Evaluation of IR systems

some slides courtesy James Allan@umass

statistical language model

D = { One fish, two fish, red fish, blue fish.
Black fish, blue fish, old fish, new fish. }

len(D) = 16

P(fish|D) = 8/16 = 0.5
P(blue|D) = 2/16 = 0.125
P(one|D) = 1/16 = 0.0625
…
P(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

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

Independence assumption (Naïve Bayes)

statistical language model

D_1 = { This one, I think, is called a Yink.
He likes to wink, he likes to drink.

D_2 = { He likes to drink, and drink, and drink.
The thing he likes to drink is ink.

D_2 = { The ink he likes to drink is pink.
He links to wink and drink pink ink.

Query "drink"
- $P(drink|D_1) = 1/16$
- $P(drink|D_2) = 4/16$
- $P(drink|D_3) = 2/16$

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

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

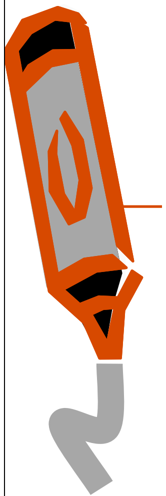
- TREC includes five disks, so has numerous subsets
- The TDT corpora are also well-known (though small)
  - In English, Arabic, and Chinese
  - Both text, television audio, and radio audio

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

evaluation



Web | Images | Groups | News | Froogle | Local | Scholar | more »

cikm 2003 [Search] Advanced Search Preferences

Web

12. **CIKM 2003**: New Orleans, Louisiana, USA — **R**
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 — **R**
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 — **R**
CIKM'2003 highlights. 12th ACM International Conference on Information and Knowledge Management, 3-8 November, New Orleans ...
smi.ucd.ie/~rinat/papers/**cikm**03_rep.html - 22k - Cached - Similar pages

Collaborative Filtering Mailing List Archive: [collab@sims] CFP — **N**
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 — **N**
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

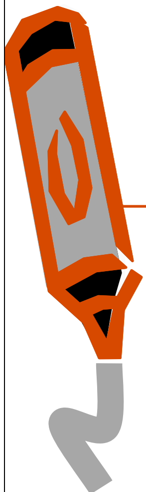
[Asis-l] **CIKM 2003** — **R**
[Asis-l] **CIKM 2003**. Padmini Srinivasan padmini@lakshmi.info-science.uiowa.edu Mon, 29 Sep **2003** 12:59:36 -0500. Previous message: [Asis-l] Re: ...
mail.asis.org/pipermail/ asis-l/**2003**-September/001024.html - 17k - Cached - Similar pages

[PDF] **CIKM 2003** — **N**
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 — **R**
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**) — **N**
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** — **N**
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_**cikm**03&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

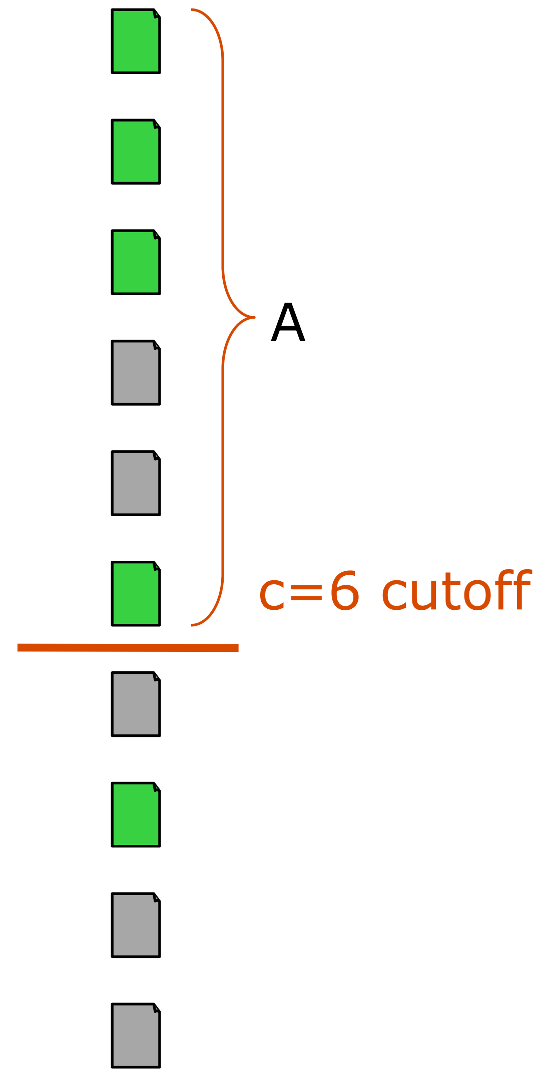
- for large collections that is not practical
  - millions of documents x tens of queries

- 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
  - issues
    - fairness
    - accuracy
    - how to treat unjudged documents ?

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 = $|\text{relevant} \cap \text{retrieved}| \div |\text{retrieved}|$

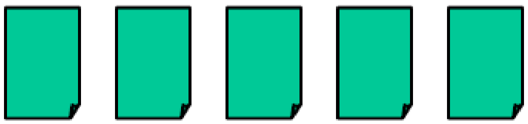
    $= P(\text{ relevant } | \text{ retrieved })$
- Recall
  - proportion of all relevant documents in the collection included in the retrieved set
  - Recall = $|\text{relevant} \cap \text{retrieved}| \div |\text{relevant}|$

    $= P(\text{ retrieved } | \text{ relevant })$

- Precision and recall are well-defined for sets

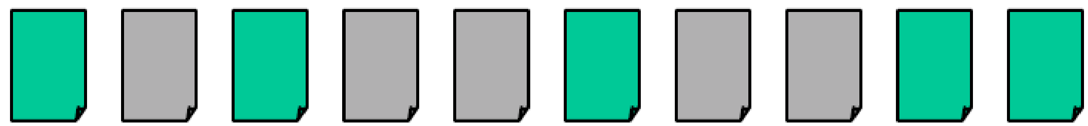
- For ranked retrieval
  - Compute a P/R point for each relevant document
  - Compute value at fixed recall points (e.g., precision at 20% recall)
  - Compute value at fixed rank cutoffs (e.g., precision at rank 20)

list precision and recall



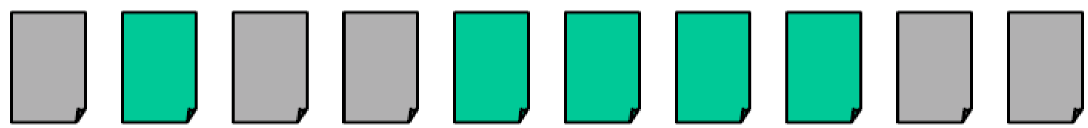
= the relevant documents

Ranking #1



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

Ranking #2



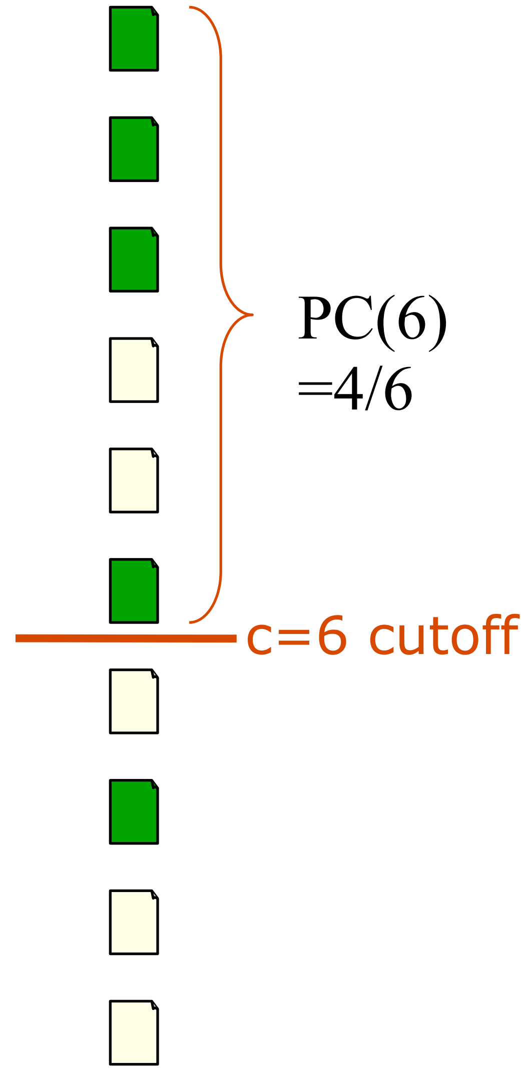
Recall	0.0	0.2	0.2	0.2	0.4	0.6	0.8	1.0	1.0	1.0
Precis.	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.63	0.55	0.5

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

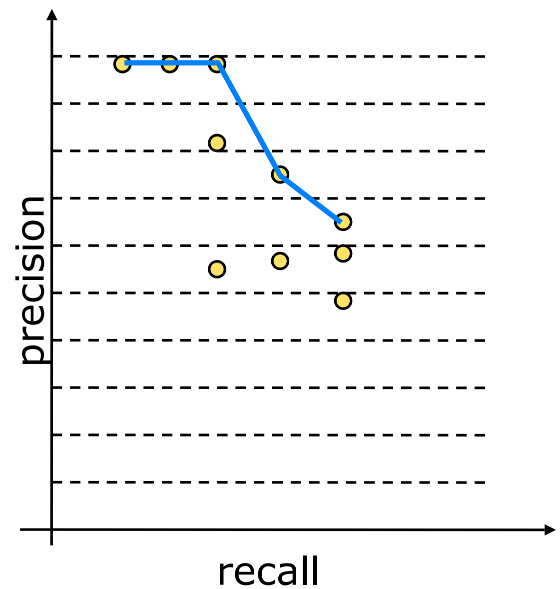
-breakeven point
  -at cutoff R prec = recall

-empirically shown to be effective

-related with average precision

precision-recall curves

precision	recall
1/1	1/7
2/2	2/7
3/3	3/7
3/4	3/7
3/5	3/7
4/6	4/7
4/7	4/7
5/8	5/7
5/9	5/7
5/10	5/7



average precision (AP)

- one number that reflects the quality of entire list
- average precisions at relevant ranks
- divide by R when average

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%

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 enphasize 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	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)·3 +(3/10)·4 + (2/10)·5 + (1/10)·6 = 4

b-pref

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

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

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

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<T0

  where T0 is found in a table

A	B	DIFF	RANK	SIGNED RANK
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