UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

Offline Evaluation

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

One problem

$f(q, d)$

Many solutions

- Similarity-based models
- Probabilistic models
- Extended models
- Machine-learned models

Why evaluate?

- Lots of alternative solutions
  - Which one to choose?
  - How to improve upon them?
- Evaluation enables an informed choice
  - Rigor of science
  - Efficiency of practice

Performing experiments

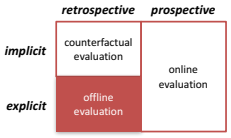
- Key components
  - Experimental setup
  - Analysis of results
- Key concern: ***reproducibility***
  - Must specify each and every detail needed for reproducing our method and the experiment

Experimental setup

- Research questions
- Evaluation methodology
- Evaluation benchmarks
- Reference comparisons
- Parameter tuning

Evaluation methodology

- Feedback
  - Implicit
  - Explicit
- Mode
  - Retrospective
  - Prospective



Test collection-based evaluation

- Three core components
  - A corpus of documents
  - A set of users' queries
  - A map of users' relevance assessments

TREC Topic Example

<top>
<num> Number: 794

<title> pet therapy

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

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

</top>

Evaluation metrics

- General form: $\Delta(R, G)$
  - R: ranking produced by model f for query q
  - G: ground-truth produced for query q
- Metrics should be chosen according to the task

Classification Metrics

A is set of relevant documents,
B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Type I Error

(false positive)

Type II Error

(false negative)

Classification Errors

- *False Positive* (Type I error)
  - a non-relevant document is retrieved

$$Fallout = \frac{|\overline{A} \cap B|}{|\overline{A}|}$$

- *False Negative* (Type II error)
  - a relevant document is not retrieved
  - MissRate = 1 - *Recall*
- *Precision* is used when probability that a positive result is correct is important

F Measure

- *Harmonic mean* of recall and precision

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} = \frac{2RP}{(R+P)}$$

  - harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large
- More general form

$$F_\beta = (\beta^2 + 1)RP/(R + \beta^2 P)$$

  - β is a parameter that determines relative importance of recall and precision

Ranking Effectiveness

= the relevant documents

Ranking #1

Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

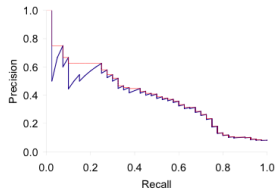
Ranking #2

Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

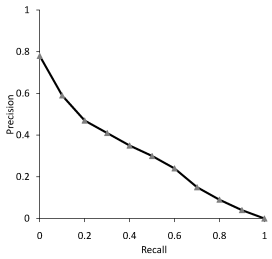
Summarizing a Ranking

- Calculating recall and precision at fixed rank positions
  - E.g., Precision@10, Recall@10
- Calculating precision at standard recall levels, from 0.0 to 1.0
  - E.g., Precision@Recall=30%
  - requires *interpolation*

Precision vs recall graph (interpolation)



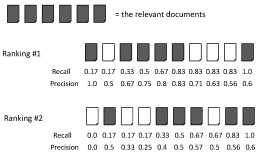
Precision vs recall graph (avg. over 50 queries)



Summarizing a Ranking

- Calculating recall and precision at fixed rank positions
  - E.g., Precision@10, Recall@10
- Calculating precision at standard recall levels, from 0.0 to 1.0
  - E.g., Precision@Recall=30%
  - requires *interpolation*
- Averaging the precision values from the rank positions where a relevant document was retrieved

Average Precision



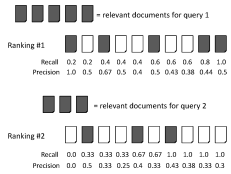
$$\text{Ranking } \#1\text{: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

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

Averaging

- *Mean Average Precision* (MAP)
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers
  - assumes user is interested in finding many relevant documents for each query
  - requires many relevance judgments in text collection

MAP



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

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

Focusing on Top Documents

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

Focusing on Top Documents

- Precision at Rank R
  - R typically 5, 10, 20
  - easy to compute, average, understand
  - not sensitive to rank positions less than R
- Reciprocal Rank
  - reciprocal of the rank at which the first relevant document is retrieved
  - *Mean Reciprocal Rank (MRR)* is the average of the reciprocal ranks over a set of queries
  - very sensitive to rank position

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/*log (rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

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

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0
  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:
  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Normalized DCG

- DCG numbers are averaged across a set of queries at specific rank values
  - e.g., DCG at rank 5 is 6.89 and at rank 10 is 9.61
- DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking*
  - makes averaging easier for queries with different numbers of relevant documents

NDCG Example

- Perfect ranking:
  3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- ideal DCG values:
  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10
- NDCG values (divide actual by ideal):
  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88
  - NDCG $\leq$ 1 at any rank position

Significance Tests

- Given the results from a number of queries, how can we conclude that ranking algorithm A is better than algorithm B?
- A significance test enables us to reject the *null hypothesis* (no difference) in favor of the *alternative hypothesis* (B is better than A)

Significance Tests

1. Compute the effectiveness measure for every query for both rankings.
2. Compute a *test statistic* based on a comparison of the effectiveness measures for each query. The test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide whether or not the null hypothesis should be rejected.
3. The test statistic is used to compute a *P-value*, which is the probability that a test statistic value at least that extreme could be observed if the null hypothesis were true. Small P-values suggest that the null hypothesis may be false.
4. The null hypothesis (no difference) is rejected in favor of the alternate hypothesis (i.e., B is more effective than A) if the P-value is $\leq \alpha$, the *significance level*. Values for α are small, typically .05 and .1, to reduce the chance of a Type I error.

type I error = incorrect rejection of a true null hypothesis (a "false positive")

Example Experimental Results

Query	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25

t-Test

- Parametric assumption is that the difference between the effectiveness values is a sample from a normal distribution
- Null hypothesis is that the mean of the distribution of differences is zero
- Test statistic

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} \cdot \sqrt{N}$$

– for the example,

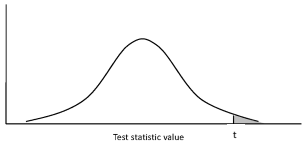
$$\overline{B-A} = 21.4,\ \sigma_{B-A} = 29.1,\ t = 2.33,\ \text{p-value}=.02$$

One-sided vs. two-sided tests

- One-sided: p-value computed from one tail
  - Useful when testing whether $A > B$, if the consequences of the opposite outcome ($A < B$) are negligible
- Two-sided: p-value computed from both tails
  - Useful when testing whether $A \neq B$, if you want to allow for the possibility that your treatment is innocuous of even worse
  - Most typical in ranking evaluation

One-Sided Test ($A > B$)

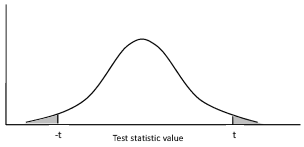
- Distribution for the possible values of a test statistic assuming the null hypothesis



- shaded area is *region of rejection*

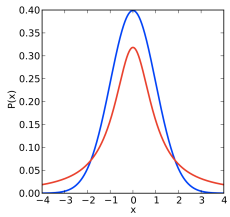
Two-Sided Test ($A \neq B$)

- Distribution for the possible values of a test statistic assuming the null hypothesis

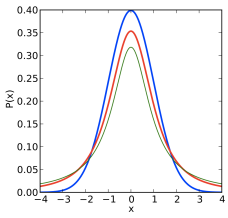


- shaded area is *region of rejection*

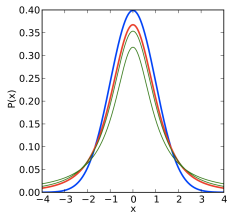
t-distribution ($\nu = 1$)



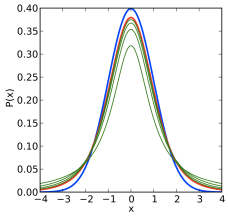
t-distribution ($\nu = 2$)



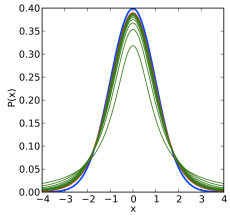
t-distribution ($\nu = 3$)



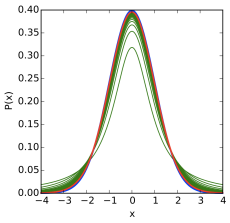
t-distribution ($\nu = 5$)



t-distribution ($\nu = 10$)



t-distribution ($\nu = 30$)



T-table

One-sided	75%	80%	85%	90%	95%	97.5%	99%	
Two-sided	50%	60%	70%	80%	90%	95%	98%	
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	...
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	
...								

T-table

One-sided	75%	80%	85%	90%	95%	97.5%	99%	
Two-sided	50%	60%	70%	80%	90%	95%	98%	
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	...
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	
...								

$n = 5 \rightarrow \nu = 4, t = 2.132$

One-sided: $P(T < t) = 0.95 \rightarrow P(T \geq t) = 0.05$

Two-sided: $P(-t < T < t) = 0.90 \rightarrow P(T \leq -t) + P(T \geq t) = 0.10$

Criticisms

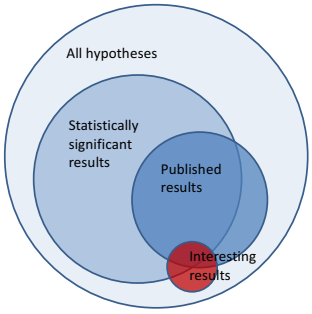
- t statistic revisited

  $t = \frac{\overline{B-A}}{\sigma_{B-A}}.\sqrt{N}$ larger *t* (more extreme), smaller *p*-value

- t can be large for two reasons
  - Large effect size: $\overline{B-A}/\sigma_{B-A}$
  - Large sample size: *N*
- How to easily make your results significant?
  - P-hacking **#DONT**

Criticism

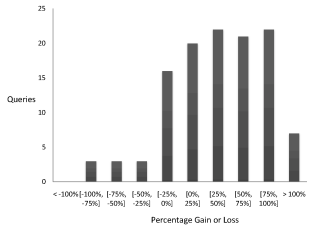
- "A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important" [Ellis, 2010]



Summary

- No single measure is the correct one for any application
  - choose measures appropriate for task
  - use a combination
  - shows different aspects of the system effectiveness
- Use significance tests (t-test)
  - Also report effect sizes!
- Analyze performance of individual queries

Query Summary



References

- Search Engines: Information Retrieval in Practice, Ch. 8
  Croft et al., 2009
- Introduction to Information Retrieval, Ch. 8
  Manning et al., 2008
- Test collection based evaluation of IR systems
  Sanderson, FnTIR 2010

References

- Statistical significance testing in theory and in practice
  Carterette, SIGIR 2017
- The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation
  Sakai, SIGIR 2017
- Statistical reform in information retrieval?
  Sakai, SIGIR Forum 2014

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Information-Theoretic Models

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br