Intern Presentation

A/B Testing by Interleaving

Sida Wang
My Project

• Evaluating search relevance by interleaving results and collecting user data
  – Interleaving Framework
    • Generic, Extensible
  – Experiments to evaluate relevance by interleaving
    • Based on the paper [How Does Clickthrough Data Reflect Retrieval Quality?](https://example.com) by F. Radlinski et al
Evaluating Search Relevance

• Without Interleaving
  - Full time human judges -> precision, recall, NDCG
  - Compare Search
Compare Search
Issues

• But do Microsoft people pick O14 Search or Google Mini?
• Maybe people tend to pick the left?
• **Alters the search experience**
  – Can never collect a lot of data using this method
By Interleaving
By Interleaving
Considerations

• Minimize impact to UX
  – So no demo, it looks exactly like normal search
• Minimize Bias
  – Summary normalization
  – Interleaving algorithms
• Reliability / performance / and the usual
Experiments I did

• Automated random clicks
• Automated clicks according to relevance judgments
• Clicks from real people
Random Clicks

Control Using Automated Random Clicks

% of Votes Received vs Clicks

- Betaa
- MSW
- Ties
A Lot of Random Clicks

Control Using Automated Random Clicks

% of Votes Received vs Clicks graph with lines representing different categories: ac11, a86f, and ties.
Experiments I did

- Automated random clicks
- Automated clicks according to relevance judgments
- Clicks from real people
O12 vs. O14
Automated Clicks Using Relevance Judgments

% of Votes Received

Clicks

-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1000
2000
3000
4000

Acing05 Degraded
Acing05
Ties
Experiments I did

• Automated random clicks
• Automated clicks according to relevance judgments
• Clicks from real people
O12 vs. O14
O12 vs. O14 in BSG ALL

% of Votes Received

Clicks

O12
O14
Tie
Method of Analysis (election)

- Vote by query, by user, by session etc.
- query = person, user = state
## Summary of Results

**Method of Voting**  
O12 vs. O14

<table>
<thead>
<tr>
<th>Method</th>
<th>O12 vs. O14</th>
</tr>
</thead>
<tbody>
<tr>
<td>by queries (direct election)</td>
<td>12 vs. 24</td>
</tr>
<tr>
<td>by users (1 vote per state)</td>
<td>4 vs. 9</td>
</tr>
<tr>
<td>by sessions (~electoral votes)</td>
<td>5 vs. 11</td>
</tr>
</tbody>
</table>

- System does not seem to matter much, but too little clicks (85) to draw significant conclusion
What Logically Follows

• Google Mini vs. O14 (after fixing Google Mini)
• FAST vs. O14 (after fixing RSS in fssearchoffice)
• I’d love to see the results
What can interleaving do?

- Give relevance team more confidence
- Use interleaving for displaying results
- Use interleaving to automatically tune the search engine
Add Confidence

• In addition to very traditional measures like NDCG, Precision and Recall. It is nice to have another independent metric.

• Automatic
  – Does not require human judgments

• Scalable
  – Small impact to UX
What can interleaving do?

• Give relevance team more confidence
• **Use interleaving for displaying results**
• Use interleaving to automatically tune the search engine
Display

Dmitriy Meyerzon

Dmitriy Meyerzon - This is the personal space of Dmitriy Meyerzon. ... This is the personal space of Dmitriy Meyerzon.
http://my/sites/dmitriym - Dmitriy Meyerzon - 6/15/2009

Dmitriy Meyerzon.htm
From: Dmitriy Meyerzon Sent: Tuesday, October 07, 2003 6:59 PM To: Heather Newman (Wadeware LLC) Subject: RE: FY1: PDC Staff & Speakers for SharePoint & FrontPage to LAX, Alaska Airlines ...

RE: Change the ranking weight of a managed property in a FullTextSqlQuery-9.EML
Patrice Manach; Dmitriy Meyerzon; Ismaël René Nzouetom; Office 12 Servers/WSS 3 Tech Discussions ... Dmitriy Meyerzon; Ismaël René Nzouetom; Office 12 Servers/WSS 3 Tech Discussions ... Dmitriy Meyerzon
[View duplicates]
Display

Jimi Hendrix – Discover music, videos, concerts, & pictures at Last.fm
Watch videos & listen to Jimi Hendrix: All Along the Watchtower, Purple Haze & more, plus 210 pictures. James Marshall Hendrix (November 27, 1942 – September ...
www.last.fm/music/jimi+hendrix - Cached - Similar - ☰ ☰ ☰

Image results for jimi hendrix - Report images

Video results for jimi hendrix

Jimi Hendrix - All Along The Watchtower Live ...
4 min 31 sec
www.youtube.com

Jimi Hendrix Purple Haze
2 min 30 sec
www.youtube.com

Jimi Hendrix - Live at Woodstock
57 min
video.google.com

Jimi Hendrix: The Uncut Story - Episode 1 ...
52 min
video.google.com

News results for jimi hendrix

Erotic Madonna tapes, Hendrix contract in NY sale - 2 days ago
By Christine Kearney NEW YORK (Reuters) - Rock legend Jimi Hendrix's first recording contract worth $1 and erotic audio and video tapes sent by Madonna to ...
Reuters - 496 related articles »

Jimi Hendrix hailed as Flying V legend - Gear4music.com - 2 related articles »

A Jimi Hendrix Experience
A Jimi Hendrix Experience, with pictures, lyrics, and music from the greatest guitar player to ever grace the earth.
www.musicfanclubs.org/jimihendrix/ - Cached - Similar - ☰ ☰ ☰
What can we do?

• Give relevance team more confidence
• Use interleaving for displaying results
• **Use interleaving to automatically tune the search engine**
Automatic Tuning

• Many relevance models, each is good for a particular type of corpora (specs, user data, academic articles, product catalog, websites)

• Use interleaving in 10% of searches

• Use user click data to:
  – Automatically and dynamically decide on the best model, or tweak model parameters
Thank you!

• Dmitriy, Eugene, Puneet
• Jamie, Jessica, Ping, Victor, Relevance Team
• Russ, Jon
• Search Team

• Hope to see you again in the future!
Extra Slides
Automatic Tuning – Pair wise?

- Pair wise comparisons scales poorly
- But there seems to be “strong stochastic transitivity”
  - Given locations A, B, C
  - If A > B > C then $\Delta_{AC} > \text{Max}(\Delta_{AB}, \Delta_{BC})$
How to Interleave

• Balanced
• Team Draft
Balanced Interleaving

Algorithm 1 Balanced Interleaving

Input: Rankings $A = (a_1, a_2, \ldots)$ and $B = (b_1, b_2, \ldots)$

$I \leftarrow (); k_a \leftarrow 1; k_b \leftarrow 1;$
$A\text{First} \leftarrow \text{RandBit}() \ldots$ decide which ranking gets priority

while $(k_a \leq |A|) \land (k_b \leq |B|)$ do ...

if $(k_a < k_b) \lor ((k_a = k_b) \land (A\text{First} = 1))$ then

if $A[k_a] \notin I$ then $I \leftarrow I + A[k_a]$ .. append next $A$ result
$k_a \leftarrow k_a + 1$

else

if $B[k_b] \notin I$ then $I \leftarrow I + B[k_b]$ .. append next $B$ result
$k_b \leftarrow k_b + 1$

end if

end while

Output: Interleaved ranking $I$

More formally, denote $A = (a_1, a_2, \ldots)$, $B = (b_1, b_2, \ldots)$, $I = (i_1, i_2, \ldots)$, and let $c_1, c_2, \ldots$ be the ranks of the clicks w.r.t. $I$. To estimate $l$, [13] proposes to use the lowest ranked click, namely $l \approx c_{\text{max}} = \max\{c_1, c_2, \ldots\}$. Furthermore, to derive a preference between $A$ and $B$, one compares the number of clicks in the top

$$k = \min\{j : (i_{c_{\text{max}}} = a_j) \lor (i_{c_{\text{max}}} = b_j)\} \quad (1)$$

results of $A$ and $B$. In particular, the number $h_a$ of clicks attributed to $A$ and the number $h_b$ of clicks attributed to $B$ is computed as

$$h_a = |\{c_j : i_{c_j} \in (a_1, \ldots, a_k)\}| \quad (2)$$
$$h_b = |\{c_j : i_{c_j} \in (b_1, \ldots, b_k)\}|. \quad (3)$$

If $h_a > h_b$ we infer a preference for $A$, if $h_a < h_b$ we infer a preference for $B$, and if $h_a = h_b$ we infer no preference.
Team Draft

1st pick: LeBron James

2nd pick: Kobe Bryant

3rd pick: Tim Duncan

1st pick: John Smith

2nd pick: LeBron James

3rd pick: Tim Duncan
A/B Testing By Interleaving