



Evaluation of IR systems



statistical language model

$D = \left\{ \begin{array}{l} \text{One fish, two fish, red fish, blue fish.} \\ \text{Black fish, blue fish, old fish, new fish.} \end{array} \right.$

$$\text{len}(D) = 16$$

$$P(\text{fish}|D) = 8/16 = 0.5$$

$$P(\text{blue}|D) = 2/16 = 0.125$$

$$P(\text{one}|D) = 1/16 = 0.0625$$

...

$$P(\text{eggs}|D) = 0/16 = 0$$

...


} A “topic”



statistical language model

- Document came from a topic
- Did query come from *this* document's topic?

- For each document, find probability its topic could have generated the query

$$\begin{aligned} P(Q|T_D) &\approx P(Q|D) \\ &= P(q_1, \dots, q_t|D) \\ &= \prod_{i=1}^t P(q_i|D) \end{aligned}$$


Independence assumption
(Naïve Bayes)

statistical language model



$D_1 = \left\{ \begin{array}{l} \text{This one, I think, is called a Yink.} \\ \text{He likes to wink, he likes to drink.} \end{array} \right.$

$D_2 = \left\{ \begin{array}{l} \text{He likes to drink, and drink, and drink.} \\ \text{The thing he likes to drink is ink.} \end{array} \right.$

$D_3 = \left\{ \begin{array}{l} \text{The ink he likes to drink is pink.} \\ \text{He links to wink and drink pink ink.} \end{array} \right.$

Query “drink”

• $P(\text{drink}|D_1) = 1/16$

• $P(\text{drink}|D_2) = 4/16$

• $P(\text{drink}|D_3) = 2/16$

Query “pink ink”

• $P(Q|D_1) = 0 \cdot 0 = 0$

• $P(Q|D_2) = 0 \cdot 1/16 = 0$

• $P(Q|D_3) = 2/16 \cdot 2/16 = 0.016$

Query “wink drink”

• $P(Q|D_1) = 0.004$

• $P(Q|D_2) = 0$

• $P(Q|D_3) = 1/16 \cdot 2/16 = 0.008$



does it work ?

- Highly artificial examples suggested model is “OK”
- Our intuition says (?) model is OK
- Some thought should point up obvious problems
 - Thoughts?
- Is it really any good?
 - How can we find out?
 - How can we know if changes make it better?



evaluation of IR systems

- many things to evaluate
- test collections
- relevance
- system effectiveness
- significance tests
- TREC conference
- comments

IR evaluation overview

docs
85K

queries
25

IR system

Math Eval

- ranking
- sets
- metrics
- conf matrix
- AP

trec_eval(HWS)

- store data → hash before computation

hash[?][?] = ?
key1 key2 val

HWS

qrel file
= "ground truth"

docid	qid	1/0
Docid	qid	1/0

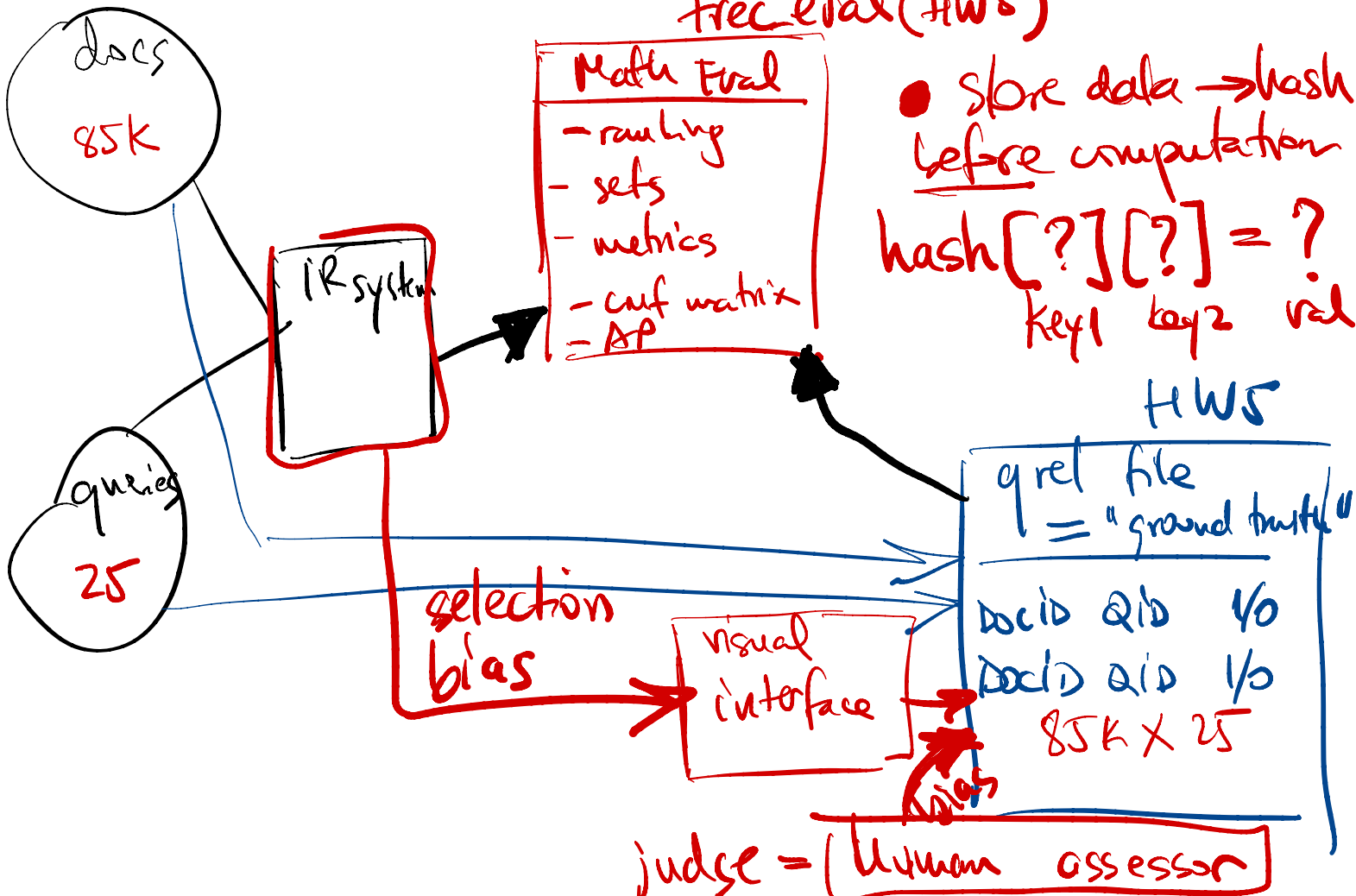
85K x 25

selection bias

visual interface

judges

judge = (human assessor)



5) IR Eval min requirements

- prec/recall, R-prec, AP, NDCG, RR

- Evaluation Pipeline (above)

- Need for scale. → wly 25q and not 1q? or not 1000q?

- Need for Stat. testing (true for any evaluation)

- ^{need} advance metrics (~~stat. testing~~): diversity, time-complex

- need for user studies (any evaluation)

- HWS
aka AP formula ← math-metrics are with user perception of quality.

- query type: { navigational (prec), informational (rec)
open(?), close (know expect), QA, product



evaluations


- IR system often component of larger system
- Might evaluate several aspects
 - Assistance in formulating queries
 - Speed of retrieval
 - Resources required
 - Presentation of documents
 - Ability to find relevant documents
 - Appealing to users (market evaluation)
- Evaluation generally comparative
 - System A vs. B
- Cost-benefit analysis possible
- Most common evaluation: retrieval effectiveness



test collections

- Compare retrieval performance using a test collection
 - set of documents
 - set of queries
 - set of relevance judgments (which docs relevant to each query)
- To compare the performance of two techniques:
 - each technique used to evaluate test queries
 - results (set or ranked list) compared using some performance measure
 - most common measures - precision and recall
- Usually use multiple measures to get different views of performance
- Usually test with multiple collections - performance is collection dependent

test collections



Collection Characteristics	Cranfield	CACM	ISI	West	TREC2
Collection size (docs)	1,400	3,204	1,460	11,953	742,611
Collection size (Mb)	1.5	2.3	2.2	254	2,162
Year created	1968	1983	1983	1990	1991
Unique stems	8,226	5,493	5,448	196,707	1,040,415
Stem occurrences	123,200	117,578	98,304	21,798,833	243,800,000
Max within document frequency		27	27	1,309	
Mean document length (words)	88	36.7	67.3	1,823	328
Number of queries	225	50	35	44	100

- TREC includes five disks, so has numerous subsets
- The TDT corpora are also well-known (though small)
 - In English, Arabic, and Chinese
 - Both text, television audio, and radio audio

About 60K stories



relevance

- difficult to define
- relevant doc = judged “useful” in the context of a query
 - who judges ?
 - humans not very consistent
 - judgments depend on more than doc and query
- with real collections, never know full set of relevant documents
- retrieval model incorporates some notion of relevance
- individuals may disagree occasionally but they agree on average

Web

[12. CIKM 2003: New Orleans, Louisiana, USA](#)

12. **CIKM 2003**: New Orleans, Louisiana, USA. Proceedings of the **2003 ACM CIKM** International Conference on Information and Knowledge Management, New Orleans, ...
www.informatik.uni-trier.de/~ley/db/conf/cikm/cikm2003.html - 56k - [Cached](#) - [Similar pages](#)

[CIKM](#)

Proceedings of the **2003 ACM CIKM** International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, **2003**. ...
www.informatik.uni-trier.de/~ley/db/conf/cikm/ - 10k - Jun 25, 2005 - [Cached](#) - [Similar pages](#)

[CIKM'2003 review](#)

CIKM'2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans ...
smi.ucd.ie/~rinat/papers/cikm03_rep.html - 22k - [Cached](#) - [Similar pages](#)

[Collaborative Filtering Mailing List Archive: \[collab@sims\] CFP](#)

ACM **CIKM 2003** Call For Papers. 12th International Conference on Information and Knowledge ... caliber papers submitted to **CIKM 2003** will be accepted. ...
www.pdesigner.net/1996/0697.html - 17k - [Cached](#) - [Similar pages](#)

[TOC](#)

Proceedings of the twelfth international conference on Information and knowledge management citation. **2003**, New Orleans, LA, USA November 03 - 08, **2003** ...
portal.acm.org/toc.cfm?id=956863&type=proceeding - [Similar pages](#)

[\[Asis-\] CIKM 2003](#)

[Asis-] **CIKM 2003**. Padmini Srinivasan padmini@lakshmi.info-science.uiowa.edu Mon, 29 Sep 2003 12:59:36 -0500. Previous message: [Asis-] Re: ...
mail.asis.org/pipermail/asis-l/2003-September/001024.html - 17k - [Cached](#) - [Similar pages](#)

[\[PDF\] CIKM 2003](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)
CIKM 2003. Jacob Kogan. Charles Nicholas. Marc Teboulle. -means and beyond - p.1/53. Page 2. Outline of the talk. how to build a partition ...
www.csee.umbc.edu/~nicholas/clustering/jacob.pdf - [Similar pages](#)

[Tutorial on Document Clustering](#)

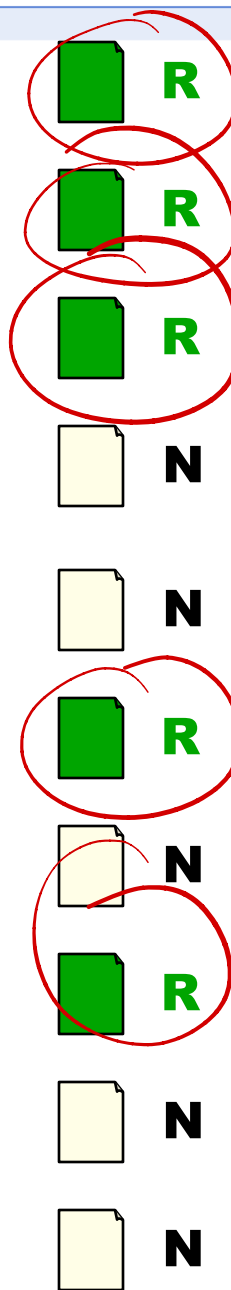
CIKM 2003 Tutorial. Clustering Large and High-Dimensional Data ... Katya Pelekhov and Daniela Rus, "Using Star Clusters for Filtering", **CIKM 2000**, (pdf) ...
www.csee.umbc.edu/~nicholas/clustering/ - 9k - [Cached](#) - [Similar pages](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)

CIKM has a strong tradition of workshops devoted to emerging areas of database ... The **CIKM 2004** web page; The **CIKM 2003** Web Page; The **CIKM 2002** Web Page ...
www.cikm.org/ - 7k - [Cached](#) - [Similar pages](#)

[CIKM 2003, New Orleans, USA, November 2003](#)

Home. **CIKM 2003**, New Orleans, USA, November 2003. << Bild 6 | Bild 7/80 | Bild 8 >>. Miniaturansicht.
www.torsten-priebe.de/showpics.php?folder=2003-11a_cikm03&picture=7 - 2k - [Cached](#) - [Similar pages](#)



find/judge relevant docs

- did the system find all relevant docs ?
 - need complete judgments
 - i.e. a “R” or “N” for all query-doc pairs

- for large collections that is not practical
 - millions of documents x tens of queries

*docs not selected
for assessment?*

- partial set of judgments

- pooling

- judge top n documents from each system
- use judgments across systems (union)

- sampling

- possibly estimate size of relevant set
- design sampling technique from measure

- search based

- use manually guided search
- until convinced all relevance found

selection bias

*⇒ implies
not relevant*

Consider the query "civil war battles in South Carolina," meaning "Which civil war battles were fought in South Carolina?"

Please assign a grade to this document: [\[link\]](#)

doc

- 3 Key - This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.
- 2 Highly Relevant - The content of this page provides substantial information on the topic.
- 1 Relevant - The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page.
- 0 Non-relevant - The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query.

non-binary ⇒ change in metrics.

creativity = query

Questions: Find ways of measuring creativity.

Relevant items include definitions of creativity, descriptions of characteristics associated with creativity, and factors linked to creativity.

↓ This One

★ They're Equally Good ★

* They're Equally Bad *

This One ↓

AGENCY: National Institute of Standards and Technology Commerce.
SUMMARY: The inventions listed below are owned by the U.S. Government, as represented by the Department of Commerce, and are available for licensing in accordance with 35 U.S.C. 207 and 37 CFR Part 404 to achieve expeditious commercialization of results of federally funded research and development. FOR FURTHER INFORMATION CONTACT: Technical and licensing information on these inventions may be obtained by writing to: Marcia Salkeld, National Institute of Standards and Technology, Office of Technology Commercialization, Physics Building, Room B&hyph;256, Gaithersburg, MD 20899; Fax 301&hyph;869&hyph;2751. Any request for information should include the NIST Docket No. and Title for the relevant invention as indicated below.
SUPPLEMENTARY INFORMATION: The inventions available for licensing area: NIST Docket No. 90&hyph;030D Title: Monomers For Double Ring-Opening Polymerization With Expansion Description: NIST researchers have created a new class of monomers that undergo double ring-opening polymerization with an expansion in volume. When used in resinous compositions the result is a volume neutral curing process at ambient temperature, and a final product that exhibits high adhesive strength. NIST Docket No. 90&hyph;036 Title: Epitaxial Iron Films Exhibiting Large Polar Kerr Rotation Description: The invention is a magneto-optic iron film that greatly enhances Kerr rotation compared with conventional iron films. The material could be utilized in magneto-optic data storage media. NIST Docket No. 93&hyph;028C

her father's death when she was 8; by attempted suicide and four months of treatment in a mental hospital while she was in college; by her husband's faithlessness, both imagined and real, which caused her to end their marriage; and finally by her suicide on Feb. 11, 1963.

In trying to account for an ending one can only perceive as tragic, Stevenson has chosen to read Plath's life in conventional psychological terms, although her "diagnosis" remains murky. Repeated references to "divided being" recall R. D. Laing's work on schizophrenia, but just as often she mentions "depressed" and "manic" extremes, suggesting bipolar depression, a diagnosis apparently supported by the doctor who was treating Plath when she ended her life. Stevenson makes this bipolarity her controlling metaphor. As early as high school, "Sylvia had a rare, infectious capacity for exultation -- as great a gift for rapture as she had for misery." Indeed, in Stevenson's view, there were two Sylvia Plaths: "the outer Sylvia, characterized by Robert Lowell as 'a brilliant tense presence, embarrassed by restraint,' and the inner woman, fraught with fears and aggressions." And although at one point Stevenson claims that "the writer was beginning to identify with the woman, the woman with the writer; there could be no true distinction between them," her later evaluation is more sinister: that Plath had projected "the 'desired image' (the required image) of herself as Eve -- wife, mother, homemaker, protector of the wholesome, the good, and the holy, an identity that both her upbringing and her own instinctive physical being had fiercely aspired to. Now her submerged and subversive self, utterly true to itself, utterly detached, completely the artist, turned on the Eve scenario and



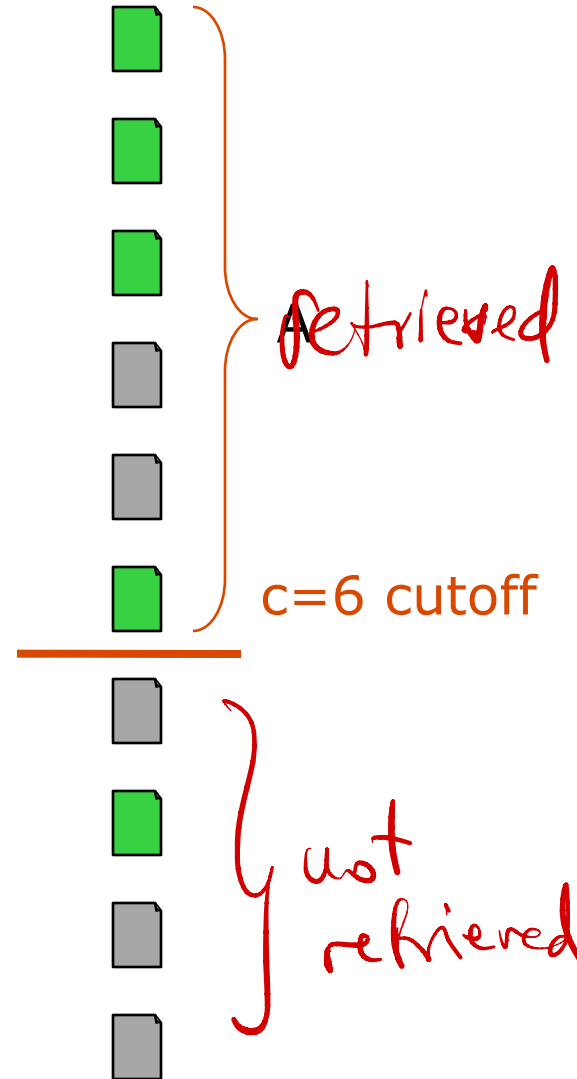
evaluation of IR systems

- many things to evaluate
- test collections
- relevance
- system effectiveness
- significance tests
- TREC conference
- comments



ranked lists

- with respect to a given query
- R = number of relevant documents in the entire corpus (collection)
- treat A as a set
- how many relevant documents ?
- at what rate ?





precision and recall

- Precision

- Proportion of a retrieved set that is relevant
- Precision = $\frac{|\text{relevant} \cap \text{retrieved}|}{|\text{retrieved}|}$
= $P(\text{relevant} | \text{retrieved})$

- Recall

- proportion of all relevant documents in the collection included in the retrieved set
- Recall = $\frac{|\text{relevant} \cap \text{retrieved}|}{|\text{relevant}|}$
= $P(\text{retrieved} | \text{relevant})$

- Precision and recall are well-defined for sets

- For ranked retrieval

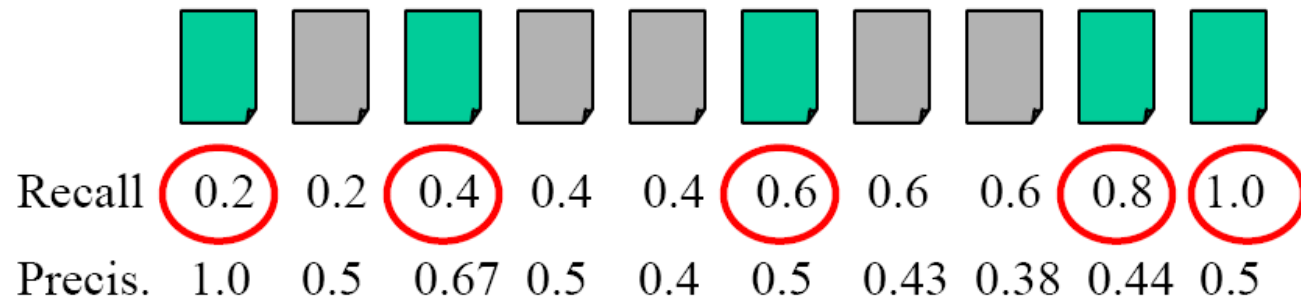
- Compute a P/R point for each relevant document
- Compute value at fixed recall points (e.g., precision at 20% recall)
- Compute value at fixed rank cutoffs (e.g., precision at rank 20)

list precision and recall

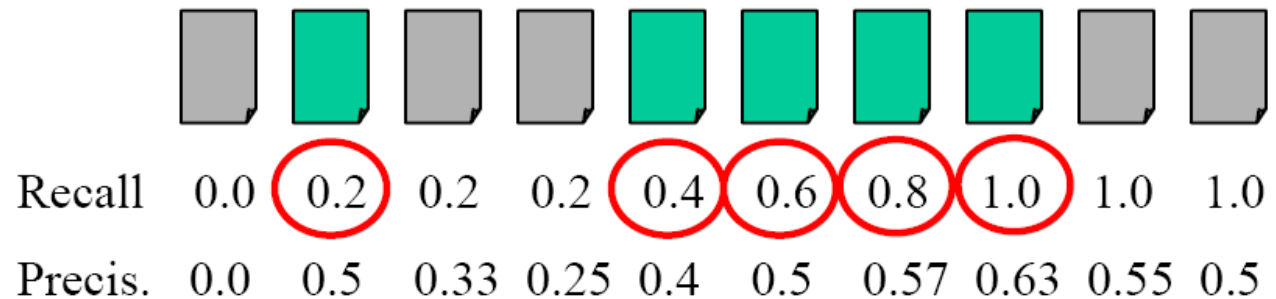


= the relevant documents

Ranking #1



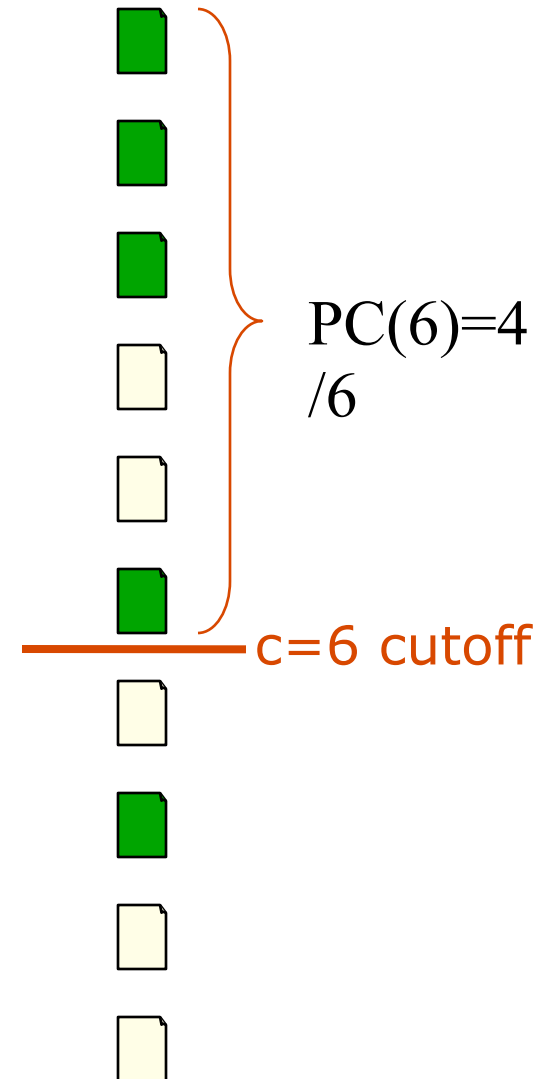
Ranking #2



precision at cutoff (PC)



- high cutoff: “I am feeling lucky”
- P10 motivated by web search
- low cutoff: comprehensive search





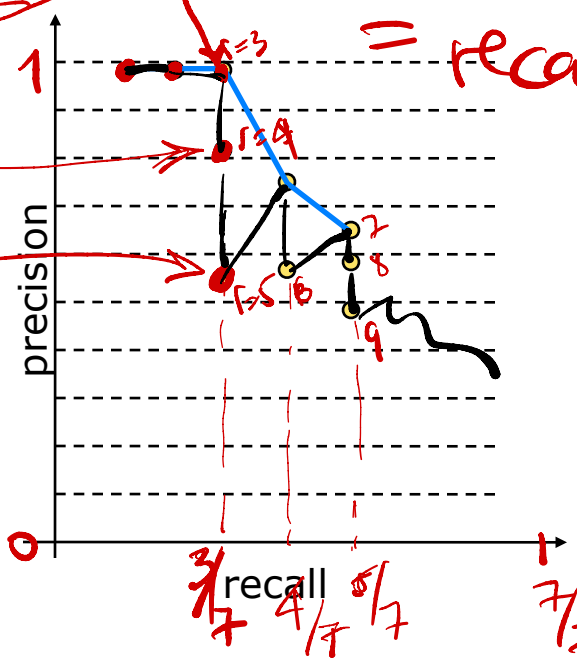
R-precision (RP)

- i.e. precision at cutoff R
- breakeven point
 - at cutoff R $\text{prec} = \text{recall}$
- empirically shown to be effective
- related with average precision

precision-recall curves

total number of relevant = R
 $= 7$

	precision	recall
→ [green]	1/1	1/7
→ [green]	2/2	2/7
→ [green]	3/3	3/7
→ [yellow]	3/4	3/7
→ [yellow]	3/5	3/7
→ [green]	4/6	4/7
→ [yellow]	4/7	4/7
→ [green]	5/8	5/7
→ [yellow]	5/9	5/7
→ [yellow]	5/10	5/7



$R\text{-prec} = \text{prec at rank/cutoff}$
 $= \text{recall at rank } R$

$$AP = \frac{1}{7} \left[\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{6} + \frac{5}{8} + \dots \right]$$

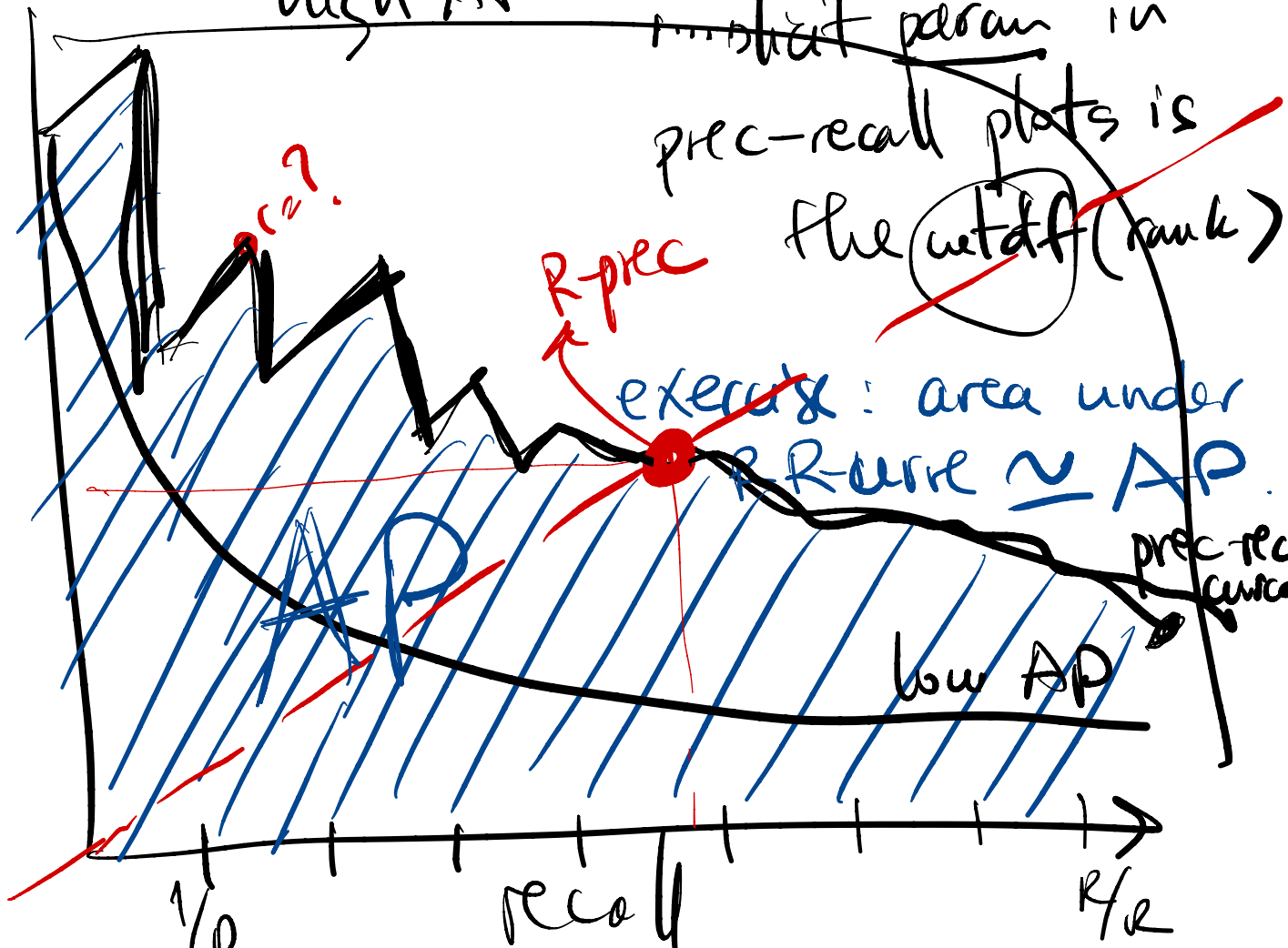


high AP

important param in
prec-recall plots is

the width (rank)

prec



example: area under
R-prec \approx AP

1/2

recall

R/r

average precision (AP)



- one number that reflects the quality of entire list

$$AP = \text{AVG}(\text{prec@r}) = \frac{1}{R} \sum_{r=1}^{\%L} \text{rel}(r) \cdot \text{prec@r}$$

all r: relevant docs

r=1:end

- average precisions at relevant ranks

- divide by R when average

*all relevant
(not only the ones displayed)*

Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precis.	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

AvgPrec= 62.2%

Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

AvgPrec= 52.0%

high AP: relevant docs are retrieved/ordered
 at top of the list
 binary relevance (0,1)



$$AP = \frac{1}{R} [1 + 1 + 1 \dots + 1] = 1$$

$$= 1000000$$

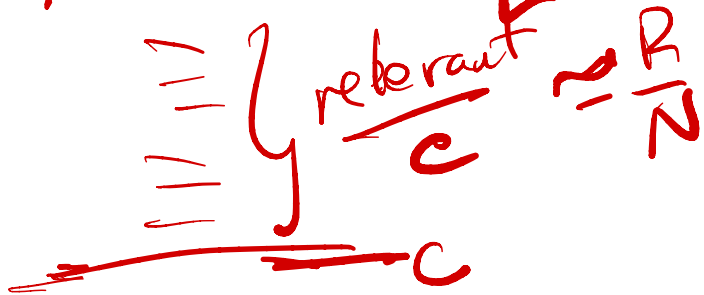
$$= 100$$

random list N docs, R relevant
 unif distributed $\frac{1}{N}$ relevant roughly

every $\frac{N}{R}$ ranks

10,000

$$AP \approx \frac{1}{R} \left[\frac{R}{N} + \frac{R}{N} + \dots + \frac{R}{N} \right] \approx \frac{R}{N}$$





interpolation

- as a trend, precision decreases
- and recall increases
- but it is not always so
- how to handle recall zero
- how to average graphs



interpolated AP

- average precision at standard recall points
- for a given query, compute P/R point for every relevant doc.
- interpolate precision at standard recall levels
 - 11-pt is usually 100%, 90, 80, ..., 10, 0% (yes, 0% recall)
 - 3-pt is usually 75%, 50%, 25%
- average over all queries to get average precision at each recall level
- average interpolated recall levels to get single result
 - called “interpolated average precision”
 - not used much anymore; “mean average precision” more common
 - values at specific interpolated points still commonly used

trec-eval demo

```
14:17>> bin/Buckley/trec_eval trec8/qrels/qrel.trec8 trec8/input/input.READWARE
```

```
Queryid (Num):      50
Total number of documents over all queries
  Retrieved:        3060
  Relevant:           4728
  Rel_ret:          2019
Interpolated Recall - Precision Averages:
  at 0.00           0.9528
  at 0.10           0.8255
  at 0.20           0.7527
  at 0.30           0.6307
  at 0.40           0.4919
  at 0.50           0.2905
  at 0.60           0.2652
  at 0.70           0.1772
  at 0.80           0.1351
  at 0.90           0.0731
  at 1.00           0.0175
Average precision (non-interpolated) for all rel docs (averaged over queries)
  0.4001
Precision:
  At 5 docs:        0.8400
  At 10 docs:       0.7740
  At 15 docs:       0.7427
  At 20 docs:       0.6840
  At 30 docs:       0.6100
  At 100 docs:      0.3474
  At 200 docs:      0.2016
  At 500 docs:      0.0808
  At 1000 docs:     0.0404
R-Precision (precision after R (= num_rel for a query) docs retrieved):
  Exact:            0.4481
```



E measure

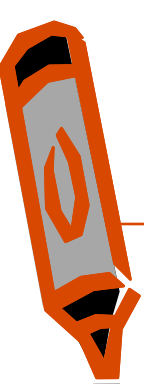
- p =recision, r = recall
- $E = 1 - \frac{1}{\alpha \frac{1}{p} + (1-\alpha) \frac{1}{r}}$
- good results mean small values of E
- E is a set measure
- α = parameter to emphasize p or r
- use $\alpha = \frac{1}{\beta^2 + 1}$, then $E = 1 - \frac{(\beta^2 + 1)pr}{\beta^2 p + r}$
- related to set symmetric difference



F measure

- $F = 1 - E = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}$
- good results mean large values of E
- F also is a set measure
- $F1$ measure is popular : F with $\beta = 1$
- $$F1 = \frac{2pr}{p+r}$$
- $F1$ is in fact the harmonic mean of p and r
- heavily penalizes low values of p or r

expected search length



1 2

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevance	N	Y	N	Y	Y	Y	Y	N	Y	N	N	N	Y	N	Y	N	N	N	N	N

For type 2 query with n=2, search length is 2

For query with n=6, search length is 3


Rank	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4	
Relevance	N	N	Y	Y	N	Y	Y	Y	N	Y	Y	N	N	N	N	N	N	N	N	Y	N
			1	2	3	4	5	X	X	X	X	X									

For type 2 query with n=6, possible search lengths are 3,4,5 or 6 depending on ordering in level 3.

Of the 10 ways in which 2 relevant docs could be distributed in 5, 4 would have search length 3, 3 have search length 4, 2 have search length 5, and 1 has search length 6.

Expected Search Length is $(4/10) \cdot 3 + (3/10) \cdot 4 + (2/10) \cdot 5 + (1/10) \cdot 6 = 4$

b-pref


$$\text{bpref} = \frac{1}{R} \sum_r 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

$$\text{bpref-10} = \frac{1}{R} \sum_r 1 - \frac{|n \text{ ranked higher than } r|}{10 + R}$$

→ effort (# of nonrelevant docs) to go through till we find next relevant one.

<http://www.itl.nist.gov/iad/IADpapers/2004/p102-buckley.pdf>



out of possib
MAX

by rank / effort / across doc

Normalized Discounted Cumulative Gain = benefit / doc

Normalized Discounted Cumulative Gain

- Gain : usefulness of a document, depends on relevance = function of relevance
- Cumulative : add the gain at all ranks (up to a cutoff)

- r=1 ⇒ gain
- r=2 ⇒ gain
- r=3 ⇒ gain

$$CG(c) = \sum_{k=1}^c gain(k)$$

high r = high gain

effort = cost = discount

- Discounted : weight the ranks with a discounting function
- Normalized : normalize so that the result is between 0 and 1

popular webic

$$MS-NDCG(c) = Z_c \cdot \sum_{k=1}^c d(k) gain(k)$$

length of doc?

= dot product < cost · benefits >

Normalized Discounted Cumulative Gain



• Microsoft version

$$gain(k) = 2^{rel(k)} - 1 = \exp_2(relevant)$$

$rel = \{0, 1, 2, 3, 4\} \Rightarrow gain = \{1, 2, 4, 8, 16\}$

domain/task specific

$$d(k) = \frac{1}{\log(1+k)}$$

$k=15 \Rightarrow d(k) = \frac{1}{4}$

$$NDCG(c) = Z_c \cdot \sum_{k=1}^{rank k} \frac{2^{rel(k)} - 1}{\log(1+k)}$$

specific to format ranked list

- task type (navigational vs informational)

norm [$\sum_{ranks} gain \cdot discount$]

Reciprocal Rank = 1

$\frac{1}{N}$
 $\frac{1}{N}$
 $\frac{1}{R}$
 $\frac{1}{N}$
 $\frac{1}{N}$
 $\frac{1}{N}$

RR = $\frac{1}{3}$

first rank with relevant doc.

single good result.

• for navigational queries:
~ how many docs I have to read until the answer.

general $\frac{1}{k}$

rank in list where I stop. (I am happy)

typical $\frac{1}{2}$ good

$\frac{1}{8} - \frac{1}{20} \rightarrow$ bad

avg over k queries/searches
 $MRR = \frac{1}{K} \sum_{k=1}^K RR(k)$



evaluation of IR systems

- many things to evaluate
- test collections
- relevance
- system effectiveness
- significance tests
- TREC conference
- comments



significance tests

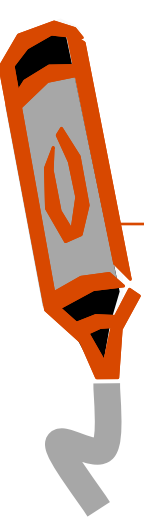
- System A beats System B on one query
 - Is it just a lucky query for System A?
 - Maybe System B does better on some other query
 - Need as many queries as possible
- Empirical research suggests 25 is minimum needed
- TREC tracks generally aim for at least 50 queries
- System A and B identical on all but one query
 - If System A beats System B by enough on that one query, average will make A look better than B
- As above, could just be a lucky break for System A
 - Need A to beat B frequently to believe it is really better
- System A is only 0.00001% better than System B
 - Even if it's true on every query, does it mean much?



significance tests


- Are observed differences statistically different?
- Generally can't make assumptions about underlying distribution
 - Most significance tests do make such assumptions
- Single-valued measures are easier to use, but R/P is possible
- Sign test or Wilcoxon signed-ranks test are typical
 - Do not require that data be normally distributed
 - Sign test answers how often
 - Wilcoxon answers how much
 - Sign test is crudest but most convincing
- Are observed differences detectable by users?

sign test



- For techniques A and B, compare average precision for each pair of results generated by queries in test collection
- If difference is large enough, count as + or -, otherwise ignore
- Use number of +'s and the number of significant differences to determine significance level
- For example, for 40 queries...
 - Technique A produced a better result than B 12 times
 - B was better than A 3 times
 - And 25 were “the same”...
 - $p < 0.035$ and technique A is significantly better than B at the 5% level
 - If $A < B$ 18 times and $B > A$ 9 times...
 - $p < 0.122$ and A is not significantly better than B at the 5% level

Wilcoxon test

- 
- compute diff
 - rank diff by absolute value
 - sum separately +ranks and – ranks
 - two tailed test
 - $T = \min(+ranks, -ranks)$
 - reject null hypothesis if $T < T_0$
where T_0 is found in a table

A	B	DIFF	RANK	SIGNEDRANK
97	96	-1	1.5	-1.5
88	86	-2	3	-3
75	79	4	4	4
90	89	-1	1.5	-1.5
85	91	6	6.5	6.5
94	89	-5	5	-5
77	86	9	8	8
89	99	10	9	9
82	94	12	10	10
90	96	6	6.5	6.5

+ranks = 44

-ranks = 11

$T = 11$

$T_0 = 8$ (from table)

conclusion : not significant



TREC conference

- Text Retrieval Conference
- Established in 1992 to evaluate large-scale IR
 - Retrieving documents from a gigabyte collection
- Run by NIST's Information Access Division
 - Initially sponsored by DARPA as part of Tipster program
 - Now supported by many, including DARPA, ARDA, and NIST
- Probably most well known IR evaluation setting
 - Started with 25 participating organizations in 1992 evaluation
 - In 2003, there were 93 groups from 22 different countries
- Proceedings available on-line (<http://trec.nist.gov>)
 - Overview of TREC 2003 at <http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf>



TREC conference

- TREC consists of IR research tracks
 - Ad-hoc retrieval, routing, cross-language, scanned documents, speech recognition, query, video, filtering, Spanish, question answering, novelty, Chinese, high precision, interactive, Web, database merging, NLP, ...
- Each track works on roughly the same model
 - November: track approved by TREC community
 - Winter: track's members finalize format for track
 - Spring: researchers train system based on specification
 - Summer: researchers carry out formal evaluation
 - Usually a “blind” evaluation: researchers do not know answer
 - Fall: NIST carries out evaluation
 - November: Group meeting (TREC) to find out:
 - How well your site did
 - How others tackled the problem
 - Many tracks are run by volunteers outside of NIST (e.g., Web)
- “Coopetition” model of evaluation
 - Successful approaches generally adopted in next cycle