Modern Retrieval Evaluations

Hongning Wang
CS@UVa
What we have known about IR evaluations

• Three key elements for IR evaluation
  – A document collection
  – A test suite of information needs
  – A set of relevance judgments

• Evaluation of unranked retrieval sets
  – Precision/Recall

• Evaluation of ranked retrieval sets
  – P@k, MAP, MRR, NDCG

• Statistic significance
  – Avoid randomness in evaluation
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

• Are traditional IR evaluations qualified for this purpose?
  – What is missing?
Do user preferences and evaluation measures line up? [Sanderson et al. SIGIR’10]

• Research question
  1. Does effectiveness measured on a test collection predict user preferences for one IR system over another?
  2. If such predictive power exists, does the strength of prediction vary across different search tasks and topic types?
  3. If present, does the predictive power vary when different effectiveness measures are employed?
  4. When choosing one system over another, what are the reasons given by users for their choice?
Experiment settings

• User population
  – Crowd sourcing
    • Mechanical Turk
    • 296 ordinary users

• Test collection
  – TREC’09 Web track
    • 50 million documents from ClueWeb09
  – 30 topics
    • Each included several sub-topics
    • Binary relevance judgment against the sub-topics
Experiment settings

- IR systems
  - 19 runs of submissions to the TREC evaluation

Query: espn sports

Aspect: Take me to the ESPN Sports home page.

If you are a user requiring documents about the required aspect above, which result would you choose?

- Left result is better
- Results are equally good
- Right result is better
- None of the results are relevant

Please mention your reason below (incomplete answers will not be accepted):

The right had more relevant information.

Users need to make side-by-side comparison to give their preferences over the ranking results.
Experimental results

• User preferences v.s., retrieval metrics

<table>
<thead>
<tr>
<th>Users</th>
<th>nDCG</th>
<th>MRR</th>
<th>P(10)</th>
<th>ERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>160</td>
<td>159</td>
<td>131</td>
<td>164</td>
</tr>
<tr>
<td>Rnk eql</td>
<td>21</td>
<td>21</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>Disagree</td>
<td>66</td>
<td>57</td>
<td>61</td>
<td>62</td>
</tr>
</tbody>
</table>

- Metrics generally match users’ preferences, no significant differences between metrics
Experimental results

• Zoom into nDCG
  – Separate the comparison into groups of small differences and large differences

<table>
<thead>
<tr>
<th>Users</th>
<th>nDCG</th>
<th>Small Δ</th>
<th>Large Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>160</td>
<td>96</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>65%</td>
<td>62%</td>
<td>70%</td>
</tr>
<tr>
<td>Rank equal</td>
<td>21</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td>10%</td>
<td>5%</td>
</tr>
<tr>
<td>Disagree</td>
<td>66</td>
<td>43</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>27%</td>
<td>28%</td>
<td>25%</td>
</tr>
<tr>
<td>247</td>
<td>155</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

– Users tend to agree more when the difference between the ranking results is large
Experimental results

- What if when one system did not retrieve anything relevant

<table>
<thead>
<tr>
<th>Users</th>
<th>nDCG</th>
<th>MRR</th>
<th>P(10)</th>
<th>ERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>88</td>
<td>88</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>Rnk eql</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Disagree</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>122</td>
<td>122</td>
<td>122</td>
</tr>
</tbody>
</table>

- All metrics tell the same and mostly align with the users
Experimental results

• What if when both systems retrieved something relevant at top positions

<table>
<thead>
<tr>
<th>Users</th>
<th>nDCG</th>
<th>MRR%</th>
<th>P(10)</th>
<th>ERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>72</td>
<td>56%</td>
<td>43</td>
<td>76</td>
</tr>
<tr>
<td>Rnk eql</td>
<td>11</td>
<td>9%</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Disagree</td>
<td>42</td>
<td>33%</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Ties</td>
<td>3</td>
<td>2%</td>
<td>40</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>128</td>
<td>128</td>
<td>128</td>
</tr>
</tbody>
</table>

– P@10 cannot distinguish the difference between systems
Conclusions of this study

• IR evaluation metrics measured on a test collection predicted user preferences for one IR system over another
• The correlation is strong when the performance difference is large
• Effectiveness of different metrics vary
How does clickthrough data reflect retrieval quality [Radlinski CIKM’08]

• User behavior oriented retrieval evaluation
  – Low cost
  – Large scale
  – Natural usage context and utility

• Common practice in modern search engine systems
  – A/B test
A/B test

• Two-sample hypothesis testing
  – Two versions (A and B) are compared, which are identical except for one variation that might affect a user's behavior
    • E.g., indexing with or without stemming
  – Randomized experiment
    • Separate the population into equal size groups
      – 10% random users for system A and 10% random users for system B
    • Null hypothesis: no difference between system A and B
      – Z-test, t-test
Behavior-based metrics

CRITERIA

1. Number of relevant results/documents
2. # of times the search query is refined
3. # of documents that were browsed
4. Amount of scrolling
5. # of clicks # of queries
6. # of documents returned/ unit of time
7. \[
\frac{\sum_{i=1}^{n} t_i}{t}
\]
8. How clean is the visualization
9. % of results coming from reliable websites
10. % of top results
Behavior-based metrics

• Abandonment Rate
  – Fraction of queries for which no results were clicked on

• Reformulation Rate
  – Fraction of queries that were followed by another query during the same session

• Queries per Session
  – Mean number of queries issued by a user during a session
Behavior-based metrics

• Clicks per Query
  – Mean number of results that are clicked for each query

• Max Reciprocal Rank
  – Max value of $1/r$, where $r$ is the rank of the highest ranked result clicked on

• Mean Reciprocal Rank
  – Mean value of $\sum_i 1/r_i$, summing over the ranks $r_i$ of all clicks for each query

• Time to First Click
  – Mean time from query being issued until first click on any result

• Time to Last Click
  • Mean time from query being issued until last click on any result
Behavior-based metrics

When search results become **worse**:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Change as ranking gets worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment rate</td>
<td>Increase (more bad result sets)</td>
</tr>
<tr>
<td>Reformulation rate</td>
<td>Increase (more need to reformulate)</td>
</tr>
<tr>
<td>Queries per session</td>
<td>Increase (more need to reformulate)</td>
</tr>
<tr>
<td>Clicks per query</td>
<td>Decrease (fewer relevant results)</td>
</tr>
<tr>
<td>Max recip. rank</td>
<td>Decrease (top results are worse)</td>
</tr>
<tr>
<td>Mean recip. rank</td>
<td>Decrease (more need for many clicks)</td>
</tr>
<tr>
<td>Time to first click</td>
<td>Increase (good results are lower)</td>
</tr>
<tr>
<td>Time to last click</td>
<td>Decrease (fewer relevant results)</td>
</tr>
</tbody>
</table>
Experiment setup

• Philosophy
  – Given systems with known relative ranking performance
  – Test which metric can recognize such difference

Reverse thinking of hypothesis testing
• In hypothesis testing, we choose system by test statistics
• In this study, we choose test statistics by systems
Constructing comparison systems

• Orig > Flat > Rand
  – Orig: original ranking algorithm from arXiv.org
  – Flat: remove structure features (known to be important) in original ranking algorithm
  – Rand: random shuffle of Flat’s results

• Orig > Swap2 > Swap4
  – Swap2: randomly selects two documents from top 5 and swaps them with two random documents from rank 6 through 10 (the same for next page)
  – Swap4: similar to Swap2, but select four documents for swap
Result for A/B test

- 1/6 users of arXiv.org are routed to each of the testing systems in one month period

<table>
<thead>
<tr>
<th></th>
<th>$H_1$</th>
<th>Orig</th>
<th>Flat</th>
<th>Rand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate (Mean)</td>
<td>&lt;</td>
<td>0.680 ± 0.021</td>
<td>0.725 ± 0.020</td>
<td>0.726 ± 0.020</td>
</tr>
<tr>
<td>Reformulation Rate (Mean)</td>
<td>&lt;</td>
<td>0.247 ± 0.021</td>
<td>0.257 ± 0.021</td>
<td>0.260 ± 0.021</td>
</tr>
<tr>
<td>Queries per Session (Mean)</td>
<td>&lt;</td>
<td>1.925 ± 0.098</td>
<td>1.963 ± 0.100</td>
<td>2.000 ± 0.115</td>
</tr>
<tr>
<td>Clicks per Query (Mean)</td>
<td>&gt;</td>
<td>0.713 ± 0.091</td>
<td>0.556 ± 0.081</td>
<td>0.533 ± 0.077</td>
</tr>
<tr>
<td>Max Reciprocal Rank (Mean)</td>
<td>&gt;</td>
<td>0.554 ± 0.029</td>
<td>0.520 ± 0.029</td>
<td>0.518 ± 0.030</td>
</tr>
<tr>
<td>Mean Reciprocal Rank (Mean)</td>
<td>&gt;</td>
<td>0.458 ± 0.027</td>
<td>0.442 ± 0.027</td>
<td>0.439 ± 0.028</td>
</tr>
<tr>
<td>Time (s) to First Click (Median)</td>
<td>&lt;</td>
<td>31.0 ± 3.3</td>
<td>30.0 ± 3.3</td>
<td>32.0 ± 4.0</td>
</tr>
<tr>
<td>Time (s) to Last Click (Median)</td>
<td>&gt;</td>
<td>64.0 ± 19.0</td>
<td>60.0 ± 14.0</td>
<td>62.0 ± 9.0</td>
</tr>
</tbody>
</table>
Result for A/B test

- 1/6 users of arXiv.org are routed to each of the testing systems in one month period

<table>
<thead>
<tr>
<th></th>
<th>$\text{H}_1$</th>
<th>ORIG $\geq$ Swap2 $\geq$ Swap4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abandonment Rate (Mean)</strong></td>
<td>$&lt;$</td>
<td>$0.704 \pm 0.021 \geq 0.680 \pm 0.021 \geq 0.698 \pm 0.021$</td>
</tr>
<tr>
<td><strong>Reformulation Rate (Mean)</strong></td>
<td>$&lt;$</td>
<td>$0.248 \pm 0.021 \geq 0.250 \pm 0.021 \geq 0.248 \pm 0.021$</td>
</tr>
<tr>
<td><strong>Queries per Session (Mean)</strong></td>
<td>$&lt;$</td>
<td>$1.971 \pm 0.110 \geq 1.957 \pm 0.099 \geq 1.884 \pm 0.091$</td>
</tr>
<tr>
<td><strong>Clicks per Query (Mean)</strong></td>
<td>$&gt;$</td>
<td>$0.720 \pm 0.098 \geq 0.760 \pm 0.127 \geq 0.734 \pm 0.125$</td>
</tr>
<tr>
<td><strong>Max Reciprocal Rank (Mean)</strong></td>
<td>$&gt;$</td>
<td>$0.538 \pm 0.029 \geq 0.559 \pm 0.028 \geq 0.488 \pm 0.029$</td>
</tr>
<tr>
<td><strong>Mean Reciprocal Rank (Mean)</strong></td>
<td>$&gt;$</td>
<td>$0.444 \pm 0.027 \geq 0.467 \pm 0.027 \geq 0.394 \pm 0.026$</td>
</tr>
<tr>
<td><strong>Time (s) to First Click (Median)</strong></td>
<td>$&lt;$</td>
<td>$28.0 \pm 2.2 \geq 28.0 \pm 3.0 \geq 32.0 \pm 3.5$</td>
</tr>
<tr>
<td><strong>Time (s) to Last Click (Median)</strong></td>
<td>$&gt;$</td>
<td>$71.0 \pm 19.0 \geq 56.0 \pm 10.0 \geq 66.0 \pm 15.0$</td>
</tr>
</tbody>
</table>
Result for A/B test

• Few of such comparisons are significant

<table>
<thead>
<tr>
<th></th>
<th>weak</th>
<th>signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate (Mean)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Reformulation Rate (Mean)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Queries per Session (Mean)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Clicks per Query (Mean)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Max Reciprocal Rank (Mean)</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Mean Reciprocal Rank (Mean)</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Time (s) to First Click (Median)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Time (s) to Last Click (Median)</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
Interleave test

• Design principle from sensory analysis
  – Instead of asking for absolute ratings, ask for relative comparison between alternatives
    • E.g., is A better than B?
  – Randomized experiment
    • Interleave results from both A and B
    • Giving interleaved results to the same population and ask for their preference
    • Hypothesis test over preference votes
Coke v.s. Pepsi

• Market research
  – Do customers prefer coke over pepsi, or they do not have any preference
  – Option 1: A/B Testing
    • Randomly find two groups of customers and give coke to one group and pepsi to another, and ask them if they like the given beverage
  – Option 2: Interleaved test
    • Randomly find a group of users and give them both coke and pepsi, and ask them which one they prefer
Interleave for IR evaluation

• Team-draft interleaving

Input: Rankings $A = (a_1, a_2, \ldots)$ and $B = (b_1, b_2, \ldots)$
Init: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset$
while $(\exists i : A[i] \notin I) \land (\exists j : B[j] \notin I)$ do
  if $(|TeamA| < |TeamB|) \lor$
    $((|TeamA| = |TeamB|) \land (RandBit() = 1))$ then
    $k \leftarrow \min_i \{i : A[i] \notin I\}$ ...... top result in $A$ not yet in $I$
    $I \leftarrow I + A[k]$; ......................... append it to $I$
    $TeamA \leftarrow TeamA \cup \{A[k]\}$ ...... clicks credited to $A$
  else
    $k \leftarrow \min_i \{i : B[i] \notin I\}$ ...... top result in $B$ not yet in $I$
    $I \leftarrow I + B[k]$ ............................... append it to $I$
    $TeamB \leftarrow TeamB \cup \{B[k]\}$ ...... clicks credited to $B$
  end if
end while
Output: Interleaved ranking $I$, $TeamA$, $TeamB$
Interleave for IR evaluation

• Team-draft interleaving

Ranking A: 2 3 1 4 5 7 8 6

Ranking B: 1 2 5 3 6 8 7 4

RND = 0

Interleaved ranking 1 2 3 5 4 6
Result for interleaved test

- 1/6 users of arXiv.org are routed to each of the testing systems in a one month period
  - Test which group receives more clicks

<table>
<thead>
<tr>
<th>Comparison Pair</th>
<th>Query Based</th>
<th>User Based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A wins</td>
<td>B wins</td>
</tr>
<tr>
<td>Orig ⇒ Flat</td>
<td>47.7%</td>
<td>37.3%</td>
</tr>
<tr>
<td>Flat ⇒ Rand</td>
<td>46.7%</td>
<td>39.7%</td>
</tr>
<tr>
<td>Orig ⇒ Rand</td>
<td>55.6%</td>
<td>29.8%</td>
</tr>
<tr>
<td>Orig ⇒ Swap2</td>
<td>44.4%</td>
<td>40.3%</td>
</tr>
<tr>
<td>Swap2 ⇒ Swap4</td>
<td>44.2%</td>
<td>40.3%</td>
</tr>
<tr>
<td>Orig ⇒ Swap4</td>
<td>47.7%</td>
<td>37.8%</td>
</tr>
</tbody>
</table>
Conclusions

• Interleaved test is more accurate and sensitive
  – 9 out of 12 experiments follow our expectation

• Only click count is utilized in this interleaved test
  – More aspects can be evaluated
    • E.g., dwell-time, reciprocal rank, if leads to download, is last click, is first click

• Interleave more than two systems for comparison
Comparing the sensitivity of information retrieval metrics [Radlinski & Craswell, SIGIR’10]

• How sensitive are those IR evaluation metrics?
  – How many queries do we need to get a confident comparison result?
  – How quickly it can recognize the difference between different IR systems?
Experiment setup

• IR systems with known search effectiveness
• Large set of annotated corpus
  – 12k queries
  – Each retrieved document is labeled into 5-grade level
• Large collection of real users’ clicks from a major commercial search engine
• Approach
  – Gradually increase evaluation query size to investigate the conclusion of metrics
Sensitivity of NDCG@5

System effectiveness: A>B>C
Sensitivity of P@5

System effectiveness: A>B>C
Sensitivity of interleaving

Freq. of better ranking winning by interleaving

Number of impressions sampled

1k 2k 5k 10k 20k 50k 100k 200k

Experiment
- majorAC
- majorBC
- majorAB
- minorD
- minorE
Correlation between IR metrics and interleaving

<table>
<thead>
<tr>
<th>Inter’l Scoring</th>
<th>IR Metric</th>
<th>Correlation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per impression</td>
<td>NDCG@5</td>
<td>0.882</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>MAP@10</td>
<td>0.689</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>P@5</td>
<td>0.662</td>
<td>0.223</td>
</tr>
<tr>
<td>Per query</td>
<td>NDCG@5</td>
<td>0.910</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>MAP@10</td>
<td>0.776</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>P@5</td>
<td>0.733</td>
<td>0.159</td>
</tr>
</tbody>
</table>
How to assess search result quality?

• Query-level relevance evaluation
  – Metrics: MAP, NDCG, MRR, CTR

• Task-level satisfaction evaluation

Goal: find existing work for "action-level search satisfaction prediction"
Example of search task

• Information need: *find out what metal can float on water*

<table>
<thead>
<tr>
<th>Search Actions</th>
<th>Engine</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: metals float on water</td>
<td>Google</td>
<td>10s</td>
</tr>
<tr>
<td>SR: wiki.answers.com</td>
<td></td>
<td>2s</td>
</tr>
<tr>
<td>BR: blog.sciseek.com</td>
<td></td>
<td>3s</td>
</tr>
<tr>
<td>Q: which metals float on water</td>
<td>Google</td>
<td>31s</td>
</tr>
<tr>
<td>Q: metals floating on water</td>
<td>Google</td>
<td>16s</td>
</tr>
<tr>
<td>SR: <a href="http://www.blurtit.com">www.blurtit.com</a></td>
<td></td>
<td>5s</td>
</tr>
<tr>
<td>Q: metals floating on water</td>
<td>Bing</td>
<td>53s</td>
</tr>
<tr>
<td>Q: lithium sodium potassium float on water</td>
<td>Google</td>
<td>38s</td>
</tr>
<tr>
<td>SR: <a href="http://www.docbrowm.info">www.docbrowm.info</a></td>
<td></td>
<td>15s</td>
</tr>
</tbody>
</table>
Beyond DCG: User Behavior as a Predictor of a Successful Search [Ahmed et al. WSDM’10]

- Modeling users’ sequential search behaviors with Markov models
  - A model for successful search patterns
  - A model for unsuccessful search patterns

ML for parameter estimation on annotated data set
Predict user satisfaction

- Choose the model that better explains users’ search behavior

\[
P(S = 1|B) = \frac{p(S=1)}{p(S=0) + p(S=0)}
\]

Likelihood: how well the model explains users’ behavior
Prior: difficulty of this task, or users’ expertise of search

Prediction performance for search task satisfaction
What you should know

• IR evaluation metrics generally align with users’ result preferences
• A/B test v.s. interleaved test
• Sensitivity of evaluation metrics
• Direct evaluation of search satisfaction