Multimedia Information Retrieval and Technology

Lecture 9 Evaluation II

By : Fangyu Wu

Room: SD555

Xi'an Jiaotong-Liverpool University

西交利物浦大学

Evaluation of unranked retrieval results

Evaluation of ranked retrieval results

- ⍰Eleven-point interpolated average precision
- ⍰MAP (Mean Average Precision)
- ⍰Precision@K
- ⍰Normalized Discounted Cumulative Gain

Average Precision

For any information need $q_j \in Q$, we denote the set of relevant documents as $\{d_1, \cdots, d_{m_j}\}$

Precision(d_k) is the precision of retrieved results when we reach the document d_k.

Average Precision

For a single information need, Average Precision is the average of the precision value obtained for the set of top *k* documents existing after **each relevant document** is retrieved.

$$AP = \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(d_k)$$

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Average Precision

For an information need $q_j \in Q$, the IR system to be evaluated has returned the following documents:

$$AP = \frac{1+1+0.6}{3} \approx 0.87$$

		P	R	
1	R	1	0.01	
2	R	1	0.02	
3	N	2/3	0.02	
4	N	2/4	0.02	
5	R	3/5	0.03	
6	N	3/6	0.03	

Here, m_j=3, AP=?

Average Precision

■ ■ ■ ■ ■ ■ = the relevant documents

Ranking #1	■	□	■	■	■	■	□	□	□	■
Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6
Ranking #2	□	■	□	□	■	■	■	□	■	■

6 relevant docs in total.

Recall? Precision?

Average Precision

■ ■ ■ ■ ■ ■ = the relevant documents

Ranking #1

	■	□	■	■	■	■	□	□	□	■
Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

	□	■	□	□	■	■	■	□	■	■
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

AP=?

AP$_{\#1}$ = $\frac{1+0.67+0.75+0.8+0.8}{}$

Average Precision

■ ■ ■ ■ ■ ■ = the relevant documents

Ranking #1	■	□	■	■	■	■	□	□	□	■
Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2	□	■	□	□	■	■	■	□	■	■
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

Exercise I

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1   R N R N N   N N N R R

System 2   N R N N R   R R N N N

What is the AP of each system? Which has a better performance?

Solution

AP(System 1) = (1/4)*(1+(2/3)+(3/9)+(4/10)) = 0.6

AP(System 2) =(1/4)*(1/2 + 2/5 + 3/6 + 4/7) = 0.493

System1 has a higher average precision

Mean Average Precision

Then:

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

MAP is Average Precision across multiple queries and rankings

Exercise II: MAP

■ ■ ■ ■ ■ = relevant documents for query 1

Ranking #1	■	□	■	□	□	■	□	□	■	■
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

$AP_1 = \frac{1+0.6}{}$

$= 0$

■ ■ ■ = relevant documents for query 2

Ranking #2	□	■	□	□	■ ✓	□	■ ✓	□	□	□
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

✓ ✓ ✓

$AP_2 = \frac{0.5}{}$

$=$

Xi'an Jiaotong-Liverpool University
西交利物浦大学

✓ ✓ ✓

Solution

■ ■ ■ ■ ■ = relevant documents for query 1

Ranking #1	■	□	■	□	□	■	□	□	■	■
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

■ ■ ■ = relevant documents for query 2

Ranking #2	□	■	□	□	■	□	■	□	□	□
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$

$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Mean average precision

MAP is macro-averaging: each query counts equally

Now perhaps most commonly used measure in research papers. Among evaluation measures, MAP has been shown to have especially good discrimination and stability.

Using MAP, fixed recall levels are not chosen, and there is no interpolation.

Mean average precision

MAP assumes user is interested in finding many relevant documents for each query

MAP requires many relevance judgments in text collection

Good for web search?

What matters is rather how many good results there are on the first page or the first three pages.

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Evaluation of unranked retrieval results

Evaluation of ranked retrieval results

- Eleven-point interpolated average precision
- MAP
- Precision@K
- Normalized Discounted Cumulative Gain

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Precision@K

Measuring precision at fixed low levels of retrieved results, such as 10 or 30 documents.

Ignores documents ranked lower than K

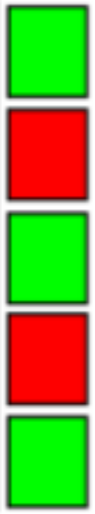
Precision@K

Example:

Prec@3 of 2/3

Prec@4 of 2/4

Prec@5 of 3/5



In similar fashion we have Recall@K

Exercise II

The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N N N N R N R N N N R N N N N R

e. Assume that these 20 documents are the complete result set of the system. What is the AP of the system based on the result above?

Solution

a) 0.3

b) 3/7

c) 0.33 or 0.3

d) 0.364

e) 0.555

Evaluation of unranked retrieval results

Evaluation of ranked retrieval results

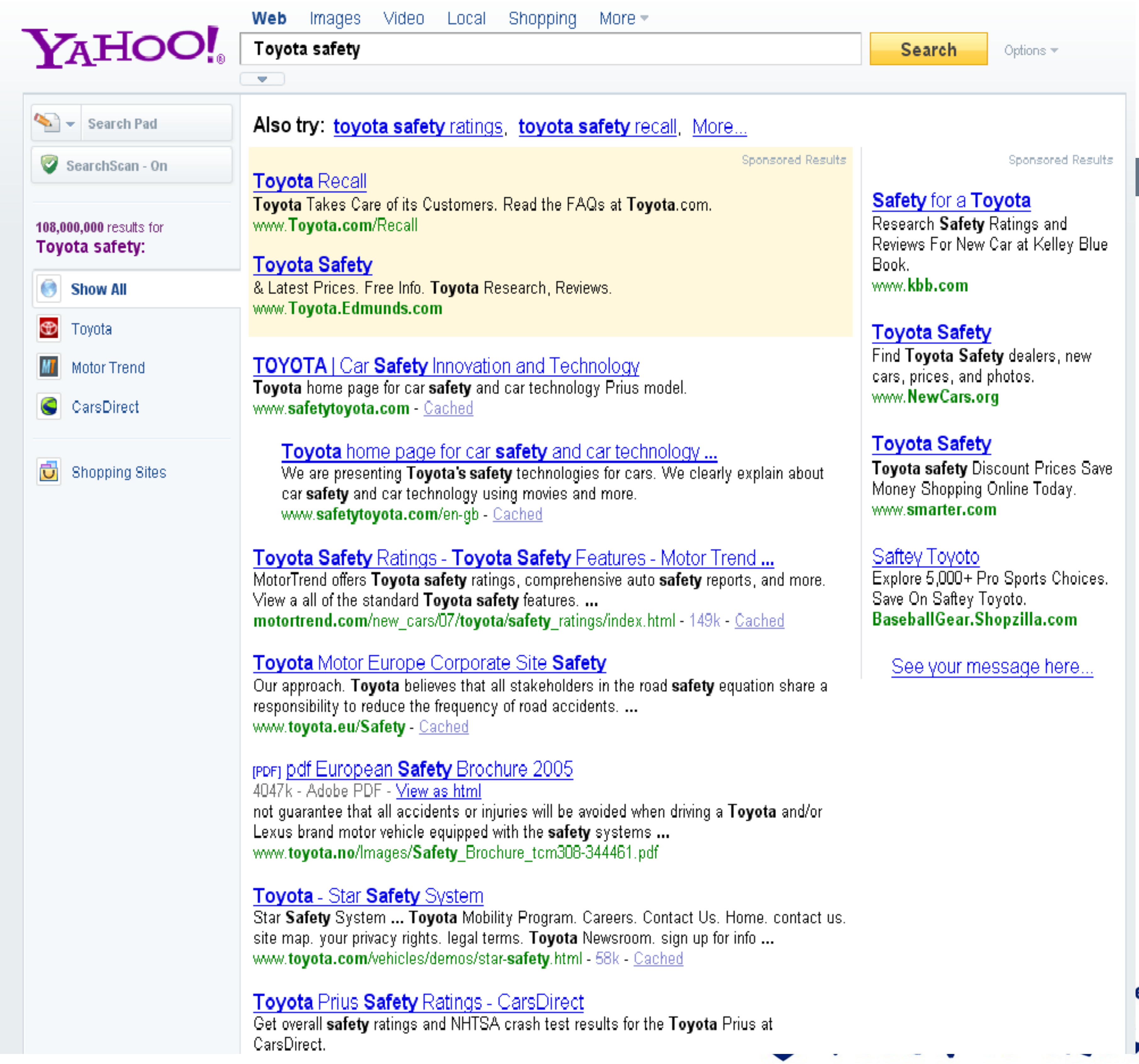
- Eleven-point interpolated average precision
- MAP
- Precision@K
- Normalized Discounted Cumulative Gain

Discounted Cumulative Gain

A Popular measure for evaluating web search and related tasks. A measure of ranking quality.

Two assumptions:

- Highly relevant documents are more useful than marginally relevant documents;
- the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined.



Toyota safety

108,000,000 results for **Toyota safety:**

Also try: toyota safety ratings, toyota safety recall, More...

Sponsored Results

Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at **Toyota**.com.
www.**Toyota.com**/Recall

Toyota Safety
& Latest Prices. Free Info. **Toyota** Research, Reviews.
www.**Toyota.Edmunds.com**

TOYOTA | Car **Safety** Innovation and Technology
Toyota home page for car **safety** and car technology Prius model.
www.**safetytoyota.com** - Cached

Toyota home page for car **safety** and car technology **...**
We are presenting **Toyota's safety** technologies for cars. We clearly explain about car **safety** and car technology using movies and more.
www.**safetytoyota.com**/en-gb - Cached

Toyota Safety Ratings - **Toyota Safety** Features - Motor Trend **...**
MotorTrend offers **Toyota safety** ratings, comprehensive auto **safety** reports, and more. View a all of the standard **Toyota safety** features. **...**
motortrend.com/new_cars/07/**toyota**/**safety**_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site **Safety**
Our approach. **Toyota** believes that all stakeholders in the road **safety** equation share a responsibility to reduce the frequency of road accidents. **...**
www.**toyota.eu**/**Safety** - Cached

[PDF] pdf European **Safety** Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a **Toyota** and/or Lexus brand motor vehicle equipped with the **safety** systems **...**
www.**toyota.no**/Images/**Safety**_Brochure_tcm308-344461.pdf

Toyota - Star **Safety** System
Star **Safety** System **...** **Toyota** Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. **Toyota** Newsroom. sign up for info **...**
www.**toyota.com**/vehicles/demos/star-**safety**.html - 58k - Cached

Toyota Prius **Safety** Ratings - CarsDirect
Get overall **safety** ratings and NHTSA crash test results for the **Toyota** Prius at CarsDirect.

Sponsored Results

Safety for a **Toyota**
Research **Safety** Ratings and Reviews For New Car at Kelley Blue Book.
www.**kbb.com**

Toyota Safety
Find **Toyota Safety** dealers, new cars, prices, and photos.
www.**NewCars.org**

Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.**smarter.com**

Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com

See your message here...

ersity

Discounted Cumulative Gain

Uses ***graded relevance*** as a measure of usefulness, or *gain,* from examining a document

Eg: 5 ranked documents judged on 0-3 relevance scale

D1	3
D2	2
D3	3
D4	0
D5	1

D1	3
D3	3
D4	0
D2	2
D5	1

Summarize a Ranking: DCG

Let the relevance ratings of the top n documents be rel_1, rel_2,…, rel_n (in ranked order)

Cumulative Gain (CG) at rank n

CG = $rel_1 + rel_2 + \ldots + rel_n$

total gain accumulated at a particular rank *n*

Discounted Cumulative Gain

Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks.

Typical discount is:

$$1/\log(rank)$$

With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

Discounted Gain:

$$DG = \frac{rel_i}{\log_2 i}$$

DCG is the total discounted gain accumulated at a particular rank *n*:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

DCG Example

Docs	D1	D2	D3	D4	D5	D6	D 7	D8	D9	D10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3									
DCG	3									

DCG Example

Docs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3	2/1 =2	3/1.59 =1.89	0	0	1/2.59 =0.39				
DCG	3	5	6.89	6.89	6.89	7.28				

DCG Example

Ranking Perfecto (Idctual) D_1 D_3 D_9 D_2 D_7 D_8 D_6 D_4 D_5 D_{10}

Docs	D1	D2	D3	D4	D5	D6	D 7	D8	D9	D10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3	2/1 =2	3/1.59 =1.89	0	0	1/2.59 =0.39	0. 71	0.67	0.95	0
DCG	3	5	6.89	6.89	6.89	7.28	7. 99	8.66	9.61	9.61

Recap: Relevance judgments

Relevance can reasonably be thought of as a scale, with some documents highly relevant and others marginally so..

This decision is referred to as the ***gold standard*** or ***ground truth*** judgment of relevance.

NDCG (Normalized DCG)

The ideal (perfect) ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc

The ideal (perfect) ranking is based on the ***gold standard*** or ***ground truth*** judgment of relevance

Summarize a Ranking: NDCG

Normalized Discounted Cumulative Gain (NDCG) at rank *n:*

Normalize DCG at rank *n* by the DCG value at rank *n* of the ideal ranking

$$NDCG_{rank\ n} = \frac{Actual\ DCG_{\ rank\ n}}{Ideal\ DCG_{\ rank\ n}}$$

- Normalization is useful for contrasting queries with varying numbers of relevant results
- Popular in evaluating web search.

Example

Docs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3									
DCG	3									

Perfect (ideal) ranking:

3, 3, 3, 2, 2, 2, 1, 0, 0, 0

ideal DCG values:

Perfect (ideal) ranking:

3, 3, 3, 2, 2, 2, 1, 0, 0, 0

ideal DCG values:

3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

Actual DCG(3, 2, 3, 0, 0, 1, 2, 2, 3, 0):

3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Docs	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3	2/1=2	3/1.59=1.89	0	0	1/2.59=0.39	0.71	0.67	0.95	0
DCG	3	5	6.89	6.89	6.89	7.28	7.99	8.66	9.61	9.61

Example

NDCG at rank 8?

$$NDCG_{rank\ n} = \frac{Actual\ DCG_{\ rank\ n}}{Ideal\ DCG_{\ rank\ n}}$$

NDCG values (divide actual by ideal):

1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

NDCG - Exercise

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0

Please compare the NDCG at rank 4 for ranking function 1 and 2.

Solution

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4} \right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Solution

	$NDCG_{RF1}$=1.00	$NDCG_{RF2}$=0.9203