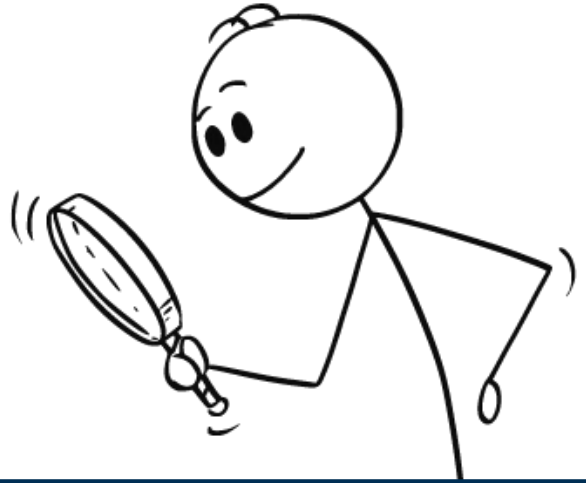


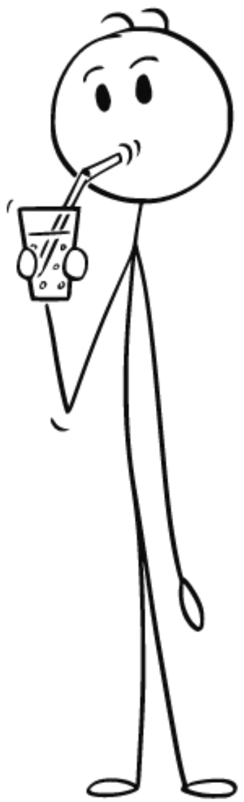
CSE 4/535 Information Retrieval

Sayantan Pal
PhD Student, Department of CSE
338Z Davis Hall

University at Buffalo
Department of CSE

Before we start

1. Project 2 will be released today
2. Any doubts related to midterm or project 1 grading, join office hours
3. Today's lecture - Evaluation Methodology Result Summaries
4. First 20 mins project 2 discussion



Recap - Previous Class

1. Efficient Scoring in a Complete Search System
2. Speeding up vector space ranking

Today’s Lecture…

- Results summaries:
    - Making our good results usable to a user
- How do we know if our results are any good?
    - Evaluating a search engine
        - Benchmarks
        - Precision and recall

Result Summaries

- Having ranked the documents matching a query, we wish to present a results list
- Most commonly, a list of the document titles plus a short summary, aka “10 blue links”

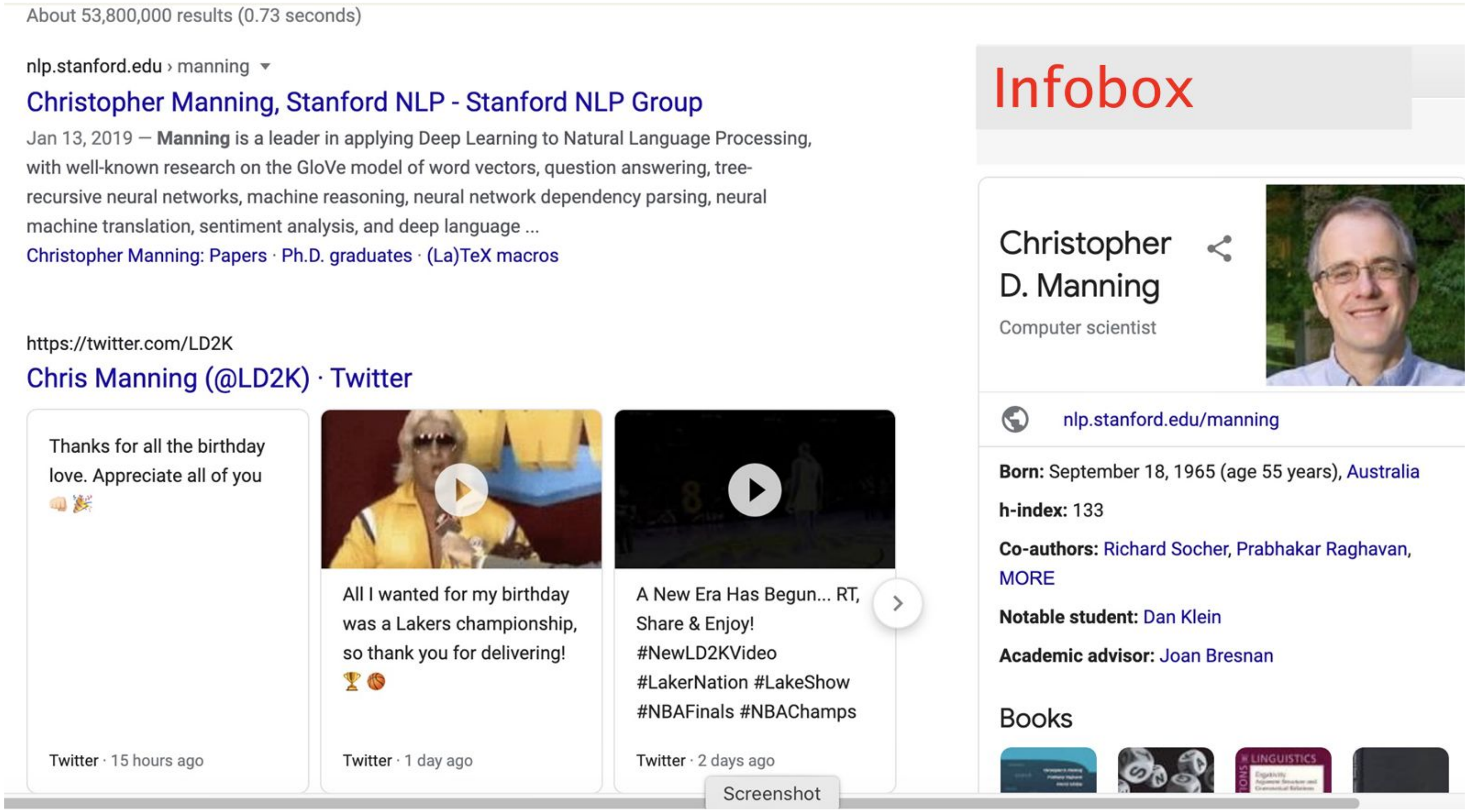
John McCain
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com · Cached page

JohnMcCain.com - **McCain**-Palin 2008
John McCain 2008 - The Official Website of **John McCain**'s 2008 Campaign for President ... African American Coalition; Americans of Faith; American Indians for **McCain**; Americans with ...
www.**johnmccain**.com/Informing/Issues · Cached page

John McCain News- msnbc.com
Complete political coverage of **John McCain**. ... Republican leaders said Saturday that they were worried that Sen. **John McCain** was heading for defeat unless he brought stability to ...
www.msnbc.msn.com/id/16438320 · Cached page

John McCain | Facebook
Welcome to the official Facebook Page of **John McCain**. Get exclusive content and interact with **John McCain** right from Facebook. Join Facebook to create your own Page or to start ...
www.facebook.com/**johnmccain** · Cached page

Search Results Today: “Chris Manning”



Google Knowledge Graph

Learn more

Knowledge Graph

From Wikipedia, the free encyclopedia

A request that this article title be changed to *Google Knowledge Graph* is under discussion. Please **do not move** this article until the discussion is closed.

This article is about Google's implementation of a knowledge graph. For the general concept in information science, see Knowledge graph.

The **Google Knowledge Graph** is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources. The information is presented to users in an infobox next to the search results. These infoboxes were added to Google's search engine in May 2012, starting in the United States, with international expansion by the end of the year.[1] Google has referred to these infoboxes, which appear to the right (top on mobile) of search results, as "knowledge panels".[2]

The information covered by Google's Knowledge Graph grew quickly after launch, tripling its size within seven months (covering 570 million entities and 18 billion facts[3]). By mid-2016, Google reported that it held 70 billion facts[4] and answered "roughly one-third" of the 100 billion monthly searches they handled. By May 2020, this had grown to 500 billion facts on 5 billion entities.[5]

There is no official documentation of how the Google Knowledge Graph is implemented.[6] According to Google, its information is retrieved from many sources, including the *CIA World Factbook*, Wikidata, and Wikipedia.[1][7] It is used to answer direct spoken questions in Google Assistant[8][9] and Google Home voice queries.[10] It has been criticized for providing answers without source attribution or citation.[11]

Contents [hide]

Screenshot

Summaries

- The title is typically automatically extracted from document metadata. What about the summaries?
  - This description is **crucial**.
  - User can identify **good/relevant hits based** on **description**.
- Two basic kinds:
  - **Static**
  - **Dynamic**
- A static summary of a document is **always the same**, regardless of the query that hit the doc
- A dynamic summary is a **query-dependent** attempt to explain why the document was retrieved for the query at hand

Static summaries

- In typical systems, the static summary is a subset of the document
- Simplest heuristic: the first 50 (or so –this can be varied) words of the document
  - Summary cached at indexing time
- More sophisticated: extract from each document a set of “key”sentences
  - Simple NLP heuristics to score each sentence
  - Summary is made up of top-scoring sentences.
- Most sophisticated: NLP used to synthesize a summary
  - Seldom/rarely used in IR

Dynamic summaries

- Present one or more “windows” within the document that contain several of the query terms
  - **“KWIC” snippets: Keyword in Context presentation**
- Generated in conjunction with scoring
  - If query **found as a phrase**, all or some occurrences of the phrase in the doc
  - If not, document windows that **contain multiple query terms**
- The summary itself gives the entire content of the window –all terms, not only the query terms –how?

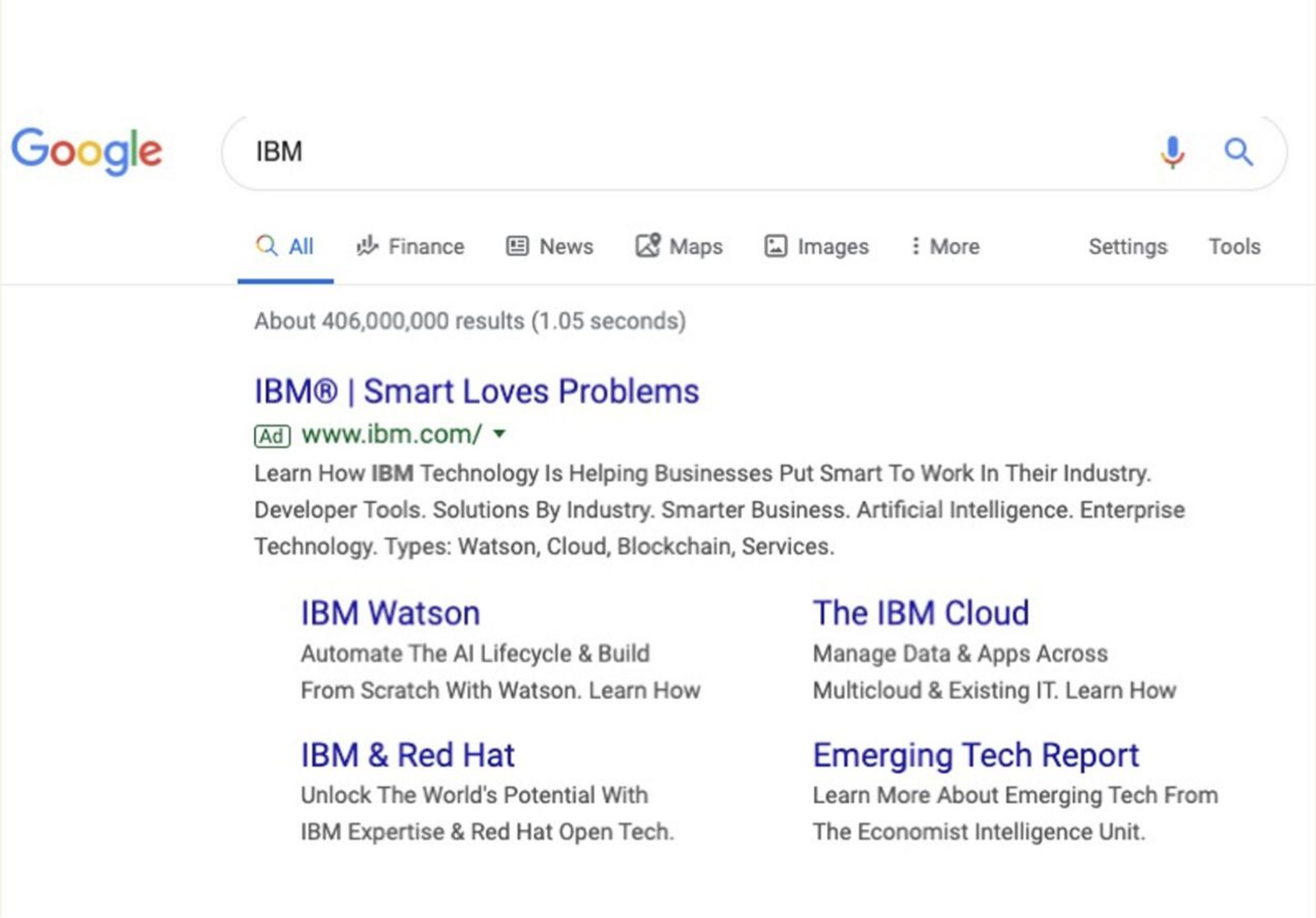
Generating dynamic summaries

- If we have only a positional index, we cannot (easily) reconstruct context window surrounding hits
- If we **cache the documents at index time**, can find windows in it, cueing from hits found in the positional index
  - E.g., positional index says “the query is a phrase in position 4378” so we go to this position in the cached document and stream out the content
- Most often, cache only a fixed-size prefix of the doc
  - Note: Cached copy can be outdated

Dynamic summaries

- Producing good dynamic summaries is a tricky optimization problem
  - The real estate for the summary is normally small and fixed
  - Want short item, so show as many KWIC matches as possible, and perhaps other things like title
  - Want snippets to be long enough to be useful
  - Want linguistically well-formed snippets: users prefer snippets that contain complete phrases
  - Want snippets maximally informative about doc
- But users really like snippets, even if they complicate IR system design

Helpful Result Summaries



Evaluating search engines

Situation

- Thanks to your stellar performance in CSE 435/535, you quickly rise to VP of Search at internet retail giant nozama.com. Your boss brings in her nephew Sergey, who claims to have built a better search engine for nozama. Do you
  - Laugh and send him to rival Tramlaw Labs?
  - Counsel Sergey to take CSE 435/535?
  - Try a few queries on his engine and say “Not bad”?
  - … ?

What could you ask Sergey?

- How fast does it index?
- Number of documents/hour
- Incremental indexing –nozama adds 10K products/day
- How fast does it search?
- Latency and CPU needs for nozama’s 5 million products
- Does it recommend related products?
- This is all good, but it says nothing about the quality of Sergey’s search
- You want nozama’s 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
  - Pioneered by Cyril Cleverdon in the Cranfield Experiments

- But how do you measure relevance?

Measuring relevance

- Three elements:
  - A **benchmark document collection**
  - A **benchmark suite of queries**
  - An assessment of either Relevant or Non-relevant for each query and each document

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

- Benchmark documents –nozama's 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 relevance levels also used(0, 1, 2, 3 …)
- What are some issues already?

Relevance judgments

- Binary (relevant vs. non-relevant) in the simplest case
  - More nuanced relevance levels also used(0, 1, 2, 3 …)
- 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 crowdsourcing platforms
  - Hope that this is cheaper than hiring qualified assessors
- Lots of literature on using crowdsourcing for such tasks
  - You get fairly good signal, but the variance in the resulting judgments is quite high

Link: https://www.mturk.com/

What else?

- Still need test queries
  - Must be germane to docs available
  - Must be representative of actual user needs
  - Random query terms from the documents are 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”

Early public test Collections (20^{th} C)

TABLE 4.3 Common Test Corpora

Collection	NDocs	NQrys	Size (MB)	Term/Doc	Q-D RelAss
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

Recent datasets: 100s of million web pages (GOV, ClueWeb, …)

Now we have the basics of a benchmark

1. Let's review some evaluation measures
   a. Precision
   b. Recall
   c. DCG
   d. …

Evaluating an IR system

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

Unranked retrieval evaluation: Precision and Recall

- **Binary assessments**

Precision: fraction of retrieved docs that are relevant = P(relevant|retrieved)

Recall: fraction of relevant docs that are retrieved = P(retrieved|relevant)

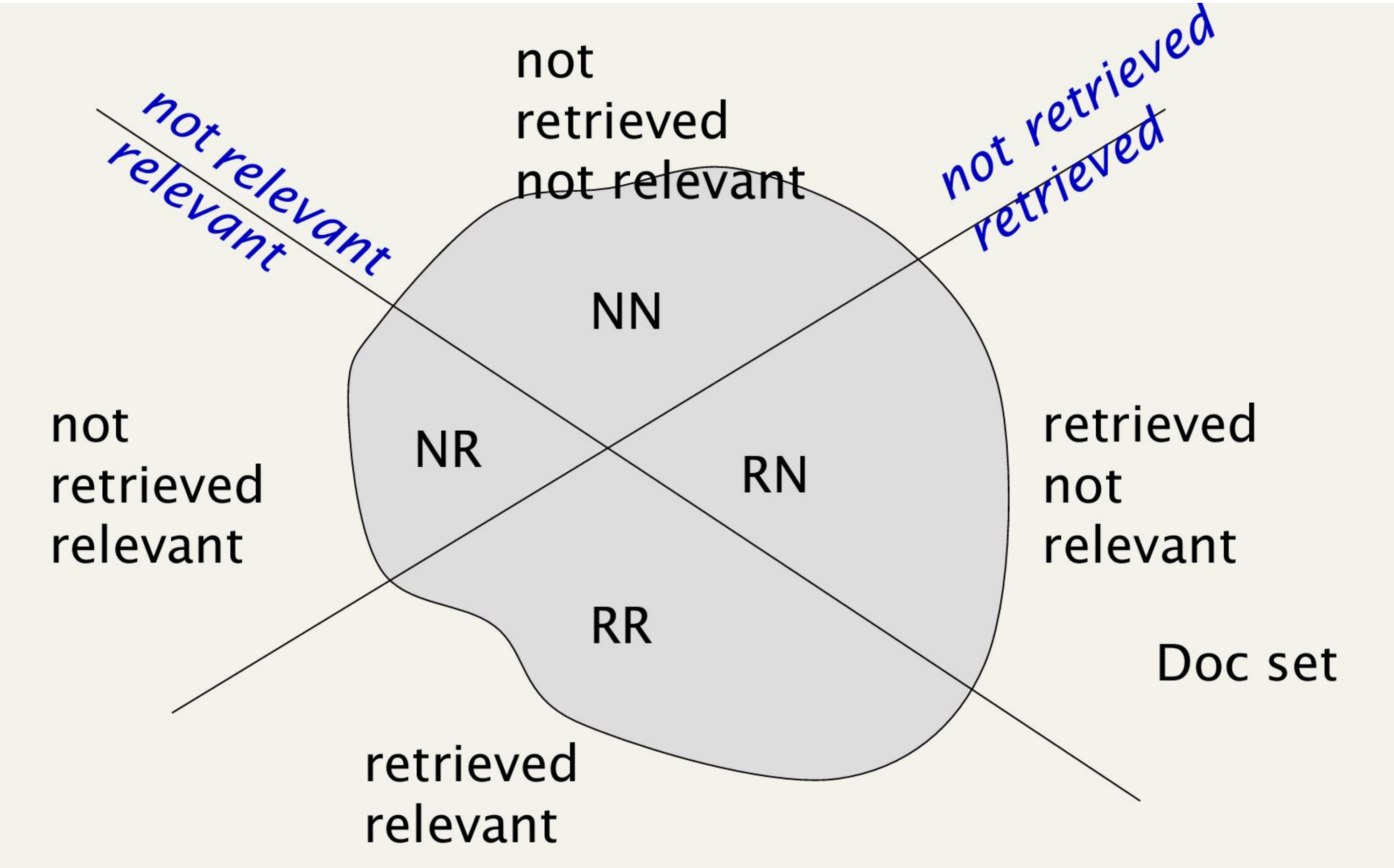
	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

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

Standard relevance benchmarks

- TREC -National Institute of Standards and Technology (NIST) has run a large IR test bed for many years
- Reuters and other benchmark doc collections used
- “Retrieval tasks” specified
  - sometimes as queries
- Human experts mark, for each query and for each doc, Relevant or Non-relevant
  - or at least for subset of docs that some system returned for that query

Measures based on relevance



How important is accuracy for IR systems

- Given a query, an engine classifies each doc as “Relevant” or “Non-relevant”
- The accuracy of an engine: the fraction of these classifications that are correct
- Accuracy is a commonly used evaluation measure in machine learning classification work
- Why is this not a very useful evaluation measure in IR?

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

snoogle.com

Search for:

Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

Precision or Recall?
Which one will you choose?

A combined measure: F

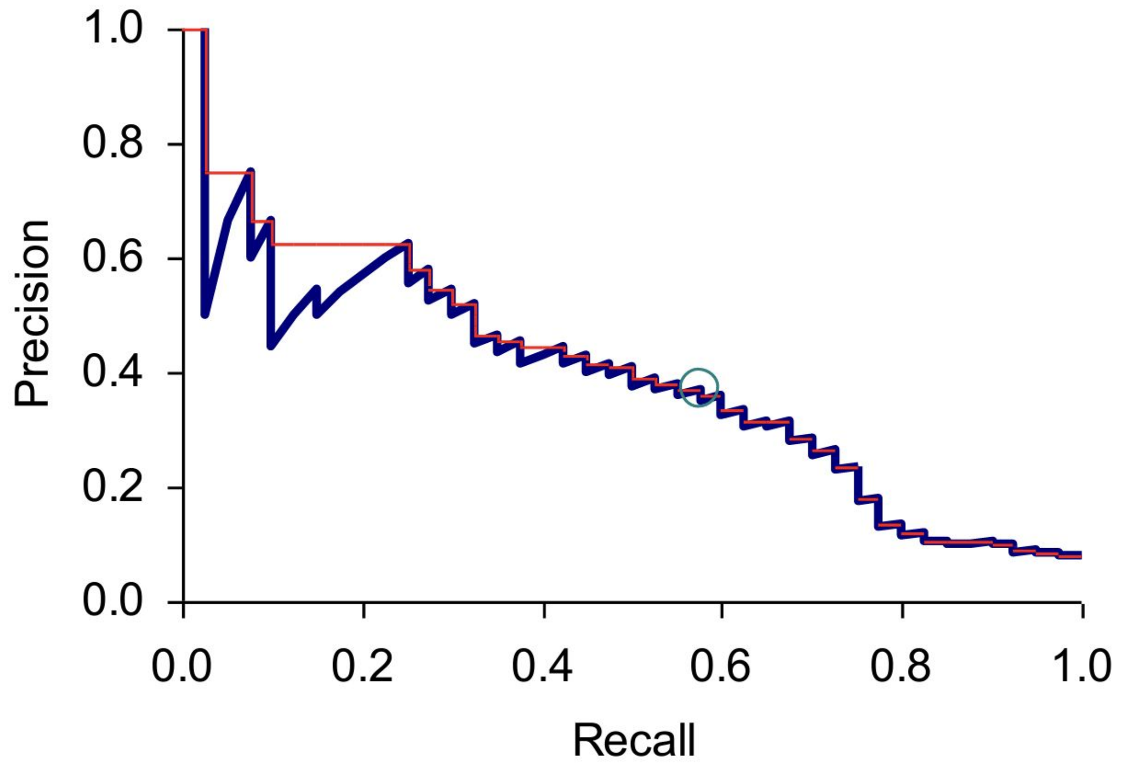
A combined measure: *F*

- Combined measure that assesses precision/recall tradeoff is **F measure** (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F_1 measure
  - i.e., with $\beta = 1$ or $\alpha = ½$ 2PR/(P+R)
- Harmonic mean is a conservative average

A precision-recall curve



If the (k+1)th doc retrieved is non-rel, recall same as for kth doc, but prec decreases

Want to remove jags (blue) by interpolated precision (red)

11-point Interpolated Avg Precision

Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

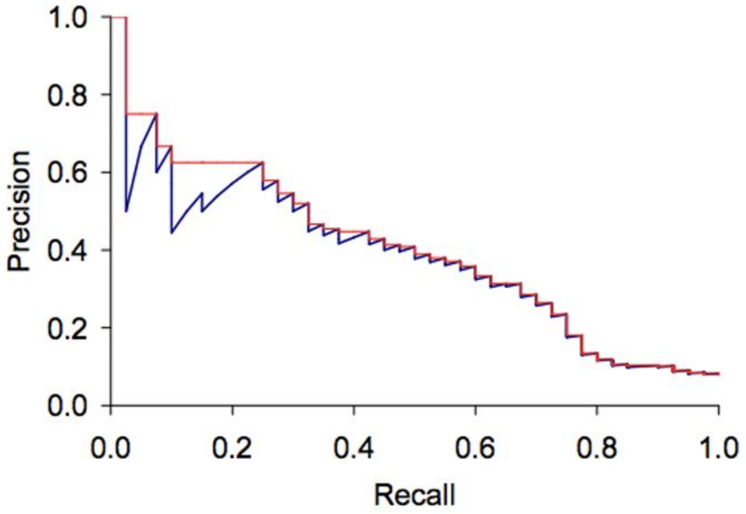
▶ **Table 8.1** Calculation of 11-point Interpolated Average Precision. This is for the precision-recall curve shown in Figure 8.2.

Precision/Recall @rank

Rank	Doc
1	d_{12}
2	d_{123}
3	d_4
4	d_{57}
5	d_{157}
6	d_{222}
7	d_{24}
8	d_{26}
9	d_{77}
10	d_{90}

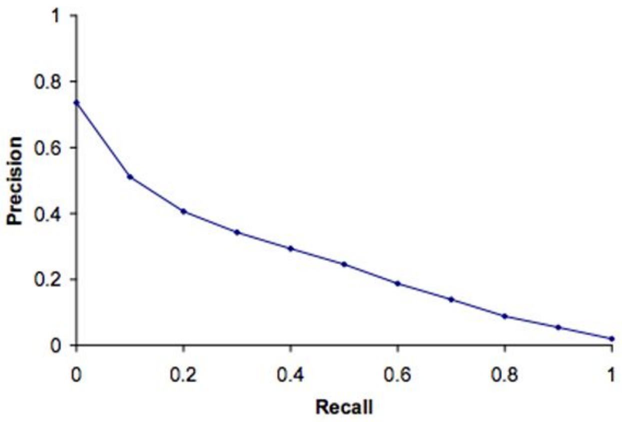
- Blue documents are relevant
- P@n: P@3=0.33, P@5=0.2, P@8=0.25
- R@n: R@3=0.33, R@5=0.33, R@8=0.66

A precision-recall curve



- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \ldots$)
- Interpolation (in red): Take maximum of all future points
- Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.

Averaged 11-point precrecall



- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, . . .
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- The curve is typical of performance levels at TREC (more later).

Algorithm for 11-point precision

$$P_{11\_pt} = \frac{1}{11} \sum_{j=0}^{10} \frac{1}{N} \sum_{i=1}^{N} \tilde{P}_i(r_j)$$

with $\tilde{P}_i(r_j)$ the precision at the jth recall point in the ith query (out of N)

- Define 11 standard recall points $r_j = \frac{j}{10}$: $r_0 = 0,\ r_1 = 0.1\ \ldots\ r_{10} = 1$
- To get $\tilde{P}_i(r_j)$, we can use $P_i(R = r_j)$ directly if a new relevant document is retrieved exacty at r_j
- Interpolation for cases where there is no exact measurement at r_j:

$$\tilde{P}_i(r_j) = \begin{cases} max(r_j \leq r < r_{j+1}) P_i(R = r) & \text{if } P_i(R = r) \text{ exists} \\ \tilde{P}_i(r_{j+1}) & \text{otherwise} \end{cases}$$

- Note that $P_i(R = 1)$ can always be measured.
- Worked avg-11-pt prec example for supervisions at end of slides.

Interpolated Precision: example

Interpolated means that for each recall value from 0.0, 0.1, 0.2 … to 1.0 (11 values) find the maximum precision at Recall Precision table where the recall value is greater than or equal to recall level.

For instance for recall level of 0.2, we need to find the maximum precision where recall is greater than or equal to 0.2 in original recall precision table.

Total number of relevant documents for query 2 is 4.

Recall – Precision Table :

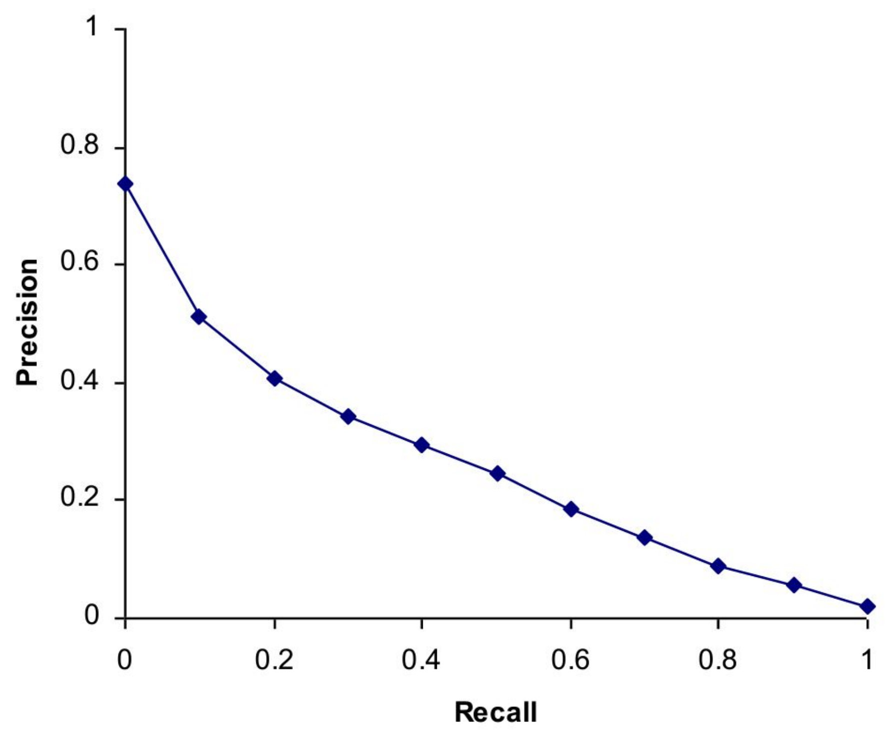
Rank	1	2	3	4	5	6	7	8	9	10
Relevance	1	0	1	0	1	0	0	0	1	0
Precision	1	1/2	2/3	2/4	3/5	3/6	3/7	3/8	4/9	4/10
Recall	1/4	1/4	2/4	2/4	3/4	3/4	3/4	3/4	4/4	4/4

Interpolated Recall – Precision Table :

Precision	1	1	1	2/3	2/3	2/3	3/5	3/5	4/9	4/9	4/10
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Typical (good) 11 point precisions

- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



Macro and Micro Averaging

- Micro – average over each point
  - Calculated over all decisions and then averaged
  - E.g. micro-averaged precision
  - Tends to overemphasize performance on largest categories
- Macro – average of averages per query
  - Statistics calculated for each query and then averaged
  - E.g. macro-averaged precision
  - Over-emphasizes performance on the smallest

References

1. Slides provided by Sougata Saha (Instructor, Fall 2022 - CSE 4/535)
2. Materials provided by Dr. Rohini K Srihari
3. https://nlp.stanford.edu/IR-book/information-retrieval-book.html