Auditing Search Engines for Differential Satisfaction across Demographics

Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, Emine Yilmaz

University College London
Microsoft Research New York
Fairness across demographics

• Online services - advertised as being available to any user

• Ethical
  • Equal access to everyone

• Practical
  • Equal access helps attract a large and diverse population of users
  • Service providers are scrutinized for seemingly unfair behavior [1,2,3]

• Onus on us
  • develop fair systems

Auditing services for fairness

We offer methods for **auditing a system’s** performance for detection of differences in user satisfaction across demographics.
From public libraries to search engines

- Modern analogue of public libraries
- Dominant role in information access
- Fairness in *performance*!
Are Search Engines Fair?
From public libraries to search engines

Search Engines:
- Rely on ML models to optimize for user satisfaction
- Make use of implicit signals
- **Metric** driven development

... not easy to audit
Tricky: straightforward optimization can lead to differential performance

**Goal:** estimate difference in user satisfaction between two demographic groups.

- Search engine uses a standard metric: **time spent** on clicked result page as an indicator of satisfaction.

- Suppose older users issue more of **“retirement planning”** queries

<table>
<thead>
<tr>
<th>Age: &lt;30 years</th>
<th>Age: &gt;50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% users</td>
<td>10% users</td>
</tr>
</tbody>
</table>
1. Aggregate Metrics can be misleading

- Overall metrics can hide differential satisfaction

- **Average user satisfaction for “retirement planning” may be high.**

But,

- Average satisfaction for younger users=0.7
- Average satisfaction for older users=0.2
2. Query-level metrics can hide differential satisfaction

Assuming same user satisfaction for “retirement planning” for both older and younger users = 0.7

What if average satisfaction for <query-X> = 0.9? (e.g. <query-X> = “facebook”)  

Older users still receive more of lower-quality results than younger users.
3. More critically, even individual-level metrics can also hide differential satisfaction

**Metric itself could be confounded with demographics**

**Consider:** Reading time for the same webpage result for the same user satisfaction

Younger Users → Time spent on a webpage

Older Users → Time spent on a webpage
We must control for natural demographic variation to meaningfully audit for differential satisfaction.
Outline

1 Background

2 Data & metrics

3 Proposed approaches:
   1 Context Matching
   2 Hierarchical Multi-level model

4 From metrics to satisfaction

5 Discussion
Data: Demographic characteristics of search engine users

- Internal logs from Bing.com for two weeks
- 4 M users | 32 M impressions | 17 M sessions
- Demographics: Age & Gender
- Age:
  - post-Millenial: <18
  - Millenial: 18-34
  - Generation X: 35-54
  - Baby Boomer: 55-74

... also perform external auditing using comScore data
Metrics Considered

1. Graded Utility (GU)
   • based on search outcome and user effort

2. Reformulation Rate (RR)
   • fraction of queries that were reformulated

3. Successful Click Count (SCC)
   • clicks with significant dwell times

4. Page Click Counts (PCC)
   • total no of clicks on SERP

Goal: estimate difference in user satisfaction between demographic groups

Obvious solution: demographic binning!
Overall metrics across Demographics

- Substantial differences in performance across age
- Gender – not so much

... how true are these?
Pitfalls with Overall Metrics

Conflates two separate effects:

• natural demographic variation caused by the differing traits among the different demographic groups e.g.
  • Different queries issued
  • Different information need for the same query
  • Even for the same satisfaction, demographic A tends to click more than demographic B

• Systemic difference in user satisfaction due to the search engine

... we need to disentangle them!
Utilize work from causal inference
Utilize work from causal inference
Utilize work from causal inference
Utilize work from causal inference
Outline

1 Motivation
2 Problems with naïve auditing
3 Data & Metrics
4 Proposed approaches:
   1 Context Matching
   2 Hierarchical Multi-level model
5 From metrics to satisfaction
6 Discussion
Proposed Approaches

1) Context Matching
2) Multi-level model

Extremely restrictive
More robust

Generalizable
Less Robust
I. Context Matching: selecting for activity with near-identical context

For any two users from different demographics,

1. **Same Query**
2. **Same Information Need:**
   1. Control for user intent: same final SAT click
   2. Only consider navigational queries
3. **Identical top-8 Search Results**

1.2 M impressions
19K unique queries
617K users
Age-wise differences in metrics disappear

- General auditing tool: robust
- Very low coverage across queries
  - Did we control for too much? – lose over 60% of data!
Proposed Approaches

1) Context Matching
   Extremely restrictive
   More robust

2) Multi-level model
   Generalizable
   Less Robust
Query-level Multilevel Model

• A **hierarchical** approach that treats the data as a mixture of distributions based on demographics and queries

• Non-nested **multi-level** model
  • Users & Queries: nested within **non-nested** age and gender groups & topics
  • second level captures variation with individual query properties

• Age effects
• Gender effects
• Topic effects
• <age, gender, topic> interaction effects

$$E(Y) = f^{-1}(\alpha_{agt} + \beta_{agt}X)$$

$$\begin{pmatrix} \alpha_{agt} \\ \beta_{agt} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} + \alpha_a + \alpha_g + \alpha_t + \alpha_{a \times g \times t} + \beta_a + \beta_g + \beta_t + \beta_{a \times g \times t}$$

$$\begin{pmatrix} \alpha_k \\ \beta_k \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_k \right) \quad k \in \{a, g, t\}$$

Specific example: $$GU_i \sim \mathcal{N}(\alpha_{agt} + \beta_{agt}X_i, \sigma_y^2)$$
Age-wise differences appear again: bigger differences for harder queries
Outline

1  Motivation
2  Problems with naïve auditing
3  Data & Metrics
4  Proposed approaches:
   1  Context Matching
   2  Hierarchical Multi-level model
5  From metrics to satisfaction
6  Discussion
From Metric to Satisfaction

• Estimating absolute satisfaction is non-trivial

• We estimate relative satisfaction by considering pairs of impressions:
  • which impression led to a higher satisfaction

• Construct a conservative “high-precision, low-recall” proxy for pairwise satisfaction
  • by only considering “big” differences in observed metric for the same query

• Logistic regression model for estimating probability of impression i being more satisfied than impression j:

\[ P(S_i > S_j) = \logit^{-1}(\beta_0 + \beta_{a_i} a_i + \beta_{a_j} a_j + \beta_{g_i} g_i + \beta_{g_j} g_j + \beta_{i,j} a_i a_j g_i g_j) \]
Again, see a small age-wise difference in satisfaction

<table>
<thead>
<tr>
<th>Age i</th>
<th>Age j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>0.56</td>
</tr>
</tbody>
</table>

- Older users are slightly more satisfied than younger users

\[ P(S_i \succ S_j) \]
Discussion

- Auditing is more nuanced than merely measuring metrics on demographically-binned traffic
  - developed techniques to auditing search engines
- We find light trend towards older users being more satisfied.
- General framework for internally auditing systems
  - Plug-in different metrics
  - Plug-in different demographics/user groups

Future Work

- Develop metrics which are not confounded with demographics
- Investigate causes of metric differences
  - Query level analysis
  - SERP level analysis
- Dwell time thresholds for SAT prediction based on demographic information
Auditing is more nuanced than merely measuring metrics on demographically-binned traffic.

General framework for auditing systems
- Plug-in different metrics
- Plug-in different demographics/user groups

Thank You!

Rishabh Mehrotra
PhD candidate @ UCL
http://www.rishabhmehrotra.com

@erishabh
r.mehrotra@cs.ucl.ac.uk
## Future Work

<table>
<thead>
<tr>
<th>Query</th>
<th>Demographics</th>
<th>Metric Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>essential oils guide</td>
<td>Female Age 2</td>
<td>vs</td>
</tr>
<tr>
<td>make your own game</td>
<td>male3</td>
<td>vs</td>
</tr>
<tr>
<td>macbook pro vs macbook air</td>
<td>Female2</td>
<td>vs</td>
</tr>
<tr>
<td>editing software for youtube videos</td>
<td>Male2</td>
<td>vs</td>
</tr>
<tr>
<td>emotions</td>
<td>Male2</td>
<td>vs</td>
</tr>
<tr>
<td>avaya phone manual</td>
<td>Female3</td>
<td>vs</td>
</tr>
<tr>
<td>catholic saints</td>
<td>Male4</td>
<td>vs</td>
</tr>
<tr>
<td>futures market</td>
<td>Male3</td>
<td>vs</td>
</tr>
<tr>
<td>medal of honor walkthrough ps3</td>
<td>Male3</td>
<td>vs</td>
</tr>
<tr>
<td>all wheel drive cars</td>
<td>Male4</td>
<td>vs</td>
</tr>
<tr>
<td>kob tv albuquerque news 4</td>
<td>Female4</td>
<td>vs</td>
</tr>
<tr>
<td>foods high in iron</td>
<td>Female3</td>
<td>vs</td>
</tr>
<tr>
<td>478-288-1122</td>
<td>Male3</td>
<td>vs</td>
</tr>
<tr>
<td>cheeseburger dip</td>
<td>Female4</td>
<td>vs</td>
</tr>
<tr>
<td>argosy capital</td>
<td>Male3</td>
<td>vs</td>
</tr>
</tbody>
</table>
External Auditing

- Experiment on a publicly available dataset
- 2 weeks logs of comScore data
- Use PCC metric to gauge satisfaction
- Probability of impression $i$ being more satisfied than impression $j$:

$$P(S_i > S_j) = \logit^{-1}(\beta_0 + \beta_{a_i}a_i + \beta_{a_j}a_j + \beta_{g_i}g_i + \beta_{g_j}g_j + \beta_{ij}a_i a_j g_i g_j)$$
Future Work
Demographic distribution of user activity

- **fraction of users**
  - Female
  - Male

- **query frequency**
  - Female
  - Male

- **Age Groups**
  - 1
  - 2
  - 3
  - 4

- **Query Type**
  - Navigational
  - Informational

- **Body Parts**
  - Head
  - Torso
  - Tail
Characterizing Demographics: Gender

No of Users

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>5000000</td>
</tr>
</tbody>
</table>

Avg Session Length (no of Impressions)

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3</td>
</tr>
</tbody>
</table>

Avg No of Characters Per Query

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

% Head Queries

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Avg No of Sessions Per User

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3</td>
</tr>
</tbody>
</table>

Avg No of Words Per Query

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3</td>
</tr>
</tbody>
</table>

% Nav Queries

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3</td>
<td></td>
</tr>
</tbody>
</table>

% Tail Queries

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Some highly discriminating queries in terms of P(D|Q):

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>premier league</td>
<td>pinterest</td>
</tr>
<tr>
<td>bbc football</td>
<td>hautelook</td>
</tr>
<tr>
<td>watchespn</td>
<td>weight watchers</td>
</tr>
<tr>
<td>pirate bay</td>
<td>sephora</td>
</tr>
</tbody>
</table>
External Auditing

- Experiment on a publicly available dataset
- 2 weeks logs of comScore data
- Use PCC metric to gauge satisfaction
- Probability of impression $i$ being more satisfied than impression $j$:

\[
P(S_i \succ S_j) = \logit^{-1}(\beta_0 + \beta_{a_i} a_i + \beta_{a_j} a_j + \beta_{g_i} g_i + \beta_{g_j} g_j + \beta_{ij} a_i a_j g_i g_j)
\]
Characterizing Demographics:

Some highly discriminating queries in terms of $P(D|Q)$:
• Young user, Old user
• Issue same query
• See search results
• How satisfied are you?
Query level Difficulty

- $X_i$: Feature corresponding to inherent difficulty of query
- Typical methods (reformulations, dwell times) employ user behavior – correlated with demographics
- Need a measure unconfounded with demographics
- Method:
  - Per demographic order query by increasing order of avg GU score
  - Compute per demographic percentile of the query (~query’s difficulty in each demographic)
  - Mean of percentiles across demographics
Algorithm 1 Compute satisfaction label

1: if $RR_i < RR_j$ then return +1
2: if $RR_i > RR_j$ then return -1
3: if $GU_i - GU_j > \delta^1_{GU}$ then return +1
4: if $GU_j - GU_i > \delta^1_{GU}$ then return -1
5: if $SCC_i - SCC_j > \delta^1_{SCC}$ then return +1
6: if $SCC_j - SCC_i > \delta^1_{SCC}$ then return -1
7: if $GU_i - GU_j > \delta^2_{GU} \land SCC_i - SCC_j > \delta^2_{SCC}$ then return +1
8: if $GU_j - GU_i > \delta^2_{GU} \land SCC_j - SCC_i > \delta^2_{SCC}$ then return -1
9: else return 0