Djoerd Hiemstra
http://www.cs.utwente.nl/~hiemstra/
HOW DO MACHINES DETERMINE RELEVANCE ...
STATISTICS
IT DEPENDS...

- **Ad hoc**: Language models, BM25
- **Spam filter**: Naive Bayes
- **Network data**: PageRank
- **Clicks**: Learning to rank
- ...

“*It Depends*”
-Socrates
THERE IS NO “ONE SIZE FITS ALL”!

- **Web**: Web graph, anchor text
- **Videos**: Views, likes, content-based features
- **Advertisements**: Bids, click-through-rate
- **Tweets**: Retweets, likes,
- **Scientific papers**: Citations
- **Restaurants**: Geo, reviews
- **Products**: Price, nr. in stock
- …
WE NEED SEARCH SPECIALIZATION!

- CLEF eHealth
- ImageCLEF
- LifeCLEF
- Uncovering Plagiarism
- Social Book Search
- News Recommendation
- Living Labs (products, papers)
- ...

UNIVERSITY OF TWENTE.
BIG DATA FALLACY?

- If we have all data, we can learn one model that...

... participates in every CLEF lab?
SEARCH SPECIALIZATION & DELEGATION
http://search.utwente.nl
QUERY-BASED SAMPLING

“Query-based sampling is a (...) method of acquiring resource descriptions that does not require explicit cooperation from resource providers. Instead, resource descriptions are created by running queries and examining the documents that are returned.”

Jamie Callan and Margaret Connell. Query-Based Sampling of Text Databases. ACM Transactions on Information Systems 19(2), 2001
SYSTEM IDEA

- Fresh results for “alpha”
- Fresh results for “beta”
- Fresh results for “gamma”
- Fresh results for “delta”
- Fresh results for “epsilon”
SYSTEM IDEA

“CLEF”
Sample results for “CLEF”
Forward “CLEF” to Wikipedia
Fresh results for “CLEF”

“CLEF”
Fresh results for “CLEF”
DO SAMPLES RESEMBLE THE FULL INDEX?

**TREC123**

**WT2G**

- **Jensen Shannon Divergence (JSD)**
- **Bandwidth Usage (Kilobytes)**

Legend:
- Full Documents
- Snippets
DO SAMPLES RESEMBLE THE FULL INDEX?

TREC “FEDWEB” TRACK

- Large-scale federated search evaluation:
  - Resource selection
  - Result merging
  - Vertical selection
Table 1: Categorization resources

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>General web search</td>
<td>10</td>
<td>Google, Yahoo, AOL, Bing, Baidu</td>
</tr>
<tr>
<td>Multimedia</td>
<td>21</td>
<td>Hulu, YouTube, Photobucket</td>
</tr>
<tr>
<td>Q &amp; A</td>
<td>2</td>
<td>Yahoo Answers, Answers.com</td>
</tr>
<tr>
<td>Jobs</td>
<td>7</td>
<td>LinkedIn Jobs, Simply Hired</td>
</tr>
<tr>
<td>Academic</td>
<td>16</td>
<td>Nature, CiteSeerX, SpringerLink</td>
</tr>
<tr>
<td>News</td>
<td>8</td>
<td>Google News, ESPN</td>
</tr>
<tr>
<td>Shopping</td>
<td>6</td>
<td>Amazon, eBay, Discovery Channel Store</td>
</tr>
<tr>
<td>Encyclopedia/Dict</td>
<td>6</td>
<td>Wikipedia, Encyclopedia Britannica</td>
</tr>
<tr>
<td>Books &amp; Libraries</td>
<td>3</td>
<td>Google Books, Columbus Library</td>
</tr>
<tr>
<td>Social &amp; Social Sharing</td>
<td>7</td>
<td>Facebook, MySpace, Tumblr, Twitter</td>
</tr>
<tr>
<td>Blogs</td>
<td>5</td>
<td>Google Blogs, WordPress</td>
</tr>
<tr>
<td>Other</td>
<td>17</td>
<td>OER Commons, MSDN, Starbucks</td>
</tr>
</tbody>
</table>
Resource Selection Results (44 runs)
Resource Selection Results: use of sampled pages only

nDCG@20
Resource Selection Results: use of sampled snippets only
Resource Selection Results: use of external resources

nDCG@20

ecomsvz  ecomsv  ecomsv
FW14Search50  FW14Search75  FW14Search100
FW14Docs100  ICTNETR05  ICTNETR06
ICTNETR07  ICTNETR02  ICTNETR03
ICTNETR04  ICTNETR02  ICTNETR01
ITTTailG2000  UPDFW14tipsm  UPDFW14tkms
ITTTailG2000  UPDFW14tipsm  UPDFW14tkms

UNIVERSITET GENT  UNIVERSITY OF TWENTE  YAHOO LABS
QUERY-BASED SAMPLING: DISCUSSION

1. Sampling snippets is as effective as sampling full documents
2. Can be done at no extra costs(!)
FEDWEB GREATEST HITS
https://fedwebgh.intec.ugent.be

- A citable (static) dataset!
FEDWEB GREATEST HITS
https://fedwebgh.intec.ugent.be

- >50 topics (queries) per year (2013, 2014)
- Result pages for 150 resources
- Click-through for snippets
- Relevance judgments for pages
- Duplicates detected
- Results and pages for sample queries
- Additional (duplicate) judgments
INVENT YOUR OWN RESEARCH
https://fedwebgh.intec.ugent.be

- Monitoring: What changed in 1 year?
- Clicks vs. Page relevance
- Web search without web search engines
- Size estimation
- …
HOW MUCH CHANGED IN 1 YEAR?

“Big Web” search

UNIVERSITY OF TWENTE.
CHANGE RATE PER CATEGORY

![Graph showing change rate per category]
DISCUSSION (1)

- Many things change in 1 year
  - Big differences per resource
    (Jobs vs. travel)
- Challenge:
  - Remove outdated results from sample index
  - Learn change rate!

Mohammadreza Khelghati, Djoerd Hiemstra and Maurice van Keulen. Efficient Web Harvesting Strategies for Monitoring Deep Web Content. (Submitted for publication)
RELEVANT RESULTS?
CAN WE DO WITHOUT LARGE SEARCH ENGINES?

<table>
<thead>
<tr>
<th></th>
<th>Only WSE</th>
<th>Non-WSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>k = 5</td>
<td>0.328</td>
<td>0.835</td>
</tr>
<tr>
<td>k = 10</td>
<td>0.217</td>
<td>0.735</td>
</tr>
</tbody>
</table>

Table 4: Oracle experiment - Rel or better
CAN WE DO WITHOUT LARGE SEARCH ENGINES?

<table>
<thead>
<tr>
<th></th>
<th>Only WSE</th>
<th>Non-WSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>k = 5</td>
<td>0.120</td>
<td>0.891</td>
</tr>
<tr>
<td>k = 10</td>
<td>0.070</td>
<td>0.844</td>
</tr>
</tbody>
</table>

Table 5: Oracle experiment - Better than HRel
DISCUSSION (2)

- Relevant results in all resource categories
- General web search engines needed for top results
DO CLICKS IMPLY RELEVANCE?

Table 4: Overview of the relationship between page and snippet judgments, for different types of resources, and based on the page relevance level $P \geq \text{HRel}$.

<table>
<thead>
<tr>
<th>Resource</th>
<th>$S=\text{Unlikely}$</th>
<th>$S=\text{Maybe}$</th>
<th>$S=\text{Sure}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Web search</td>
<td>0.20</td>
<td>0.40</td>
<td>0.65</td>
</tr>
<tr>
<td>Multimedia</td>
<td>0.09</td>
<td>0.23</td>
<td>0.48</td>
</tr>
<tr>
<td>Q &amp; A</td>
<td>0.00</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Jobs</td>
<td>0.00</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>Academic</td>
<td>0.03</td>
<td>0.08</td>
<td>0.14</td>
</tr>
<tr>
<td>News</td>
<td>0.09</td>
<td>0.19</td>
<td>0.42</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.06</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>Encyclopedia/Dict</td>
<td>0.05</td>
<td>0.23</td>
<td>0.58</td>
</tr>
<tr>
<td>Books</td>
<td>0.12</td>
<td>0.10</td>
<td>0.18</td>
</tr>
<tr>
<td>Social &amp; Social Sharing</td>
<td>0.06</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>Blogs</td>
<td>0.12</td>
<td>0.23</td>
<td>0.40</td>
</tr>
<tr>
<td>Other</td>
<td>0.04</td>
<td>0.08</td>
<td>0.34</td>
</tr>
<tr>
<td>All</td>
<td>0.09</td>
<td>0.21</td>
<td>0.50</td>
</tr>
</tbody>
</table>
CLICKS ON U. TWENTE SEARCH

- 84.4% of clicks are on result 1 (considering resources as results)
CLICKS ON U. TWENTE SEARCH

- Skip/click pairs

Figure 1. Percentage of clicks on a search engine if the rank is larger than one (everything below 5% is in the other section)
DISCUSSION (3)

- Clicks
  - Bias might be severe in real system
  - Bias not present in test collection (by design!)
  - Clicks are noisy prediction of relevance

- Challenge
  - Study click bias
  - Learn from click data
Start

Searsia comes with a client and a server.

The client

The Searsia Web client can be downloaded as searsiaclient.zip and unzipped on your local machine or web server. To use the web client, open the file index.html in a web browser. Congratulations! You now run your own web application for federated search.

Client options

The client will automatically connect to the University of Twente search server. To connect to another server, edit the second line in the file js/searsia.js, which contains the API template of the server.

```javascript
var API_TEMPLATE = 'https://search.utwente.nl/searsia/search?q={q}&r={r}';
```

If you run a server on your local machine (see next section), you can connect to your own server by setting the API template to: 'http://localhost:16842/searsia/search?q={q}&r={r}'
CONCLUSIONS

- Federated Search as a Living Lab
  - Data for Federated Search
  - Software for Federated Search
  - Coming up: Experiments with our own interaction data
OUTLOOK

- “Green” (no need to crawl & store everything)
- “Democratic” (resources vote for results)
- “Cheap” (for Searsia; costly for NSA ;-) )
PUBLICATIONS

ACKNOWLEDGEMENTS

- Almer Tigelaar
- Mohammad Khelghati
- Rob Stortelder
- Dong Nguyen
- Thomas Demeester
- Ke Zhou
- Dolf Trieschnigg