Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia

Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck and Markus Strohmaier

Wikimedia Research Showcase - August 2020


Information Access on Wikipedia

- Search

Google
Information Access on Wikipedia

- Search

- Navigation
Information Access on Wikipedia

- Search
- Navigation

~70% of all visits

~30% of all visits

Amsterdam

From Wikipedia, the free encyclopedia

Amsterdam (ˌɑmstəˈrɑm) is the capital and most populous municipality of the Netherlands. Its status as the capital is mandated by the Constitution of the Netherlands, although it is not the seat of the government, which is The Hague. Amsterdam has a population of 851,373 within the city proper, 1,351,587 in the urban area, and 2,410,960 in the Amsterdam metropolitan area. The city is located in the province of North Holland in the west of the country but is not its capital, which is Haarlem. The metropolitan area comprises much of the northern part of the Randstad, one of the larger conurbations in Europe, with a population of approximately 7 million.

Amsterdam's name derives from "Amstelredamme," indicative of the city's origin around a dam in the river Amstel. Originating as a small fishing village in the late 12th century, Amsterdam became one of the most important ports in the world during the Dutch Golden Age (17th century), a result of its innovative developments in trade. During that time, the city was the leading centre for finance and diamonds. In the 19th and 20th centuries the city expanded, and many new neighbourhoods and suburbs were planned and built. The 17th-century canals of Amsterdam and the 19-20th century Defence Line of Amsterdam are on the UNESCO World Heritage List. Since the annexation of the municipality of Sloten in 1921 by the municipality of Amsterdam, the oldest historic part of the city lies in Sloten.

As the commercial capital of the Netherlands and one of the top financial centres in Europe, Amsterdam is considered an alpha world city by the Globalization and World Cities (GaWC) study group. The city is also the cultural capital of the Netherlands. Many large Dutch institutions have their headquarters there, and seven of the world's 500 largest companies, including Philips, AkzoNobel, TomTom and ING, are based in the city. Also, many leading technology companies have their European headquarters in Amsterdam, such as Uber, Netflix and Tesla. In 2012,
How do search and navigation interplay to shape the article traffic on Wikipedia?
How do search and navigation interplay to shape the article traffic on Wikipedia?

Which article properties are indicative of specific information access behavior?
Dataset

- Wikipedia clickstream
- Counts of referrer-resource pairs
- English Wikipedia (August, 2016)
- ~ 3.1 million articles
- ~ 4 billion views

<table>
<thead>
<tr>
<th>Referrer</th>
<th>Resource</th>
<th>Access form</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Amsterdam</td>
<td>Search</td>
<td>1340</td>
</tr>
<tr>
<td>London</td>
<td>Amsterdam</td>
<td>Navigation</td>
<td>230</td>
</tr>
<tr>
<td>Paris</td>
<td>Amsterdam</td>
<td>Navigation</td>
<td>360</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Amsterdam</td>
<td>Search</td>
<td>270</td>
</tr>
</tbody>
</table>

Methodology

Step 1
• Traffic features and article roles

Step 2
• Topic-specific heatmaps

Step 3
• Modeling access behavior

Which article properties are indicative of specific information access behavior?
What are the functional roles articles can assume for the traffic flow?
Traffic features

\[
\text{searchshare}(a) = \frac{\text{in}_{se}(a)}{\text{in}_{se}(a) + \text{in}_{nav}(a)}
\]

**Total pageviews**

**Incoming search traffic**

**Incoming navigation traffic**

*Searchshare* – the amount of views an article received by search
Traffic features

\[
\text{searchshare}(a) = \frac{\text{in}_{\text{se}}(a)}{\text{in}_{\text{se}}(a) + \text{in}_{\text{nav}}(a)}
\]

\[\text{Total pageviews}\]

**Searchshare** – the amount of views an article received by search

\[
\text{resistance}(a) = 1 - \frac{\text{out}_{\text{nav}}(a)}{\text{in}_{\text{se}}(a) + \text{in}_{\text{nav}}(a)}
\]

\[\text{Total pageviews}\]

**Resistance** – the ability of an article to relay traffic to other Wikipedia articles
Traffic features

\[
\text{searchshare}(a) = \frac{\text{in}_{se}(a)}{\text{in}_{se}(a) + \text{in}_{nav}(a)}
\]

\[
\text{resistance}(a) = 1 - \frac{\text{out}_{nav}(a)}{\text{in}_{se}(a) + \text{in}_{nav}(a)}
\]

**Searchshare** – the amount of views an article received by search

**Resistance** – the ability of an article to relay traffic to other Wikipedia articles

<table>
<thead>
<tr>
<th></th>
<th>Searchshare</th>
<th>Resistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.66</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Article roles

Search-exit – articles accessed by search that are unable to relay traffic
Article roles

Search-exit – articles accessed by *search* that are *unable* to relay traffic

Search-relay – articles accessed by *search* that are *able* to relay traffic
Article roles

- **Search-exit** – articles accessed by search that are unable to relay traffic
- **Search-relay** – articles accessed by search that are able to relay traffic
- **Navigation-relay** – articles accessed by navigation that are able to relay traffic
Article roles

Search-exit – articles accessed by *search* that are *unable* to relay traffic

Search-relay – articles accessed by *search* that are *able* to relay traffic

Navigation-relay – articles accessed by *navigation* that are *able* to relay traffic

Navigation-exit – articles accessed by *navigation* that are *unable* to relay traffic
Number of views

Wikipedia

Searchshare

Resistance

$10^7$

$10^6$
Which article properties are characteristic for specific article roles?
Article features

- **Network features**
  - In, out, and total degree
  - K-Core

- **Content and edit features**
  - Editors, revisions, age
  - Size(kB), sections, tables

- **Topics**
  - 20 topics discovered using LDA
  - Topics are labeled by five researchers using the top words and articles for each topic
Article features

- Network features
  - In, out, and total degree
  - K-Core

- Content and edit features
  - Editors, revisions, age
  - Size(kB), sections, tables

- Topics
  - 20 topics discovered using LDA
  - Topics are labeled by five researchers using the top words and articles for each topic
Are topics accessed differently?

**Searchshare**

**Resistance**
Topic-specific heatmaps

Create a topic-specific heatmap
Topic-specific heatmaps

- Create a topic-specific heatmap
- Bin-wise division of topic-specific and Wikipedia heatmaps
Topic-specific heatmaps

- Create a topic-specific heatmap
- Bin-wise division of topic-specific and Wikipedia heatmaps
- Result: heatmap showing the deviation of the topic-specific from the general access behavior
What is the role of a topic for the traffic?
What is the role of a topic for the traffic?
What is the role of a topic for the traffic?

Architectures and Military

- Overrepr.
- No diff.
- Underrepr.
What is the role of a topic for the traffic?

Query for „Architecture“, Click through „Military“
Which article properties are indicative of specific information access behavior?
Modeling access behavior

0.66

1.0

Search-heavy

Navigation-heavy

Resistance

Searchshare
Modeling access behavior

Step 3

Resistance

Searchshare

Exit points

Relay points

0.88

1.0

0

1.0
Results

![Graph showing AUC values for different categories: baseline (0.5), network (0.58), content & edit (0.61), topic (0.64), and all (0.7).]
Results

Searchshare

1. Article topic is indicative of searchshare
Results

**Searchshare**

1. Article topic is indicative of searchshare
Results

1. Article topic is indicative of searchshare
2. Content and edit features are indicative of resistance
Results

1. Article topic is indicative of searchshare
2. Content and edit features are indicative of resistance
3. Article position in the network is indicative of resistance
Implications

- Identification of search- and navigation-exit points
- Design of interfaces and new features
  - Topic-specific layouts to support search- and navigation-heavy reading behavior
  - Introduction of links to the page preview on search-heavy articles
- Detection of online misbehavior and vandalism
  - Unexpected changes in the access patterns on articles
Summary and Conclusion

- Studied the interplay between search and navigation on Wikipedia
- Wikipedia is good at converting search to navigation traffic
- Search and navigation are used to access different articles
- Both information access forms are crucial as they complement each other
Questions?

Thank you!

Contact:
E-Mail: dimitar.dimitrov@gesis.org
Twitter: @trovdimi