

Introduction to **Information Retrieval**

Lecture 7

Evaluation

How do you evaluate a search engine / algorithm [say for e-commerce]

- How fast does it index?
 - Number of documents/hour
 - Incremental indexing – site adds 10K products/day
- How fast does it search?
 - Latency and CPU needs for site's 5 million products
- Does it recommend related products?
- This is all good, but it says nothing about the *quality* of search
 - You want the users to be happy with the search experience

How do you tell if users are happy?

- Search returns products relevant to users
 - How do you assess this at scale?
- Search results get clicked a lot
 - Misleading titles/summaries can cause users to click
- Users buy after using the search engine
 - Or, users spend a lot of \$ after using the search engine
- Repeat visitors/buyers
 - Do users leave soon after searching?
 - Do they come back within a week/month/... ?

Happiness: elusive to measure

- Most common proxy: *relevance* of search results
 - But how do you measure relevance?
- Pioneered by Cyril Cleverdon in the Cranfield Experiments

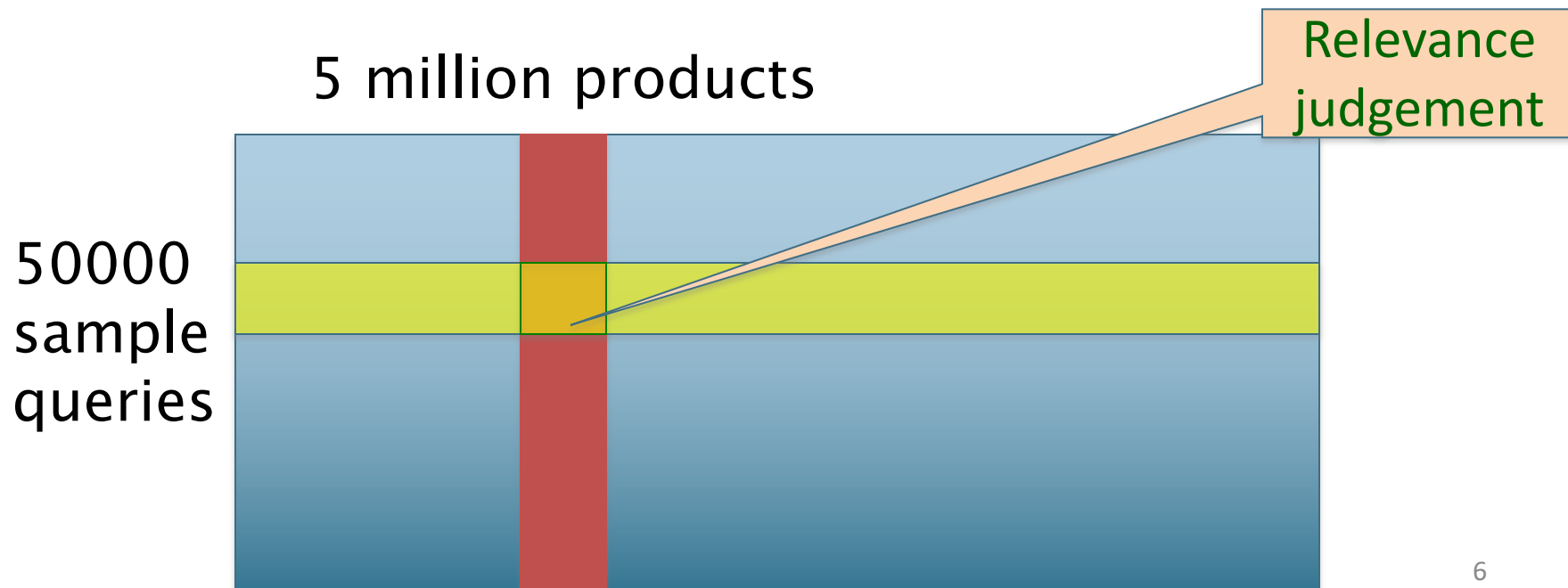


Measuring relevance

- Three elements:
 1. A benchmark document collection
 2. A benchmark suite of queries
 3. An assessment of either Relevant or Nonrelevant for each query and each document

So you want to measure the quality of a new search algorithm

- Benchmark documents – the products
- Benchmark query suite – more on this
- Judgments of document relevance for each query



Relevance judgments

- Binary (relevant vs. non-relevant) in the simplest case, more nuanced (0, 1, 2, 3 ...) in others
- What are some issues already?
- 5 million times 50K takes us into the range of a quarter trillion judgments
 - If each judgment took a human 2.5 seconds, we'd still need 10^{11} seconds, or nearly \$300 million if you pay people \$10 per hour to assess
 - 10K new products per day

Crowd source relevance judgments?

- Present query-document pairs to low-cost labor on online crowd-sourcing platforms
 - Hope that this is cheaper than hiring qualified assessors
- Lots of literature on using crowd-sourcing for such tasks
- Main takeaway – you get some signal, but the variance in the resulting judgments is very high

What else?

- Still need test queries
 - Must be appropriate to docs in corpus
 - **Must be representative of actual user needs**
 - Random query terms from the documents generally not a good idea
 - Sample from query logs if available
- Classically (non-Web)
 - Low query rates – not enough query logs
 - Experts hand-craft “user needs”

Some public test Collections

TABLE 4.3 Common Test Corpora

<i>Collection</i>	<i>NDocs</i>	<i>NQrys</i>	<i>Size (MB)</i>	<i>Term/Doc</i>	<i>Q-D RelAss</i>
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Typical
TREC

Evaluating an IR system

- Note: **user need** is translated into a **query**
- Relevance is assessed relative to the **user need**, *not* the **query**
- E.g.,
 - Information need: *My swimming pool bottom is becoming black and needs to be cleaned.*
 - Query: ***pool cleaner***
- Assess whether the doc addresses the underlying need, not whether it has these words

Now we have the basics of a benchmark

- Let's review some evaluation measures
 - *Precision*
 - *Recall*
 - DCG
 - ...

Unranked retrieval evaluation: Precision and Recall

- **Binary assessments**

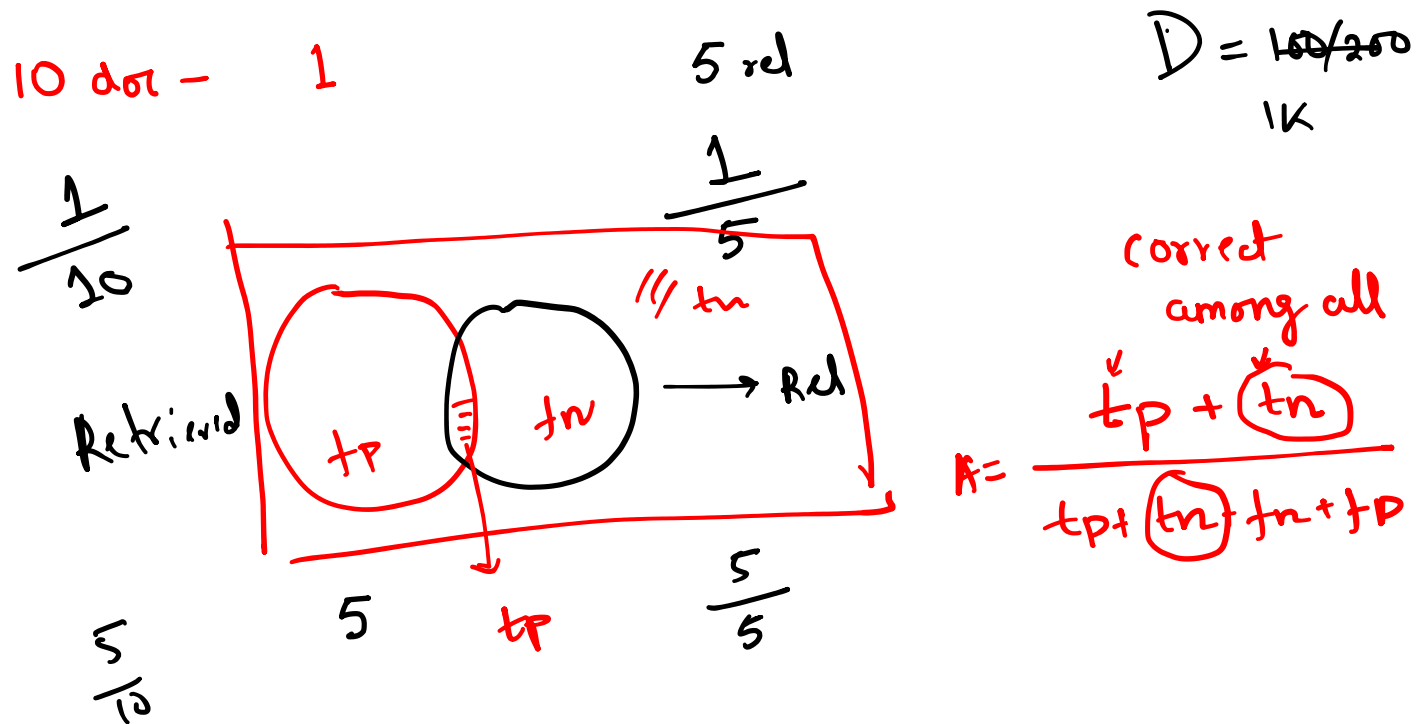
Precision: fraction of retrieved docs that are relevant =
 $P(\text{relevant} | \text{retrieved})$

Recall: fraction of relevant docs that are retrieved
= $P(\text{retrieved} | \text{relevant})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision $P = tp / (tp + fp)$
- Recall $R = tp / (tp + fn)$

Venn Diagram Based Visualization



Rank-Based Measures

- Binary relevance
 - Precision@K (P@K)
 - Mean Average Precision (MAP)
 - Mean Reciprocal Rank (MRR)
- Multiple levels of relevance
 - Normalized Discounted Cumulative Gain (NDCG)

Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K

- Ex:

- Prec@3 of ?
- Prec@4 of ?
- Prec@5 of ?

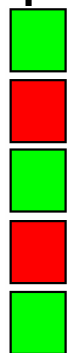


- In similar fashion we have Recall@K

Average Precision

- Consider rank position of each *relevant* doc
 - K_1, K_2, \dots, K_R
- Compute Precision@K for each $K = K_1, K_2, \dots, K_R$
- Average precision = average of P@K

■ Ex:













has AvgPrec of $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$











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

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


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

Why AP is calculated only at relevant doc positions?

Mean average precision


- MAP is Average Precision across multiple queries/rankings
- MAP is macro-averaging: each query counts equally
- Now perhaps most commonly used measure in research papers

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

 = 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

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

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

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

What if the results are not in a list?

- Suppose there's **only one Relevant Document**
- Scenarios:
 - known-item search
 - navigational queries
 - looking for a fact
- Search duration \sim Rank of the answer
 - measures a user's effort

Mean Reciprocal Rank

- Consider rank position, K , of **first relevant** doc
 - Could be – only clicked doc
- Reciprocal Rank score = $\frac{1}{K}$
- MRR is the mean RR across multiple queries

BEYOND BINARY RELEVANCE



Web Images Video Local Shopping More ▾

Toyota safety

Search

Options ▾

Search Pad

SearchScan - On

108,000,000 results for
Toyota safety:

Show All

Toyota

Motor Trend

CarsDirect

Shopping Sites

Also try: [toyota safety ratings](#), [toyota safety recall](#), [More...](#)

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 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

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](#)

Safety Toyota

Explore 5,000+ Pro Sports Choices. Save On Safety Toyota.
[BaseballGear.Shopzilla.com](#)

[See your message here...](#)

fair

fair

Good

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- 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

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of usefulness, or *gain*, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at *lower ranks*
- Typical discount is $1/\log(\textit{rank})$
 - With base 2, the discount at rank 4 is $1/2$, and at rank 8 it is $1/3$
 - Intuition: if a good document is retrieved at rank 4, system gets only half the credit that it would have got if the doc were to be retrieved at rank 1

Summarize a Ranking: DCG

- What if relevance judgments are in a scale of $[0, k]$? $k \geq 2$
- Let the ratings of the n documents be r_1, r_2, \dots, r_n (in ranked order)
- Cumulative Gain (CG) at rank n
 - $CG = r_1 + r_2 + \dots + r_n$
- Discounted Cumulative Gain (DCG) at rank n
 - $DCG = r_1 + r_2 / \log_2 2 + r_3 / \log_2 3 + \dots + r_n / \log_2 n$
 - We may use any base for the logarithm

Discounted Cumulative Gain

- *DCG* is the total gain accumulated at a particular rank p :

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

- used by some web search companies
- emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Discounted gain:
3, $2/1$, $3/1.59$, 0, 0, $1/2.59$, $2/2.81$, $2/3$, $3/3.17$, 0
= 3, 2, 1.89, 0, 0, 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
- A problem: how to compare DCG for queries having different number of relevant docs?

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**
 - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
- Normalization useful for contrasting queries with varying numbers of relevant results
- NDCG is now quite popular in evaluating Web search

NDCG for the same example

- 10 ranked documents judged on 0-3 relevance scale:
3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- Perfect 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
- Actual DCG (from two slides back):
 - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
- 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
 - $\text{NDCG} \leq 1$ at any rank position

NDCG – Another Example

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
	NDCG _{GT} =1.00		NDCG _{RF1} =1.00		NDCG _{RF2} =0.9203	

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

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

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

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