The Archive Query Log

Heinrich Sebastian Maik Lukas Harry Benno Matthias Martin



Query Logs

What are they good for?



- Query understanding
- Learning to rank
- ► SERP analysis



- ► Most are **private**

Query Logs

What are they good for?



- Query understanding
- Learning to rank
- SERP analysis

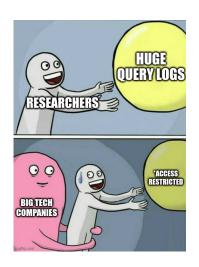
What's the issue?



- ► Most are **private**
- Public ones much smaller

Previous Logs

Query Logs



CEXCITE TO GOCL
CADAL LIDRAY
EUROPEAN LIDRAY AOL
TIMWAY Utah State Gov Bing
PubMedCiteSeer XInfoseek
EncartaStackOverflow
NOUREBelga News Agency Sogou
AOLIA
DBMS Microsoft AdCenterGAIS
Yandex AltaVistaORCAS
Baidu StartpaginaMetaSpy
OpenFindBing Videos Taobao
Dreamerparsijoo.ir MSN
Yahoo!Europeana

Previous Logs

Web Archive to the rescue!

But how?

- 1. List SERP URLs from Web Archive
- 2. Parse query
- 3. Download SERP HTML
- 4. Parse search results

AnxivOpenFindLycos
NDURE European Library
AOLIA CiteSeerX TodoCL
Yahoo!Startpagina Yandex
Timway Belga News Agency Bing
MetaspyAltaVistaBaidu
MSNMicrosoft AdCenterDBMS
Kunstmuseum.in
parsijoo.ir
Infoseek
PubMed StackOverflowTumba!
SogoUBing VideosGAIS
TaobaoEuropeanacNUIFT
AOLDreamer

Like this...

https://google.com/search?q=covid+19+usa+map

URL prefix

query

With the AQL

Meet the Archive Query Log

- webis-de/archive-query-log
 - **Q** 356 million queries
 - **≡** 166 million SERPs
 - 1.7 billion search results
 - across 25 years



And now?

- ► high diversity: topic and language
- diachronic analyses
- publically accessible via TIRA

Thank you, stay tuned for more!

Meet the Archive Query Log

- webis-de/archive-query-log
 - Q 356 million queries
 - **≡** 166 million SERPs
 - 1.7 billion search results
 - across 25 years



And now?

- high diversity: topic and language
- diachronic analyses
- publically accessible via TIRA

Thank you, stay tuned for more!

Meet the Archive Query Log

- webis-de/archive-query-log
 - Q 356 million queries
 - **≡** 166 million SERPs
 - 1.7 billion search results
 - across 25 years
 - 550 search providers



And now?

- ► high diversity: topic and language
- diachronic analyses
- publically accessible via TIRA

Thank you, stay tuned for more!