Evaluating Aggregated Search Using Interleaving

Aleksandr Chuklin - Yandex, University of Amsterdam*
Anne Schuth - University of Amsterdam
Katja Hofmann - University of Amsterdam#
Pavel Serdyukov - Yandex
Maarten de Rijke - University of Amsterdam

* now at Google
# now at Microsoft Research Cambridge

October 29, 2013, CIKM, San Francisco
Evaluating Aggregated Search Using Interleaving
Evaluating Aggregated Search Using Interleaving
Aggregated Search
Aggregated Search

Яндекс

Результаты

1. Angry Birds Chrome
   We're sorry, but it appears that your browser does not currently support the web technologies needed to play Angry Birds. To teach the pigs a lesson
   chrome.angrybirds.com

2. Welcome to Rovio.com! - Rovio Entertainment Ltd
   Rovio is an entertainment media company, and the creator of the globally successful Angry Birds franchise.
   rovio.com -> index.php?page=angry-birds/

3. Angry Birds — Википедия
   Игровой процесс, Серия игр Angry Birds, Интересные факты
   Angry Birds (рус. Сердитые птицы, Злые птицы) — видеоигра-головоломка, разработанная финской компанией Rovio Mobile, в которой игрок с помощью рогатки должен стрелять птицами по зелёным снегом, расставленным на различных конструкциях.
   ru.wikipedia.org -> wiki/Angry_Birds

4. Играй в 36 бесплатные Angry birds игры онлайн, скачай флеш игры Злые...
   У нас самая большая коллекция с ежедневным пополнением онлайн игр Angry birds для всех возрастов. Мы собрали на нашем сайте все самые новые игры Angry birds.
   OnlineGuru.ru -> Angry birds

5. Angry Birds - Android Apps on Google Play
   The survival of the Angry Birds is at stake. Dish out revenge on the greedy pigs who stole their eggs. Use the unique powers of each bird to destroy the pigs' defenses.
   https://play.google.com > store/apps/details?id=com...
Aggregated Search

**Angry Birds** Chrome
chrome.angrybirds.com
Play **Angry Birds** online in HD with Google Chrome. We're sorry, but it appears that your browser does not currently support the web technologies needed to play ...

**Angry Birds** - Home Page
shop.angrybirds.com
**Angry Birds** official online shop. Toys, Games, Apparel, Accessories, Books and Stationery, Specials.

**Angry Birds** - The Official Home of **Angry Birds**
www.angrybirds.com
The Official Home of **Angry Birds** - download games and go shopping

**Images of angry birds**
bing.com/images

**Angry Birds** Toons
www.angrybirds.com/toons
The Official Home of **Angry Birds** Toons - update your games and watch episodes

**Angry Birds** - Wikipedia
nl.wikipedia.org/wiki/Angry_Birds
Ontstaan - Spelverloop - Verschillende uitgaves - Merchandising - Trivia
**Angry Birds** is een computerspel ontwikkeld door het Finse bedrijf Rovio Mobile. De checklist is...</n
Aggregated Search

Federated Search Engine

Angry Birds - Home Page
shop.angrybirds.com
Angry Birds official online shop. Toys, Games, Apparel, Accessories, Books and Stationery, Specials.

Angry Birds - The Official Home of Angry Birds
www.angrybirds.com
The Official Home of Angry Birds - download games and go shopping

Images of angry birds
bing.com/images

Angry Birds Toons
www.angrybirds.com/toons
The Official Home of Angry Birds Toons - update your games and watch episodes

Angry Birds - Wikipedia
nl.wikipedia.org/wiki/Angry_Birds
Ontstaan - Spelverloop - Verschillende uitgaves - Merchandising - Trivia
Angry Birds is een spin-off van spelbirdquel Eline. Bowie Mobile - woods
Aggregated Search

Federated

Vertical
Evaluating Aggregated Search Using Interleaving
Evaluating Aggregated Search Using Interleaving
Why Evaluate Aggregated Search?
Why Evaluate Aggregated Search?

- Search engines evolve
Why Evaluate Aggregated Search?

• Search engines evolve
• Verticals evolve
Why Evaluate Aggregated Search?

- Search engines evolve
- Verticals evolve
- Merging verticals evolves
Why Evaluate Aggregated Search?

- Search engines evolve
- Verticals evolve
- Merging verticals evolves
- Users evolve
Why Evaluate Aggregated Search?

- Search engines evolve
- Verticals evolve
- Merging verticals evolves
- Users evolve
- A large portion of queries is answered by vertical
Why Evaluate Aggregated Search?

- Search engines evolve
- Verticals evolve
- Merging verticals evolves
- Users evolve
- A large portion of queries is answered by vertical
  - ~40% of queries have a SERP with a vertical
Why Evaluate Aggregated Search?

- Search engines evolve
- Verticals evolve
- Merging verticals evolves
- Users evolve
- A large portion of queries is answered by vertical
- ~40% of queries have a SERP with a vertical
- We cannot ignore this
Evaluating Aggregated Search Using Interleaving
Evaluating Aggregated Search Using Interleaving
Interleaving
*(TeamDraft: TDI)*
Interleaving
(TeamDraft: TDI)

System A

System B
Interleaving

*TeamDraft: TDI*
Interleaving
(\textit{TeamDraft: TDI})

System A

A loses

System B

\textit{doc 1}
\textit{doc 3}
\textit{doc 2}
\textit{doc 4}
\textit{doc 5}
\textit{doc 6}
\textit{doc 7}
\textit{doc 8}
\textit{doc 9}
\textit{doc 11}

B wins
Interleaving
Interleaving

• interleaving == evaluation

• $A > B$ or $B > A$ or $B == A$ ?
Interleaving

- interleaving == evaluation
- $A > B$ or $B > A$ or $B == A$ ?
- involves users in the decision
Interleaving

- interleaving == evaluation
- $A > B$ or $B > A$ or $B == A$ ?
- involves users in the decision
- produces relative feedback
- $A > B$
Evaluating Aggregated Search Using Interleaving
Evaluating Aggregated Search Using Interleaving
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Vertical Interleaving

System A

System B
Outline

• Introduction
• Contributions
  • Requirements
  • VA-TDI Algorithm
  • Experimental validation
• Conclusions
Requirements for VA Interleaving
Requirements for VA Interleaving

- Both systems should contribute to the interleaved list equally:
Requirements for VA Interleaving

• Both systems should contribute to the interleaved list equally:

• Their documents
Requirements for VA Interleaving

• Both systems should contribute to the interleaved list equally:
  • Their documents
  • UI aspects:
Requirements for VA Interleaving

• Both systems should contribute to the interleaved list equally:
  • Their documents
  • UI aspects:
    • vertical block size
Requirements for VA Interleaving

- Both systems should contribute to the interleaved list equally:
  - Their documents
  - UI aspects:
    - vertical block size
    - vertical block position
Requirements for VA Interleaving

• Both systems should contribute to the interleaved list equally:
  • Their documents
  • UI aspects:
    • vertical block size
    • vertical block position
  • Decision on winning system should be unbiased
Requirements for VA Interleaving

• Both systems should contribute to the interleaved list equally:
  • Their documents
  • UI aspects:
    • vertical block size
    • vertical block position
• Decision on winning system should be unbiased
• The interleaved system should not degrade user experience compared to the IR systems being interleaved
Outline

- Introduction
- Contributions
- Requirements
- VA-TDI Algorithm
- Experimental validation
- Conclusions
Algorithm for VA Interleaving
Algorithm for VA Interleaving

- We introduce VA-TDI
Algorithm for VA Interleaving

- We introduce VA-TDI
- vertical results are grouped into a single block
Algorithm for VA Interleaving

• We introduce VA-TDI
  • vertical results are grouped into a single block
  • size of vertical block is in between size of A and B
Algorithm for VA Interleaving

• We introduce VA-TDI
  • vertical results are grouped into a single block
  • size of vertical block is in between size of A and B
  • position of vertical block is in between position of A and B
Algorithm for VA Interleaving

- We introduce VA-TDI
  - vertical results are grouped into a single block
  - size of vertical block is in between size of A and B
  - position of vertical block is in between position of A and B
  - vertical block start with highest vertical document
Outline

• Introduction
• Contributions
• Requirements
• VA-TDI Algorithm
• Experimental validation
• Conclusions
Experimental Validation
Experimental Validation

• Q1: Does VA-TDI degrade the user experience?
Experimental Validation

• Q1: Does VA-TDI degrade the user experience?
• Q2: How does VA-TDI compare to A/B-testing?
Experimental Validation

- Q1: Does VA-TDI degrade the user experience?
- Q2: How does VA-TDI compare to A/B-testing?
- Q3: Can VA-TDI capture quality differences between rankings as well as TDI can?
Experimental Validation

• Q1: Does VA-TDI degrade the user experience?
• Q2: How does VA-TDI compare to A/B-testing?
• Q3: Can VA-TDI capture quality differences between rankings as well as TDI can?
• Q4: Is VA-TDI unbiased?
Experimental Validation
Experimental Validation

• Click Log Data
Experimental Validation

• Click Log Data
  • pro: captures actual user behaviour
Experimental Validation

• Click Log Data
  • pro: captures actual user behaviour
  • con: no control over collected data
Experimental Validation

- Click Log Data
  - pro: captures actual user behaviour
  - con: no control over collected data
- Simulated Clicks
Experimental Validation

- **Click Log Data**
  - pro: captures actual user behaviour
  - con: no control over collected data

- **Simulated Clicks**
  - pro: full control over collected data
Experimental Validation

• **Click Log Data**
  - pro: captures actual user behaviour
  - con: no control over collected data

• **Simulated Clicks**
  - pro: full control over collected data
  - con: simple simulated user
Click Log Data
Setup
Click Log Data Setup

- 2 months of Yandex search log
Click Log Data

Setup

- 2 months of Yandex search log
- Filter queries:
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
  - at least 4 impressions with at least a click in the top 10
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
  - at least 4 impressions with at least a click in the top 10
- Assign
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
  - at least 4 impressions with at least a click in the top 10
- Assign
  - System A: most frequent configuration
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
  - at least 4 impressions with at least a click in the top 10
- Assign
  - System A: most frequent configuration
  - System B: configuration different from A at highest rank
Click Log Data Setup

- 2 months of Yandex search log
- Filter queries:
  - one particular vertical in a result page
  - at least 2 configurations
  - at least 4 impressions with at least a click in the top 10
- Assign
  - System A: most frequent configuration
  - System B: configuration different from A at highest rank
- 5755 queries with at least one interleaving of A and B
Click Log Data Experiments
Q1: Does VA-TDI degrade the user experience?
Q1: Does VA-TDI degrade the user experience?

Log Data Experiments

<table>
<thead>
<tr>
<th>Metric</th>
<th>Normalized Metric Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clicks@1</td>
<td>A: 0.9</td>
</tr>
<tr>
<td>MaxRR</td>
<td>A: 0.9</td>
</tr>
<tr>
<td>MeanRR</td>
<td>A: 0.9</td>
</tr>
<tr>
<td>MinRR</td>
<td>A: 0.9</td>
</tr>
</tbody>
</table>
Q1: Does VA-TDI degrade the user experience?

A: No.
Q2: How does VA-TDI compare to A/B-testing?
Q2: How does VA-TDI compare to A/B-testing?
Q2: How does VA-TDI compare to A/B-testing?

<table>
<thead>
<tr>
<th></th>
<th>Clicks@1</th>
<th>MaxRR</th>
<th>MeanRR</th>
<th>MinRR</th>
<th>Interleaving</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA-TDI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/B-testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q2: How does VA-TDI compare to A/B-testing?

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clicks@1</td>
<td>B</td>
</tr>
<tr>
<td>MaxRR</td>
<td>A</td>
</tr>
<tr>
<td>MeanRR</td>
<td>A</td>
</tr>
<tr>
<td>MinRR</td>
<td>A</td>
</tr>
<tr>
<td>Interleaving</td>
<td>B</td>
</tr>
</tbody>
</table>
Q2: How does VA-TDI compare to A/B-testing?

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clicks@1</strong></td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td><strong>MaxRR</strong></td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td><strong>MeanRR</strong></td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td><strong>MinRR</strong></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td><strong>Interleaving</strong></td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
Q2: How does VA-TDI compare to A/B-testing?

A: VA-TDI is reasonably correlated with conventional absolute click metrics.

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clicks@1</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>MaxRR</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>MeanRR</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>MinRR</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Interleaving</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
Simulated Clicks
Setup
Simulated Clicks

Setup

- Generate 500 pairs A and B of synthetic rankings
Simulated Clicks

Setup

- Generate 500 pairs A and B of synthetic rankings
  - start with a set of documents, randomly assign relevance
Simulated Clicks
Setup

• Generate 500 pairs A and B of synthetic rankings
  • start with a set of documents, randomly assign relevance
  • until A dominates B
Simulated Clicks

Setup

- Generate 500 pairs A and B of synthetic rankings
  - start with a set of documents, randomly assign relevance
  - until A dominates B
  - take random permutations of documents A and B
Simulated Clicks

Setup

• Generate 500 pairs A and B of synthetic rankings
  • start with a set of documents, randomly assign relevance
  • until A dominates B
    • take random permutations of documents A and B
    • insert vertical block in both
Simulated Clicks

Setup

- Generate 500 pairs A and B of synthetic rankings
  - start with a set of documents, randomly assign relevance
  - until A dominates B
    - take random permutations of documents A and B
    - insert vertical block in both
- Simulate clicks
Simulated Clicks
Setup

- Generate 500 pairs A and B of synthetic rankings
  - start with a set of documents, randomly assign relevance
  - until A dominates B
    - take random permutations of documents A and B
    - insert vertical block in both
- Simulate clicks
  - FCM: Federated click model
Simulated Clicks
Setup

• Generate 500 pairs A and B of synthetic rankings
  • start with a set of documents, randomly assign relevance
  • until A dominates B
    • take random permutations of documents A and B
    • insert vertical block in both
• Simulate clicks
  • FCM: Federated click model
    • users are attracted by vertical
Simulated Clicks

Setup

• Generate 500 pairs A and B of synthetic rankings
  • start with a set of documents, randomly assign relevance
  • until A dominates B
    • take random permutations of documents A and B
    • insert vertical block in both

• Simulate clicks
  • FCM: Federated click model
    • users are attracted by vertical
  • RCM: Random click model
Simulated Clicks

Setup

- Generate 500 pairs A and B of synthetic rankings
  - start with a set of documents, randomly assign relevance
  - until A dominates B
    - take random permutations of documents A and B
    - insert vertical block in both
- Simulate clicks
  - FCM: Federated click model
    - users are attracted by vertical
  - RCM: Random click model
    - half the documents are clicked
Simulated Clicks
Experiments
Q1: Does VA-TDI degrade the user experience?
Q1: Does VA-TDI degrade the user experience?
Q1: Does VA-TDI degrade the user experience?

A: No. TDI does.

The graph shows the number of blocks against the original block size for both VA-TDI and TDI. The number of blocks increases with the original block size for both techniques, but the bars for TDI are generally higher, indicating a greater increase in blocks for TDI compared to VA-TDI.
Simulated Clicks Experiments
Q3: Can VA-TDI capture quality differences between rankings as well as TDI can?
Q3: Can VA-TDI capture quality differences between rankings as well as TDI can?
Q3: Can VA-TDI capture quality differences between rankings as well as TDI can?

A: VA-TDI can compare rankings as accurately as TDI.
Simulated Clicks
Experiments
Simulated Clicks

Experiments

Q4: Is VA-TDI unbiased?
Q4: Is VA-TDI unbiased?

- Using a Random Click Model

- We should not detect more than the expected amount of differences

- The expected amount is 5%
Simulated Clicks Experiments

Q4: Is VA-TDI unbiased?
Q4: Is VA-TDI unbiased?

<table>
<thead>
<tr>
<th></th>
<th>VA-TDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4.40%</td>
</tr>
<tr>
<td>200</td>
<td>3.20%</td>
</tr>
<tr>
<td>300</td>
<td>3.80%</td>
</tr>
<tr>
<td>400</td>
<td>2.60%</td>
</tr>
<tr>
<td>500</td>
<td>4.20%</td>
</tr>
</tbody>
</table>
Q4: Is VA-TDI unbiased?

<table>
<thead>
<tr>
<th></th>
<th>VA-TDI</th>
<th>TDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4.40%</td>
<td>5.80%</td>
</tr>
<tr>
<td>200</td>
<td>3.20%</td>
<td>4.40%</td>
</tr>
<tr>
<td>300</td>
<td>3.80%</td>
<td>4.00%</td>
</tr>
<tr>
<td>400</td>
<td>2.60%</td>
<td>6.80%</td>
</tr>
<tr>
<td>500</td>
<td>4.20%</td>
<td>5.80%</td>
</tr>
</tbody>
</table>
Q4: Is VA-TDI unbiased?

A: Indeed. VA-TDI, like TDI, is unbiased under random clicks.

<table>
<thead>
<tr>
<th></th>
<th>VA-TDI</th>
<th>TDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4.40%</td>
<td>5.80%</td>
</tr>
<tr>
<td>200</td>
<td>3.20%</td>
<td>4.40%</td>
</tr>
<tr>
<td>300</td>
<td>3.80%</td>
<td>4.00%</td>
</tr>
<tr>
<td>400</td>
<td>2.60%</td>
<td>6.80%</td>
</tr>
<tr>
<td>500</td>
<td>4.20%</td>
<td>5.80%</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Contributions
- Requirements
- VA-TDI Algorithm
- Experimental validation
- Conclusions
Conclusions
Conclusions

• We introduced VA-TDI
Conclusions

• We introduced VA-TDI
• VA-TDI does not hurt user experience
  • Performance in between A and B
  • No broken blocks
Conclusions

• We introduced VA-TDI
• VA-TDI does not hurt user experience
  • Performance in between A and B
  • No broken blocks
• VA-TDI reliably detects differences
  • High correlation with absolute click metrics
  • Same accuracy as TDI
Conclusions

- We introduced VA-TDI
- VA-TDI does not hurt user experience
  - Performance in between A and B
  - No broken blocks
- VA-TDI reliably detects differences
  - High correlation with absolute click metrics
  - Same accuracy as TDI
- VA-TDI is unbiased
  - No preference detected under random clicks
thank you

download Lerot from
https://bitbucket.org/ilps/lerot