Retrieval Evaluation

Hongning Wang
CS@UVa
What we have learned so far

Indexed corpus

Crawler

Doc Analyzer

Doc Rep (Index)

Indexer

Doc Rep (Index)

Index

Ranker

Ranking procedure

Feedback

Evaluation

User

Indexed corpus

Crawler

Doc Analyzer

Doc Rep (Index)

Indexer

Doc Rep (Index)

Index

Ranker

Ranking procedure

Feedback

Evaluation

User

Indexed corpus

Crawler

Doc Analyzer

Doc Rep (Index)

Indexer

Doc Rep (Index)

Index

Ranker

Ranking procedure

Feedback

Evaluation

User

Indexed corpus

Crawler

Doc Analyzer

Doc Rep (Index)

Indexer

Doc Rep (Index)

Index

Ranker

Ranking procedure

Feedback

Evaluation

User
Which search engine do you prefer: Bing or Google?

• What are your judging criteria?
  – How fast does it response to your query?
  – How many documents can it return?
Which search engine do you prefer: Bing or Google?

• What are your judging criteria?
  – Can it correct my spelling errors?
  – Can it suggest me related queries?
Retrieval evaluation

• Aforementioned evaluation criteria are all good, but not essential
  – Goal of any IR system
    • Satisfying users’ information need
  – Core quality measure criterion
    • “how well a system meets the information needs of its users.” – wiki
    • Unfortunately vague and hard to execute
Bing v.s. Google?

Who was the 34th President of the United States - The Q&A
wiki.answers.com
Mortgages > Home Equity and Refinancing
Who is the 34th President of Mexico? Pedro Lascorz was the 34th president of Mexico.
Was Dwight Eisenhower the 34th president? Ike was the 34th if you count...

Dwight D. Eisenhower - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Dwight_D._Eisenhower
Dwight David "Ike" Eisenhower was the 34th President of the United States from 1953 until 1961. He was a five-star general in the United States Army during World War II.

List of Presidents of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States
List of Presidents of the United States, sorted by previous experience; List of Vice Presidents of the United States; Presidential portrait (United States)

Who was the 34th president of the US? - Yahoo answers.yahoo.com
Politics & Government > Government
Resolved: 29 posts : 28 total answers : Published Aug 18, 2006

Dwight D. Eisenhower | The White House
www.whitehouse.gov/about/presidents/dwighteisenhower
The Presidential biographies on Whitehouse.gov are from "The Presidents of the United States of America," 4th edition, by David C. Donald, and "Presidents of the United States," by Sylvis Horwitt. For further information, see...
Quantify the IR quality measure

• Information need
  – “an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need” – wiki
  – Reflected by user query
  – Categorization of information need
    • Navigational
    • Informational
    • Transactional
Quantify the IR quality measure

• Satisfaction
  – “the opinion of the user about a specific computer application, which they use” – wiki
  – Reflected by
    • Increased result clicks
    • Repeated/increased visits
    • Result relevance
Classical IR evaluation

• Cranfield experiments
  – Pioneer work and foundation in IR evaluation
  – Basic hypothesis
    • Retrieved documents’ relevance is a good proxy of a system’s utility in satisfying users’ information need
  – Procedure
    • 1,398 abstracts of aerodynamics journal articles
    • 225 queries
    • Exhaustive relevance judgments of all (query, document) pairs
    • Compare different indexing system over such collection
Classical IR evaluation

• Three key elements for IR evaluation
  1. A document collection
  2. A test suite of information needs, expressible as queries
  3. A set of relevance judgments, e.g., binary assessment of either relevant or nonrelevant for each query-document pair
Search relevance

• Users’ information needs are translated into queries
• Relevance is judged with respect to the information need, not the query
  – E.g., Information need: “When should I renew my Virginia driver’s license?”
  Query: “Virginia driver’s license renewal”
  Judgment: whether a document contains the right answer, e.g., every 8 years; rather than if it literally contains those four words
Text REtrieval Conference (TREC)

• Large-scale evaluation of text retrieval methodologies
  – Since 1992, hosted by NIST
  – Standard benchmark for IR studies
  – A wide variety of evaluation collections
    • Web track
    • Question answering track
    • Cross-language track
    • Microblog track
    • And more...
## Public benchmarks

### TABLE 4.3 Common Test Corpora

<table>
<thead>
<tr>
<th>Collection</th>
<th>NDocs</th>
<th>NQrys</th>
<th>Size (MB)</th>
<th>Term/Doc</th>
<th>Q-D RelAss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI</td>
<td>82</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIT</td>
<td>2109</td>
<td>14</td>
<td>2</td>
<td>400</td>
<td>&gt;10,000</td>
</tr>
<tr>
<td>CACM</td>
<td>3204</td>
<td>64</td>
<td>2</td>
<td>24.5</td>
<td></td>
</tr>
<tr>
<td>CISI</td>
<td>1460</td>
<td>112</td>
<td></td>
<td>46.5</td>
<td></td>
</tr>
<tr>
<td>Cranfield</td>
<td>1400</td>
<td>225</td>
<td>2</td>
<td>53.1</td>
<td></td>
</tr>
<tr>
<td>LISA</td>
<td>5872</td>
<td>35</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medline</td>
<td>1033</td>
<td>30</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPL</td>
<td>11,429</td>
<td>93</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSHMED</td>
<td>34,8566</td>
<td>106</td>
<td>400</td>
<td>250</td>
<td>16,140</td>
</tr>
<tr>
<td>Reuters</td>
<td>21,578</td>
<td>672</td>
<td>28</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>TREC</td>
<td>740,000</td>
<td>200</td>
<td>2000</td>
<td>89-3543</td>
<td>» 100,000</td>
</tr>
</tbody>
</table>

*Table from Manning Stanford CS276, Lecture 8*
Evaluation metric

• To answer the questions
  – Is Google better than Bing?
  – Which smoothing method is most effective?
  – Is BM25 better than language models?
  – Shall we perform stemming or stopword removal?

• We need a quantifiable metric, by which we can compare different IR systems
  – As unranked retrieval sets
  – As ranked retrieval results
Evaluation of unranked retrieval sets

• In a Boolean retrieval system
  – Precision: fraction of retrieved documents that are relevant, i.e., \( p(\text{relevant} \mid \text{retrieved}) \)
  – Recall: fraction of relevant documents that are retrieved, i.e., \( p(\text{retrieved} \mid \text{relevant}) \)

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>nonrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieved</td>
<td>true positive (TP)</td>
<td>false positive (FP)</td>
</tr>
<tr>
<td>not retrieved</td>
<td>false negative (FN)</td>
<td>true negative (TN)</td>
</tr>
</tbody>
</table>

Precision: \( P = \frac{TP}{TP + FP} \)

Recall: \( R = \frac{TP}{TP + FN} \)
Evaluation of unranked retrieval sets

• Precision and recall trade off against each other
  – Precision decreases as the number of retrieved documents increases (unless in perfect ranking), while recall keeps increasing
  – These two metrics emphasize different perspectives of an IR system
    • Precision: prefers systems retrieving fewer documents, but highly relevant
    • Recall: prefers systems retrieving more documents
Evaluation of unranked retrieval sets

• Summarizing precision and recall to a single value
  – In order to compare different systems
  – F-measure: weighted harmonic mean of precision and recall, $\alpha$ balances the trade-off

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

(F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}})

– Why harmonic mean?
  • System1: P:0.53, R:0.36
  • System2: P:0.01, R:0.99

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.429</td>
<td>0.445</td>
</tr>
<tr>
<td>A</td>
<td>0.019</td>
<td>0.500</td>
</tr>
</tbody>
</table>

Equal weight between precision and recall
Evaluation of ranked retrieval results

- Ranked results are the core feature of an IR system
  - Precision, recall and F-measure are set-based measures, that cannot assess the ranking quality
  - Solution: evaluate precision at every recall point

Which system is better?
Precision-Recall curve

• A sawtooth shape curve

Interpolated precision:
\[ p_{\text{interp}}(r) = \max_{r' \geq r} p(r') \], highest precision found for any recall level \( r' \geq r \).
Evaluation of ranked retrieval results

• Summarize the ranking performance with a single number
  – Binary relevance
    • Eleven-point interpolated average precision
    • Precision@K (P@K)
    • Mean Average Precision (MAP)
    • Mean Reciprocal Rank (MRR)
  – Multiple grades of relevance
    • Normalized Discounted Cumulative Gain (NDCG)
Eleven-point interpolated average precision

- At the 11 recall levels \([0, 0.1, 0.2, \ldots, 1.0]\), compute arithmetic mean of interpolated precision over all the queries.
Precision@K

• Set a ranking position threshold $K$
• Ignores all documents ranked lower than $K$
• Compute precision in these top $K$ retrieved documents
  – E.g.,:
    
    \[
    \begin{align*}
    \text{P@3} & \text{ of } 2/3 \\
    \text{P@4} & \text{ of } 2/4 \\
    \text{P@5} & \text{ of } 3/5
    \end{align*}
    \]

• In a similar fashion we have Recall@K
Mean Average Precision

• Consider rank position of each relevant doc
  – E.g., K_1, K_2, ... K_R
• Compute P@K for each K_1, K_2, ... K_R
• Average precision = average of those P@K
  – E.g.,

\[
\text{AvgPrec} = \left( \frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) / 3
\]

• MAP is mean of Average Precision across multiple queries/rankings
AvgPrec is about one query

AvgPrec of the two rankings

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

Ranking #2: \( \frac{(0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)}{6} = 0.52 \)
MAP is about a system

Query 1, AvgPrec\(=\frac{(1.0+0.67+0.5+0.44+0.5)}{5}=0.62\)
Query 2, AvgPrec\(=\frac{(0.5+0.4+0.43)}{3}=0.44\)
MAP = \(\frac{(0.62+0.44)}{2}=0.53\)
MAP metric

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant document to be zero.
- MAP is macro-averaging: each query counts equally.
- MAP assumes users are interested in finding many relevant documents for each query.
- MAP requires many relevance judgments in text collection.
Mean Reciprocal Rank

• Measure the effectiveness of the ranked results
  – Suppose users are only looking for one relevant document
    • looking for a fact
    • known-item search
    • navigational queries
    • query auto completion

• Search duration ~ Rank of the answer
  – measures a user’s effort
Mean Reciprocal Rank

• Consider the rank position, $K$, of the first relevant document

• Reciprocal Rank = \( \frac{1}{K} \)

• MRR is the mean RR across multiple queries
Beyond binary relevance

Google daily query volume

Same P@6?!

Same MAP?!

Relevant
Nonrelevant

Excellent
Good
Fair
Bad
Bad

CS@UVa
Beyond binary relevance

• The level of documents’ relevance quality with respect to a given query varies
  – Highly relevant documents are more useful than marginally relevant documents
  – The lower the ranked position of a relevant document is, the less useful it is for the user, since it is less likely to be examined
  – *Discounted Cumulative Gain*
Discounted Cumulative Gain

• Uses graded relevance as a measure of usefulness, or gain, from examining a document

• Gain is accumulated starting at the top of the ranking and discounted at lower ranks

• Typical discount is $1/\log (\text{rank})$
  – 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 position $p$:

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

• Alternative formulation

$$ DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(1 + i)} $$

– Standard metric in some web search companies
– Emphasis on retrieving highly relevant documents
Normalized Discounted Cumulative Gain

• Normalization is useful for contrasting queries with varying numbers of relevant results
• Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  – The ideal ranking is achieved via ranking documents with their relevance labels
How about P@4, P@5, MAP and MRR?

NDCG - Example

5 documents: \(d_1, d_2, d_3, d_4, d_5\)

<table>
<thead>
<tr>
<th>i</th>
<th>Ground Truth</th>
<th>Ranking Function(_1)</th>
<th>Ranking Function(_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Document Order</td>
<td>rel(_i)</td>
<td>Document Order</td>
</tr>
<tr>
<td>1</td>
<td>d5</td>
<td>4</td>
<td>d3</td>
</tr>
<tr>
<td>2</td>
<td>d4</td>
<td>3</td>
<td>d4</td>
</tr>
<tr>
<td>3</td>
<td>d3</td>
<td>2</td>
<td>d2</td>
</tr>
<tr>
<td>4</td>
<td>d2</td>
<td>1</td>
<td>d5</td>
</tr>
<tr>
<td>5</td>
<td>d1</td>
<td>0</td>
<td>d1</td>
</tr>
</tbody>
</table>

\[
DCG_{GT} = \frac{2^4-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^2-1}{\log_2 4} + \frac{2^1-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 21.35
\]

\[
DCG_{RF1} = \frac{2^2-1}{\log_2 2} + \frac{2^3-1}{\log_2 3} + \frac{2^1-1}{\log_2 4} + \frac{2^4-1}{\log_2 5} + \frac{2^0-1}{\log_2 6} = 14.38
\]

\[
DCG_{RF2} = \frac{2^4-1}{\log_2 2} + \frac{2^2-1}{\log_2 3} + \frac{2^3-1}{\log_2 4} + \frac{2^0-1}{\log_2 5} + \frac{2^1-1}{\log_2 6} = 20.78
\]
What does query averaging hide?

Figure from Doug Oard’s presentation, originally from Ellen Voorhees’ presentation
Statistical significance tests

• How confident you are that an observed difference doesn’t simply result from the particular queries you chose?

<table>
<thead>
<tr>
<th>Query</th>
<th>Experiment 1</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>System A</td>
<td>System B</td>
</tr>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>0.22</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>0.17</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>7</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td><strong>0.20</strong></td>
<td><strong>0.40</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>Experiment 2</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>System A</td>
<td>System B</td>
</tr>
<tr>
<td>11</td>
<td>0.02</td>
<td>0.76</td>
</tr>
<tr>
<td>12</td>
<td>0.39</td>
<td>0.07</td>
</tr>
<tr>
<td>13</td>
<td>0.26</td>
<td>0.17</td>
</tr>
<tr>
<td>14</td>
<td>0.38</td>
<td>0.31</td>
</tr>
<tr>
<td>15</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>0.09</td>
<td>0.91</td>
</tr>
<tr>
<td>17</td>
<td>0.12</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td><strong>0.20</strong></td>
<td><strong>0.40</strong></td>
</tr>
</tbody>
</table>
Background knowledge

• $p$-value in statistic test is the probability of obtaining data as extreme as was observed, if the null hypothesis were true (e.g., if observation is totally random)

• If $p$-value is smaller than the chosen significance level ($\alpha$), we reject the null hypothesis (e.g., observation is not random)

• We seek to reject the null hypothesis (we seek to show that the observation is a random result), and so small $p$-values are good
Tests usually used in IR evaluations

• Sign test
  – Hypothesis: the difference median is zero between samples from two continuous distributions

• Wilcoxon signed rank test
  – Hypothesis: data are paired and come from the same population

• Paired $t$-test
  – Hypothesis: difference between two responses measured on the same statistical unit has a zero mean value

• One-tail v.s. two-tail?
  – If you aren’t sure, use two-tail
Statistical significance testing

<table>
<thead>
<tr>
<th>Query</th>
<th>System A</th>
<th>System B</th>
<th>Sign Test</th>
<th>paired t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.76</td>
<td>+</td>
<td>+0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.07</td>
<td>-</td>
<td>-0.32</td>
</tr>
<tr>
<td>3</td>
<td>0.26</td>
<td>0.17</td>
<td>-</td>
<td>-0.09</td>
</tr>
<tr>
<td>4</td>
<td>0.38</td>
<td>0.31</td>
<td>-</td>
<td>-0.07</td>
</tr>
<tr>
<td>5</td>
<td>0.14</td>
<td>0.02</td>
<td>-</td>
<td>-0.12</td>
</tr>
<tr>
<td>6</td>
<td>0.09</td>
<td>0.91</td>
<td>+</td>
<td>+0.82</td>
</tr>
<tr>
<td>7</td>
<td>0.12</td>
<td>0.56</td>
<td>+</td>
<td>+0.44</td>
</tr>
<tr>
<td>Average</td>
<td>0.20</td>
<td>0.40</td>
<td></td>
<td>p=0.9375</td>
</tr>
</tbody>
</table>

95% of outcomes

\[ p=0.9375 \quad p=0.2927 \]
Where do we get the relevance labels?

• Human annotation
  – Domain experts, who have better understanding of retrieval tasks
    • Scenario 1: annotator lists the information needs, formalizes into queries, and judges the returned documents
    • Scenario 2: given query and associated documents, annotator judges the relevance by inferring the underlying information need
Assessor consistency

• Is inconsistency of assessors a concern?
  – Human annotators are idiosyncratic and variable
  – Relevance judgments are subjective
• Studies mostly concluded that the inconsistency didn’t affect relative comparison of systems
  – Success of an IR system depends on how good it is at satisfying the needs of these idiosyncratic humans
  – Lesk & Salton (1968): assessors mostly disagree on documents at lower ranks, but measures are more affected by top-ranked documents
Measuring assessor consistency

• *kappa* statistic
  – A measure of agreement between judges
  \[ \kappa = \frac{P(A) - P(E)}{1 - P(E)} \]
  – *P(A)* is the proportion of the times judges agreed
  – *P(E)* is the proportion of times they would be expected to agree by chance
  – \( \kappa = 1 \) if two judges always agree
  – \( \kappa = 0 \) if two judges agree by chance
  – \( \kappa < 0 \) if two judges always disagree
Example of \textit{kappa} statistic

<table>
<thead>
<tr>
<th>judge 2 relevance</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>300</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>No</td>
<td>10</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>90</td>
<td>400</td>
</tr>
</tbody>
</table>

\[
P(A) = \frac{300 + 70}{400} = 0.925
\]

\[
P(E) = \left( \frac{80 + 90}{400 + 400} \right)^2 + \left( \frac{320 + 310}{400 + 400} \right)^2 = 0.2125^2 + 0.7878^2 = 0.665
\]

\[
\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.925 - 0.665}{1 - 0.665} = 0.776
\]
Prepare annotation collection

• Human annotation is expensive and time consuming
  – Cannot afford exhaustive annotation of large corpus
  – Solution: pooling
    • Relevance is assessed over a subset of the collection that is formed from the top $k$ documents returned by a number of different IR systems
Does pooling work?

• Judgments cannot possibly be exhaustive?
  – Relative rankings among the systems remain the same

• This is only a subset of people’s opinion about relevance
  – Relative rankings among the systems remain the same

• What about documents beyond top $k$?
  – Relative rankings among the systems remain the same

• A lot of research work can be done here
  – Effective pool construction
  – Depth v.s. diversity
Rethink retrieval evaluation

• Goal of any IR system
  – Satisfying users’ information need
• Core quality measure criterion
  – “how well a system meets the information needs of its users.” – wiki
What we have considered

• The ability of the system to present all relevant documents
  – Recall-driven measures

• The ability of the system to withhold non-relevant documents
  – Precision-driven measures
Challenging assumptions in classical IR evaluations

• Assumption 1
  – Queries sent to an IR system would be the same as those sent to a librarian (i.e., sentence-length request), and users want to have high recall

• Assumption 2
  – Relevance = independent topical relevance
    • Documents are independently judged, and then ranked (that is how we get the ideal ranking)
What we have not considered

• The physical form of the output
  – User interface
• The effort, intellectual or physical, demanded of the user
  – User effort when using the system
• Bias IR research towards optimizing relevance-centric metrics
What you should know

• Core criterion for IR evaluation
• Basic components in IR evaluation
• Classical IR metrics
• Statistical test
• Annotator agreement