Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City

Thou shalt not compute MRR nor ERR. Thou shalt not use MAP.
– Nobert Fuhr.

Evaluation

How to compare Search Engines?
How good is an IR system?

- Various evaluation methods
  - Precision/Recall
  - Mean Average Precision
  - Mean Reciprocal Rank
    - If first relevant doc is at kth position, RR = 1/k.
  - NDCG
    - Non-Boolean/Graded relevance scores
    - $DCG = r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$

Precision and Recall

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

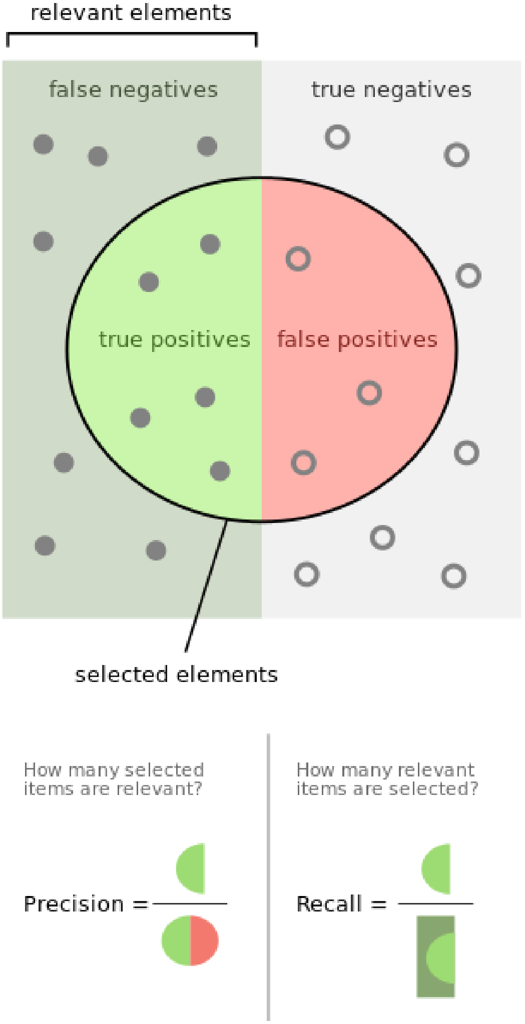


Image Source: Wikipedia

Precision and Recall

- An IR system retrieves the following 20 documents.
- There are 100 relevant documents in our collection.
- Hollow squares represent irrelevant documents.
- Solid squares with ‘R’ are relevant.

	R	R		R			R		
			R	R	R	R			

- What is Precision?
- What is Recall?

Precision and Recall

- An IR system retrieves the following 20 documents.
- There are 100 relevant documents in our collection.
- Hollow squares represent irrelevant documents.
- Solid squares with ‘R’ are relevant.

	R	R		R			R		
			R	R	R	R			

- What is Precision? Precision = 8/20.
- What is Recall? Recall = 8/100.

Can we do better?

Can we have one number to express quality?

A minor deviation ahead!

F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

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

Harmonic Mean'aaa?

Arithmetic Mean

- What is the arithmetic mean of:
  - 1,2,3
  - 1,2,3,4,5
  - 1,2,3,4,5,6,7
- What is the arithmetic mean of:
  - 1 … 99

Answer: $\frac{1}{n}\sum_{n=1}^{10} n \ = \frac{1}{n} \,.\, \frac{n(n+1)}{2} = \frac{99.100}{99.2} = 50$

Arithmetic Mean

- What is the arithmetic mean of:
  - 7,8,9 ?
  - 11,13,15?
- What is the arithmetic mean of:
  - 1, 9, 10
    - 6.7
  - 1, 8, 10
    - 6.3
  - 1, 7, 10
    - 6

Geometric Mean

- What is the geometric mean of 2 and 8 ?
- Answer: $\sqrt{2.8} = \sqrt{16}$ = 4. (Arithmetic Mean is $\frac{2+8}{2}$ = 5.)

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Geometric Mean

- What is the geometric mean of:
  - 7,8,9 ? AM=8, GM=7.96
  - 11,13,15? AM=13, GM=12.89
- What is the geometric mean of:
  - 1, 9, 10
    - AM=6.7, GM=4.48
  - 1, 8, 10
    - AM=6.3, GM=4.31
  - 1, 7, 10
    - AM=6, GM=4.1

Quiz

Which computer will you prefer?

	Computer A	Computer B	Computer C
Program 1	1	10	20
Program 2	1000	100	20

Time taken by two programs to execute on different computers.

Quiz

Which computer will you prefer?

	Computer A	Computer B	Computer C
Program 1	1	10	20
Program 2	1000	100	20

	A	B	C
Prg. 1	1	10	20
Prg. 2	1	0.1	0.02
A. Mean	**1**	**5.05**	**10.01**
G. Mean	**1**	**1**	**0.63**

	A	B	C
Prg. 1	0.1	1	2
Prg. 2	10	1	0.2
A. Mean	**5.05**	**1**	**1.1**
G. Mean	**1**	**1**	**0.63**

	A	B	C
Prg. 1	0.05	0.5	1
Prg. 2	50	5	1
A. Mean	**25.03**	**2.75**	**1**
G. Mean	**1.581**	**1.58**	**1**

Geometric Mean gives a consistent ranking
for normalized values.

Harmonic Mean

- What is the harmonic mean of 2 and 8 ?
- Answer: $\frac{2}{\frac{1}{2}+\frac{1}{8}} = 3.2$

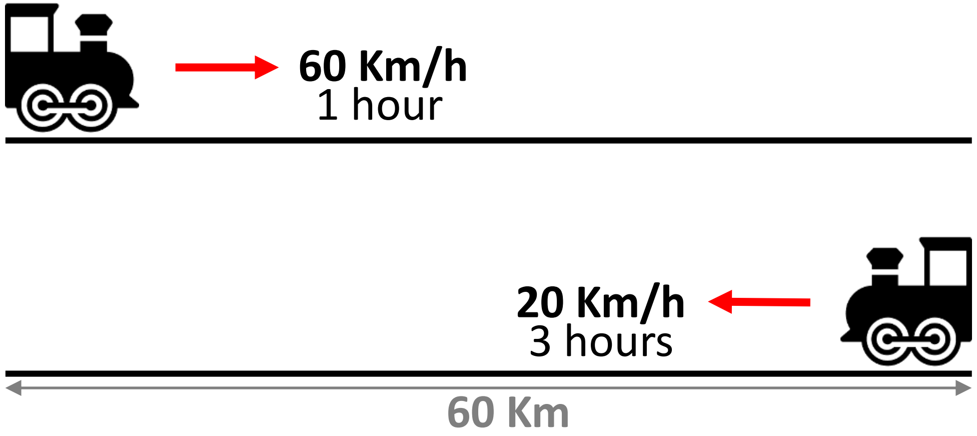
$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Harmonic Mean

- What is the harmonic mean of:
  - 7,8,9 ? AM=8, GM=7.96, HM=7.92
  - 11,13,15? AM=13, GM=12.89, HM=12.79
- What is the harmonic mean of:
  - 1, 9, 10
    - AM=6.70, GM=4.48, HM=2.48
  - 1, 8, 10
    - AM=6.30, GM=4.31, HM=2.45
  - 1, 7, 10
    - AM=6.00, GM=4.10, HM=2.41

Quiz

- Can you compute the average speed?



Compute AM, GM and HM of 60 and 20

AM = 40, GM = 63.25, HM = 30

Precision and Recall

Why Harmonic Mean for PR?

Precision and Recall

F1-Score
A Mean for Precision and Recall

$$F_1 = \frac{2\,PR}{P + R}$$

A more generalized formula:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

See “The truth of the F-measure” for a detailed discussion.
https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf

Compute Precision

- Case 1:

1	2	3	4	5	6	7	8	9	10
	R	R		R			R		
			R	R	R	R			

- Case 2:

R	R	R	R	R	R	R	R		

Compute Precision and Recall

- Case 1: Precision = 8/20, Recall = 8/100

	R	R		R			R		
			R	R	R	R			

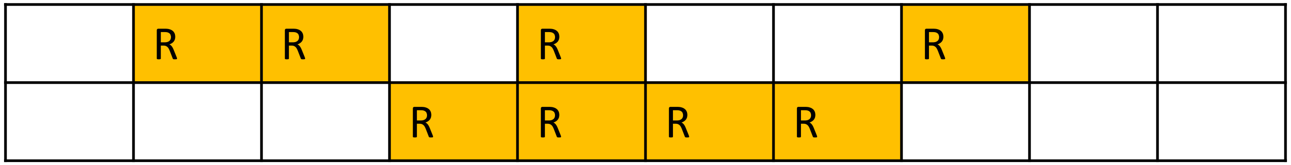
- Case 2: Precision = 8/20, Recall = 8/100

R	R	R	R	R	R	R	R		

Which IR system will you prefer?

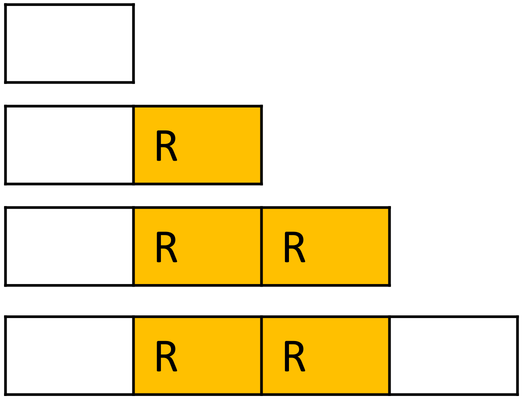
Precision@k

- We cut-off results at k and compute precision.



	R	R		R			R		
			R	R	R	R			

- P@1 = 0
- P@2 = ½
- P@3 = 2/3
- P@4 = 2/4



	R		
	R	R	
	R	R	

Average Precision

- We cut-off results at k and compute precision.

	R	R		R			R		
			R	R	R	R			

- P@1 = 0

- P@2 = ½

	R

- P@3 = 2/3

	R	R

- P@4 = 2/4

	R	R	

Average Precision = (0 + 0.5 + 0.66) / 2 = 0.58
(if we are only interested in 2 levels of relevance)

What is the Average Precision?

- Case 1:

	R		R		R		R		R

  - Average of Precision at each relevance level. Average the P@k for each relevance level. Ignore P@k at non-relevant positions.
  - Average Precision = $\frac{\tfrac{1}{2}+\tfrac{1}{2}+\tfrac{1}{2}+\tfrac{1}{2}+\tfrac{1}{2}}{5}$
- Case 2:

		R			R			R

- Average Precision = ?

What is the Average Precision?

- Case 1:

	R		R		R		R		R

  - Average Precision = $\frac{\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}}{5}$

- Case 2:

		R			R			R

- Average Precision = 1/3

Mean Average Precision

MAP computes Average Precision for all relevance levels for a set of queries.

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

What is the Average Precision?

- Case 1:

	R		R		R		R		R

  - Average Precision = $\frac{\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}}{5}$
  - If there were 10 relevant documents, and we retrieved only five,
    - MAP = $\frac{\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 0 + 0 + 0 + 0 + 0}{10}$
- Case 2:

		R			R			R

- What is AP and MAP? Assume there were only 3 relevant documents in our collection.

Compute MAP

- Query1:

	R		R		R		R		R

Only 5 relevant docs in corpus.

- Query2:

		R			R			R

- Query3:

		R			R			R

Only 3 relevant docs in corpus.

Compute MAP

- Query1:

	R		R		R		R		R

Only 5 relevant docs in corpus.

- Query2:

		R			R			R

- Query3:

		R			R			R

Only 3 relevant docs in corpus.

- Compute MAP.

$$MAP = (1/2 + 1/3 + 1/3)/3$$

Quiz

- Can you compute MAP if you do not know the total number of relevant results for any given query?
  - No! This is the case with web search. Judges may not know how many relevant documents exist.

How to compare two systems, if results are ranked and graded?

and we do not know the total number of relevant documents

Discounted Cumulative Gain

$$DCG_k = \sum_{r=1}^{k} \frac{rel_r}{\log(r+1)}$$

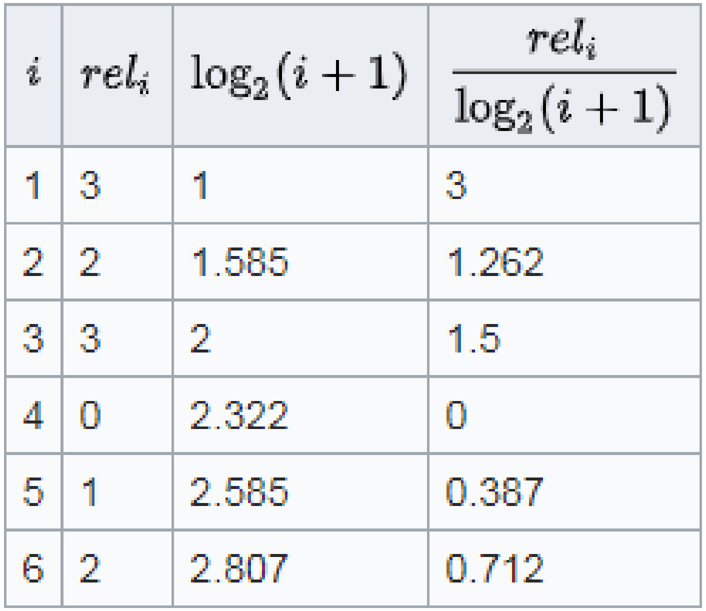
DCG_k = DCG at position k
r = rank
rel_r = graded relevance of the result at rank r

DCG Example

- Presented with a list of documents in response to a search query, an experiment participant is asked to judge the relevance of each document to the query. Each document is to be judged on a scale of 0-3 with:
  - 0 meaning not relevant,
  - 3 meaning highly relevant, and
  - 1 and 2 meaning "somewhere in between".

DCG Example

- Compute DCG



i	rel_i	$\log_2(i+1)$	$\frac{rel_i}{\log_2(i+1)}$
1	3	1	3
2	2	1.585	1.262
3	3	2	1.5
4	0	2.322	0
5	1	2.585	0.387
6	2	2.807	0.712

$$\mathrm{DCG}_6 = \sum_{i=1}^{6} \frac{rel_i}{\log_2(i+1)} = 3 + 1.262 + 1.5 + 0 + 0.387 + 0.712 = 6.861$$

Which system is better?

- 3,2,3,0,1,2 or
- 3,3,3,2,2,2

Results from System 1				Results from System 2		
rel_i	$\text{log}_2\text{(i+1)}$	$\frac{rel_i}{log_2(i+1)}$		rel_i	$\text{log}_2\text{(i+1)}$	$\frac{rel_i}{log_2(i+1)}$
3.00	1.00	3.00		3.00	1.00	3.00
3.00	1.58	1.89		2.00	1.58	1.26
3.00	2.00	1.50		3.00	2.00	1.50
2.00	2.32	0.86		0.00	2.32	0.00
2.00	2.58	0.77		1.00	2.58	0.39
2.00	2.81	0.71		2.00	2.81	0.71
		8.74				**6.86**

Which system is better?

- 3,2,3,0,1,2 or
- 3,3,3,2,2,2,1,0

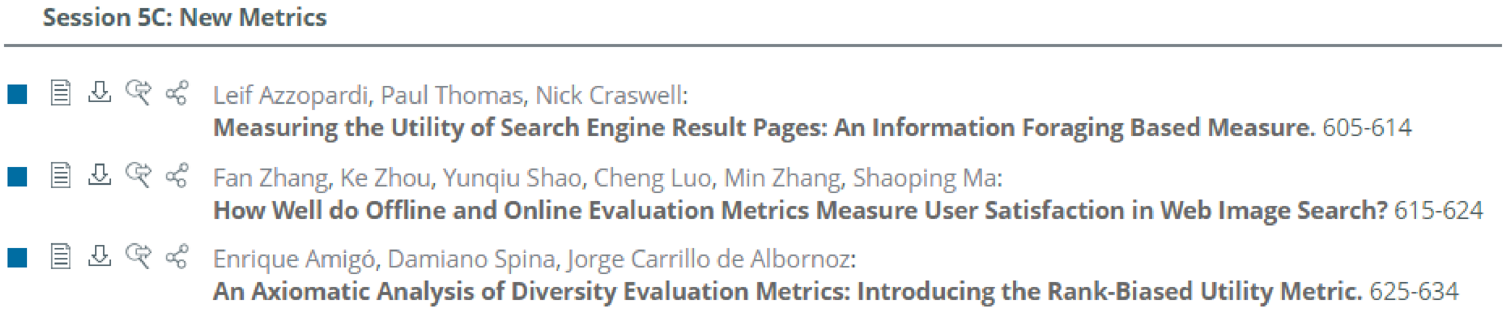
What if there are unequal number of documents?

- Ideal DCG at 6 is (the best value) DCG for 3,3,3,2,2,2
- Normalize DCG with Ideal DCG value.
- NDCG for System 1 = DCG/IDCG = 0.785.
- NDCG for System 2 = 1.

For a set of queries Q, we average the NDCG.

A Rich Area for Research

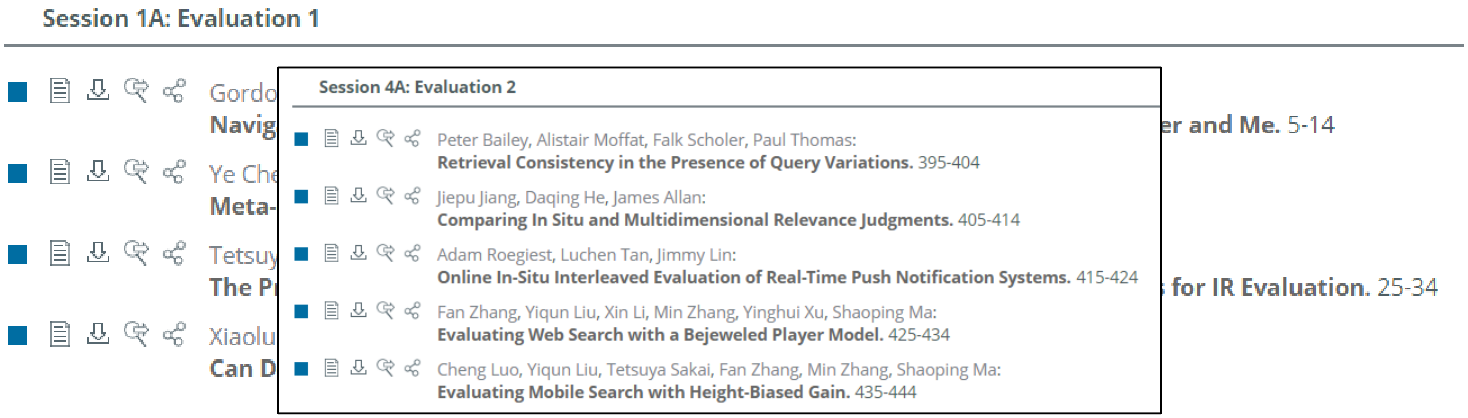
SIGIR 2018



Session 5C: New Metrics

- Leif Azzopardi, Paul Thomas, Nick Craswell:
  Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. 605-614
- Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, Shaoping Ma:
  How Well do Offline and Online Evaluation Metrics Measure User Satisfaction in Web Image Search? 615-624
- Enrique Amigó, Damiano Spina, Jorge Carrillo de Albornoz:
  An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. 625-634

SIGIR 2017



Session 1A: Evaluation 1

- Gordo
  Navig **er and Me.** 5-14
- Ye Che
  Meta-
- Tetsuy
  The P **for IR Evaluation.** 25-34
- Xiaolu
  Can D

Session 4A: Evaluation 2

- Peter Bailey, Alistair Moffat, Falk Scholer, Paul Thomas:
  Retrieval Consistency in the Presence of Query Variations. 395-404
- Jiepu Jiang, Daqing He, James Allan:
  Comparing In Situ and Multidimensional Relevance Judgments. 405-414
- Adam Roegiest, Luchen Tan, Jimmy Lin:
  Online In-Situ Interleaved Evaluation of Real-Time Push Notification Systems. 415-424
- Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, Shaoping Ma:
  Evaluating Web Search with a Bejeweled Player Model. 425-434
- Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, Shaoping Ma:
  Evaluating Mobile Search with Height-Biased Gain. 435-444

Thank You