Information Retrieval (Part II)

[DAT640] Information Retrieval and Text Mining

Krisztian Balog
University of Stavanger

September 17, 2019

Outline

- ~~Search engine architecture, indexing~~
- **Evaluation** ⇐ today
- Retrieval models
- Query modeling
- Learning-to-rank, Neural IR
- Semantic search

Evaluation

“To measure is to know.
If you can not measure it,
you can not improve it.”

—Lord Kelvin

What to measure?

- **Effectiveness** ⇐ our focus
  - How accurate are the search results?
  - I.e., the system's capability of ranking relevant documents ahead of non-relevant ones
- Efficiency
  - How quickly can a user get the results?
  - I.e., the response time of the system
- Usability
  - How useful is the system for real user tasks?

Evaluation in IR

- Search engine evaluation must rely on users!
- Core question: How we can get users involved?

Types of evaluation

- **Offline** (test collection based) $\Leftarrow$ our focus
- **Online** (live evaluation) $\Leftarrow$ our focus
- User studies
- Simulation of users
- ...

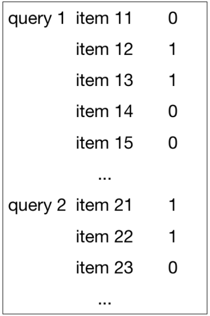
Offline evaluation

Test collection based evaluation

- *Cranfield evaluation methodology*
- Basic idea: Build reusable test collections
- Ingredients of an IR test collection
  - Dataset (corpus of documents or *information objects*)
  - Test queries (set of *information needs*)
  - Relevance assessments
  - Evaluation measures

Relevance assessments

- Ground truth labels for query-item pairs
- **Binary**
  - 0: non-relevant
  - 1: relevant
- **Graded**, for example,
  - -1: spam / junk
  - 0: non-relevant
  - 1: somewhat relevant
  - 2: relevant
  - 3: highly relevant / perfect match



query 1	item 11	0
	item 12	1
	item 13	1
	item 14	0
	item 15	0
	...	
query 2	item 21	1
	item 22	1
	item 23	0
	...	

ground truth with binary assessments

Obtaining relevance assessments

- Obtaining relevance judgments is an expensive, time-consuming process
  - Who does it?
  - What are the instructions?
  - What is the level of agreement?
- Two approaches
  - Expert judges
  - Crowdsourcing

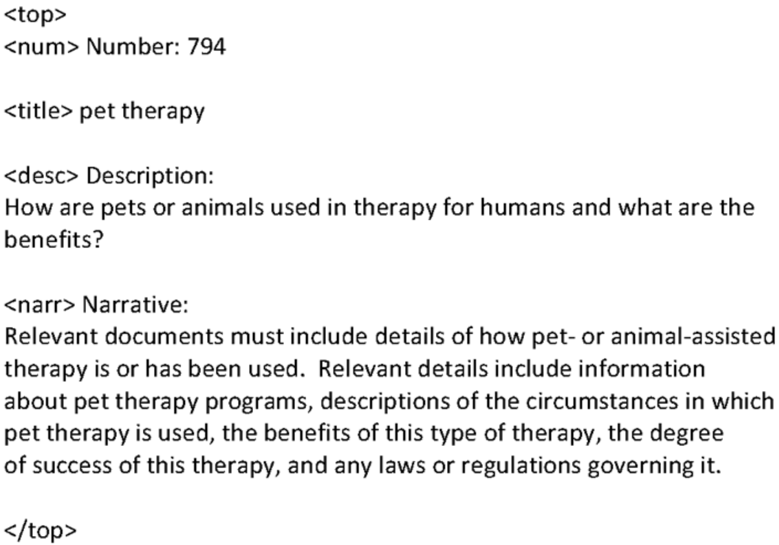
Text Retrieval Conference (TREC)

- Organized by the US National Institute of Standards and Technology (NIST)
- Yearly benchmarking cycle
- Developing test collections for various information retrieval tasks
- Relevance judgments created by expert judges, i.e., retired information analysts (CIA)

Examples of TREC document collections

Name	#Documents	Size
CACM	3k	2.2 MB
AP	242k	0.7 GB
GOV2	25M	426 GB
ClueWeb09	1B	25 TB

TREC topic example



```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Crowdsourcing

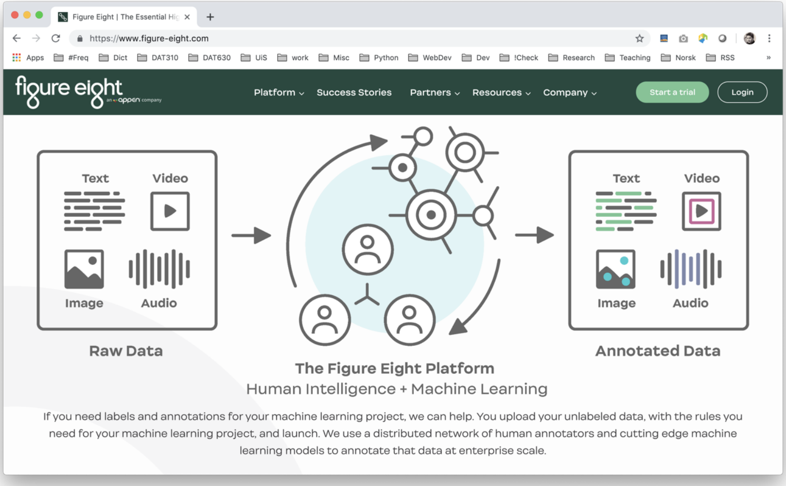
- Obtain relevance judgments on a crowdsourcing platform
    - Often branded as “human intelligence platforms”
- “Microtasks” are performed in parallel by large, paid crowds

amazon mechanical turk

figure eight

CrowdTruth

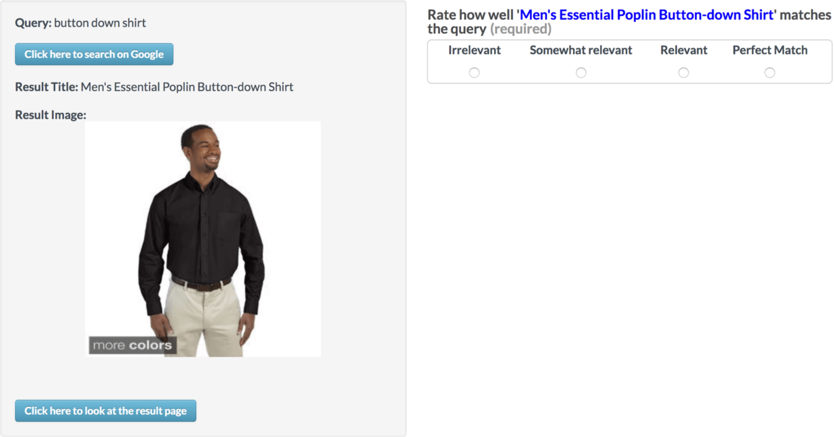
The framework for crowdsourcing ground truth data



The Figure Eight Platform
Human Intelligence + Machine Learning

If you need labels and annotations for your machine learning project, we can help. You upload your unlabeled data, with the rules you need for your machine learning project, and launch. We use a distributed network of human annotators and cutting edge machine learning models to annotate that data at enterprise scale.

Example microtask



Query: button down shirt

Result Title: Men's Essential Poplin Button-down Shirt

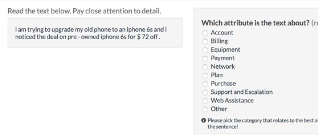
Result Image:

Rate how well 'Men's Essential Poplin Button-down Shirt' matches the query (required)

Irrelevant	Somewhat relevant	Relevant	Perfect Match
○	○	○	○

Click here to search on Google
more colors
Click here to look at the result page

Other search related annotation tasks



Read the text below. Pay close attention to detail.

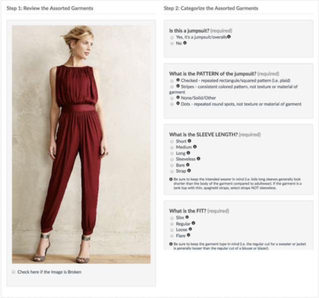
i am trying to upgrade my old phone to an iphone 6s and i noticed the deal on pre - owned iphone 6s for $ 72 off .

Which attribute is the text about? (re

○ Account
○ Billing
○ Equipment
○ Payment
○ Network
○ Plan
○ Purchase
○ Support and Escalation
○ Web Assistance
○ Other

Please pick the category that relates to the best o the sentence!

Intent classification



Step 1: Review the Assorted Garments

□ Check here if the Image is Broken

Step 2: Categorize the Assorted Garments

Is this a jumpsuit? (required)

○ Yes, it's a jumpsuit/overalls
○ No

What is the PATTERN of the jumpsuit? (required)

○ Checked - repeated rectangule/squared pattern (i.e. plaid)
○ Stripes - consistent colored pattern, not texture or material of garment
○ None/Solid/Other
○ Dots - repeated round spots, not texture or material of garment

What is the SLEEVE LENGTH? (required)

○ Short
○ Medium
○ Long
○ Sleeveless
○ Bare
○ Strap

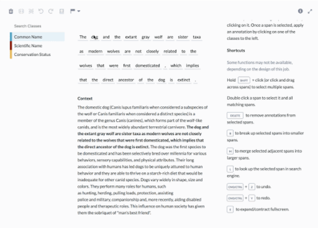
Be sure to keep the intended wearer in mind (i.e. kids long sleeves generally look shorter than the body of the garment compared to adultwear). If the garment is a tank top with thin, spaghetti straps, select straps NOT sleeveless.

What is the FIT? (required)

○ Slim
○ Regular
○ Loose
○ Flare

Be sure to keep the garment type in mind (i.e. the regular cut for a sweater or jacket is generally looser than the regular cut of a blouse or blazer).

Content categorization



The dog and the extant gray wolf are sister taxa as modern wolves are not closely related to the wolves that were first domesticated , which implies that the direct ancestor of the dog is extinct .

Context

The domestic dog (Canis lupus familiaris when considered a subspecies of the wolf or Canis familiaris when considered a distinct species) is a member of the genus Canis (canines), which forms part of the wolf-like canids, and is the most widely abundant terrestrial carnivore. **The dog and the extant gray wolf are sister taxa as modern wolves are not closely related to the wolves that were first domesticated, which implies that the direct ancestor of the dog is extinct.** The dog was the first species to be domesticated and has been selectively bred over millennia for various behaviors, sensory capabilities, and physical attributes. Their long association with humans has led dogs to be uniquely attuned to human behavior and they are able to thrive on a starch-rich diet that would be inadequate for other canid species. Dogs vary widely in shape, size and colors. They perform many roles for humans, such as hunting, herding, pulling loads, protection, assisting police and military, companionship and, more recently, aiding disabled people and therapeutic roles. This influence on human society has given them the sobriquet of "man's best friend".

clicking on it. Once a span is selected, apply an annotation by clicking on one of the classes to the left.

Shortcuts

Some functions may not be available, depending on the design of this job.

Hold SHIFT + click (or click and drag across spans) to select multiple spans.

Double click a span to select it and all matching spans.

DELETE to remove annotations from selected spans.

B to break up selected spans into smaller spans.

M to merge selected adjacent spans into larger spans.

L to look up the selected span in search engine.

CMD/CTRL + Z to undo.

CMD/CTRL + Y to redo.

E to expand/contract fullscreen.

Text annotation

Expert judges vs. crowdsourcing

- Expert judges
    - Each query-item pair is commonly assessed by a single person
    - Agreement is good because of “narrative”
- Crowdsourcing
    - Assessments are more noisy
    - Commonly, majority vote is taken
        - The number of labels collected for an item may be adjusted dynamically such that a majority decision is reached
- **Data is only as good as the guidelines!**

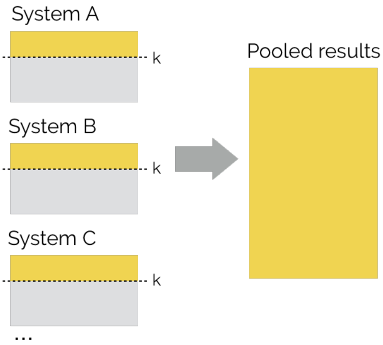
Discussion

Question

How can the relevance of all items be assessed in a large dataset for a given query?

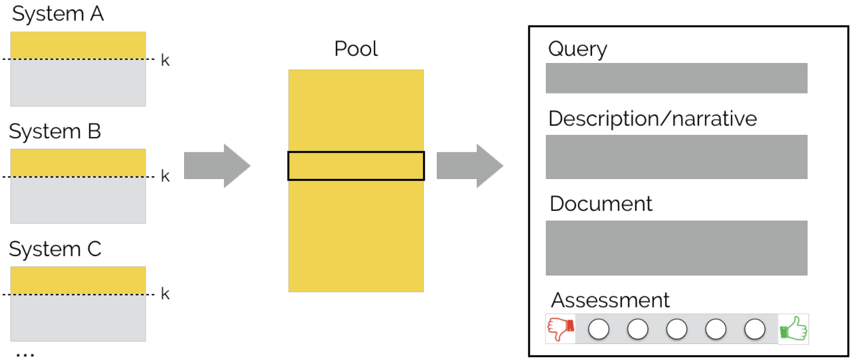
Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Top-k results from different systems (algorithms) are merged into a pool
  - Duplicates are removed
  - Item order is randomized
- Produces a large number of relevance judgments for each query, although still incomplete
  - Not assessed items are assumed to be non-relevant



Pooling

- Relevance assessments are collected for all document in the pool
    - Either using expert judges or crowd workers



Test collection based evaluation

- Ingredients of an IR test collection
  - ~~Dataset (corpus of documents or *information objects*)~~
  - ~~Test queries (set of *information needs*)~~
  - ~~Relevance assessments~~
  - **Evaluation measures**

IR evaluation measures

- Assessing the quality of a ranked list against the ground truth relevance labels
  - Commonly, a real number between 0 and 1
- **Important**: All measures are based on a (simplified) model of user needs and behavior
  - That is, the right measure depends on the particular task

Effectiveness measures

- A is the set of **relevant** documents
- B is the set of **retrieved** documents

	Relevant	Non-relevant
Retrieved	$\lvert A \cap B \rvert$	$\lvert \overline{A} \cap B \rvert$
Not retrieved	$\lvert A \cap \overline{B} \rvert$	$\lvert \overline{A} \cap \overline{B} \rvert$

Precision and recall analogously to before:

$$P = \frac{|A \cap B|}{|B|} \qquad R = \frac{|A \cap B|}{|A|}$$

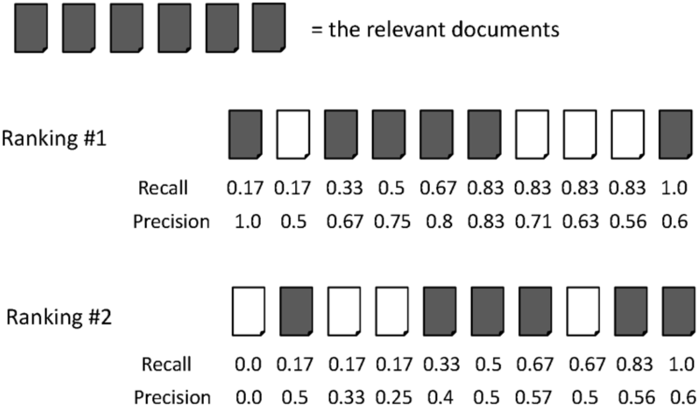
Discussion

Question

Precision and Recall are set-based metrics. How can we use them to evaluate ranked lists?

Evaluating rankings

Calculate recall and precision values at every rank position



■ ■ ■ ■ ■ ■ = 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

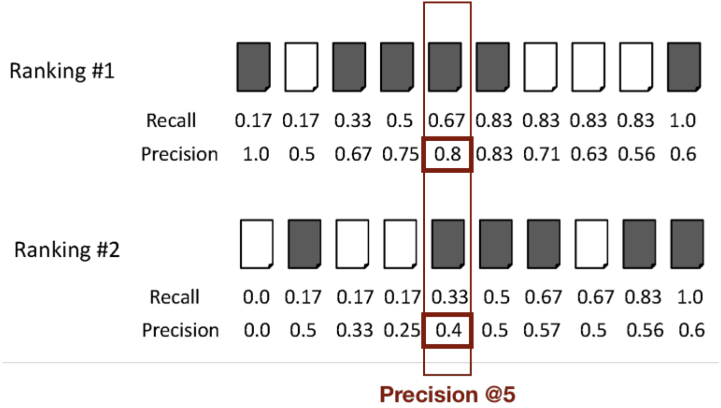
Evaluating rankings

- Calculating recall and precision values at every rank position produces a long list of numbers (see previous slide)
- Need to **summarize** the effectiveness of a ranking
- Various alternatives
  - Calculate recall and precision at fixed rank positions (P@k, R@k)
  - Calculate precision at standard recall levels, from 0.0 to 1.0 (requires interpolation)
  - Averaging the precision values from the rank positions where a relevant document was retrieved (AP)

Fixed rank positions

Compute precision/recall at a given rank position k (P@k, R@k)

- This measure does not distinguish between differences in the rankings at positions 1 to k



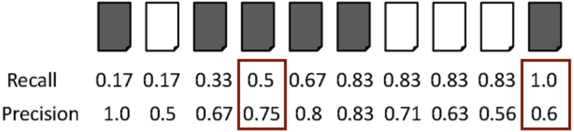
Ranking #1	■	□	■	■	■	■	□	□	□	■
Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6
Ranking #2	□	■	□	□	■	■	■	□	■	■
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

Precision @5

Standard recall levels

Calculate precision at standard recall levels, from 0.0 to 1.0

- Each ranking is then represented using 11 numbers
- Values of precision at these standard recall levels are often not available, for example:



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

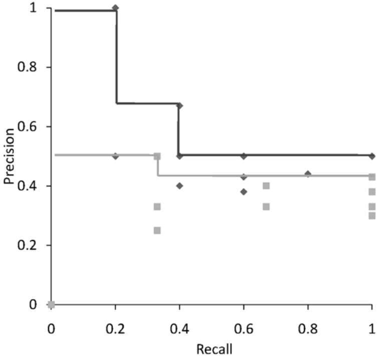
- *Interpolation* is needed

Interpolation

- To average graphs, calculate precision at standard recall levels:

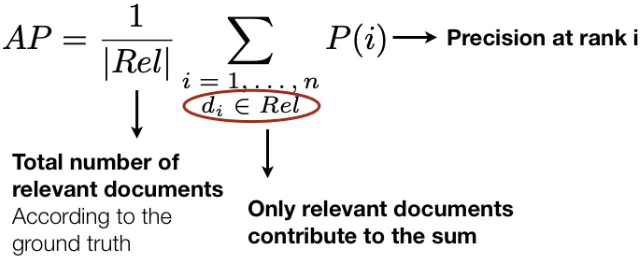
$$P(R) = \text{max}\{P' : R' \geq R \wedge (R', P') \in S\}$$

  - where S is the set of observed (R, P) points
- Defines precision at any recall level as the maximum precision observed in any recall-precision point at a higher recall level
- Produces a step function



Average Precision

- Average the precision values from the rank positions where a relevant document was retrieved



$$AP = \frac{1}{|Rel|} \sum_{\substack{i = 1, \dots, n \\ d_i \in Rel}} P(i) \longrightarrow$$

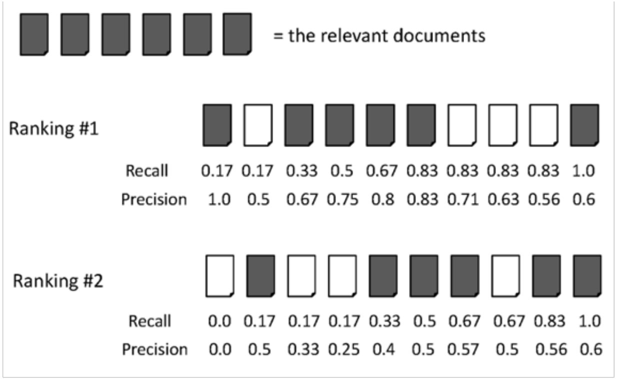
Precision at rank i

Total number of relevant documents
According to the ground truth

Only relevant documents contribute to the sum

- If a relevant document is not retrieved (in the top k ranks, e.g, $k = 1000$) then its contribution is 0.0
- AP is single number that is based on the ranking of all the relevant documents
- The value depends heavily on the highly ranked relevant documents

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

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

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

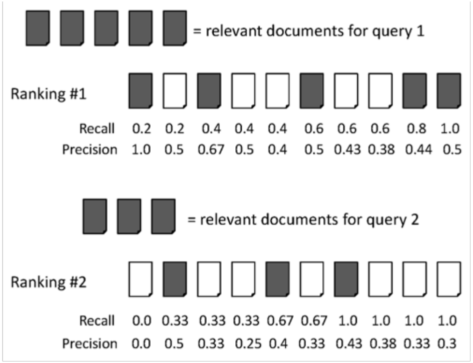
Averaging across queries

- So far: measuring ranking effectiveness on a **single query**
- Need: measure ranking effectiveness on a **set of queries**
- Average is computed over the set of queries

Mean Average Precision (MAP)

- Summarize rankings from multiple queries by averaging Average Precision
- Very succinct summary
- Most commonly used measure in research papers
- Assumes user is interested in finding many relevant documents for each query
- Requires many relevance judgments

Mean Average Precision



■ ■ ■ ■ ■ = 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

$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

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

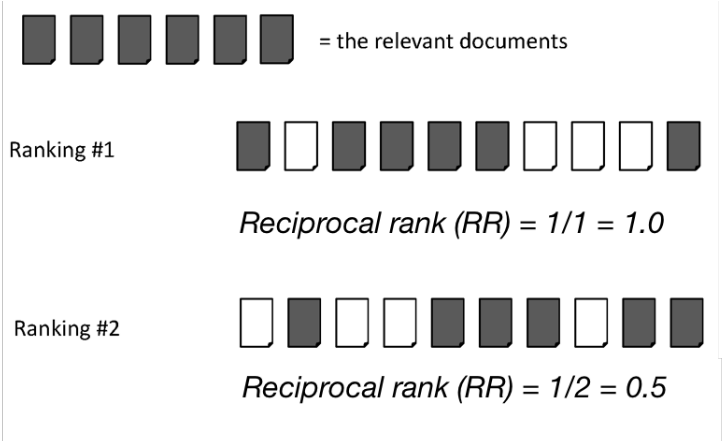
Focusing on top documents

- Users tend to look at only the top part of the ranked result list to find relevant documents
- Some search tasks have only one relevant document
  - E.g., navigational search, question answering
- Recall in those cases is not appropriate
  - Instead need to measure how well the search engine does at retrieving relevant documents at very high ranks

Focusing on top documents

- Precision at rank k (P@k)
  - k is typically 5, 10, 20
  - Easy to compute, average, understand
  - Not sensitive to rank positions less than k
- Reciprocal Rank (RR)
  - Reciprocal of the rank at which the first relevant document is retrieved
  - Mean Reciprocal Rank (MRR) is the average of the reciprocal ranks over a set of queries
  - Very sensitive to rank position

Mean Reciprocal Rank



Reciprocal rank (RR) = 1/1 = 1.0

Reciprocal rank (RR) = 1/2 = 0.5

Mean reciprocal rank (MRR) = (1.0 + 0.5) /2 = 0.75

Exercise #1 (paper-based)

Compare the retrieval effectiveness of two systems in terms of P@5, P@10, Average Precision, and Reciprocal Rank.

Exercise #2 (coding)

- Implement the computation of P@5, P@10, Average Precision, and Reciprocal Rank
- Code skeleton on GitHub: `exercises/lecture_08/exercise_2.ipynb` (make a local copy)

Graded relevance

- So far: relevance in binary
- What about graded relevance levels?

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - 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

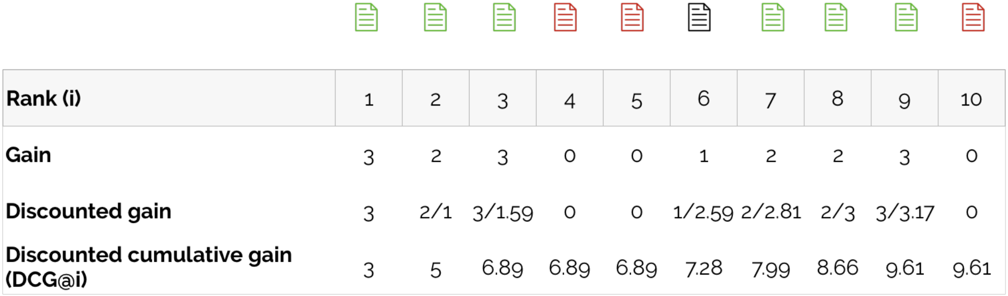
Discounted Cumulative Gain (DCG)

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

$$DCG_p = rel_1 + \sum_{i=1}^{p} \frac{rel_i}{log_2 i}$$

  - rel_i is the graded relevance level of the item retrieved at rank i
- Gain is accumulated starting at the top of the ranking and discounted by 1/log (rank)
  - E.g., discount at rank 4 is 1/2, and at rank 8 it is 1/3
- Average over the set of test queries
- Note: search engine companies have their own (secret) variants

Discounted Cumulative Gain

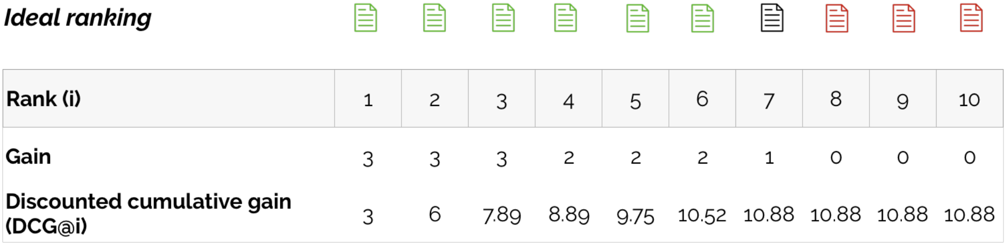


Rank (i)	1	2	3	4	5	6	7	8	9	10
Gain	3	2	3	0	0	1	2	2	3	0
Discounted gain	3	2/1	3/1.59	0	0	1/2.59	2/2.81	2/3	3/3.17	0
Discounted cumulative gain (DCG@i)	3	5	6.89	6.89	6.89	7.28	7.99	8.66	9.61	9.61

How good is a DCG@10 value of 9.61?

Normalized Discounted Cumulative Gain (NDCG)

- DCG values are often normalized by comparing the DCG at each rank with the DCG value for the perfect (ideal) ranking
  - I.e., divide DCG@i value with the ideal DCG value at rank i
  - Yields value between 0 and 1



Ideal ranking 🟩 🟩 🟩 🟩 🟩 🟩 ⬛ 🟥 🟥 🟥

Rank (i)	1	2	3	4	5	6	7	8	9	10
Gain	3	3	3	2	2	2	1	0	0	0
Discounted cumulative gain (DCG@i)	3	6	7.89	8.89	9.75	10.52	10.88	10.88	10.88	10.88

Exercise #3 (paper-based)

Evaluate the retrieval effectiveness of a systems in terms of NDCG@5 and NDCG@10.

Exercise #4 (coding)

- Implement the computation of NDCG
- Code skeleton on GitHub: `exercises/lecture_08/exercise_4.ipynb` (make a local copy)

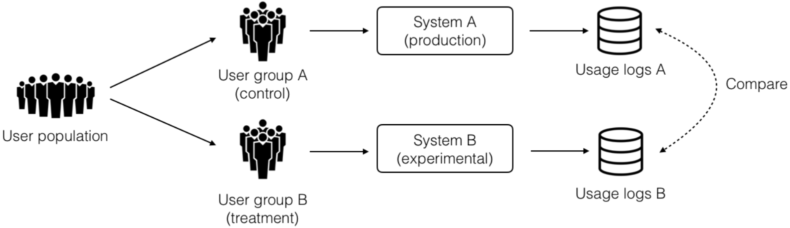
Online evaluation

Online evaluation

- **Idea**: See how normal users interact with a live retrieval system ("living lab") when just using it
- Observe implicit behavior
  - Clicks, skips, saves, forwards, bookmarks, likes, etc.
- Try to infer differences in behavior from different flavors of the live system
  - A/B testing, interleaving

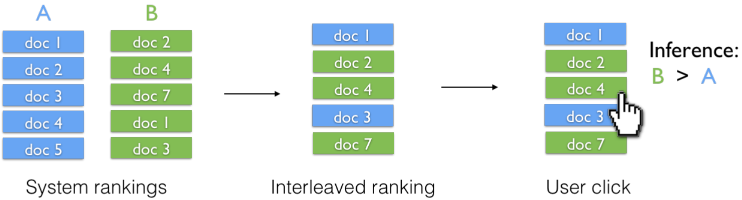
A/B testing

- Users are divided into two control (**A**) and treatment (**B**) groups
  - **A** uses the production system
  - **B** uses an experimental system
- Measure relative system performance based on usage logs



Interleaving

- Combine two rankings (A and B) into a single list
- Determine a winner on each query impression
  - Can be a draw too
- Aggregate wins on a large number of impressions to determine which ranker is better

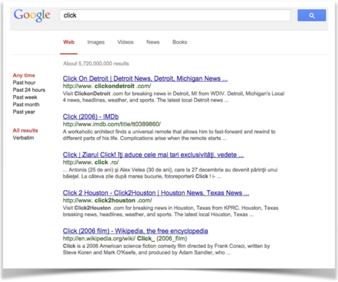


A/B testing vs. interleaving

- A/B testing
  - Between subject design
  - Can be used for evaluating any feature (new ranking algorithms, new features, UI design changes, etc.)
- Interleaving
  - Within subject design
  - Reduces variance (same users/queries for both A and B)
  - Needs 1 to 2 orders of magnitude less data
    - $\sim$100K queries for interleaving in a mature web search engine ($\gg$1M for A/B testing)
  - Limited to evaluating ranked lists

Measures in online evaluation

- **Inferred from observable user behavior**
- Clicks
- Mouse movement
- Browser action
  - Bookmark, save, print, …
- Time
  - Dwell time, time on SERP, …
- Explicit judgment
  - Likes, favorites, …
- Query reformulations
- …



About 5,720,000,000 results

Click On Detroit | Detroit News, Detroit, Michigan News ...
http://www. **clickondetroit** .com/
Visit **ClickonDetroit** .com for breaking news in Detroit, MI from WDIV. Detroit, Michigan's Local 4 news, headlines, weather, and sports. The latest local Detroit news ...

Click (2006) - IMDb
http://www.imdb.com/title/tt0389860/
A workaholic architect finds a universal remote that allows him to fast-forward and rewind to different parts of his life. Complications arise when the remote starts ...

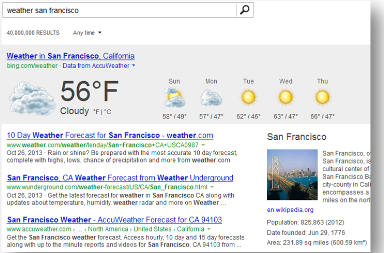
Click | Ziarul Click! îţi aduce cele mai tari exclusivităţi, vedete ...
http://www. **click** .ro/
... Antonia (25 de ani) şi Alex Velea (30 de ani), care la 27 decembrie au devenit părinţii unui băieţel. La câteva zile după marea bucurie, fotoreporterii **Click** ! i- ...

Click 2 Houston - Click2Houston | Houston News, Texas News ...
http://www. **click2houston** .com/
Visit **Click2Houston** .com for breaking news in Houston, Texas from KPRC. Houston, Texas breaking news, headlines, weather, and sports. The latest local Houston, Texas ...

Click (2006 film) - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/ **Click_** (2006_film)
Click is a 2006 American science fiction comedy film directed by Frank Coraci, written by Steve Koren and Mark O'Keefe, and produced by Adam Sandler, who ...

Challenges in online evaluation

- **Simple measures break!**



weather san francisco

40,000,000 RESULTS Any time

Weather in San Francisco, California
bing.com/weather · Data from AccuWeather

56°F
Cloudy °F | °C

Sun	Mon	Tue	Wed	Thu
58° / 49°	57° / 47°	62° / 46°	63° / 47°	66° / 47°

10 Day Weather Forecast for San Francisco - weather.com
www.weather.com/weather/tenday/San+Francisco+CA+USCA0987
Oct 26, 2013 · Rain or shine? Be prepared with the most accurate 10 day forecast, complete with highs, lows, chance of precipitation and more from weather.com

San Francisco, CA Weather Forecast from Weather Underground
www.wunderground.com/weather-forecast/US/CA/San_Francisco.html
Oct 26, 2013 · Get the latest forecast for weather in San Francisco CA along with updates about temperature, humidity, weather radar and more on Weather …

San Francisco Weather - AccuWeather Forecast for CA 94103
www.accuweather.com › … › North America › United States › California
Get the San Francisco weather forecast. Access hourly, 10 day and 15 day forecasts along with up to the minute reports and videos for San Francisco, CA 94103 from …

San Francisco

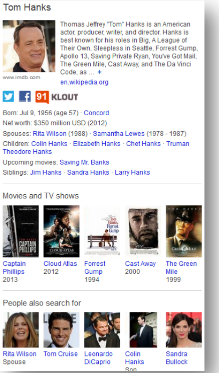
San Francisco, o
San Francisco, is
cultural center of
San Francisco Ba
city-county in Cali
encompasses a
miles on the nort

en.wikipedia.org

Population: 825,863 (2012)
Date founded: Jun 29, 1776
Area: 231.89 sq miles (600.59 km²)

Instant answers

(satisfaction not observable)



Tom Hanks

Thomas Jeffrey "Tom" Hanks is an American actor, producer, writer, and director. Hanks is best known for his roles in Big, A League of Their Own, Sleepless in Seattle, Forrest Gump, Apollo 13, Saving Private Ryan, You've Got Mail, The Green Mile, Cast Away, and The Da Vinci Code, as … +

www.imdb.com

en.wikipedia.org

91 KLOUT

Born: Jul 9, 1956 (age 57) · Concord
Net worth: $350 million USD (2012)
Spouses: Rita Wilson (1988) · Samantha Lewes (1978 - 1987)
Children: Colin Hanks · Elizabeth Hanks · Chet Hanks · Truman Theodore Hanks
Upcoming movies: Saving Mr. Banks
Siblings: Jim Hanks · Sandra Hanks · Larry Hanks

Movies and TV shows

Captain Phillips 2013
Cloud Atlas 2012
Forrest Gump 1994
Cast Away 2000
The Green Mile 1999

People also search for

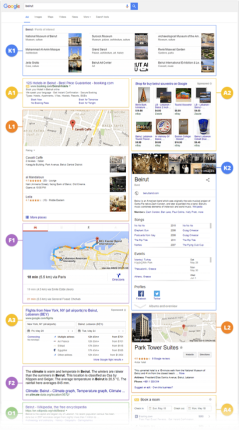
Rita Wilson Spouse
Tom Cruise
Leonardo DiCaprio
Colin Hanks Son
Sandra Bullock

Exploration

(more time/queries is not necessarily bad effort)

Challenges in online evaluation

- **Whole page relevance**
- Page is composed by a layered stack of modules
  - ○ Web result ranking
  - ○ ⇒ Result caption generation
  - ○ ⇒ Answer triggering/ranking
  - ○ ⇒ Knowledge panel composition
  - ○ ⇒ Whole page composition
- Changes in modules lower in the stack have upstream effects



Pros and cons of online evaluation

- Advantages
  - No need for expensive dataset creation
  - Perfectly realistic setting: (most) users are not even aware that they are guinea pigs
  - Scales very well: can include millions of users
- Disadvantages
  - Requires a service with lots of users
  - Can be highly nontrivial how to interpret implicit feedback signals
  - Experiments are difficult to repeat

Offline vs. online evaluation

	Offline	**Online**
Basic assumption	Assessors tell you what is relevant	Observable user behavior can tell you what is relevant
Quality	Data is only as good as the guidelines	Real user data, real and representative information needs
Realisticity	Simplified scenario, cannot go beyond a certain level of complexity	Perfectly realistic setting (users are not aware that they are guinea pigs)
Assessment cost	Expensive	Cheap
Scalability	Doesn't scale	Scales very well
Repeatability	Repeatable	Not repeatable
Throughput	High	Low
Risk	None	High

Assignment 2A

Reading

- Text Data Management and Analysis (Zhai&Massung), Chapter 9