

Understanding user behavior from search logs: a metadata-level approach

Tessel Bogaard, Information Access

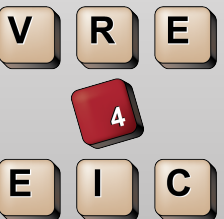
Laura Hollink, Jan Wielemaker,

Jacco van Ossenbruggen, Lynda Hardman

Acknowledgments:

Under a strict confidentiality agreement the National Library of the Netherlands has provided us with around 200M log records collected from October 2015 until March 2016. We thank the National Library for their support.

This research was partially supported by the VRE4EIC project, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 676247.



Motivation

Better understand user search behavior in a digital library, so to evaluate and improve:

- Search algorithms
- Search interfaces
- Identify gaps in the collection

Problem

How to understand user search behavior from **low-level HTTP server logs?**

- Analysis of search logs in combination with the collection
- Explicit and transparent exploration, processing, and analysis of data needed

Research based on use case in collaboration with the **National Library of the Netherlands**

How to understand user behavior from search logs?

Query-level analysis:

- Uncontrolled vocabulary
- Patterns hard to find in unique queries
- Users enter privacy-sensitive information

Query: “Oudkerk”

- politician?
- family name?
- village?

How to understand user behavior from search logs?

Digital libraries and archives **differ** from open web search:

- Collection known and available
- High quality professionally-curated metadata
- Facets based on metadata help navigate through query results

The screenshot shows the Delfher search interface. At the top, the search bar contains 'batavia' and the results are sorted by 'relevantie'. The search results show 3,424,342 newspaper articles. On the left, there are two facets: 'Periode' and 'Verspreidingsgebied'. The 'Periode' facet lists four options: 17e eeuw (250), 18e eeuw (14712), 19e eeuw (1171459), and 20e eeuw (2237921). The 'Verspreidingsgebied' facet lists five options: Landelijk (690904), Nederlands-Indië / Indonesië (2019087), Nederlandse Antillen (3970), and Regionaal/lokaal (696497). On the right, there is a 'metadata' section for a specific article. The metadata includes: Krantentitel: De locomotief : Samarangsch handels- en advertentie-blad; Datum: 04-04-1901; Soort bericht: Advertentie; Editie: Dag; Uitgever: De Groot, Kolff & Co; Plaats van uitgave: Semarang; Verspreidingsgebied: Nederlands-Indië / Indonesië; Nummer: 78; Jaargang: 50.

facets

Periode

- 17e eeuw (250)
- 18e eeuw (14712)
- 19e eeuw (1171459)
- 20e eeuw (2237921)

Verspreidingsgebied

- Landelijk (690904)
- Nederlands-Indië / Indonesië (2019087)
- Nederlandse Antillen (3970)
- Regionaal/lokaal (696497)

metadata

Krantentitel	De locomotief : Samarangsch handels- en advertentie-blad
Datum	04-04-1901
Soort bericht	Advertentie
Editie	Dag
Uitgever	De Groot, Kolff & Co
Plaats van uitgave	Semarang
Verspreidingsgebied	Nederlands-Indië / Indonesië
Nummer	78
Jaargang	50

How to understand user behavior from search logs?

Query-level analysis:

- Uncontrolled vocabulary
- Patterns hard to find in unique queries
- Users enter privacy-sensitive information

Try metadata-level analysis:

- Vocabulary controlled
- Group search interactions by shared facet-use
- Focus shifted from privacy-sensitive query

Query: “Oudkerk”

- politician?
- family name?
- village?

Facets: • announcement
• local newspaper

Click: • announcement
• Rotterdamse courant

Linking search logs and collection

Delfher
Kranten

/kranten/results
f
/kranten/view
f
/kranten/view
fa
/kranten/view
facet= 18th century
facet= article

Periode

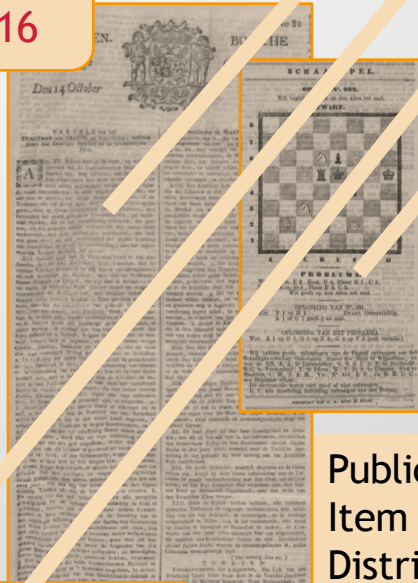
- 17e eeuw (68698)
- 18e eeuw (716272)
- 19e eeuw (14359929)
- 20e eeuw (105762505)

~200M log records
Oct.2015-Mar.2016

>100M documents
1618-1995

clicks

Enriched dataset



Metadata:

Publication date: 24-09-1866
Item type: caption
Distribution zone: national

Publication date: 14-10-1774
Item type: article
Distribution zone: local

Publication date: 02-01-1866
Item type: advert
Distribution zone: Indonesia

Napels den 8 July. De Hertog van S. Donato, is overleden.

Publication date: 31-07-1727
Item type: announcement
Distribution zone: local

Understanding usage patterns from enriched dataset

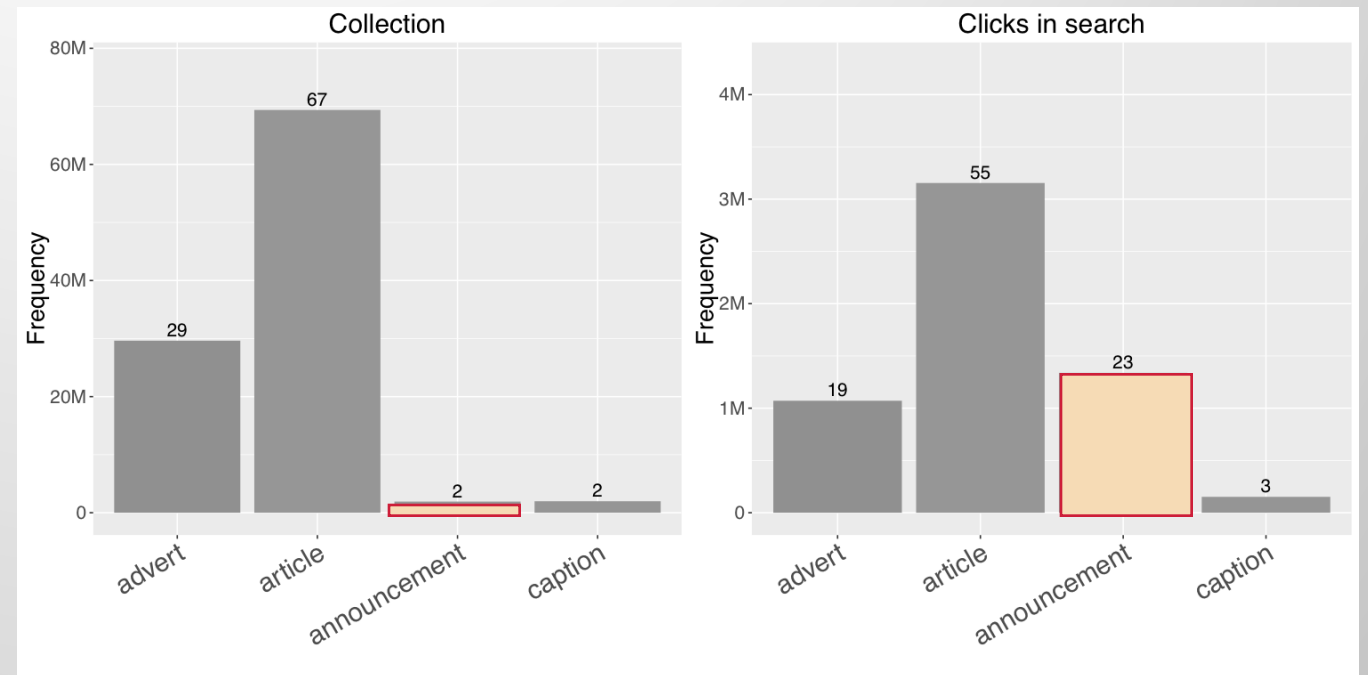
Is search for family announcements different from search for articles?

Announcements are popular:

- Announcements are 2% of the collection but receive **23%** of all clicks

Search is efficient:

- Same number of search interactions in shorter time, fewer clicks and hardly any downloads



Preliminary results

Metadata-level analysis:

- Insights into user behavior and information needs
- Starting point for inter-collection comparison of user behavior
- First step towards more privacy-preserving method of analysis
- Data, code and results shared through **SWISH DataLab**
github.com/SWI-Prolog/swish



SWI Prolog

Future work and open questions

- How to evaluate this method of analysis?
- Next step: predictive analytics based on metadata instead of query
- Continued collaboration with National Library: development usage data analytics dashboard

