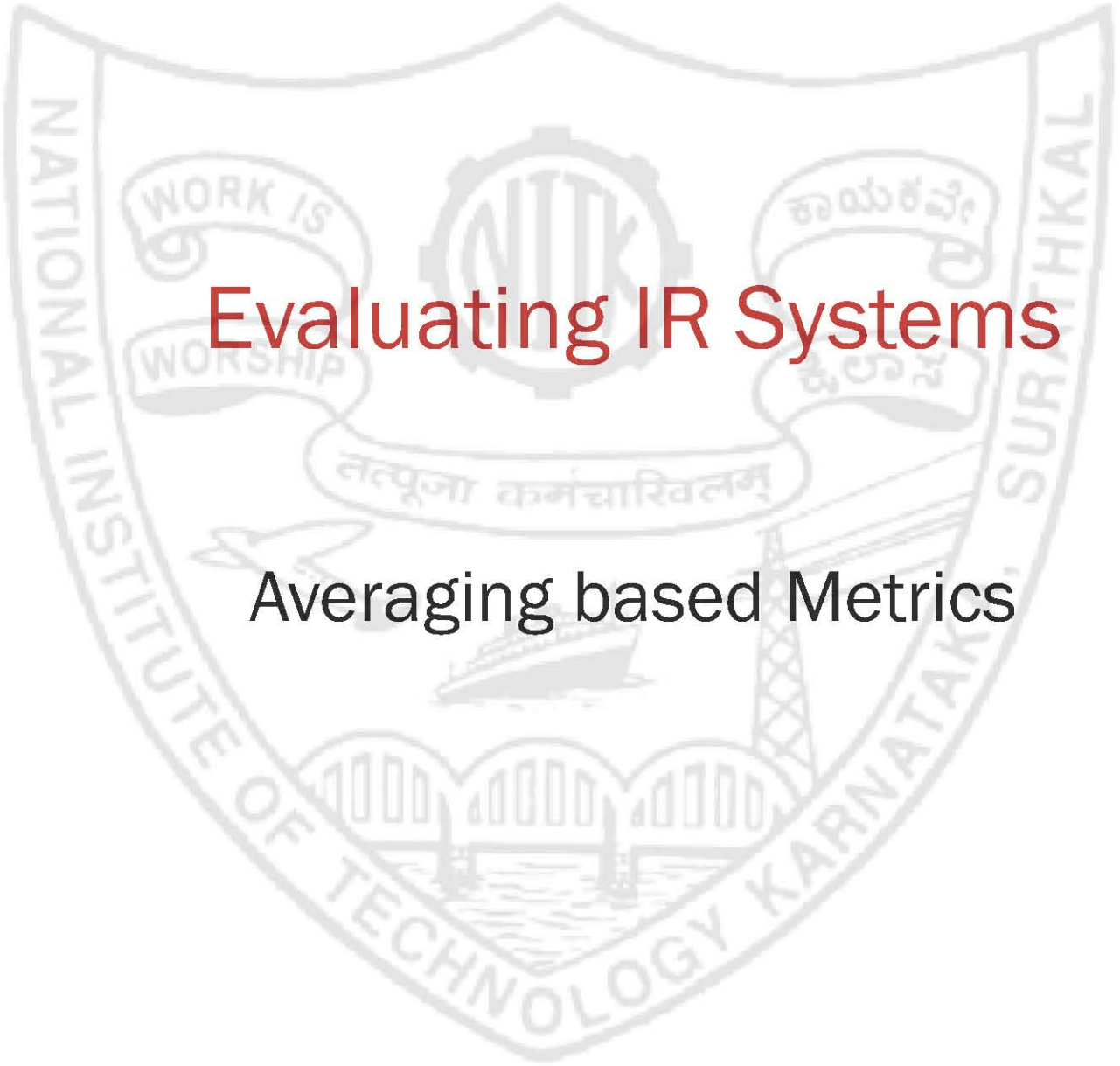
- At K = 4, 5, 6, 7, 8, 9, 10

K value	Precision@K	Recall@K
1	1/1 = 1.0	1/20 = 0.05
2	½ = 0.5	1/20 = 0.05
3	2/3=0.67	2/20 = 0.01
4	3/4= 0.75	3/20 = 0.15
5	4/5 = 0.8	4/20 = 0.2
6	5/6 = 0.83	5/20 = 0.25
7	6/7 = 0.86	6/20 = 0.3
8	6/8 = 0.75	6/20 = 0.3
9	7/9 = 0.78	7/20 = 0.35
10	7/10 = 0.7	7/20 = 0.35



P/R @K - Observations

- Forthe IR algorithm being tested,
  - Top-5 ranking performance is worser than ranking in the range 6-10 in the ranked list.
  - Best performance is in the mid ranges of the ranked list.
  - Overall recall is promising as 7 relevant docs were in the top-10 range.
  - ....



Evaluating IR Systems

Averaging based Metrics

Averaging Precision/Recall

- accounts for precision and recall without having to set the value K.
- roughly equal to the **average Area under the PR curve (AUPRC)** plotted for a predefined set of queries.

Average Precision/Recall

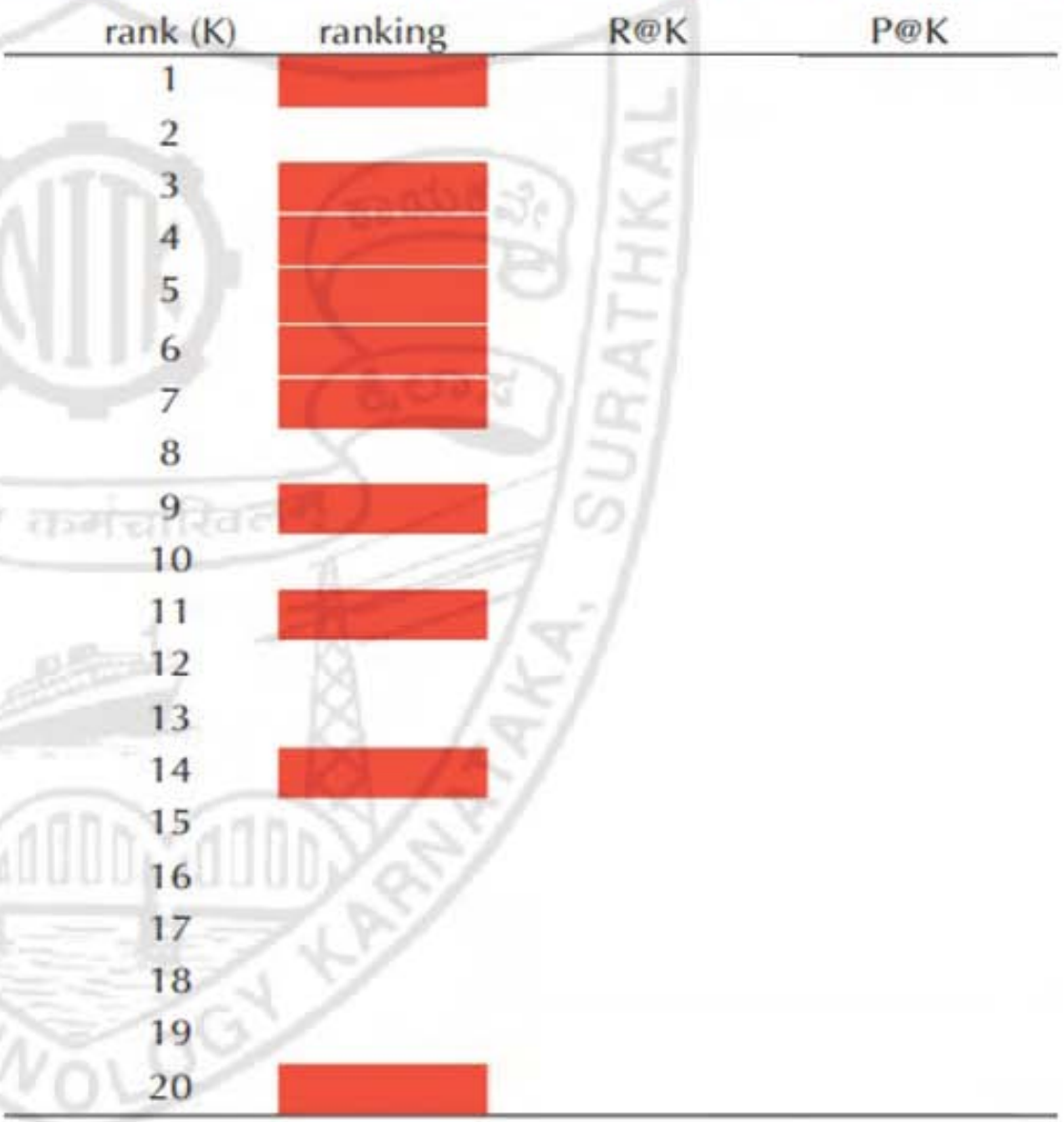
- *Process:*
  - Examine the generated ranking one-rank-at-a-time
  - If the document at rank *K* is relevant, measure P@*K*
    - Gives fraction of top-*K* documents that are relevant
  - Finally, when recall=100%, take an average of all P@*K* values
    - the number of P@*K* values will equal the number of relevant documents.

Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

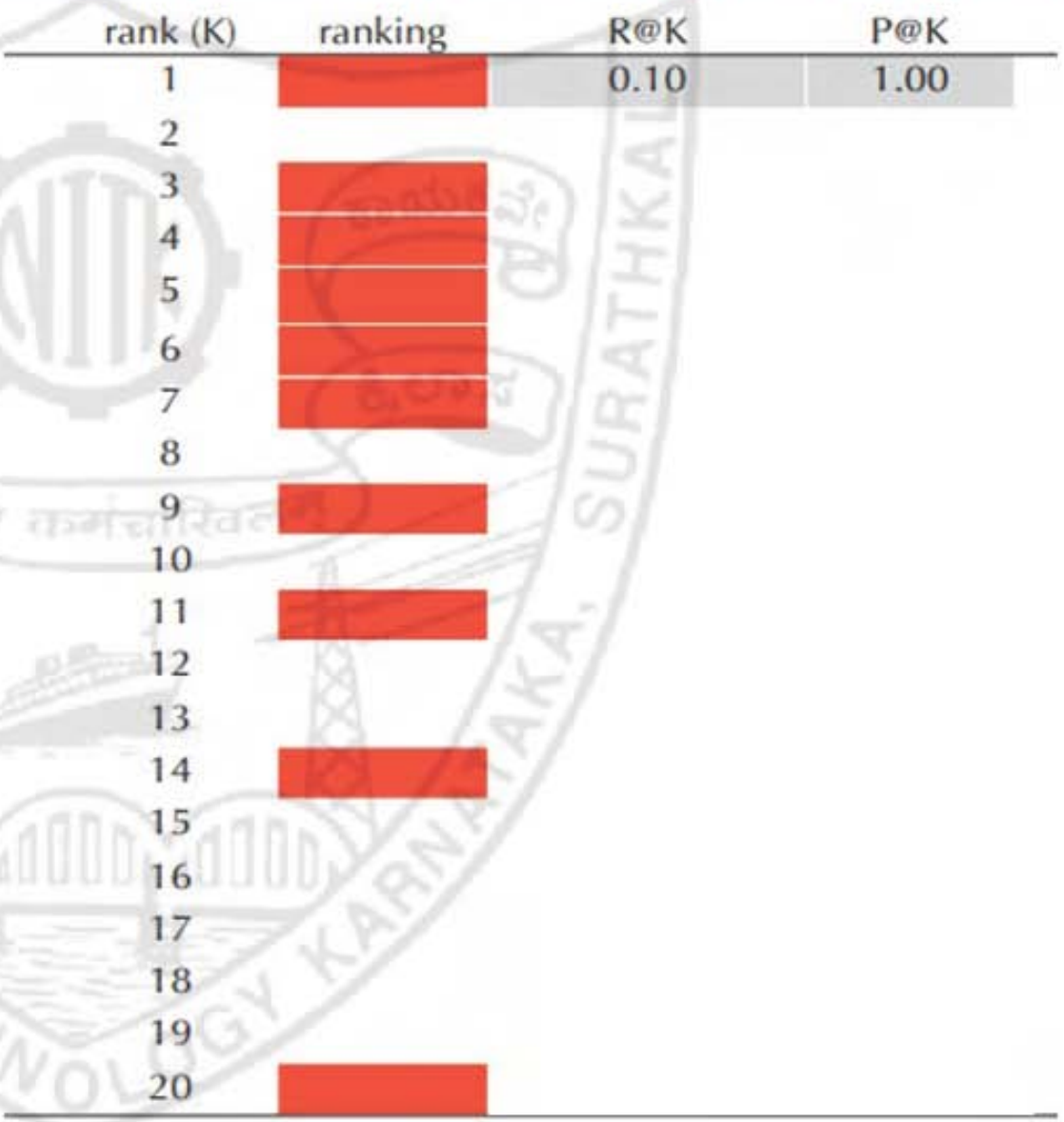


Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

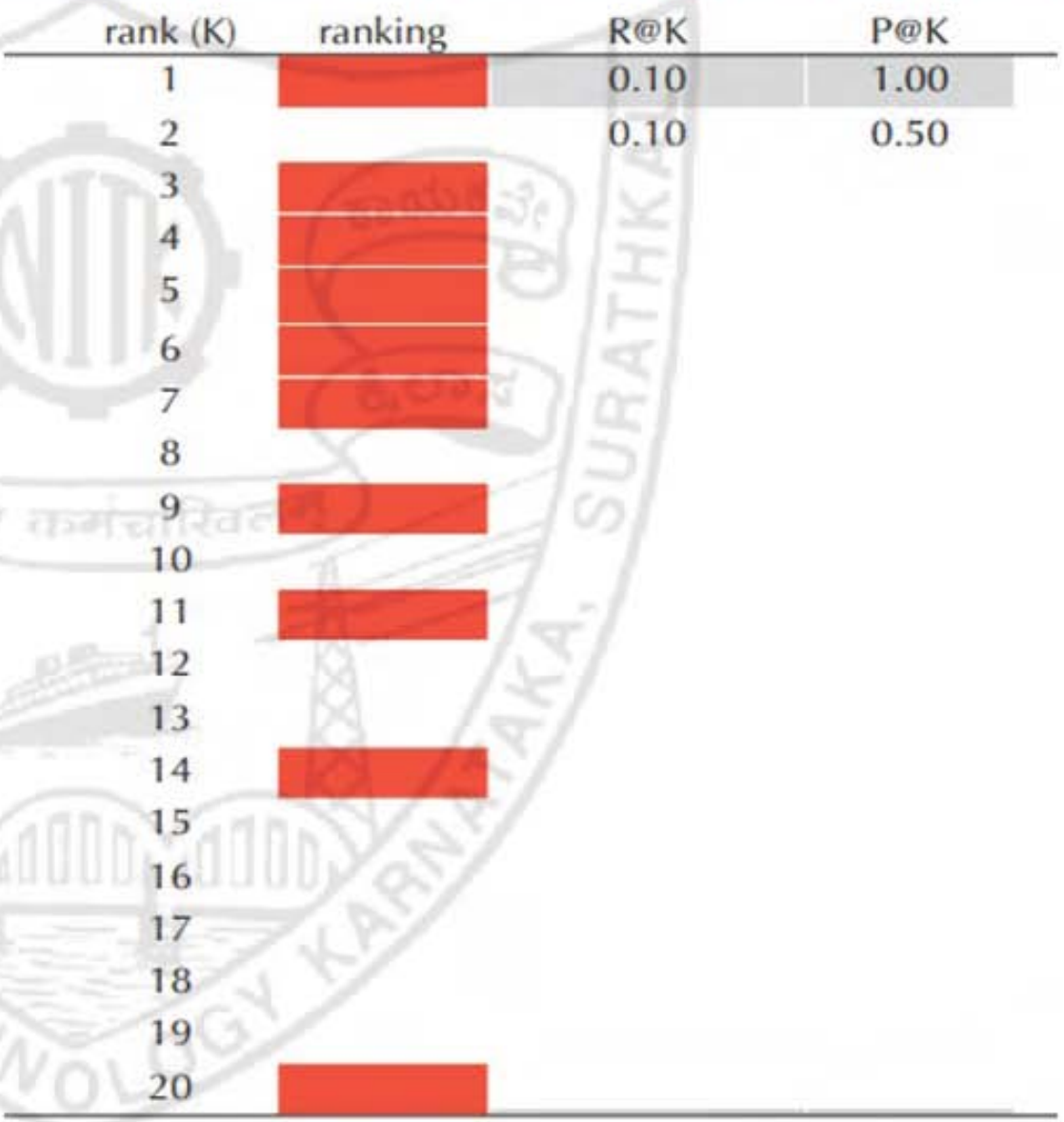


Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

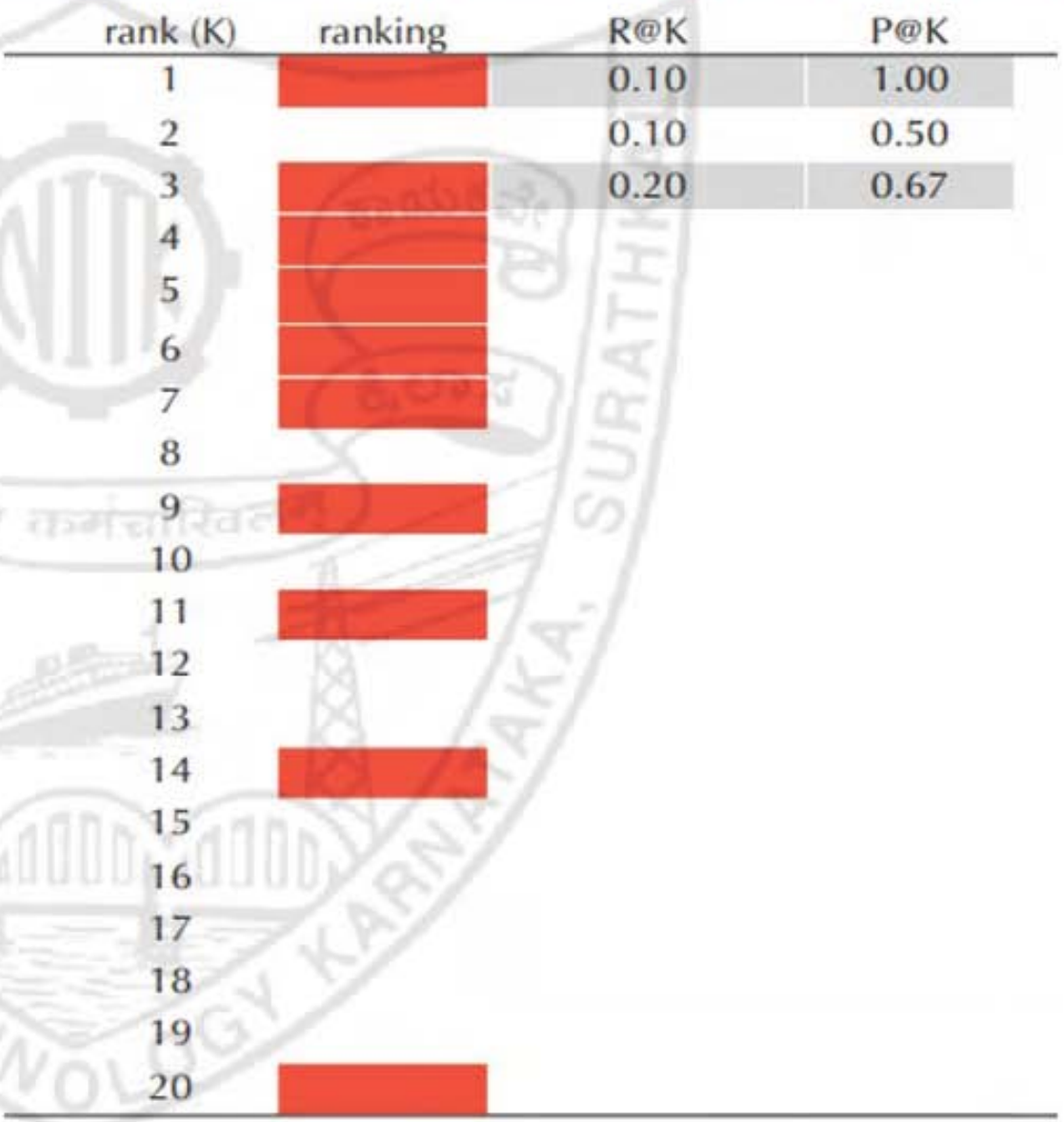


Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

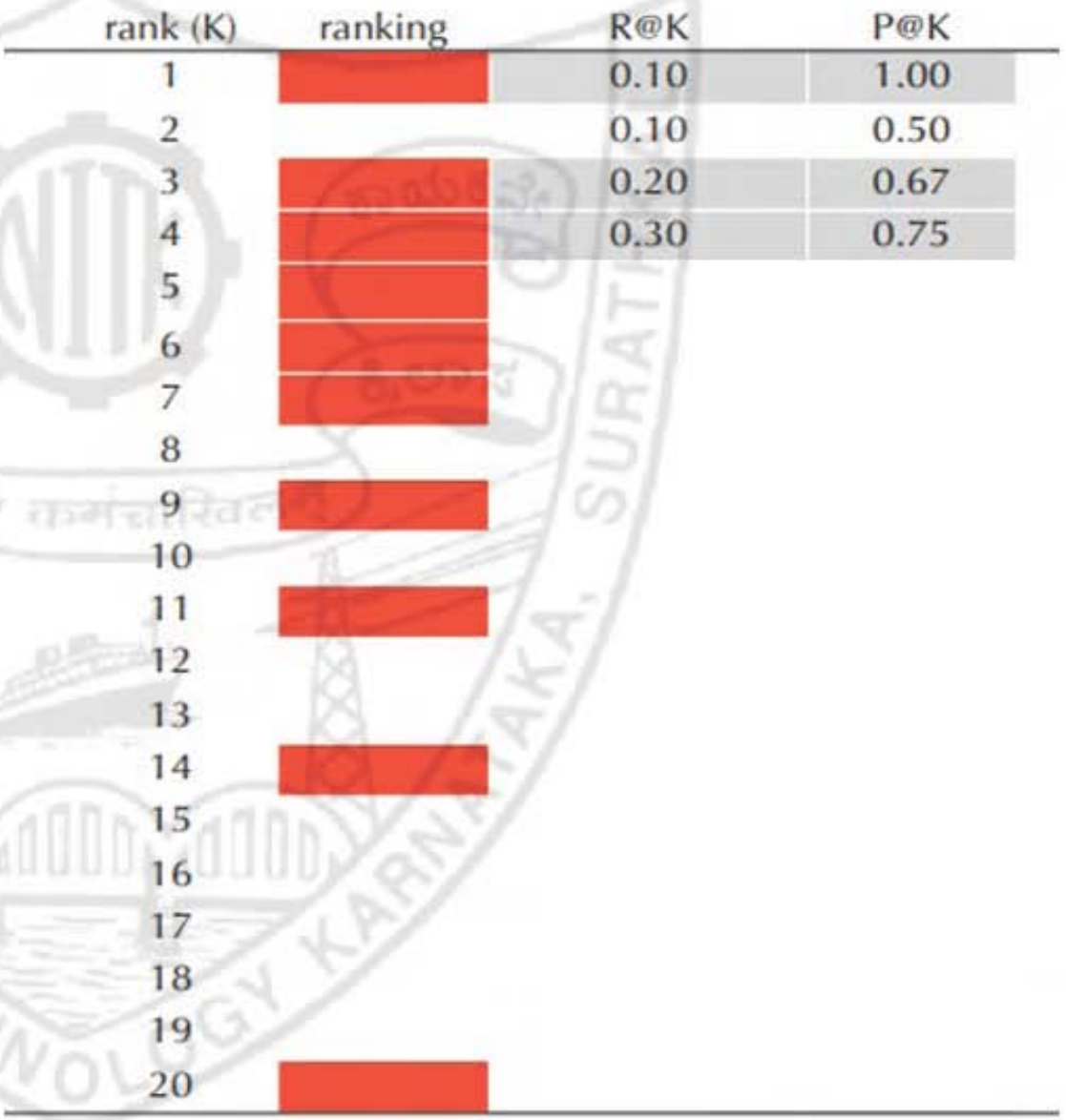


Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values



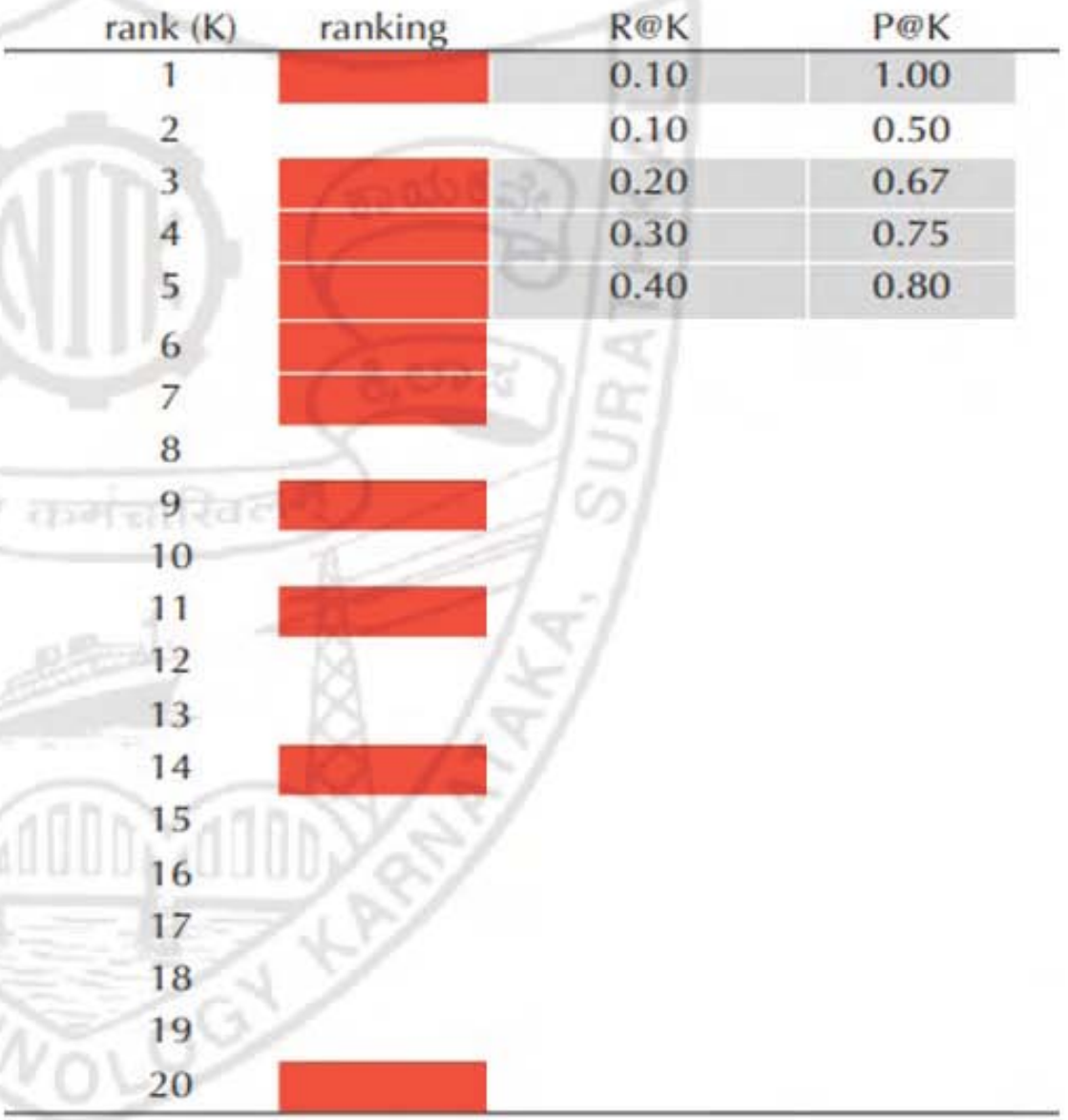
rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Average Precision - Example

- No.of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values



rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Average Precision - Example

- No. of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50

Average Precision - Example

- No. of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@*K* at that rank *K*
3. When recall = 1.0, average the P@*K* values

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50

Average Precision - Example

- No. of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50

Average Precision - Example

- No. of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@K at that rank K
3. When recall = 1.0, average the P@K values

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50

Average Precision - Example

- No. of relevant docs = 10

- Process -

1. Examine the ranking one-rank-at-a-time
2. If recall goes up, calculate P@*K* at that rank *K*
3. When recall = 1.0, average the P@*K* values

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.76

Average Precision - Insights

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.30	1.00
4		0.40	1.00
5		0.50	1.00
6		0.60	1.00
7		0.70	1.00
8		0.80	1.00
9		0.90	1.00
10		1.00	1.00
11		1.00	0.91
12		1.00	0.83
13		1.00	0.77
14		1.00	0.71
15		1.00	0.67
16		1.00	0.63
17		1.00	0.59
18		1.00	0.56
19		1.00	0.53
20		1.00	0.50
total	10.00	average-precision	1.00

Average Precision - Insights

rank (K)	ranking	R@K	P@K
1		0.00	0.00
2		0.00	0.00
3		0.00	0.00
4		0.00	0.00
5		0.00	0.00
6		0.00	0.00
7		0.00	0.00
8		0.00	0.00
9		0.00	0.00
10		0.00	0.00
11		0.10	0.09
12		0.20	0.17
13		0.30	0.23
14		0.40	0.29
15		0.50	0.33
16		0.60	0.38
17		0.70	0.41
18		0.80	0.44
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.33

Average Precision - Insights

- Effect on AP if ranks 2 and 3 are swapped

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.20	1.00
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.60	0.75
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.79

Average Precision - Insights

- Effect on AP if ranks 8 and 9 are swapped

rank (K)	ranking	R@K	P@K
1		0.10	1.00
2		0.10	0.50
3		0.20	0.67
4		0.30	0.75
5		0.40	0.80
6		0.50	0.83
7		0.60	0.86
8		0.70	0.88
9		0.70	0.78
10		0.70	0.70
11		0.80	0.73
12		0.80	0.67
13		0.80	0.62
14		0.90	0.64
15		0.90	0.60
16		0.90	0.56
17		0.90	0.53
18		0.90	0.50
19		0.90	0.47
20		1.00	0.50
total	10.00	average-precision	0.77

Average Precision - Summary

- Observations:
  - no need to choose K
  - accounts for both precision and recall
  - ranking mistakes at the top of the ranking are more influential
  - ranking mistakes at the bottom of the ranking are still accounted for
- Issues
  - not quite as easy to interpret as P/R@K

R-Precision

- useful for observing the behaviour of an algorithm for individual queries.
  - requires knowing all documents that are relevant to a query.

R-Precision

- Let R be the total number of relevant docs for a given query.
  - Examine the top ***R*** results of a system (R = relevant docs for a query)
  - Find those ***r*** docs which were returned as relevant.
  - Then,

$$\text{R-precision}^* = \frac{r}{R}$$

* Also gives the **recall** of this result set

R-Precision – An example

- Consider again the IR algorithm with the following ranking for q_1.

$$R_{q1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

01. d_{123} ●	06. d_9 ●	11. d_{38}
02. d_{84}	07. d_{511}	12. d_{48}
03. d_{56} ●	08. d_{129}	13. d_{250}
04. d_6	09. d_{187}	14. d_{113}
05. d_8	10. d_{25} ●	15. d_3 ●

- For the query q1, R = 10

  r in 10 = 4

  - R-Precision for q1 = 0.4

R-Precision – Another example

- Consider again the IR algorithm with the following ranking for q_2.

$$R_{q_2} = \{d_3, d_{56}, d_{129}\}$$

01. d_{425}	06. d_{615}	11. d_{193}
02. d_{87}	07. d_{512}	12. d_{715}
03. d_{56} ●	08. d_{129} ●	13. d_{810}
04. d_{32}	09. d_4	14. d_5
05. d_{124}	10. d_{130}	15. d_3 ●

- For the query q2, R = 3

$$r \text{ in } 3 = 1$$

  - R-Precision for q2 = 0.33