Evaluation in Information Retrieval & Text Categorization

J. Savoy
Université de Neuchâtel

Voorhees, E.M. & Harman, D.K. (2005). *TREC Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge (MA).

Slides from D. Harman, E.M. Voorhees, & J. Callan

Outline

- **Introduction**
- Evaluation campaigns
- Test-collections
- Tasks
- Evaluation measures
- Some warnings
- Conclusion

Why testing?

"In our testing, the success rate for people using external search engines was good. The success rate for people using internal search engines was atrocious. Yet Search on a specific site should actually function better than Web-wide Search."

J. Nielsen, H. Loranger: *Prioritizing Web Usability*. New Riders, 2006

- You can justify your proposition from a theoretical point of view…
- But empirical evaluation is also important

Why TREC-style evaluations?

- Evaluation is essential to learn
- Does your system perform as expected?
  Is this solution better than the baseline?
  Compare your results with previous studies
- But
  All commercial search engines inspect hard queries.
  Don't look stupid
- Why do we need test-collections (or TREC-style evaluations)?
- Testing only with real users…

Users are not sheep!

- Users are not always around
- Users are hard to control
- Users are unpredictable, making it hard to design an experiment that actually measures what you expected
- They are not homogeneous; they come to a task with different levels of knowledge, and they work/learn at different speeds, making user variation a major statistical problem requiring lots of users.
- They are **expensive**!

When are they important?

- For interface design
- To identify critical issues in an information access task
- For operational system testing, such as pinpointing the needs for training
- To verify results from user simulation studies

Outline

- Introduction
- **Evaluation campaigns**
- Test-collections
- Tasks
- Evaluation measures
- Some warnings
- Conclusion

Evaluation Campaigns Philosophy

- TREC is a modern example of the Cranfield tradition
  (C. Cleverdon, head of the librarian at the college of Aeronautics, Cranfield (UK), objective: what makes a good set of indexing terms)
  - system evaluation based on test-collections
  - we have a user model behind! Want to retrieve all relevant items
- Emphasis on advancing the state of the art from evaluation results
  - TREC's primary purpose is not competitive benchmarking
  - experimental workshop: sometimes experiments fail!

Evaluation Campaigns

- TREC (Text REtrieval Conference, trec.nist.gov/)
- CLEF (Cross Language Evaluation Forum, Europe, www.clef-campaign.org/)
- NTCIR (NII Test Collections for IR Systems, Japan, ntcir.nii.ac.jp/)
- INEX (Initiative for the Evaluation of XML Retrieval, Europe, inex.mmci.uni-saarland.de/)
- FIRE (Forum for Information Retrieval Evaluation, India, www.isical.ac.in/~clia/)

TREC conference

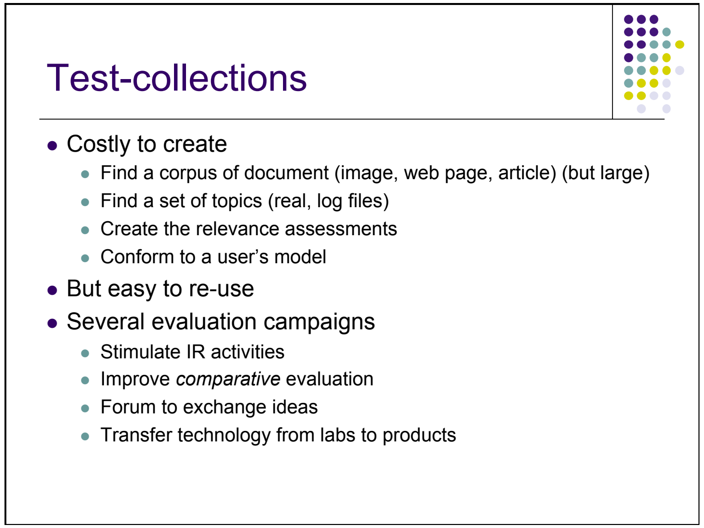
- Text REtrieval Conference
- Established in 1992 to evaluate large-scale IR
  - Retrieving documents from a gigabyte collection
- Has run continuously since then
  - TREC: meeting is in November
- Run by NIST's Information Access Division
  - Started with 25 participating organizations in 1992 evaluation
  - More than 100 groups from around 25 different countries
- Proceedings available on-line (http://trec.nist.gov)
  - Track overviews are good entry points
- NTCIR follows the same pattern (every 18 months)
- CLEF has followed a similar way

TREC general format

- TREC consists of IR research tracks
- 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)
- Building test-collections

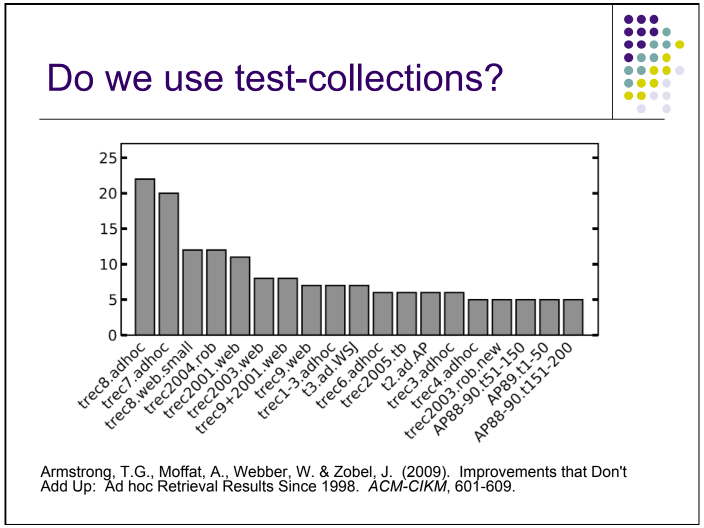
Outline

- Introduction
- Evaluation campaigns
- **Test-collections**
- Tasks
- Evaluation measures
- Some warnings
- Conclusion



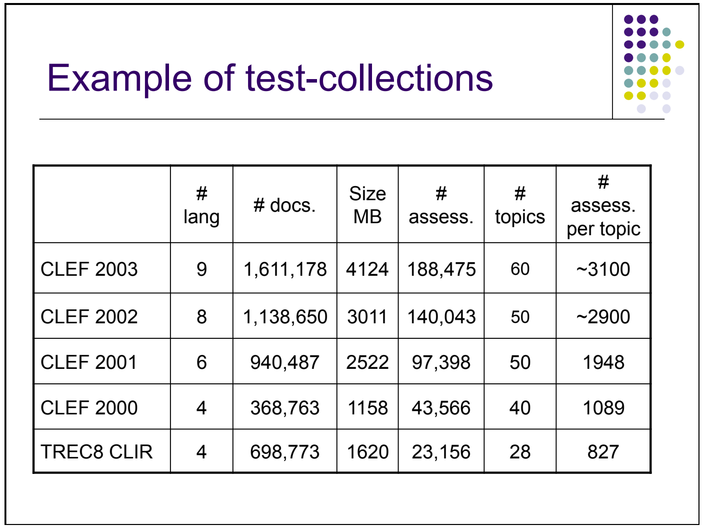
Test-collections

- Costly to create
  - Find a corpus of document (image, web page, article) (but large)
  - Find a set of topics (real, log files)
  - Create the relevance assessments
  - Conform to a user's model
- But easy to re-use
- Several evaluation campaigns
  - Stimulate IR activities
  - Improve *comparative* evaluation
  - Forum to exchange ideas
  - Transfer technology from labs to products



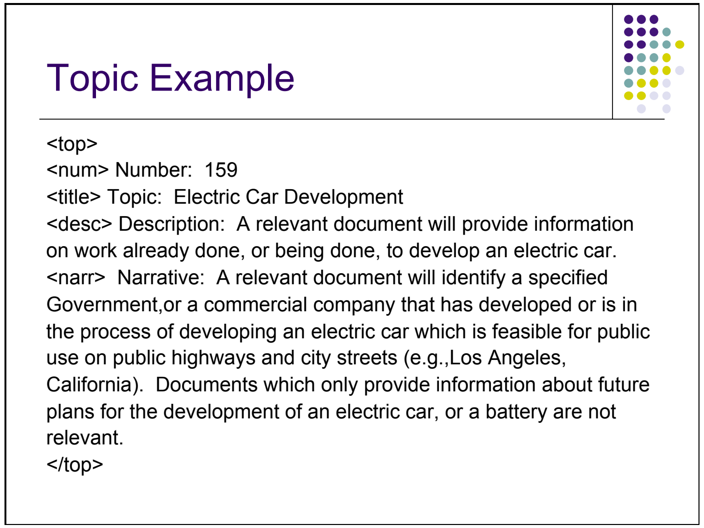
Do we use test-collections?

Armstrong, T.G., Moffat, A., Webber, W. & Zobel, J. (2009). Improvements that Don't Add Up: Ad hoc Retrieval Results Since 1998. *ACM-CIKM*, 601-609.



Example of test-collections

	# lang	# docs.	Size MB	# assess.	# topics	# assess. per topic
CLEF 2003	9	1,611,178	4124	188,475	60	~3100
CLEF 2002	8	1,138,650	3011	140,043	50	~2900
CLEF 2001	6	940,487	2522	97,398	50	1948
CLEF 2000	4	368,763	1158	43,566	40	1089
TREC8 CLIR	4	698,773	1620	23,156	28	827



Topic Example

<top>
<num> Number: 159
<title> Topic: Electric Car Development
<desc> Description: A relevant document will provide information on work already done, or being done, to develop an electric car.
<narr> Narrative: A relevant document will identify a specified Government,or a commercial company that has developed or is in the process of developing an electric car which is feasible for public use on public highways and city streets (e.g.,Los Angeles, California). Documents which only provide information about future plans for the development of an electric car, or a battery are not relevant.
</top>

Document Example

<DOC> <PMID> 7589468 </PMID> …

<TI> Alcohol dehydrogenase of class III: consistent patterns of structural and functional conservation in relation to class I and other proteins. </TI>

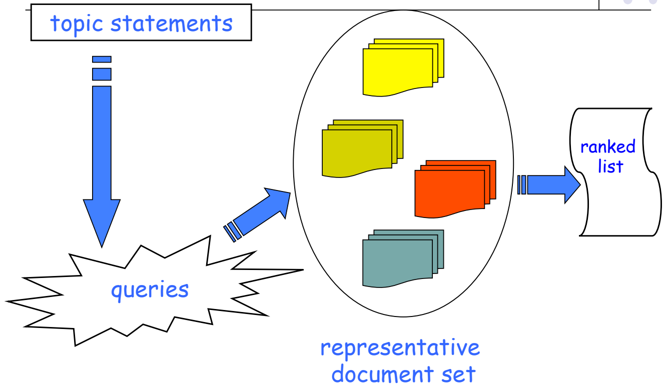
<AB> Class III alcohol dehydrogenase from the lizard Uromastix hardwickii has been characterized. This non-mammalian, gnathostomatous vertebrate class III form allows correlations of structures and functions of this class, the traditional class I alcohol dehydrogenase, and other well-studied proteins. Catalytically, results show similar recoveries and activities of all vertebrate class III forms independent of source, similar activities also in invertebrates but in lower amounts, and considerably higher specific activities in microorganisms. …. </AB> …

<SI> GENBANK/P80467 </SI> <SI> SWISSPROT/P80467 </SI>

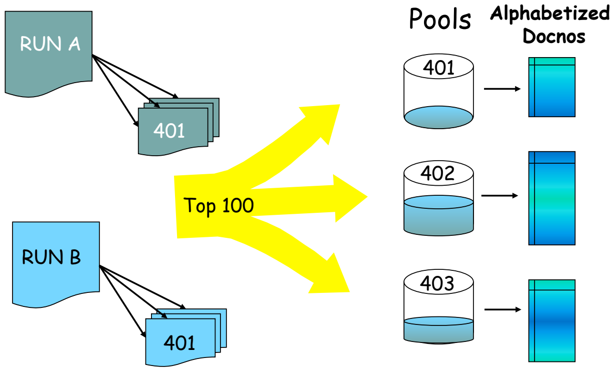
<RN> EC 1.2.1.1 (formaldehyde dehydrogenase (glutathione)) </RN>

<MH> Alcohol Dehydrogenase/chemistry/genetics </MH> <MH> Aldehyde Oxidoreductases/*chemistry/genetics/*metabolism </MH> <MH> Amino Acid Sequence</MH> <MH> Animals </MH> <MH> Conserved Sequence </MH> <MH> Evolution, Molecular </MH> <MH> Liver/enzymology </MH> …</DOC>

Pooling



Pooling



Relevancy

- Do we judge all documents?
  - Documents not judged are declared non relevant
  - Is this a bias against new systems? No (Zobel, SIGIR 98)
- Relevant?
  - A document is relevant if you would use it in a report in some manner
  - At the limit, one sentence in a document could be enough
  - Duplicates are also relevant
- Adding documents in the pool? (pool depth)
  - New relevant document were found (TREC-3)
  - Does not really affect the system ranking
- Relevancy is time and user dependent!

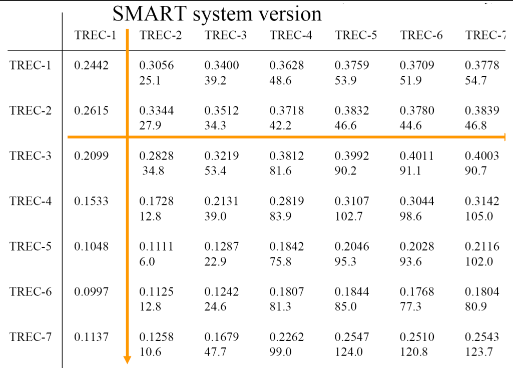
Test-collections

- Consistency
  - Idiosyncratic nature of relevance judgments does not affect comparative results
- Comparative
  - Must used the same test-collection
- Statistical testing
- Incompleteness
  - Relevance judgments must be unbiased
    Pooling is adequate (not perfect)
- Reuse the test-collections!

TREC: Pros and Cons

- "Competition" model of evaluation
  - Successful approaches generally adopted in next cycle
- Widely recognized, premier annual IR evaluation
- What is good
  - Brings together a wide range of active researchers
  - Huge distributed resources applied to common task
  - Substantial gains on tasks rapidly
  - Valuable evaluation corpora usually available after track completes
- What is less good
  - Annual evaluation can divert resources from research
    - Evaluations often require significant engineering effort
  - Recently, an explosion of tracks
    - Means less energy applied to individual tasks
- Do we do some progress?

Improvements over the years



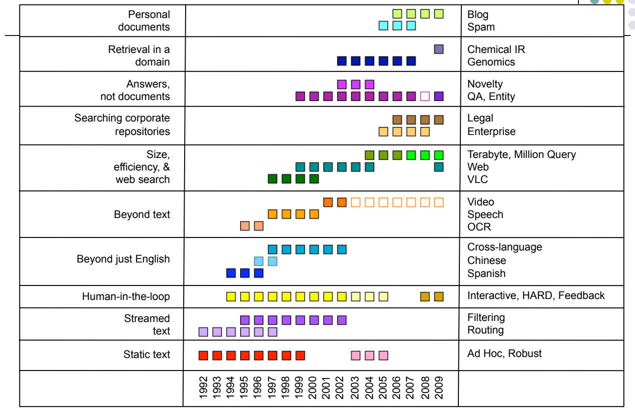
SMART system version

	TREC-1	TREC-2	TREC-3	TREC-4	TREC-5	TREC-6	TREC-7
TREC-1	0.2442	0.3056 25.1	0.3400 39.2	0.3628 48.6	0.3759 53.9	0.3709 51.9	0.3778 54.7
TREC-2	0.2615	0.3344 27.9	0.3512 34.3	0.3718 42.2	0.3832 46.6	0.3780 44.6	0.3839 46.8
TREC-3	0.2099	0.2828 34.8	0.3219 53.4	0.3812 81.6	0.3992 90.2	0.4011 91.1	0.4003 90.7
TREC-4	0.1533	0.1728 12.8	0.2131 39.0	0.2819 83.9	0.3107 102.7	0.3044 98.6	0.3142 105.0
TREC-5	0.1048	0.1111 6.0	0.1287 22.9	0.1842 75.8	0.2046 95.3	0.2028 93.6	0.2116 102.0
TREC-6	0.0997	0.1125 12.8	0.1242 24.6	0.1807 81.3	0.1844 85.0	0.1768 77.3	0.1804 80.9
TREC-7	0.1137	0.1258 10.6	0.1679 47.7	0.2262 99.0	0.2547 124.0	0.2510 120.8	0.2543 123.7

Outline

- Introduction
- Evaluation campaigns
- Test-collections
- **Tasks**
- Evaluation measures
- Some warnings
- Conclusion

The TREC Tracks



Ad Hoc Search

- The *classical* task
- Static collection of documents
  - Written in the same language (monolingual IR)
  - Request in one language, documents in another (BLIR, CLIR)
  - Request in one language, documents in multiple languages (MLIR)
- Noisy Text (OCR)
- INEX ad hoc task (Can the logical structure help the search?)
- Patents (NTCIR, CLEF, TREC)
  - Prior art search, novelty search, invalidity search
  - Text classification (IPC, US classification system)
- Robust
  How can we improve the system effectiveness when facing with difficult queries
- Novelty (sentence-based)

Routing & Filtering

- Retrieve documents from a steam (newswire), user profile(s)
- Adaptive filtering with positive examples (topic + 2-4 relevant items)
  - One document at the time, relevance assessment immediately available
  - No backtracking, no temporal caching
  - Evaluation on set retrieval
- Adaptive filtering, text topic only
- Batch-adaptive filtering (topic + complete relevance on training corpus)
  - One document at the time (+ feedback), evaluation on set retrieval
- Batch-filtering (nonadaptive)
  - One document at the time (without feedback), evaluation on set retrieval
- Routing
  - Ad hoc seach, ranked output

Web Search

- Search in another environment
  - Dynamic Collection
  - Various qualities (various formats, mirror, etc.)
  - Hyperlinks
  - Anchor Texts
  - High Precision
- Different Web test-collections
  - VLC, Very Large Collection
  - WT2g, WT10g
  - .GOV
  - Terabyte
  - In Japanese (NTCIR)

Web Search

- Ad hoc search

  “<num> Number: 511
  <title> diseases caused by smoking?
  <desc> Description: What diseases does smoking cause?
  <narr> Narrative:
  A relevant document must describe smoking tobacco products as a cause of a disease.
  Diseases caused by second-hand smoke and smokeless tobacco are not relevant.
- Named page searching (know-item search)

  “Patuxent Wildlife Research Center”
- Homepage finding task

  “English Server at Carnegie Mellon University”
- Topic distillation

  <num> Number: TD3
  <title> Lewis and Clark expedition</title>
  <desc> Description: What are some useful sites containing information about the historic Lewis and Clark expedition?

Question/Answering

- Answer (information) not documents
- Mainly in TREC, also in NTCIR, CLEF
- Around 3 Gb of newspapers / newswires
- Return the answer (250 bytes, 50 bytes)
- Mainly factoid questions
  - The correct answer could be nil
  - Yes/No question
- Question in one language, answer in another (CLEF)
- What is the exact answer (string)
  - Correct
  - Unsupported
  - Incorrect
- Eval: MRR (single answer)

Question/Answering

- Factoid questions
  - What does the Peugeot company manufacture?
  - Who was President Cleveland's wife?
  - Name the highest mountain
  - Where is Logan International located?
    Where is Logan Airport?
    What city is Logan Airport in?
    Logan International serves what city?
    What city's airport is named Logan International?
    What city is served by Logan International Airport?
- List questions
  - What woman have worn Chanel clothing to award ceremonies?
- Definition questions
  - What in a neutron?

Image Search

- Importance of non-textual media
  - TREC-video
  - Cross-Language Image Retrieval (ImageCLEF)
- Using both text and image matching techniques
  - bilingual ad hoc retrieval task (ES/FR/DE/IT/NL)
  - an interactive search task
  - a medical image retrieval task
  - annotation task
- Important group in CLEF

Image Search

Ad hoc task

Example topics

Pictures of Rome taken in April 1908

Text Classification

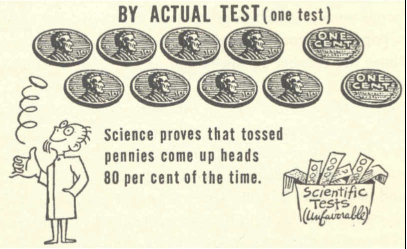
- Assign the right label to a given document
  - Having a set of c_i categories, for each d_j (document) assign the correct single/multiple c_i.
  - Usually using machine learning techniques
- Various forms
  - Extract GeneRIF in Genomics
  - Triage: find article showing experimental evidence requiring GO annotation
  - Automatic assignment of subject-heading
- Reuse collection to do your own experiment
  - Authorship attribution (newspapers)
  - Subject-heading (GIRT, Genomics)

News Tasks / Tracks

- CLEF – IP (patents), difficult task, recall oriented
- CLEF-Image
  - Photo annotation
  - Medical image retrieval
  - Wikipedia image retrieval
- CLEF PAN
- TREC: Session track
- TREC ClueWeb09 (25Tb web pages): 1MQ track
- NTCIR: more NLP-based tasks
- FIRE: ad hoc, MLaF (classification), SMS

Outline

- Introduction
- Evaluation campaigns
- Test-collections
- Tasks
- **Evaluation measures**
- Some warnings
- Conclusion


BY ACTUAL TEST (one test)
ONE CENT
Science proves that tossed pennies come up heads 80 per cent of the time.
Scientific Tests (Unfavorable)

Underlying hypothesis

"It is an error to entertain any proposition with greater assurance than the proofs it is built upon will warrant."
John Locke : *Essay on Human Understanding*, 1690

“First, all documents can be judged as either relevant or not relevant to the query.
Second, each relevant document is equally important and the value of a relevant document does not depend on how many other relevant documents are available, meaning that there is no diminishing marginal return.”

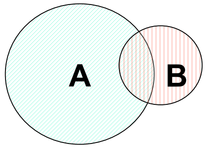
[D. Hull: Using Statistical Testing in the Evaluation of Retrieval Experiments. Proceedings ACM-SIGIR'93, June 1993, 329-338]

Evaluation

- Mainly *binary* judgments in IR practice, but some efforts are conducted to deal with graded relevance
- To compare *relatively* 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*
  - AP for one query, MAP for many queries
  - Precision at a given cutoff value (P@10)
  - MRR
- Statistical testing
- Query-by-query analysis

Precision and recall

- Precision
  - Proportion of a retrieved set (A) that is relevant (A∩B)
  - Precision = |relevant ∩ retrieved| ÷ |retrieved| = |A∩B| ÷ |A|
- Recall
  - Proportion of all relevant documents in the collection included in the retrieved set
  - Recall = |relevant ∩ retrieved| ÷ |relevant| = |A∩B| ÷ |B|
- Precision and recall are well-defined for *sets*
- For *ranked* retrieval?



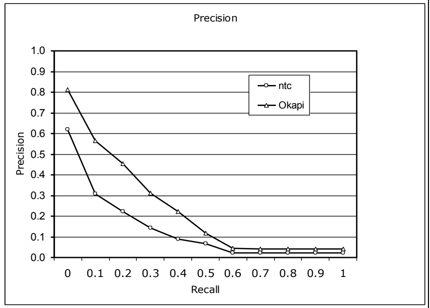
Average Precision

Rank	System A		System B	
1	R	1/1	nR	
2	R	2/2	R	1/2
3	nR		R	2/3
…	nR		nR	
35	nR		R	3/35
…	nR		nR	
108	R	3/108	nR	
	AP =	0.6759	AP =	0.4175
				-38.2%

Precision / Recall graphs

But a picture worth 1000 words!

However the precise values are unknown



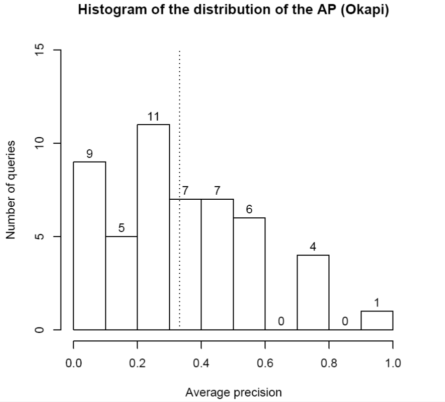
Mean average precision (MAP)

A single value MAP: 0.3321 or an histogram?

Here, for one query, the perfect answer

For 9 queries, Okapi “fails”

(ZH, NTCIR-5, indexing unigram & bigram)



Other measures

- Geometric MAP
  - Used in the robust track (to penalize poor answers)
    Replace AP=0 by a small value (e.g., 0.0001)

$$\text{MAP} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \ AP_i$$

$$\text{GMAP} = \sqrt[n]{\textstyle\prod_{i=1}^{n} \ AP_i} = e^{\frac{1}{n} \cdot \sum_{i=1}^{n} \ log(AP_i)}$$

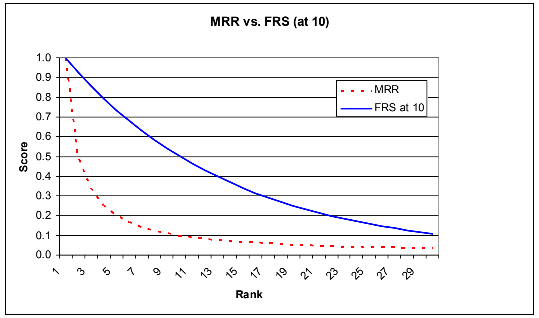
- Need to add an ε for each AP_i
- Many others
  - e.g. utility-based measure by associating cost to each cell in the contingency table

MRR

- Know-item search (only one correct answer)
- MRR (= 1 / rank) penalizes a false answer in the first position
- FRS@10 = $1.08^{(1\text{-rank})}$ or the First Relevance Score at 30 (= $1.24^{(1\text{-rank})}$) (Tomlinson, SIGIR 2006)

Rank	MRR	FRS@10
1	1.000	1.000
2	0.500	0.926
3	0.333	0.857
4	0.250	0.794
5	0.200	0.735
6	0.167	0.681
7	0.143	0.630
8	0.125	0.583
9	0.111	0.540
10	0.100	0.500

MRR and FRS@10



Text categorization

- Contingency table for each category c_i (i = 1, 2, …, |C|)
- Correct answer: TP (true positive) & TN (true negative)
  Incorrect: FP (false positive) & FN (false negative)
- Accuracy = # correct answers / # tests
- Precision $p_i = TP_i / (TP_i + FP_i)$
- Recall $r_i = TP_i / (TP_i + FN_i)$

Category C_i		**Human Expert**	
		Yes	No
Classifier	Yes	TP_i	FP_i
	No	FN_i	TN_i

Text categorization (Binary)

- Precision $p_i = TP_i / (TP_i + FP_i)$
- Recall $r_i = TP_i / (TP_i + FN_i)$
- If your system answers always “yes”?
  FN = 0, then recall = 100%
- If your system says always “no”?
  FP = 0, then precision = 100%
- If you have more than one category?

Category C_i		**Human Expert**	
		Yes	No
Classifier	Yes	TP_i	FP_i
	No	FN_i	TN_i

Text categorization

- With |C| categories?
- Micro-averaging: one document = one decision

$$P = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FP_i}$$

$$R = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i}$$

- Macro-averaging: one category = one decision

$$P = \frac{\sum_{i=1}^{|C|} p_i}{|C|} \qquad R = \frac{\sum_{i=1}^{|C|} r_i}{|C|}$$

- Need a single measure?

Single-valued P/R measures

- F measure
  - Good results mean larger values of F

$$F_\beta = 1 - E = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

- F_1 measure is popular: F with β=1
  - Particularly popular with classification researchers

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

Why 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 0.01% 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 (non-parametric or parametric test)
  - 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) is typical
  - Do not require that data be normally distributed
  - Sign test answers how often (binomial test)
  - Wilcoxon answers how much
  - Sign test is crudest but most convincing
- Are observed differences detectable by users?

Sign test example

- For System A and B, compare AP for each pair of results generated by queries in test collection
- If difference is large enough, count as + or -, otherwise ignore ties
- Use number of +' s and the number of significant differences to determine significance level
- H_o: Similar performance
  System A = System B
  the number of + = number of –
  H_1: Dissimilar performance
  System A ≠ System B

Sign test example

- For example, for 40 queries…
  - System A produced a better result than B 12 times but B was better than A 3 times (n = 15)
  - And 25 were “the same”…
  - *p*-value = 0.03516 and System A *is* significantly better than B at the 5% level

  - If A>B 18 times and B>A 9 times…
  - *p*-value = 0.1221 and A is *not* significantly better than B (at the 5% significance level)

Parametric test

- You can use the *t*-test (Student)
  - Comparing two means
  - Applied the paired t-test
  - Based on the amplitudes of the differences (AP) between two system for a set of *n* queries
  - H_o: Similar performance
MAP A = MAP B
H_1: Dissimilar performance
MAP A ≠ MAP B
  - *t*-test is also available in Excel

Example

Which system is better than the other?

Query	System A	System B
1	0.50	0.99
2	0.40	0.39
3	0.50	0.49
4	0.30	0.29
5	0.40	0.39
6	0.40	0.39
7	0.45	0.44
8	0.35	0.34
9	0.40	0.39
10	0.30	0.29
Mean AP	0.40	0.44

Need to do some progress: Failure Analysis

- Why your system is not perfect?
- What do you learn?
How can you improve your system?
- Analysis query-by-query the result
- Spelling error
«Innondationeurs en Hollande et en Allemagne»
«Flooding in Holland and Germany»
other examples: Irak or Iraq, Oscar or Oskar
- Stopword list
«IT engineer» → it engineer → engineer

Failure Analysis

- Stemmer
  «Bankruptcy of Barings» → «bankruptcy bare»
- «AI in Latin America» → not Artificial Intelligence!
  Need to specify the country name
- The target is not really specified
  «World Soccer Championship» → only the final result
- More complex
  «Chinese currency devaluation»
  → in relevant docs, we have
  (“china”, “currency”) or (“china”) or (“devaluation”)
  “china” in 1,090 docs,
  “currency” in 2,475 docs,
  “devaluation” in 552 docs

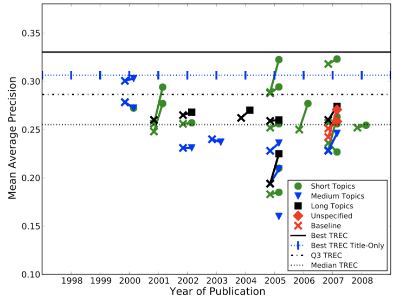
Outline

- Introduction
- Evaluation campaigns
- Test-collections
- Tasks
- Evaluation measures
- **Some warnings**
- Conclusion

Choice of a baseline

- You select a weak baseline (as often do), and try your system to improve on it
  - Obtain statistically significant improvement
  - More easy to publish your work!
- You select a strong baseline
  (e.g., one of best system in the evaluation campaign)
  - Difficult to achieve significant improvement
  - Difficult to publish negative results
  - If so, you have a real impact on the domain
- However, the first solution is clearly more frequent!

TREC-8 Baseline (22 papers)



Armstrong, T.G., Moffat, A., Webber, W. & Zobel, J. (2009). Improvements that Don't Add Up: Ad hoc Retrieval Results Since 1998. *ACM-CIKM*, 601-609.

Does IR reach a plateau?

- Half of the papers has a baseline lower than the TREC median (achieved in 1999?)
- No trend toward better MAP
- Using the P@10, we obtain similar conclusions
- New solutions are more efficient, distributed, use less memory, can work with noisy data, …
- The improvement of a technique depends on the combination of other options
- Test a new technique using a range of possible configurations (showing that it generally improves)

Text Classification

- In Machine Learning, we can detect a clear trend towards more complex classifiers.
- Published results tend to show an improvement
  - Do we need to ignore simple approach?
- New solutions show only small improvements (law of diminishing returns) (simple classifiers tend to represent 85% to 99% of the performance)
- Complex solution tend to overfit the data
- Evolution of the underlying distributions
- Errors in category labels, data

Hand, D.J.. (2006). Classifier Technology and the Illusion of Progress. *Statistical Science, 21(1), 1-14.*

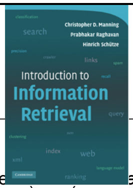
Outline

- Introduction
- Evaluation campaigns
- Test-collections
- Tasks
- Evaluation measures
- Some warnings
- **Conclusion**

User perceptions

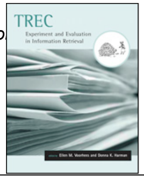
- The final user sees only his/her query (not the mean)
- «The *unhappy* customer, on average, will tell 27 other people about their experience. With the use of the internet, whether web pages or e-mail, that number can increase to the thousands …»
- «*Dissatisfied* customers tell an average of 10 other people about their bad experience. 12% tell up to 20 people.»
- On the other hand, *satisfied* customers will tell an average of 5 people about their positive experience.
- There is a clear practical interest of having a search system performing relatively good for all submitted queries.

References



The books

- R. Baeza-Yates, & B. Ribiero-Ne ern *Information Retrieval*. Addison-Wesley, Reading (MA).
- M. Boughanem & J. Savoy, Eds. (2009). *Recherche d'information : Etat des lieux et perspectives*. Lavoisier, Paris.
- S. Büttcher, C.L.A. Clarke, & G.V. Cormack (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, Cambridge (MA).
- B. Croft, D. Metzler, & T. Strohman (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley, Reading (MA).
- C.D. Manning, P. Raghavan & H. Schütze (2008). *Introductio Retrieval*. Cambridge University Press, Cambridge (UK).
- E.M. Voorhees, & D.K. Harman (2005). *TREC Experiment and Evaluation in Information Retrieval*. The MIT Press, Cambridge (MA).



References

The scientific journals

- IR Information Retrieval, Springer
- IPM Information Processing & Management, Elsevier
- JASIST Journal of the American Society for Information Science & Technology, ASIS
- ACM digital library (portal.acm.org/)
- Several special issues on different tracks

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY

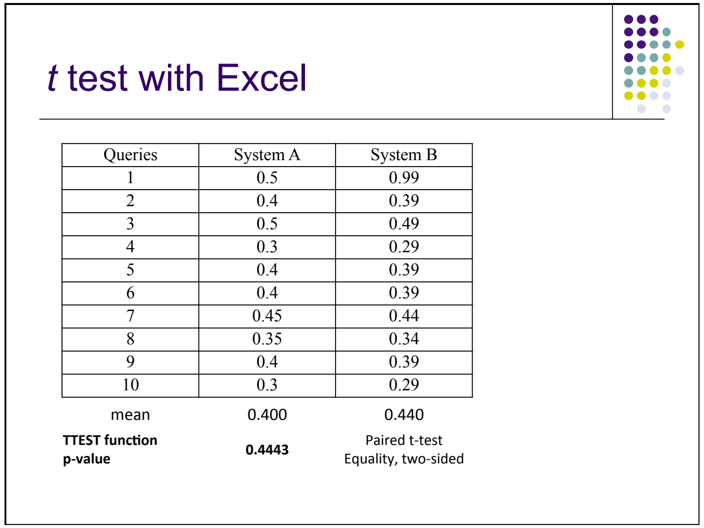
References

Scientific conferences

- ACM-SIGIR
- ACM-CIKM
- ECIR
- Asian AIRS
- CORIA (in Neuchatel in 2013!)

Sign test example

```
> binom.test(3, 15, p=0.5)
                         Exact binomial test
data:  3 and 15
number of successes = 3, number of trials = 15, p-value = 0.03516
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.04331201 0.48089113
sample estimates:   probability of success   0.2
> binom.test(3, 15, p=0.5, alternative="less”)
                         Exact binomial test
data:  3 and 15
number of successes = 3, number of trials = 15, p-value = 0.01758
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4397844
sample estimates:  probability of success   0.2
```



t test with Excel

Queries	System A	System B
1	0.5	0.99
2	0.4	0.39
3	0.5	0.49
4	0.3	0.29
5	0.4	0.39
6	0.4	0.39
7	0.45	0.44
8	0.35	0.34
9	0.4	0.39
10	0.3	0.29
mean	0.400	0.440
TTEST function p-value	**0.4443**	Paired t-test Equality, two-sided