Sensitive Online Search Evaluation

Filip Radlinski
Bing / Microsoft Research
Cambridge, UK
Online Search Evaluation Goals

Original Search System

Personalized Search
Personalized Search → Personalized Web Search Personalized Web → Data Integration in Web Data Extraction System Personalized Web Search JIR-ONG ...
research.microsoft.com/pubs/79334/publishedversion.pdf - PDF file

A personalized search research based on vocabulary semantic.net
Along with the fast developing of network technology, the number of Web page and user of network search become very enormous. In order to solve the problem of ...
portal.acm.org/citation.cfm?id=1794768

Zakta – Personalized Social Search Engine - edit sec to refox
Zakta, unlike other social search engines, can be considered as its ability to dig deeper to get the required information and ...
techpp.com/2009/10/15/zakta-personalized-social-search

Related Searches for personalized search results
Ontology-based Personalized S... Disable Personalized Search Personalized Search Results Personalization Business

New Search System

ACM SIGIR Special Interest Group on Information Retrieval Home Page
Welcome to the ACM SIGIR Web site. ACM SIGIR addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, and...
www.sigir.org

Personalizing Search via Automated Analysis of Interests and...
Personalizing Search via Automated Analysis of interests and Activities Jaime Teven MIT, CSAIL 32 Vassar Street, 60-32 Cambridge, MA 02138 USA lee.van@csail.mit.edu Susan T ...

Research framework to utilize folksonomy for ... SIGIR '08 Proceedings of the 31st Annual ACM SIGIR conference on Research ...
10.1145/1375424.1375483

Personali\ized search - Wikipedia, the free encyclopedia
Personalized search refers to search experiences ... specific groups of people. personalized search depends on a user profile that is unique to the individual. Research ...
en.wikipedia.org/wiki/Personalized_search

Research from Microsoft: Personalized Search, Determining a Query ...
The other day I posted about a paper presented at the SIGIR conference a few weeks ago. Apparently, that get Findory CEO, Greg Linden, looking for other ...
bloginsearchenginewatch.com/blog/050926-121640

Adapting SEO for Personalized Search
Ok, but seriously, the last round of personalized search research we did here on it seems to suggest that a lot of the personalization, in relatively new query ...
www.searchenginejournal.com/adapting-seo-for-personalized-search/22207

Which is better?

Goals: Correctness, Practicality, Efficiency
Retrieval Evaluation Goals

• Correctness
  – If my evaluation says the new method is better, would users really agree?
  – Would the users really notice?
  – If my evaluation says the new method isn’t better, is that true?

• Practicality
  – The metrics are as simple and intuitive as possible

• Efficiency / Sensitivity
  – I want to make the best use of my resources: How do I best trade off time/cost and sensitivity to changes?
  – Want to avoid “I’m not sure”
Evaluation

Two general types of retrieval evaluation:

• “Offline evaluation” : Manual judgments
  Ask experts or users to explicitly evaluate your retrieval system.

• “Online evaluation” : Observing users
  See how normal users interact with your retrieval system when just using it.
  - Measurement can be passive, or active
Offline Evaluation

• Offline evaluation of a search system usually involves these steps:
  1. Select queries to evaluate on
  2. Get results for those queries
  3. Assess the relevance of those results to the queries
  4. Compute your offline metric
## Offline Evaluation

<table>
<thead>
<tr>
<th>Query</th>
<th>Document</th>
<th>Relevant?</th>
</tr>
</thead>
<tbody>
<tr>
<td>bcs</td>
<td><a href="http://www.bcs.org/">http://www.bcs.org/</a></td>
<td>✓</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Another offline approach

System 1

Search Solution Group Recruiter
www.searchsolutiongroup.com
Search Solution group is a national executive search firm that delivers the best talent for prominent client companies

Search All Positions
Search through all open job positions. ... CONTACT OFFICES. North ... Contact

Become a Client
Search Solution Group finds the best talent for clients with speed, ... Contact If ...

Our Team
Meet our talented team members at Search Solution Group!

Candidates
New career opportunities through Search Solution Group for high ...

Submit a Resume
Search Solution Group | Executive Recruiting Firm ... CONTACT ...

See results only from searchsolutiongroup.com

Search Solutions International, Inc.
searchsolutionscorp.com
Search Solutions International, Inc. 1050 Sweet Briar Place Wellington, Florida 33414 Phone: 661-283-4205 FAX: 954-416-5172 11/18/2013 New Job Postings

Search Solutions - Search Solutions is a recruiting / staffing ...
www.thesearchsolutions.com
Introduction. The Search Solutions. Search Solutions, a recruiting and staffing firm specializing in information technology and engineering professionals.

Related searches for search solutions
Legal Search Solutions Boca Raton Search Solutions Charlotte
Job Search Solutions Tony Beshara Search Solution Group Inc
Executive Search Solutions Inc Search Solutions Group Charlotte
Adobe Marketing Cloud

Search marketing service offering search engine optimisation ...
searchsolutionsuk.com
Search engine optimisation specialists offering online marketing ... Welcome to Search Solutions : Our team has been active in the Search Engine Optimisation market ...

Our
www.search-solutions.co.uk
Search. Home Page; Our Approach; Registration; Latest News; Contact Us; We have 30 Years + Experience in the Professional Qualified Accountancy Recruitment Industry

Job Search & Recruitment Agency | Search Consultancy
www.searchconsultancy.co.uk
Get the latest job search results from the UK's leading recruitment agency ... We can provide you with the complete range of fully integrated recruitment solutions ...

Related Searches for search solutions
Search Solutions Leeds Search Solutions Recruitment
Enterprise Search Solutions

Web Search Solutions - Website Development, Search Engine ...
www.websearchsolutions.co.uk
Web Search Solutions optimise your website to be at the forefront of your business' industry.

Search Solutions 2011
irsg.bcs.org/SearchSolutions2011/ssc2011.php
Wednesday 16th November, Covent Garden, London Innovations in Web & Enterprise Search. Search Solutions is a special one-day event dedicated to the latest innovations ...

Business Solutions – Google
www.google.co.uk/services
Business Solutions: Skip to content. Get your business online. Google+ pages give your
Online Evaluation

Assumption:

Observable user behavior reflects relevance

• This assumption gives us “high fidelity”
  
  Real users replace the judges: No ambiguity in information need; Users actually want results; Measures performance on real queries

• But introduces a major challenge
  
  We can’t train the users: How do we know when they are happy? Real user behavior requires careful design and evaluation
Implicit Feedback

• A variety of data captures online search behavior:

  – Search Queries
    • The sequence of queries issued to the search engine
  – Results and Clicks
    • The results shown, and which results were clicked on
  – Mouse movement, selections and hovering, scrolling, dwell time, bookmarking, ...
  – Potentially what the user does after searching
    • Sequence of URLs visited
Online Evaluation Designs

• We have some key choices to make:

  1. Document Level or Ranking Level?

<table>
<thead>
<tr>
<th>Document Level</th>
<th>Ranking Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to know about the <em>documents</em></td>
<td>I am mostly interested in the <em>rankings</em></td>
</tr>
<tr>
<td>Similar to the offline approach, I’d like to find out the quality of each document.</td>
<td>I’m trying to evaluate retrieval functions. I don’t need to be able to drill down to individual documents.</td>
</tr>
</tbody>
</table>

  2. Absolute or Relative?

<table>
<thead>
<tr>
<th>Absolute Judgments</th>
<th>Relative Judgments</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want a score on an absolute scale</td>
<td>I am mostly interested in a comparison</td>
</tr>
<tr>
<td>Similar to the offline approach, I’d like a number that I can use to compare many methods, over time.</td>
<td>It’s enough if I know which document, or which ranking, is better. Its not necessary to know the absolute value.</td>
</tr>
</tbody>
</table>
Absolute Ranking-Level Evaluation

• Document-level feedback requires converting judgments to evaluation metric (of a ranking)

• Ranking-level judgments directly define such a metric

<table>
<thead>
<tr>
<th>Some Absolute Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
</tr>
<tr>
<td>Reformulation Rate</td>
</tr>
<tr>
<td>Queries per Session</td>
</tr>
<tr>
<td>Clicks per Query</td>
</tr>
<tr>
<td>Click rate on first result</td>
</tr>
<tr>
<td>Max Reciprocal Rank</td>
</tr>
<tr>
<td>Time to first click</td>
</tr>
<tr>
<td>Time to last click</td>
</tr>
<tr>
<td>% of viewed documents skipped (pSkip)</td>
</tr>
</tbody>
</table>

[Radlinski et al. 2008; Wang et al. 2009]
Monotonicity Assumption

• Consider two sets of results: A & B
  – A is high quality
  – B is medium quality

• Which will get more clicks from users, A or B?
  – A has more good results: Users may be more likely to click when presented results from A.
  – B has fewer good results: Users may need to click on more results from ranking B to be satisfied.

• Need to test with real data
  – If either direction happens consistently, with a reasonable amount of data, we can use this to evaluate online
Example Evaluation

• Experiments performed on the arXiv.org e-print archive.
  – Index of research articles in physics, maths, computer science, etc.
  – The users are mostly scientific users.

• Each article has rich meta-data:
  – Title, authors, abstract, full text, article identifier, a few others.
Original Ranking Function

• Start with something reasonable - sum of:
  – Similarity between query and title
  – Similarity between query and abstract
  – Similarity between query and authors
  – ...

• Text in title, author list and abstract particularly important for good matches.
Degradation Type 1

• Degraded results in two steps:

  1. “**FLAT**”: Ignored all the meta-data except for full text, author list and article id.
  2. “**RAND**”: Randomized the top 11 results returned by FLAT.

• Subjective impression:

  **ORIG** is substantially better than **RAND**, and evaluation should be able to see this difference.
Degradation Type 2

• Degraded results in two different steps:

  1. “SWAP2”: Randomly swap two documents between ranks 1 and 5 with two between 7 and 11.
  2. “SWAP4”: Randomly swap four documents between ranks 1 and 5 with four between 7 and 11.

• Subjective impression:
  Difference smaller than before; top 11 documents always include all the same results.
What we now have

• We have two triplets of ranking functions.
• It is reasonable to assume that we know the relative quality of the rankings:
  
  \[ \text{ORIG} \succ \text{FLAT} \succ \text{RAND} \]
  \[ \text{ORIG} \succ \text{SWAP2} \succ \text{SWAP4} \]

• This gives us 6 pairs of ranking functions that we can compare.
• We’ll see if there is any difference in behaviour.
## Absolute Metrics

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Hypothesized Change as Quality Falls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
<td>% of queries with no click</td>
<td>Increase</td>
</tr>
<tr>
<td>Reformulation Rate</td>
<td>% of queries that are followed by reformulation</td>
<td>Increase</td>
</tr>
<tr>
<td>Queries per Session</td>
<td>Session = no interruption of more than 30 minutes</td>
<td>Increase</td>
</tr>
<tr>
<td>Clicks per Query</td>
<td>Number of clicks</td>
<td>Decrease</td>
</tr>
<tr>
<td>Clicks @ 1</td>
<td>Clicks on top results</td>
<td>Decrease</td>
</tr>
<tr>
<td>pSkip [Wang et al ’09]</td>
<td>Probability of skipping</td>
<td>Increase</td>
</tr>
<tr>
<td>Max Reciprocal Rank*</td>
<td>1/rank for highest click</td>
<td>Decrease</td>
</tr>
<tr>
<td>Mean Reciprocal Rank*</td>
<td>Mean of 1/rank for all clicks</td>
<td>Decrease</td>
</tr>
<tr>
<td>Time to First Click*</td>
<td>Seconds before first click</td>
<td>Increase</td>
</tr>
<tr>
<td>Time to Last Click*</td>
<td>Seconds before final click</td>
<td>Decrease</td>
</tr>
</tbody>
</table>

(*) only queries with at least one click count
### Evaluation of Absolute Metrics on ArXiv.org

- How well do statistics reflect the known quality order?

<table>
<thead>
<tr>
<th>Evaluation Metric</th>
<th>Consistent (weak)</th>
<th>Inconsistent (weak)</th>
<th>Consistent (strong)</th>
<th>Inconsistent (strong)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks per Query</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks @ 1</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>pSkip</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Max Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mean Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Time to First Click</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time to Last Click</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

[Radlinski et al. 2008; Chapelle et al. 2012]
Evaluation of Absolute Metrics on ArXiv.org

• How well do statistics reflect the known quality order?

<table>
<thead>
<tr>
<th>Evaluation Metric</th>
<th>Consistent</th>
<th>Inconsistent</th>
<th>Consistent</th>
<th>Inconsistent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks per Query</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks @ 1</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>pSkip</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Max Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mean Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Time to First Click</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time to Last Click</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Absolute Metric Summary**

• None of the absolute metrics reliably reflect expected order.

• Most differences not significant with thousands of queries.

➤ (These) absolute metrics not suitable for ArXiv-sized search engines with these retrieval quality differences.

[Radlinski et al. 2008; Chapelle et al. 2012]
Comparing Rankings Efficiently

• Suppose you want to compare two rankings

  Ranking A
  http://www.profootballhof.com/
  http://www.nfl.com/halloffame
  http://www.mahalo.com/pro-football-hall-of-fame-game

  Ranking B
  http://www.profootballhof.com/
  http://en.wikipedia.org/wiki/NFL_Hall_Of_Fame_Game
  http://en.wikipedia.org/wiki/NFL_Hall_Of_Fame

• So far, we assumed some users see A, others B.
• We measure a metric on both, and compare
  – But we really just want to know which is better
• What if we can show something different?
Taste-test analogy

• Suppose we conduct taste experiment: vs  
  – Want to maintain a natural usage context

• Experiment 1: absolute metrics
  – Each participant’s refrigerator randomly stocked
    • Either Pepsi or Coke (anonymized)
  – Measure how much participant drinks

• Issues:
  – Calibration (person’s thirst, other confounding variables...)
  – Higher variance
Taste-test analogy

• Suppose we conduct taste experiment: vs
  – Want to maintain natural usage context

• Experiment 2: relative metrics
  – Each participant’s refrigerator randomly stocked
    • Some Pepsi (A) and some Coke (B)
  – Measure how much participant drinks of each
    • (Assumes people drink rationally!)

• Issues solved:
  – Controls for each individual participant
  – Lower variance
Online Evaluation with Interleaving

- A within-user online ranker comparison
  - Presents results from both rankings to every user
Why might mixing rankings help?

• Suppose results are worth money. For some query:
  – Ranker A: 
    ➔ User clicks
  – Ranker B: 
    ➔ User also clicks

• Users of A may not know what they’re missing
  – Difference in behaviour is small

• But if we can mix up results from A & B
  ➔ Strong preference for B

• Challenge: Mix in a way to avoid biases
Online Evaluation with Interleaving

• A within-user online ranker comparison
  – Presents results from both rankings to every user

- The ranking that gets more of the clicks wins
Team Draft Interleaving

**Ranking A**
1. Napa Valley – The authority for lodging...
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...
   www.napavalley.com/wineries
3. Napa Valley College
   www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley
   www.ivebeenthere.co.uk
5. Napa Valley Wineries and Wine
   www.napavintners.com
6. Napa Country, California
   en.wikipedia.org/wiki/Napa_Valley

**Ranking B**
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...
   www.napavalley.com
3. Napa: The Story of an American Eden...
   books.google.co.uk/books?isbn=
4. Napa Valley Hotels – Bed and Breakfast...
   www.napalinks.com
5. NapaValley.org
   www.napavalley.org
6. The Napa Valley Marathon
   www.napavalleymarathon.org

**Presented Ranking**
### Team Draft Interleaving

#### Ranking A
1. Napa Valley – The authority for lodging...  
   www.napavalley.com
2. Napa Valley Wineries - Plan your wine...  
   www.napavalley.com/wineries
3. Napa Valley College  
   www.napavalley.edu/homex.asp
4. Been There | Tips | Napa Valley  
   www.ivebeenthere.co.uk
5. Napa Valley Wineries and Wine  
   www.napavintners.com
6. Napa Country, California  
   en.wikipedia.org/wiki/Napa_Valley

#### Ranking B
   en.wikipedia.org/wiki/Napa_Valley
2. Napa Valley – The authority for lodging...  
   www.napavalley.com
3. Napa: The Story of an American Eden...  
   books.google.co.uk/books?isbn=...
4. Napa Valley Hotels – Bed and Breakfast...  
   www.napalinks.com
5. NapaValley.org  
   www.napavalley.org
6. The Napa Valley Marathon  
   www.napavalleymarathon.org

#### Presented Ranking
1. Napa Valley – The authority for lodging...  
   www.napavalley.com
   en.wikipedia.org/wiki/Napa_Valley
3. Napa: The Story of an American Eden...  
   books.google.co.uk/books?isbn=...
4. Napa Valley Wineries – Plan your wine...  
   www.napavalley.com/wineries
5. Napa Valley Hotels – Bed and Breakfast...  
   www.napalinks.com
6. Napa Valley College  
   www.napavalley.edu/homex.asp
7. NapaValley.org  
   www.napavalley.org

[Tie!][Radlinski et al. 2008]
Quantitative Analysis

• Can we quantify how well interleaving performs?
  – Compared to Absolute Ranking-level Metrics
  – Compared to Offline Judgments

• How reliable is it?
  – Does Interleaving correctly identify the better retrieval function?

• How sensitive is it?
  – How much data is required to achieve a target confidence level (p-value)?

[Radlinski et al. 2008; Chapelle et al. 2012]
Experimental Setup

• Selected 4-6 pairs of ranking functions to compare in different settings
  – Known retrieval quality, by construction or by judged evaluation

• Observed user behavior in two experimental conditions
  – Randomly used one of the two individual ranking functions
  – Presented an interleaving of the two ranking functions

• Evaluation performed on three different search platforms
  – arXiv.org
  – Bing Web search
  – Yahoo! Web search

[Radlinski et al. 2008; Radlinski & Craswell 2010; Chapelle et al. 2012]
Comparison with Offline Judgments

- Experiments on Bing (large scale experiment)
- Plotted interleaving preference vs NDCG@5 difference
- Good calibration between expert judgments and interleaving

[Radlinski & Craswell 2010; Chapelle et al. 2012]
Comparison with Absolute Metrics (Online)

- Experiments on Yahoo!
  (very small differences in relevance)
- Large scale experiment

[Chapelle et al. 2012]
## Comparative Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Consistent (weak)</th>
<th>Inconsistent (weak)</th>
<th>Consistent (strong)</th>
<th>Inconsistent (strong)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks per Query</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Clicks @ 1</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>pSkip</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Max Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mean Reciprocal Rank</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Time to First Click</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time to Last Click</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Interleaving</strong></td>
<td><strong>6</strong></td>
<td><strong>0</strong></td>
<td><strong>6</strong></td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>

- Comparison on arXiv.org experiments
- Results on Yahoo! qualitatively similar

[Radlinski et al. 2008; Chapelle et al. 2012]
When to use Interleaving

• Benefits
  – A direct way to elicit user preferences
  – More sensitive than many other online metrics
  – Deals with issues of position bias and calibration
  – Roughly 10 clicked queries =~ 1 judged query (on Bing)

• Drawbacks
  – Reusability: Not easy to reuse judgment data collected
  – Benchmark: No absolute number for benchmarking
  – Interpretation: Unable to interpret much at the document-level, or about user behavior
Thanks! Questions?

filiprad@microsoft.com

Acknowledgments
Joint work with Olivier Chapelle, Nick Craswell, Thorsten Joachims, Madhu Kurup, Yisong Yue