Quantifying Bias in Search Engines and beyond

Prof.dr.ir. Arjen P. de Vries
arjen@acm.org
PI Lab Quarterly Event
Nijmegen, June 29th, 2018
Outline

- Bias on the Web
- Bias in Search Engines
  - Quantifying Bias using Retrievability
- Discussion
  - Role PI.Lab?
Any remedy for bias starts with awareness of its existence.

Bias on the Web reflects biases within ourselves, manifested in subtler ways.

We must consider and account for bias in the design of Web-based systems that truly address the needs of users.
Bias on the Web – Highlights

- **Activity Bias**
  - In many collections, a few contributors create half the data:
    - Two Twitter datasets: 0.05% of the most popular people attracted almost 50% of the participants in the 1st and 2% of the users generated 50% of the tweets in the 2nd dataset.
    - Facebook dataset: 7% of active users produced 50% of the posts.
    - Amazon reviews dataset: 4% of active users produced 50% of reviews (*fraud?*)
    - Wikipedia: 2000 people (0.04% of registered editors) wrote half of the entries of English Wikipedia.

- A “Digital Desert” of Web content no-one ever sees.
Bias on the Web – Highlights

- Data Bias
  - 50% of popular web sites are in English…
    … but percentage of native English speakers is only 5%, all English speakers 13% (Disclaimer: stated source is Wikipedia.)

- Own example:
  POI recommendation using geo-tagged Flickr photos: everyone visiting Paris visits the Eiffel Tower…

- Consequence: optimizing for a data driven performance measure like accuracy forces the approach to recommend the Eiffel Tower even if this is a useless recommendation!
Bias on the Web – Highlights

▪ UI/UX influence on collected data:
  ▪ Presentation Bias
    ▪ You can only click on what you see!
  ▪ Position Bias
    ▪ You only see where you look!

▪ Click distributions only useful when de-biased!

▪ Second-order bias originating in personalization: biasing content to our own pre-existing selection biases?
Question:

- How to quantify the ranking bias in search engines?
Retrievability

- Measures bias in the access of documents in a collection

Relationship bias and effectiveness?

- Prior work indicates that high effectiveness (measured using P@10 and MAP) correlates with low “retrievability bias”

Retrievability

- Measure the accessibility of all documents in a collection given a set of queries
  - Retrievability score $r(d)$ measures how often document $d$ is retrieved by a given set of queries
Retrievability

- Measure the accessibility of all documents in a collection given a set of queries
  - Retrievability score \( r(d) \) measures how often document \( d \) is retrieved by a given set of queries

- “Evaluate” a retrieval method by the distribution of retrievability scores generated:
  - Lorenz curve visualizes the inequality
  - Gini coefficient quantifies the inequality
Lorenz curve for n=5

- 0, 0, 0, 0, 1
- 1, 1, 1, 1, 1
- 0, 0, 1, 1, 2

% of accumulated \( r(d) \)

% of documents
Retrievability & Ranked Retrieval

- Every document is always retrieved!
Retrievability & Ranked Retrieval

- Every document is always retrieved!

- Consider a cut-off, that corresponds intuitively to the amount of effort we expect a user to be willing to invest in learning the answer to their query
  - E.g., cut-off of 10:
    retrievability score equals the number of times this document is retrieved in the top 10 of the results for each of the queries in the (given) query set
“... by a given set of queries”

Q: How do we get the set of queries for which to measure retrievability?

Note:
Choice of query set matters, see:
Simulation

- Original study:
  - Draw at random a large set of single terms and bigrams (from the documents in the collection)
  - Inspired by “Query Based Sampling” for resource description in non-cooperative federated search

- Applied in our study on the Dutch Web Archive:
  - 2M most frequent terms (frequencies ranging from 5 to 200M)
  - 2M most frequent bigrams (frequencies ranging from 20 to 35M)
Alternative “Query” Set

- Our study, on the Dutch Web Archive:
  - Anchor Text from external links
  - De-duplicated for year of crawl
    *(Most sites crawled once a year, but a subset more frequently.)*

- Side note:
Retrievability Score Inequality ("All")

Lorenz curve

Cumulative Normalized $r(d)$

Documents Ordered by $r(d)$

$Q_s, c = 10$
Zero Retrievability

- Original approach:
  - If a document is never retrieved by any query, $r(d) = 0$

- Alternative (“union”):
  - Consider only the documents that are retrieved by at least one of the methods under consideration
    - Fair comparison across retrieval methods for a given cut-off
    - Reduces the impact of the high number of documents that has a retrievability score of 0
Retrievability Score Inequality ("Union")
Retrievability = Findability?

- Divide collection in 4 equi-sized bins
  - Using wealth distribution (area under the curve)
- Can a document be found if the user queries the collection using a query sampled from that document?
  - Need to take care to create a reasonable artificial known-item query…
- Documents that are “highly retrievable” are significantly easier to find (using MRR to evaluate effectiveness)
Conclusions

- Analysis of retrievability helps understand the behaviour of retrieval systems, and makes explicit the inherent biases that affect the retrieved results.

- Knowing which documents are particularly hard to find allows the institutions to improve their retrieval systems and the users to adapt their search strategies and be aware of the retrieval bias and the source of that bias.
Quantifying retrieval bias in Web archive search

Thaer Samar, Myriam C. Traub, Jacco van Ossenbruggen, Lynda Hardman & Arjen P. de Vries

http://dx.doi.org/10.1007/s00799-017-0215-9
Quantifying Bias is not Enough!

- Retrievability compares ranking algorithms, so it can quantify bias – we can choose the least biased algorithm, but cannot reduce the bias observed

- Very recent work by Singh and Joachims *corrects* bias in the exposure of items (due to presentation bias) associated to user groups; to ensure a *fair* exposure of those groups
Discussion

▪ How to determine the impact of the bias inherent in any sample of the Web that you are exposed to ("you" being a person or an algorithm)?
Legal Context

GDPR recital 71: the data controller should take measures to “[prevent] discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation ...”
Discussion I: Awareness

- Bias on the Dutch Web / Social Media
  - Case studies on CommonCrawl and/or KB Web Archive

- Methods and techniques
  - Explore and extend current / develop new techniques
  - Tools for the average Web user

- My research: *Personal* Web Archives, to help you understand and reflect on biases you are exposed to
Figure 2: Search interface for WASP: ① shortcuts for usual time settings; ② selected query time interval; ③ date and time picker; ④ query box; ⑤ description of current result page; ⑥ title of result with links to archived and live version; ⑦ URL of the result; ⑧ archive time of the result; ⑨ snippet for the result.

Figure 3: Screenshot of a web page reproduced from the archive. A small black banner is inserted at the bottom right of the browser viewport to remind users that they are viewing an archived page.
Discussion II: Exploitation vs. Exploration

- Measures that do not discourage exploration
  - Countering the Popularity Bias

- Can we somehow quantify future rewards, not just current rewards (?)

- Upcoming SIGIR paper:
  - Should I Follow the Crowd? A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems
    ir.ii.uam.es/pubs/sigir2018.pdf
  - Properties of measures informing better experimental designs
Discussion III: Explanations

- Insight in how classifiers etc. decide – not trivial for tools like deep networks over, say, character sequences

- Challenge:
  - How to create a Ground Truth?

*When is an explanation of an algorithmic decision a good explanation?*
“More importantly, a good explanation method should not reflect what humans attend to, but what task methods attend to. For instance, the family name “Kolstad” has 11 out of its 13 appearances in the 20 newsgroups corpus in sci.electronic posts. Thus, task methods probably learn it as a sci.electronics indicator. Indeed, the explanation method in Fig 1 (top) marks “Kolstad” as relevant, but the human annotator does not.”
Discussion IV: De-biasing data?

- Compare measures on social media or web data to those from more traditional data (e.g., demographics, CBS)

- Integrate observations from different sources; e.g., use population density as well as geo-located Tweet frequency to counter geographical bias

- De-bias representations derived from biased data (?)
  - E.g., word embeddings (e.g. Bolukbasi et al. “Man is to computer as woman is to homemaker?”)

See also:

- Transparency and accountability of algorithms:
  - **Awareness**
  - Access and redress
  - Accountability
  - **Explanation**
  - Data provenance
  - Auditability
  - **Validation and testing**

- [fairness-measures.org/](http://fairness-measures.org/)