university of groningen

Retrieval Evaluation

Reference – Chap 04: Retrieval Evaluation, Baeza-Yates & Ribeiro-Neto, Modern Information Retrieval, 2nd Edition

Guru Swaroop Bennabhaktula

Postdoctoral researcher

27th Sep, 2023

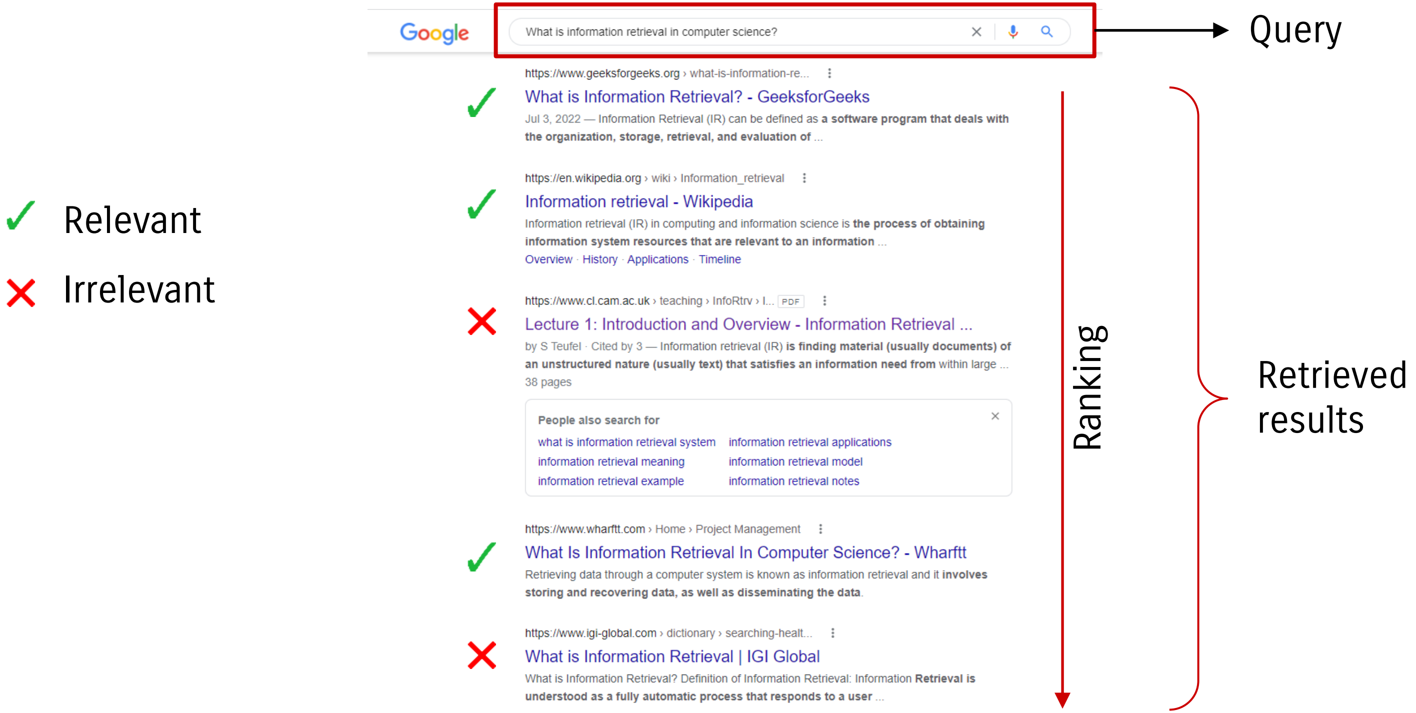
Information Retrieval - WBCS040-05

INFORMATION SYSTEMS GROUP

University of Groningen, The Netherlands

Introduction

› An example of an Information Retrieval System - Google search



Introduction

Characteristics of the Metrics for Retrieval

- Quantitative
- Interpretable
  - A high value → high relevance to the user
  - A low value → low relevance to the user
- Repeatable

Systematic evaluation of the IR system allows answering questions:

- A modification to the ranking function is proposed, do we go ahead and launch it?
- A new probabilistic function has just been devised, is it superior to the vector model and the BM25 rankings?

Reference Collections

- It is a labelled dataset required to evaluate an IR system

- A reference collection consists of
  - A set $D = \{d_j\}$ of pre-selected documents
  - A set $I = \{i_m\}$ of information need descriptors
  - A set of binary relevance judgements associated with each pair (i_m, d_j)
    - $R(i_m, d_j) = \begin{cases} 1 & \text{if } \ d_j \text{ is relevant to } i_m \\ 0 & \text{otherwise} \end{cases}$

      where,

      $i_m \in I$

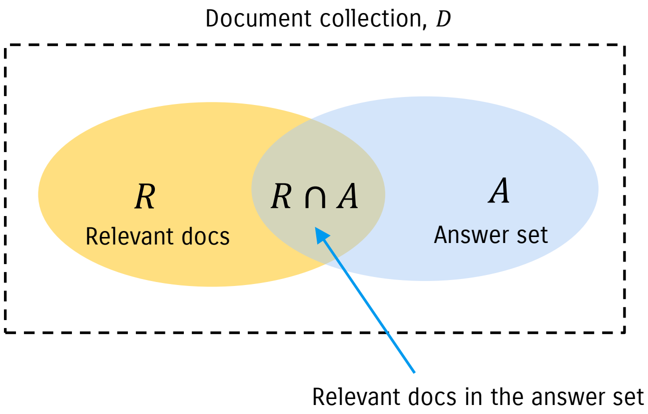
      $d_j \in D$
    - These judgements are produced by a human specialist

Retrieval Metrics

- Precision and Recall
- Single Value Summaries
  - P@n, MAP, R-Precision, F
- User Oriented Measures
- DCG - Discounted Cumulative Gain
- Rank Correlation Metrics

Precision & Recall

- Consider a reference collection, D
- For a given query, $i_m \in I$
  - R – Set of relevant documents
  - A – Answer set (set of retrieved documents)
  - $R \cap A$ – Relevant documents in the answer set



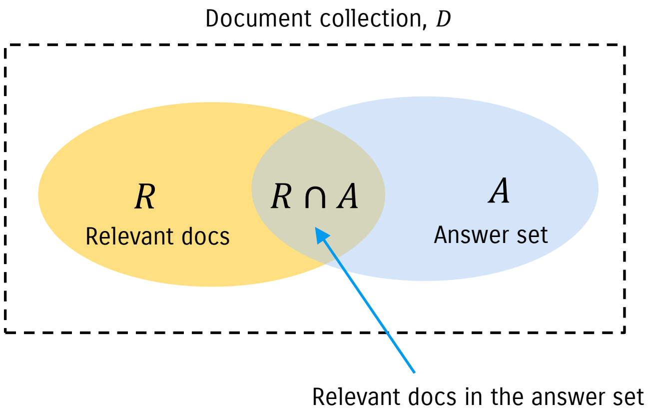
- Precision
  - Fraction of retrieved documents that are relevant
  - $P = \frac{|R \cap A|}{|A|}$

- Recall
  - Fraction of relevant documents that are retrieved
  - $R = \frac{|R \cap A|}{|R|}$

Notation:
For a set X, |X| denotes its cardinality
i.e., the number of elements in the set

Precision & Recall

- Issues
  - The definition of precision and recall assume that all documents in answer set A have been examined.
  - In practice, users are presented with a ranked set of documents.
  - Users examine the documents one by one starting from the top
  - Precision and recall vary as the user proceeds with the examination of answer set A
- Possible solution
  - It is more appropriate to plot Precision v/s Recall



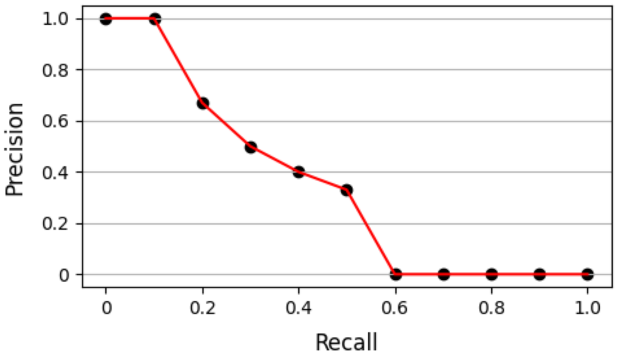
Precision, $P = \frac{|R \cap A|}{|A|}$

Recall, $R = \frac{|R \cap A|}{|R|}$

Precision & Recall

- Consider a reference collection
- Given a query - q_1
- Relevant docs - R_{q_1}
  - 10 relevant documents
  - $R_{q_1} = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

- PR plot for 11 standard recall values r_j



Recall r_j	Precision @ R $P(r_j)$
0.0	1.00
0.1	1.00
0.2	0.67
0.3	0.50
0.4	0.40
0.5	0.33
0.6	0
0.7	0
0.8	0
0.9	0
1.0	0

Ranking produced by a retrieval algorithm

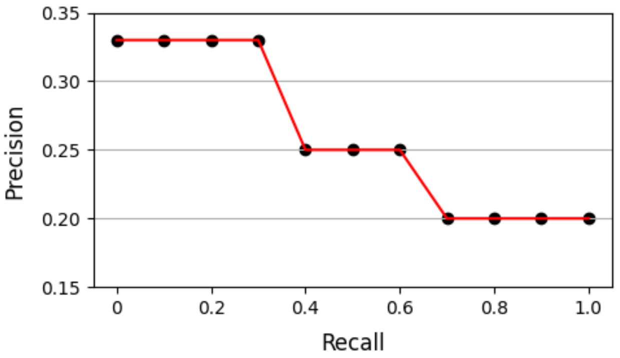
Rank	Document	Recall r	Precision @ R $P(r)$
1	d_{123}	0.1	1.00
2	d_{84}		
3	d_{56}	0.2	0.67
4	d_6		
5	d_8		
6	d_9	0.3	0.50
7	d_{511}		
8	d_{129}		
9	d_{187}		
10	d_{25}	0.4	0.40
11	d_{38}		
12	d_{48}		
13	d_{250}		
14	d_{113}		
15	d_3	0.5	0.33

Precision & Recall

- Let's consider another query
- Given a query - q_2
- Relevant docs - R_{q_2}
  - 3 relevant documents
  - $R_{q_2} = \{d_3, d_{56}, d_{129}\}$

$$P(r_j) = \max_{r_j \leq r} P(r)$$

- PR plot for 11 standard recall values r_j



Recall r_j	Precision @ R $P(r_j)$
0.0	0.33
0.1	0.33
0.2	0.33
0.3	0.33
0.4	0.25
0.5	0.25
0.6	0.25
0.7	0.20
0.8	0.20
0.9	0.20
1.0	0.20

Ranking produced by a retrieval algorithm

Rank	Document	Recall r	Precision @ R $P(r)$
1	d_{425}		
2	d_{87}		
3	d_{56}	0.33	0.33
4	d_{32}		
5	d_{124}		
6	d_{615}		
7	d_{512}		
8	d_{129}	0.66	0.25
9	d_4		
10	d_{130}		
11	d_{193}		
12	d_{715}		
13	d_{810}		
14	d_5		
15	d_3	1.0	0.20

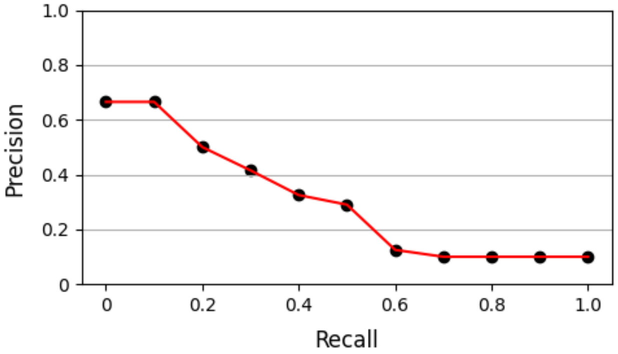
Precision & Recall

- In the examples so far, the PR figures have been computed only for a single query
- Usually, however, the retrieval algorithms are evaluated by considering several distinct test queries
- Let's say we have N_q test queries, we average the precision and recall levels as:

$$\bar{P}(r_j) = \sum_{i=1}^{N_q} \frac{P_i(r_j)}{N_q}$$

where, $\bar{P}(r_j)$ is the average precision at recall level r_j
$P_i(r_j)$ is the precision at recall level r_j for the i^{th} query

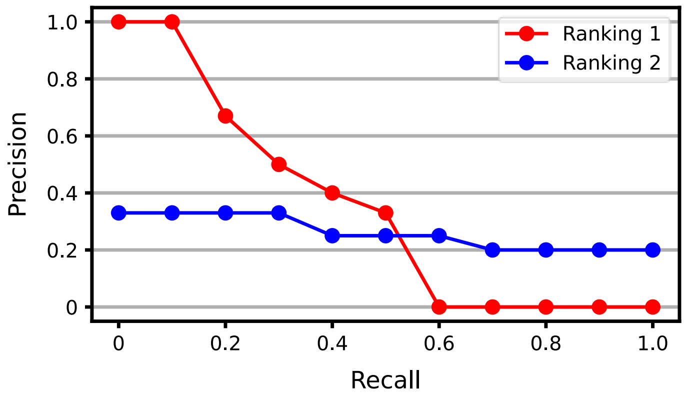
- The figure on the right illustrates the PR plot averaged over queries q_1 and q_2



Recall r_j	P @ R $P_1(r_j)$	P @ R $P_2(r_j)$	Precision @ R $\bar{P}(r_j)$
0.0	1.00	0.33	0.66
0.1	1.00	0.33	0.66
0.2	0.67	0.33	0.50
0.3	0.50	0.33	0.41
0.4	0.40	0.25	0.32
0.5	0.33	0.25	0.29
0.6	0	0.25	0.12
0.7	0	0.20	0.10
0.8	0	0.20	0.10
0.9	0	0.20	0.10
1.0	0	0.20	0.10

Precision & Recall

- Averaged PR curves can be used to compare the performance of distinct IR algorithms
- For instance,



- Ranking 1 – Preferred for web search
- Ranking 2 – Preferred for medical / legal domains

Precision & Recall

Issues with P & R measures:

- Estimation of maximum recall for a query is infeasible / impractical
  - Example – In web-based applications, it is not possible to identify all relevant documents for a given query.
- Averaging precision over many queries might disguise important anomalies
  - Example – Outliers may go undetected
- Interpreting a PR plot can become cumbersome. Need a simplistic approach.
  - Single valued score

Single Valued Summaries

- Average Precision at n
- MAP – Mean Average Precision
- R-Precision
    - Precision Histograms
- Mean reciprocal rank
- E-measure
- F-measure

Single Valued Summaries

Average Precision @ n

- In the case of web-search, the majority of searches do not require a high recall
- Let's revisit the example query - q_1
  - For this query we have
    - P@5 = 0.40
    - P@10 = 0.40
- These metrics provide an early indication of which algorithm might be preferable in the eyes of the users
  - Question - Can we use **Precision @ R** as an early indication to evaluate web search results?

Rank n	Document	Precision @ n $P(n)$
1	d_{123}	1.00
2	d_{84}	0.50
3	d_{56}	0.67
4	d_{6}	0.50
5	d_{8}	0.40
6	d_{9}	0.50
7	d_{511}	0.43
8	d_{129}	0.38
9	d_{187}	0.33
10	d_{25}	0.40
11	d_{38}	0.36
12	d_{48}	0.33
13	d_{250}	0.31
14	d_{113}	0.29
15	d_{3}	0.33

Single Valued Summaries

Mean Average Precision

- The idea here is to generate a single-valued summary of the "entire ranking" by averaging the precision values after each new relevant document is observed

- Avg. precision for query q_1
  - $AP_1 = \frac{1 + 0.67 + 0.5 + 0.4 + 0.33 + 0 + 0 + 0 + 0 + 0}{10} = 0.28$

- Avg. precision for query q_2
  - $AP_2 = \frac{0.33 + 0.25 + 0.20}{3} = 0.26$

- Mean average precision for query set $\{ q_1 , q_2 \}$
  - $MAP = \frac{AP_1 + AP_2}{2} = 0.27$

Ranking for query q_1

Rank n	Document	Precision @ n $P(n)$
1	d_{123}	1.00
2	d_{84}	
3	d_{56}	0.67
4	d_{6}	
5	d_{8}	
6	d_{9}	0.50
7	d_{511}	
8	d_{129}	
9	d_{187}	
10	d_{25}	0.40
11	d_{38}	
12	d_{48}	
13	d_{250}	
14	d_{113}	
15	d_{3}	0.33

Single Valued Summaries

R - Precision

- R – The number of relevant docs for a given query
- This is a special case of Average Precision @ n, with n = R
- Revisiting the example query - q_1
  - For this query we have
    - R = 10 (number of relevant docs)
    - R-Precision, P@R = P@10 = 0.40

	Rank n	Document	Precision @ n $P(n)$
	1	d_{123}	1.00
	2	d_{84}	0.50
	3	d_{56}	0.67
	4	d_6	0.50
	5	d_8	0.40
	6	d_9	0.50
	7	d_{511}	0.43
	8	d_{129}	0.38
	9	d_{187}	0.33
R =	10	d_{25}	0.40
	11	d_{38}	0.36
	12	d_{48}	0.33
	13	d_{250}	0.31
	14	d_{113}	0.29
	15	d_3	0.33

Precision Histograms

- This is a natural extension of R-Precision
- Useful for comparing 2 different retrieval algorithms
  - $RP_A(i)$: R-Precision for algorithm A, wrt the i^{th} query
  - $RP_B(i)$: R-Precision for algorithm B, wrt the i^{th} query
  - $RP_{A/B}(i) = RP_A(i) - RP_B(i)$

A precision histogram for 10 hypothetical queries

Observations:

- A performs well for 8 queries
- B performs well only for 2 queries

Single Valued Summaries

The E-Measure

- It is a measure that combines both precision and recall
- It allows a user to specify relative importance of precision and recall
- $E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$ $\qquad 0 \le E(j) \le 1$

where,

$P(j)$ is the precision at the j^{th} position in the ranking

$r(j)$ is the recall at the j^{th} position in the ranking

$b \ge 0$ is a user specified parameter (relative importance of precision and recall)

- If $b == 0$ then $\quad E(j) = 1 - P(j)$
- If $b \to \infty$ then $\quad \lim_{b \to \infty} E(j) = 1 - r(j)$
- If $b == 1$ then $\quad E(j) = 1 - \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} = 1 - F(j)$

where, $F(j)$ is the harmonic mean of $P(j)$ and $r(j)$ – also known as the F-measure

Single Valued Summaries

The F-Measure

- It is the harmonic mean precision and recall
- $F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$

Properties

- Assumes values in the interval $[0, 1]$
- $F = 0$ indicates that no relevant documents were retrieved
  $F = 1$ indicates that the set of relevant documents equal the set of retrieved documents
- F assumes a high value only when both precision and recall are high
- Maximizing F implies finding the best possible compromise between precision and recall
- F-measure and E-measure are related by the equation $F(j) = 1 - E(j)$

User Oriented Measures

Issue - Assume set of relevant documents is the same for all users

Solution - User Oriented Measures

- **Coverage Ratio** – The fraction of documents known and relevant that are in the answer set.

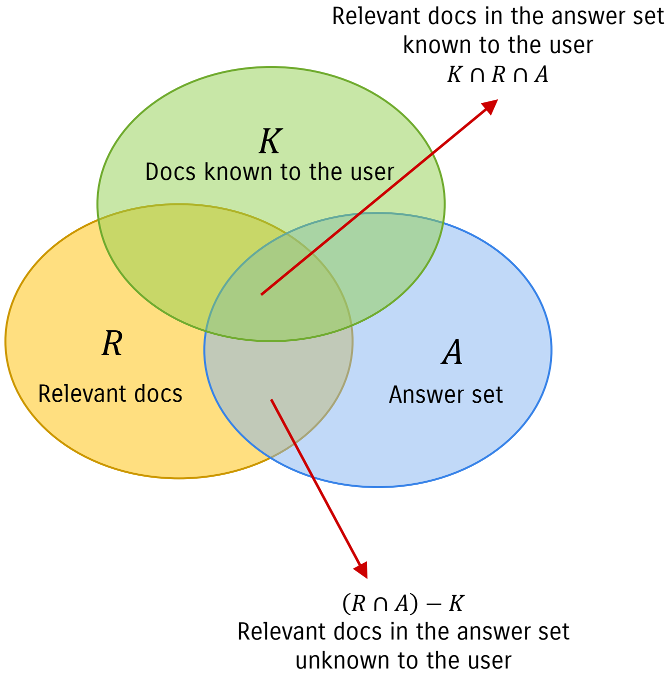
$$coverage\ \ ratio = \frac{|K \cap R \cap A|}{|K \cap R|}$$

- **Novelty Ratio** – The fraction of relevant documents in the answer set that unknown to the user.

$$novelty\ \ ratio = \frac{(R \cap A) - K}{|R \cap A|}$$

- A high coverage ratio indicates that the system is finding most relevant documents to the user
- A high novelty ratio indicates that the system is revealing many unknown yet relevant documents to the user

For a given reference collection,
an information request, and a retrieval algorithm



Discounted Cumulative Gain

- Are all relevant documents equally important? Is it always the case?
  - Precision and recall allow only binary relevance assessments
    - They do not distinguish between highly relevant docs or mildly relevant ones
- These issues can be addressed by a technique known as DCG (or) discounted cumulative gain
  - DCG assigns multi-grade relevance scores for documents
    - 0 : non-relevant
    - 1 : mildly relevant
    - 2 : moderately relevant
    - 3 : highly relevant
  - DCG metric is designed such that
    - Highly relevant documents are preferred at the top of the ranking
    - Relevant documents appearing at the end of the ranking are less valuable

Discounted Cumulative Gain

Consider the graded relevance scores, assigned by specialists, for queries q_1 and q_2

- $R_{q_1} = \{[d_3, 3], [d_5, 3], [d_9, 3], [d_{25}, 2], [d_{39}, 2], [d_{44}, 2], [d_{56}, 1], [d_{71}, 1], [d_{89}, 1], [d_{123}, 1]\}$
- $R_{q_2} = \{[d_3, 3], [d_{56}, 2], [d_{129}, 1]\}$

For a query q_j

1. Compute Gain Vector, G_j, based on relevance scores
2. Compute Cumulative Gain, CG_j

$$CG_j[i] = \begin{cases} G_j[1] & \text{if } i = 1 \\ G_j[i] + CG_j[i-1] & \text{otherwise} \end{cases}$$

3. Compute Discounted Cumulative Gain, DCG

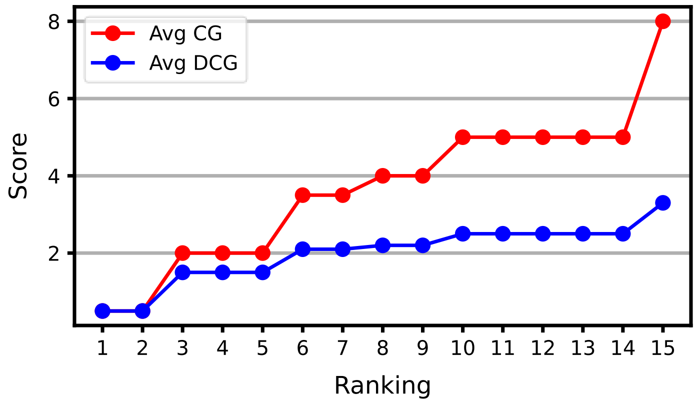
$$DCG_j[i] = \begin{cases} G_j[1] & \text{if } i = 1 \\ \frac{G_j[i]}{\log_2 i} + DCG_j[i-1] & \text{otherwise} \end{cases}$$

Rank	Doc	G_1	CG_1	DCG_1
1	d_{123}	1	1	1.0
2	d_{84}	0	1	1.0
3	d_{56}	1	2	1.6
4	d_6	0	2	1.6
5	d_8	0	2	1.6
6	d_9	3	5	2.8
7	d_{511}	0	5	2.8
8	d_{129}	0	5	2.8
9	d_{187}	0	5	2.8
10	d_{25}	2	7	3.4
11	d_{38}	0	7	3.4
12	d_{48}	0	7	3.4
13	d_{250}	0	7	3.4
14	d_{113}	0	7	3.4
15	d_3	3	10	4.2

Discounted Cumulative Gain

Average DCG

- Given a set of N_q queries, average $\overline{CG}[i]$ and $\overline{DCG}[i]$ can be computed as
  - $\overline{CG}[i] = \sum_{j=1}^{N_q} \frac{CG_j[i]}{N_q}$
  - $\overline{DCG}[i] = \sum_{j=1}^{N_q} \frac{DCG_j[i]}{N_q}$
- For queries q_1 and q_2 these averages are given by:



Rank	G_1	G_2	$\overline{CG}$	$\overline{DCG}$
1	1	0	0.5	0.5
2	0	0	0.5	0.5
3	1	2	2	1.5
4	0	0	2	1.5
5	0	0	2	1.5
6	3	0	3.5	2.1
7	0	0	3.5	2.1
8	0	1	4	2.2
9	0	0	4	2.2
10	2	0	5	2.5
11	0	0	5	2.5
12	0	0	5	2.5
13	0	0	5	2.5
14	0	0	5	2.5
15	3	3	8	3.3

Discounted Cumulative Gain

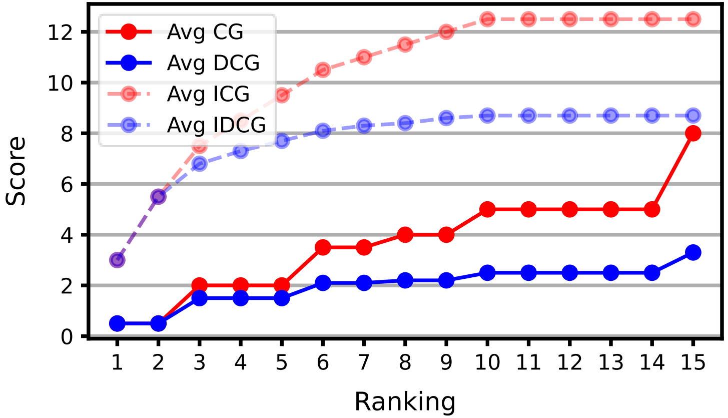
Ideal DCG

- Need ideal DCG scores to determine how much room for improvement there is
- For a given test query q, lets assume the relevance assessment contains
  - n_3 documents with a relevance score of 3
  - n_2 documents with a relevance score of 2
  - n_1 documents with a relevance score of 1
  - n_0 documents with a relevance score of 0
- Then, the ideal gain vector for the query q, is created by sorting all relevance scores in descending order:
  - $IG = [3, \ \dots , \ 3, \ 2, \ \dots , \ 2, \ 1, \ \dots , \ 1, \ 0, \ \dots , \ 0]$
  - As before we can compute:
    - Ideal cumulative gain – ICG
    - Ideal discounted cumulative gain - $IDCG$
- Finally, for a set of queries we can compute:
  - Average Ideal cumulative gain – $\overline{ICG}$
  - Average Ideal discounted cumulative gain - $\overline{IDCG}$

Discounted Cumulative Gain

Ideal DCG

- For instance, consider the example queries q_1 and q_2, with the graded relevance scores
  - $R_{q_1} = \{[d_3, 3], [d_5, 3], [d_9, 3], [d_{25}, 2], [d_{39}, 2], [d_{44}, 2], [d_{56}, 1], [d_{71}, 1], [d_{89}, 1], [d_{123}, 1]\}$
  - $R_{q_2} = \{[d_3, 3], [d_{56}, 2], [d_{129}, 1]\}$

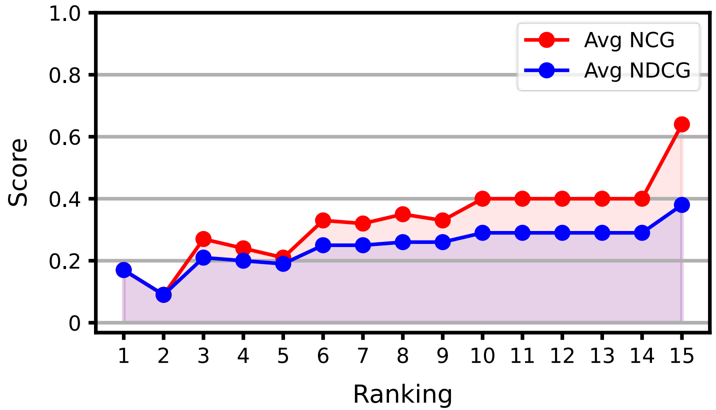


Rank	G_1	G_2	$\overline{CG}$	$\overline{DCG}$	IG_1	IG_2	$\overline{ICG}$	$\overline{IDCG}$
1	1	0	0.5	0.5	3	3	3.0	3.0
2	0	0	0.5	0.5	3	2	5.5	5.5
3	1	2	2	1.5	3	1	7.5	6.8
4	0	0	2	1.5	2	0	8.5	7.3
5	0	0	2	1.5	2	0	9.5	7.7
6	3	0	3.5	2.1	2	0	10.5	8.1
7	0	0	3.5	2.1	1	0	11.0	8.3
8	0	1	4	2.2	1	0	11.5	8.4
9	0	0	4	2.2	1	0	12.0	8.6
10	2	0	5	2.5	1	0	12.5	8.7
11	0	0	5	2.5	0	0	12.5	8.7
12	0	0	5	2.5	0	0	12.5	8.7
13	0	0	5	2.5	0	0	12.5	8.7
14	0	0	5	2.5	0	0	12.5	8.7
15	3	3	8	3.3	0	0	12.5	8.7

Discounted Cumulative Gain

Normalized DCG

- In order to directly compare two DCG curves, it is necessary to normalize them
- Given a set of N_q queries, $NCG[i] = \frac{\overline{CG}[i]}{\overline{ICG}[i]}$; $NDCG[i] = \frac{\overline{DCG}[i]}{\overline{IDCG}[i]}$
- The AUC represents the quality of the ranking algorithm
  - Useful for comparing two distinct ranking algorithms



Rank	$\overline{CG}$	$\overline{DCG}$	$\overline{ICG}$	$\overline{IDCG}$	NCG	$NDCG$
1	0.5	0.5	3.0	3.0	0.17	0.17
2	0.5	0.5	5.5	5.5	0.09	0.09
3	2	1.5	7.5	6.8	0.27	0.21
4	2	1.5	8.5	7.3	0.24	0.20
5	2	1.5	9.5	7.7	0.21	0.19
6	3.5	2.1	10.5	8.1	0.33	0.25
7	3.5	2.1	11.0	8.3	0.32	0.25
8	4	2.2	11.5	8.4	0.35	0.26
9	4	2.2	12.0	8.6	0.33	0.26
10	5	2.5	12.5	8.7	0.40	0.29
11	5	2.5	12.5	8.7	0.40	0.29
12	5	2.5	12.5	8.7	0.40	0.29
13	5	2.5	12.5	8.7	0.40	0.29
14	5	2.5	12.5	8.7	0.40	0.29
15	8	3.3	12.5	8.7	0.64	0.38

Discounted Cumulative Gain

Key Takeaways

- Account for multiple level of relevance assessments
  - Advantage – Distinguish between highly relevant and mildly relevant
  - Disadvantage – It is hard to get such labeled data
- Allows to systematically combine document ranks with relevance scores
- CG provides a single metric of retrieval performance at any position in the ranking
- DCG makes the metric more immune to outliers

Rank Correlation Metrics

- Precision and Recall allow comparing the *relevance* of the results produced by the two ranking functions
- However, there are situations in which
  - We cannot directly measure the relevance
  - We are more interested in measuring how differently a ranking function varies when compared to another well known ranking function
- In these cases we are interested in measuring the relative ordering produced by the two ranking functions
- This can be measured by using statistical functions known as the rank correlation metrics
  - The Spearman Coefficient
  - The Kendall Tau Coefficient

Rank Correlation Metrics

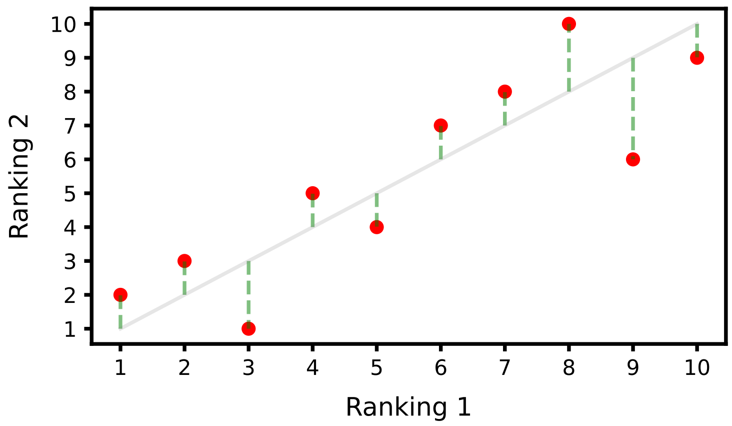
Properties

- Consider the rankings R_1 and R_2
- A rank correlation metric yields a correlation coefficient $C(R_1, R_2)$ with the following properties:
  - $-1 \leq C(R_1, R_2) \leq 1$ i.e., values are bound
  - If $C(R_1, R_2) == 1$ then, the agreement between the two rankings is perfect i.e., they are the same
  - If $C(R_1, R_2) == -1$ then, the disagreement between the two rankings is perfect i.e., they are the reverse of each other
  - If $C(R_1, R_2) == 0$ then, the two rankings are completely independent
  - Increase in the values of $C(R_1, R_2)$ implies increase in the agreement between the two rankings

The Spearman Coefficient

- It is one of the most widely used rank correlation metric
- Let's derive this coefficient
  - Let $s_{1,j}$ be the position of the document d_j in ranking R_1
  - Let $s_{2,j}$ be the position of the document d_j in ranking R_2
- Consider 10 example documents retrieved by the rankings R_1 and R_2

Docs	$s_{1,j}$	$s_{2,j}$	$s_{1,j} - s_{2,j}$	$(s_{1,j} - s_{2,j})^2$
d_{123}	1	2	-1	1
d_{84}	2	3	-1	1
d_{56}	3	1	+2	4
d_{6}	4	5	-1	1
d_{8}	5	4	+1	1
d_{9}	6	7	-1	1
d_{511}	7	8	-1	1
d_{129}	8	10	-2	4
d_{187}	9	6	+3	9
d_{25}	10	9	+1	1
		Sum of squared distances		24



The Spearman Coefficient

Let there be K documents

- Sum of squared distances between the rankings is given by
  - $\sum_{j=1}^{K}(s_{1,j} - s_{2,j})^2$
- The maximum value of sum of squares of ranking differences is given by
  - $\frac{K \times (K^2 - 1)}{3}$
- In order to get bounded scores (in the range $[0, 1]$), compute the fraction
  - $\frac{\sum_{j=1}^{K}(s_{1,j} - s_{2,j})^2}{\frac{K \times (K^2 - 1)}{3}}$
- The value of the above fraction is
  - 0 – when the two rankings are in perfect agreement
  - 1 – when the rankings are in perfect disagreement
- If we multiply the fraction with 2, its value shifts to the range $[0, 2]$
- If we further subtract the result from 1, the resultant value shifts to the range $[-1, +1]$

The Spearman Coefficient

Formally, we define the *Spearman rank correlation coefficient* as

$$S(R_1, R_2) = 1 - \frac{6 \times \sum_{j=1}^{K} (s_{1,j} - s_{2,j})^2}{K \times (K^2 - 1)}$$

where,

- R_1, R_2 - two distinct rankings
- K – number of ranked documents
- $s_{1,j}$ - Rank of document d_j in R_1
- $s_{2,j}$ - Rank of document d_j in R_2

For the two rankings in the adjacent Table we have

$$S(R_1, R_2) = 1 - \frac{6 \times 24}{10 \times (10^2 - 1)}$$

$$= 1 - \frac{144}{990}$$

$$= 0.854$$

Docs	$s_{1,j}$	$s_{2,j}$	$s_{1,j} - s_{2,j}$	$(s_{1,j} - s_{2,j})^2$
d_{123}	1	2	-1	1
d_{84}	2	3	-1	1
d_{56}	3	1	+2	4
d_6	4	5	-1	1
d_8	5	4	+1	1
d_9	6	7	-1	1
d_{511}	7	8	-1	1
d_{129}	8	10	-2	4
d_{187}	9	6	+3	9
d_{25}	10	9	+1	1
			Sum of squared distances	24

The Kendall Tau Coefficient

- Kendall Tau coefficient has a simple algebraic structure with an intuitive interpretation
- When we think of rank correlations, we think of how two rankings tend to vary in similar ways, i.e., how they tend to move in the same direction
- To illustrate, consider two documents d_j and d_k and their positions in the rankings R_1 and R_2
  - Further, consider the differences in rank positions for these two documents in each ranking, i.e.,

$$s_{1,k} - s_{1,j}$$
$$s_{2,k} - s_{2,j}$$

    - If these differences have the same sign, we say that the document pair $[d_k,\ d_j]$ is concordant in both the rankings
    - If these differences have a different sign, we say that the document pair $[d_k,\ d_j]$ is discordant in both the rankings

The Kendall Tau Coefficient

- Consider the top 5 documents in rankings R_1 and R_2
- The ordered document pairs in R_1

  $[d_{123}, d_{84}]$, $[d_{123}, d_{56}]$, $[d_{123}, d_6]$, $[d_{123}, d_8]$
  $[d_{84}, d_{56}]$, $[d_{84}, d_6]$, $[d_{84}, d_8]$
  $[d_{56}, d_6]$, $[d_{56}, d_8]$
  $[d_6, d_8]$

  A total of 10 ordered pairs (Note - For K docs there are $K * (K-1)/2$ ordered pairs)

- Similarly, the ordered document pairs in R_2

  $[d_{56}, d_{123}]$, $[d_{56}, d_{84}]$, $[d_{56}, d_8]$, $[d_{56}, d_6]$
  $[d_{123}, d_{84}]$, $[d_{123}, d_8]$, $[d_{123}, d_6]$
  $[d_{84}, d_8]$, $[d_{84}, d_6]$
  $[d_8, d_6]$

Docs	$s_{1,j}$	$s_{2,j}$	$s_{1,j} - s_{2,j}$
d_{123}	1	2	-1
d_{84}	2	3	-1
d_{56}	3	1	+2
d_6	4	5	-1
d_8	5	4	+1

The Kendall Tau Coefficient

- On comparing the two sets of ordered pairs for R_1 and R_2 we can compute the concordant (C) and discordant (D) pairs
- For ranking R_1 we have

  C, D, C, C
  D, C, C
  C, C
  D

- Similarly, for R_2

  D, D, C, C
  C, C, C
  C, C
  D

- There are a total of 20 pairs, 14 concordant, and 6 discordant pairs

Docs	$s_{1,j}$	$s_{2,j}$	$s_{1,j} - s_{2,j}$
d_{123}	1	2	-1
d_{84}	2	3	-1
d_{56}	3	1	+2
d_6	4	5	-1
d_8	5	4	+1

The Kendall Tau Coefficient

The Kendall Tau coefficient is defined as

$$\tau\,(R_1, R_2) = P(R_1 = R_2) - P(R_1 \neq R_2)$$

where,

$P(R_1 = R_2)$ is the probability that the rankings are concordant

$P(R_1 \neq R_2)$ is the probability that the rankings are discordant

Docs	$s_{1,j}$	$s_{2,j}$	$s_{1,j} - s_{2,j}$
d_{123}	1	2	-1
d_{84}	2	3	-1
d_{56}	3	1	+2
d_6	4	5	-1
d_8	5	4	+1

In our example,

$$\tau\,(R_1, R_2) = P(R_1 = R_2) - P(R_1 \neq R_2)$$

$$= \frac{14}{20} - \frac{6}{20}$$

$$= 0.4$$

Note - Kendall Tau coefficient is defined only for ranking over same set of documents

- In case a different set of documents are retrieved by the two rankings then ignore the non-common documents

User based evaluation

- Side-by-side panels
- A/B Testing
- Crowdsourcing
- Evaluation with clickthrough data

User based evaluation

Side-by-side panels

- Top 5 answers produced by two retrieval algorithms for the query, “information retrieval evaluation”

[PDF] Pharmaceutical **Information** Flyer
PDF/Adobe Acrobat
PHARMACEUTICAL **INFORMATION RETRIEVAL** AND **EVALUATION** SERVICE. Future Solutions Now ... **information** need, • **retrieval** of the appropriate documents, • **evaluation** ...
www.**uiowa.edu**/~idis/Pharm_Info_Flyer.pdf

ROMIP: Russian **Information Retrieval Evaluation** Seminar
Russian **information retrieval evaluation** initiative was launched in 2002 with ... a basis for independent **evaluation** of **information retrieval** methods, aimed to be ...
romip.ru/en

[PDF] Reflections on **Information Retrieval Evaluation** Mei-Mei Wu & Diane ...
PDF/Adobe Acrobat
Reflections on **Information Retrieval Evaluation**. Mei-Mei Wu ... Research and **evaluation** in **information retrieval**. Journal of Documentation , 53 (1), 51-57. ...
pnclink.org/annual/annual1999/1999pdf/wu-mm.pdf

Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (IR) is the science of searching for ... that was needed for **evaluation** of text **retrieval** methodologies on a very large text collection. ...
en.wikipedia.org/wiki/**Information_retrieval**

The Music **Information Retrieval Evaluation** eXchange (MIREX)
The 2005 Music **Information Retrieval Evaluation** eXchange (MIREX 2005): Preliminary Overview. ... Music **Information Retrieval** Systems **Evaluation** Laboratory: ...
www.**dlib.org**/dlib/december06/downie/12downie.html

yahoo!

[PDF] Reflections on **Information Retrieval Evaluation** Mei-Mei Wu & Diane ...
PDF/Adobe Acrobat
digital library initiatives, **information retrieval** (IR) **evaluation** has ...... **Evaluation** of **evaluation** in **information retrieval**. Proceedings of the **...**
pnclink.org/annual/annual1999/1999pdf/wu-mm.pdf -

[PDF] **Retrieval Evaluation** with Incomplete **Information**
PDF/Adobe Acrobat
The philosophy of **information**. **retrieval evaluation**. In **Evaluation** of Cross-Language. **Information Retrieval** Systems. Proceedings of CLEF **...**
www.nist.gov/itl/iad/IADpapers/2004/p102-buckley.pdf

Evaluation criteria for **information retrieval** systems. - [Traduzir esta página]
The contrast between the value placed on discriminatory power in discussions of indexing and classification and on the transformation of a query into a set **...**
informationr.net/ir/4-4/paper62.html - 36k

Information retrieval - Wikipedia, the free encyclopedia - [Traduzir esta página]
The aim of this was to look into the **information retrieval** community by supplying the infrastructure that was needed for **evaluation** of text **retrieval ...**
en.wikipedia.org/wiki/**Information_retrieval** - 59k

[PDF] **Information Retrieval** System **Evaluation**: Effort, Sensitivity, and **...**
PDF/Adobe Acrobat
Information Retrieval System **Evaluation**:. Effort, Sensitivity, and Reliability. Mark Sanderson. Department of **Information** Studies, University of **...**
dis.shef.ac.uk/mark/publications/my_papers/SIGIR2005.pdf

Google

User based evaluation

Side-by-side panels

- In a side-by-side experiment, users are aware that they are participating in an experiment
- Further, a side-by-side experiment cannot be repeated in the same conditions of a previous execution
- Finally, side-by-side panels do not allow to measure by how much is system A better when compared to system B

User based evaluation

A / B Testing

- A/B testing consists of displaying to selected users a modification in the layout of a page
  - The group of selected users constitute a fraction of all users such as, for instance, 1%
  - The method works well for sites with large audiences
- By analyzing how the users react to the change, it is possible to analyze if the modification proposed is positive or not
- A/B testing provides a form of real-world human experimentation, without the setting of a lab

User based evaluation

Crowdsourcing

- It can be used to quickly get labelled data (query ↔ relevance documents) for a reference collection
- Crowdsourcing is a term used to describe tasks that are outsourced to a large group of people, called “workers”
- It is an open call to solve a problem or carry out a task, one which usually involves a monetary value in exchange for such service
- Example: Amazon Mechanical Turk

User based evaluation

Evaluation with clickthrough data

- Reference collections can be prepared for a relatively small number of queries
- In real-world applications (such as web search), the query log is typically composed of billions of queries.
  - It is impractical to make reference collections for such a large query set
- One very promising alternative is evaluation based on the analysis of clickthrough data
- It can be obtained by observing how frequently the users click on a given document, when it is shown in the answer set for a given query
- This is particularly attractive because the data can be collected at a low cost

User based evaluation

Evaluation with clickthrough data

A minimal example…

- A click ≠ relevance judgement. It just indicates user preference
  - It is correlated (to relevance judgement) but noisy
  - Used to generate preferences
- We could build evaluation metrics directly from user preferences
  - Consider the Kendall Tau rank correlation coefficient: $\tau = \frac{|C| - |D|}{|C| + |D|}$
  - $|C|,\ |D|$ - number of concordant and discordant pairs respectively

d_1
d_2
d_3 (clicked)
d_4

→

$d_3 > d_1$
$d_3 > d_2$
$d_3 > d_4$

user preferences

$d_1 > d_2$ $d_2 > d_3$
$d_1 > d_3$ $d_2 > d_4$
$d_1 > d_4$ $d_3 > d_4$

retrieval algorithm preferences

$$\tau = \frac{1-2}{1+2} = -0.33$$

Summary

Key Takeaway

- No single best metric for retrieval evaluation.
- Use metrics based on the context, considering its pros and cons

Assignment

- Aim - To evaluate web search results
- Part A – Retrieve web search results
  - Learn to programmatically retrieve the web search results from popular search engines such as Google search, Yahoo, etc.,
- Part B – Evaluation
  - Evaluate the retrieved results
  - Compare the retrieved results by different search engines to one another