Predicting Query Performance Using Query, Result, and User Interaction Features

Qi Guo
Emory University
Ryen White, Susan Dumais, Jue Wang, Blake Anderson
Microsoft

Presented by Tetsuya Sakai, Microsoft Research
Motivation

- Query evaluation is critical for search engines
  - Understanding the quality of search results for individual queries (or in the aggregate)

- Query evaluation often involves:
  - Time-consuming and expensive human judgments
  - User studies covering only a small fraction of queries

- Automated methods can lead to rapid and cost-effective query performance prediction

- Prior work used features of queries, results, the collection
  - *E.g.*, Query clarity (Cronen-Townsend et al., 2002); query difficulty prediction (Hauff et al., 2008)
Contribution

Our work differs from previous research:

- Investigates query, results, and interaction features
- Uses search engine logs (rather than standard IR test collections), since they reflect diversity of Web search tasks

Contributions:

- Investigate novel and rich set of interaction features in predicting query performance
- Determine which features and combinations of features are more important in predicting query quality
- Understand how accuracy varies with query frequency
Predicting Query Performance

- Measured using DCG at rank position 3 (DCG@3)
  - Captures relevance of top-ranked search results
  - Relevance of each result measured on five-point scale

\[
DCG@3 = \sum_{i=1}^{3} \frac{(2^{rel_i} - 1)}{\log_2(i + 1)}
\]

- Range of score [0, 66.1] - normalized to [0, 1]
- Goal is to develop a model that can accurately predict DCG@3
- We use DCG (rather than NDCG) because it is an absolute performance score, not normalized by the score of other rankers
Features

- Three classes:
  - **Query** from Bing search logs
    - *E.g.*, QueryLength, HasURLFragment, HasSpellCorrection
  - **Results** from Bing results pages
    - *E.g.*, AvgNumAds, AvgNumResults, MaxBM25F
    - Text-matching baseline
  - **Interaction** from Bing search logs and MSN toolbar logs
    - *E.g.*, AvgClickPos, AvgClickDwell, AbandonmentRate
    - Include search engine switching and user satisfaction estimates
      - Satisfaction estimates based on page dwell times

- Logs collected during one week in July 2009
Experiment

- 2,834 queries from randomly sampling Bing query logs
  - Mixture of common and rare queries
  - 60% training / 20% validation / 20% testing
  - Explicit relevance judgments used to generate ground truth DCG values for training and testing

- Query / Results / Interaction features generated for each query in the set
Experiment

- Prediction model
  - Regression: multiple additive regression trees (MART)
  - Advantages of MART include model interpretability, facility for rapid training and testing, and robustness

- Metrics used to evaluate performance
  - Pearson’s correlation (R), mean absolute error (MAE)
  - Compare predicted DCG@3 with ground truth (DCG@3 based on explicit human judgments)

- Five-fold cross validation to improve result reliability
Findings: All Features

- Effectively predicts DCG@3
  - $R=0.699$, $MAE=0.160$

- Correlation is sensible across the full range of DCG values

- Most predictive feature is an interaction feature
  - *Average rank of result click*

- Disagreements in prediction associated with novel result presentation
  - *E.g.*, Instant answers (likes maps and images) may influence user interaction features
### Findings: Feature Combinations

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>R</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query + Results + Interaction (full model)</td>
<td>0.699</td>
<td>0.154</td>
</tr>
<tr>
<td>Results + Interaction</td>
<td>0.698</td>
<td>0.160</td>
</tr>
<tr>
<td>Query + Interaction</td>
<td>0.678</td>
<td>0.164</td>
</tr>
<tr>
<td>Interaction only</td>
<td>0.667*</td>
<td>0.166*</td>
</tr>
<tr>
<td>Query + Results</td>
<td>0.556**</td>
<td>0.193**</td>
</tr>
<tr>
<td>Results only</td>
<td>0.522**</td>
<td>0.200**</td>
</tr>
<tr>
<td>Query only</td>
<td>0.323**</td>
<td>0.228**</td>
</tr>
</tbody>
</table>

(Diff. from full model: * = p < .05, ** = p < .01)

- **Interaction** features perform close to all features
  - Strong predictive signal in interaction behavior
- **Results** features perform reasonably well
- **Query** features perform poorly
  - Do not add much to **Results or Interaction** features
Findings: Query Frequency

- Interaction features are important, but mostly available for frequent queries
  - How well can we do on infrequent queries?

- We looked at the correlation for different frequency bins
  - Ranked queries by frequency
  - Divided queries into equally-sized bins
  - Computed correlation between predicted & actual DCG@3
Findings: Query Frequency

- Linear regression revealed very slight relationship between query frequency and prediction accuracy ($R^2 = .008$)

- This is good – we can accurately predict for non-popular queries.
Summary

- Automatically predicted search engine performance using query, results, and interaction features
- Strong correlation (R ≈ 0.7) between predicted query performance and human relevance judgments using all feature classes
- Users’ search interactions provide a strong signal of engine performance, performing well alone and adding substantially to Query and Results features
Implications and Future Work

• Accurate prediction can help search engines:
  • Know when to apply different processing / ranking / presentation methods
  • Identify poorly-performing queries
  • Sample queries of different quality

• Further research is required to understand:
  • Role of other features
  • Effects related to the nature of the document collection
  • Impact of engine settings on prediction effectiveness