Measuring personalization of Web Search

AUTHOR: ANIKO HANNAK
PRESENTED BY BOYU RAN
1/15/15
Motivation for personalization of web search

- Users usually view only the first few pages of the search results
- Problem: Relevant results beyond first few pages have a much lower chance of being visited
- Personalization aims to:
  - Tailor search results to individuals based on knowledge of their interests, rather than from web popularity
  - Identify most relevant information and put them on top of the search list.
  - Filter irrelevant information
For mobile clients
- Smaller space for display search results
- Input modes inherently limited
- Users likely to view fewer search results
- Thus relevance is crucial
Problems

- Filter Bubble Effect
- Alternative search engine
  - Duckduckgo.com
- Little scientific quantification of the basis and extent of search personalization
Google Search
- First introduced personalized search in 2004
- Merged this product into Google Search in 2005
- In 2009, Google began personalizing search results for all users
- Recently, Google started including personalized contents from Google+ social network into search results
Main Goals

- Develop a methodology for measuring personalization in web search results
- Use this methodology to measure the extent of personalization on Google Web Search today
- Investigate the causes of personalization
Seek to answer two broad questions:
- What user features influence Google’s search personalization algorithms?
- To what extent does search personalization actually affect search results?

High-level design
- Create x Google Accounts that each vary by one specific feature, execute identical queries from each account, once per day for d days.
- Save the results of each query
- Finally, compare the results of the queries to determine whether the same results are being served in the same order to each account.
- If the results vary, then the changes can be attributed to personalization linked to the given experimental feature
Source of Noise

- Updates to the Search Index
- Distributed Infrastructure
- Geolocation
- A/B Testing
- Carry-over effect
Controlling against the noise

- All the queries are performed on Google Search Webpage, rather than Search API
- All machines execute searches for the same query at the same time
  - Eliminate temporal effects
- Use static DNS entries to direct traffic to a specific Google IP address
  - Eliminate difference between datacenter
- Wait 11 minutes to avoid carry-over effects
- Send queries from the same /24 subnet
  - Geolocation
- Include a control account
Implementation

- PhantomJS
  - Executes JavaScript, manage cookies
  - More scalable
Implementation

- Search Queries
  - Breadth and impact
  - 120 queries, 12 categories, chosen from Google Zeitgeist, WebMD

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tech</td>
<td>Gadgets, Home Appliances</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Politics, News Sources</td>
<td>20</td>
</tr>
<tr>
<td>News</td>
<td>Apparel Brands, Travel Destinations,</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Home and Garden</td>
<td></td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Quirky</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Weird Environmental, What-Is?</td>
<td></td>
</tr>
<tr>
<td>Quirky</td>
<td>Humanities</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td></td>
</tr>
<tr>
<td>Humanities</td>
<td>Science</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Health, Environment</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>120</td>
</tr>
</tbody>
</table>

Table 1: Categories of search queries used in our experiments.
Collecting real-world data

Amazon’s Mechanical Turk (AMT)

Participants were required
  - To be in the US
  - Have Google account

Users were instructed to configure web browser to use an HTTP proxy

then users were directed to visit a Web page that automatically performed 80 Google searches
How often do real users receive personalized search results?

Figure 5: % of AMT and control results changed at each rank.
Most personalized queries:
Politics, companies

Least personalized queries:
Factual and health related queries

Table 2: Top 10 most/least personalized queries.
We now turn to examining which user features Google Search use to personalize results.

Two metrics for measuring personalization

- Jaccard Index
- Edit Distance
  - \([a.com,b.com,c.com]\)
  - \([c.com,b.com]\)
  - 2 (one insertion and one swap)
<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Tested Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracking</td>
<td>Cookies</td>
<td>Logged In, Logged Out, No Cookies</td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Win. XP, Win. 7, OS X, Linux</td>
</tr>
<tr>
<td>User-Agent</td>
<td>Browser</td>
<td>Chrome 22, Firefox 15, IE 6, IE 8, Safari 5</td>
</tr>
<tr>
<td>Geolocation</td>
<td>IP Address</td>
<td>MA, PA, IL, WA, CA, UT, NC, NY, OR, GA</td>
</tr>
<tr>
<td>Google Account</td>
<td>Gender</td>
<td>Male, Female, Other</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>15, 25, 35, 45, 55, 65</td>
</tr>
<tr>
<td>Search History,</td>
<td>Gender</td>
<td>Male, Female</td>
</tr>
<tr>
<td>Click History,</td>
<td>Age</td>
<td>&lt;18, 18-24, 25-34, 35-44, 45-54, 55-64, ≥65</td>
</tr>
<tr>
<td>and Browsing History</td>
<td>Income</td>
<td>$50-100K, $50-100K, $100-150K, &gt;$150K</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>No College, College, Grad School</td>
</tr>
<tr>
<td></td>
<td>Ethnicity</td>
<td>Caucasian, African American, Asian, Hispanic</td>
</tr>
</tbody>
</table>

Table 3: User features evaluated for effects on search personalization.
Basic Cookie Tracking

Figure 6: Results for the cookie tracking experiments.
Browser experiment

Figure 7: Results for the browser experiments.
Figure 8: Results for the geolocation experiments.
Figure 10: Results for the Google Profile: Gender experiments.
Three types of historical activity

- Search history
- Search-result-click history
- Browsing history
Figure 11: Results for the search history: income level experiments.
There is little difference between the controls and the test accounts.

Thus we do not observe personalization based on click history.
Browsing History

- Same as previous two
- No personalization based on browsing history

Discussion:
Temporal dynamics

Figure 12: Day-to-day consistency of results for the geolocation experiments.
Dynamics of Query Categories

Figure 13: Day-to-day consistency within search query categories for the geolocation test.
Figure 14: Differences in search results for five query categories.
Future Work

- Scope
- Incompleteness
- Generality
- Impact