A Mastodon Corpus to Evaluate Federated Microblog Search

Matti Wiegmann  Jan Heinrich Reimer  Maximilian Ernst  Martin Potthast  Matthias Hagen  Benno Stein

Bauhaus-Universität Weimar  Friedrich-Schiller-Universität Jena  Leipzig University  ScaDS.AI

webis.de
Overview

*Mastodon is a federated* and open-source microblogging service. *We want to improve the full-text search.*

Our Contribution:

- Webis Mastodon Corpus 2024:
  - 35 million unique public posts 733 million total
  - Timelines of 1,015 Mastodon nodes 10% of discoverable nodes
  - Across 61 days (Dec-Feb).

* Mastodon nodes are hosted by users and federate via the ActivityPub protocol (W3C Rec.). All ActivityPub-enabled apps form the "Fediverse".
Motivation

1. Mastodon is relevant.

Recent shift in Mastodon's policy: full-text search is now wanted. Not the case with private social media.

Search on Mastodon is interesting. Various challenges for local and federated search.

https://mastodon-analytics.com/
Motivation

1. Mastodon is relevant.

2. We as researchers can contribute directly.
   - Recent shift in Mastodon’s policy: full-text search is now wanted.
   - Not the case with private social media.

https://mastodon-analytics.com/
Motivation

1. Mastodon is relevant.

2. We as researchers can contribute directly.
   - Recent shift in Mastodon’s policy: full-text search is now wanted.
   - Not the case with private social media.

3. Search on Mastodon is interesting.
   - Various challenges for local and federated search.

https://mastodon-analytics.com/

@Wiegmann, 2024
Motivation

Challenges:

- Develop great local search.
  - Features are different from e.g. X/Twitter.
- Develop federated search.
  - Local search is poor for small instances.
  - Many nodes are specialized.
  - Efficiency vs. effectiveness trade off.
- Integrate other Fediverse apps.
  - Many ActivityPub apps are in the timeline.
  - Different media types (like PeerTube videos) can be searched.

Post

User

Local

User

Local

User

Search (status quo)

User

Local

©Wiegmann, 2024
Motivation

Challenges:

- Develop great local search.
  - Features are different from e.g. X/Twitter.

- Develop federated search.
  - Local search is poor for small instances.
  - Many nodes are specialized.
  - Efficiency vs. effectiveness trade off.

Search (status quo)
Motivation

Challenges:

- Develop great local search.
  - Features are different from e.g. X/Twitter.

- Develop federated search.
  - Local search is poor for small instances.
  - Many nodes are specialized.
  - Efficiency vs. effectiveness trade off.

- Integrate other Fediverse apps.
  - Many ActivityPub apps are in the timeline.
  - Different media types (like PeerTube videos) can be searched.

Post

Search (status quo)
Corpus Construction

1. **Node sampling**
   - Select 1,000 nodes (ca. 10%) based on 6 activity statistics.
   - Replace nodes that went dark (15).

2. **Crawling**
   - For 61 days (12. Dec to 21. Feb)
   - Discard posts with `noindex`.
   - Via streaming API with search API as backup.

3. **Store**
   - Posts (in elasticsearch)
     - ca. 6TB total / 88 GB per day index size.
Corpus Construction

1. **Node sampling**
   - Select 1,000 nodes (ca. 10%) based on 6 activity statistics.
   - Replace nodes that went dark (15).

2. **Crawling** of the local + remote timeline
   - For 61 days (12. Dec to 21. Feb)
   - Discard posts with `noindex`.
   - Via streaming API with search API as backup.

![Node Activity Chart](chart.png)
1. **Node sampling**
   - Select 1,000 nodes (ca. 10%) based on 6 activity statistics.
   - Replace nodes that went dark (15).

2. **Crawling** of the local + remote timeline
   - For 61 days (12. Dec to 21. Feb)
   - Discard posts with *noindex*.
   - Via streaming API with search API as backup.

3. **Store** the posts (in elasticsearch)
   - ca. 6TB total / 88 GB per day index size.
Insights

- Number of Posts and Centrality
  - Most posts on Mastodon are duplicates. → Querying many nodes is inefficient. duplicates, traffic, resources
  - Querying fewer nodes is more efficient but less effective.

<table>
<thead>
<tr>
<th>Node</th>
<th>Posts</th>
<th>% of unique</th>
<th>% in remote</th>
<th>% in local</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mastodon.social</td>
<td>16M</td>
<td>44%</td>
<td>80%</td>
</tr>
<tr>
<td>2</td>
<td>mastodon.online</td>
<td>10M</td>
<td>28%</td>
<td>97%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>toot.community</td>
<td>7M</td>
<td>19%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>total posts</td>
<td>733M</td>
<td>—</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>unique posts</td>
<td>35M</td>
<td>100%</td>
<td>72%</td>
</tr>
</tbody>
</table>
Insights

- **Number of Posts and Centrality**
  - Most posts on Mastodon are duplicates.
    → Querying many nodes is inefficient. duplicates, traffic, resources
    → Querying fewer nodes is more efficient but less effective.
  - Big nodes collect a large % of all unique posts.
    → Can we not just search `mastodon.social`?
    → Narrow resource selection is impolite.

<table>
<thead>
<tr>
<th>Node</th>
<th>Posts</th>
<th>% of unique</th>
<th>% in remote</th>
<th>% in local</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mastodon.social</td>
<td>16M</td>
<td>44%</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>2 mastodon.online</td>
<td>10M</td>
<td>28%</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 toot.community</td>
<td>7M</td>
<td>19%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>total posts</td>
<td>733M</td>
<td>—</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>unique posts</td>
<td>35M</td>
<td>100%</td>
<td>72%</td>
<td>28%</td>
</tr>
</tbody>
</table>

Expert communities? Principles of Decentralization?
Insights

- Number of Posts and Centrality
- Text Length

Text Length

<table>
<thead>
<tr>
<th>Node</th>
<th>Posts</th>
<th>% of unique</th>
<th>% in remote</th>
<th>% in local</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mastodon.social</td>
<td>16M</td>
<td>44%</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>2 mastodon.online</td>
<td>10M</td>
<td>28%</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 toot.community</td>
<td>7M</td>
<td>19%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>total posts</td>
<td>733M</td>
<td>—</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>unique posts</td>
<td>35M</td>
<td>100%</td>
<td>72%</td>
<td>28%</td>
</tr>
</tbody>
</table>
**Insights**

- **Number of Posts and Centrality**
- **Text Length**
- **Languages** EN (35%), JA (23%), DE (5%), ZH (3%)

![Text Length Graph]

<table>
<thead>
<tr>
<th>Node</th>
<th>Posts</th>
<th>% of unique</th>
<th>% in remote</th>
<th>% in local</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mastodon.social</td>
<td>16M</td>
<td>44%</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>2 mastodon.online</td>
<td>10M</td>
<td>28%</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 toot.community</td>
<td>7M</td>
<td>19%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>total posts</td>
<td>733M</td>
<td>—</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>unique posts</td>
<td>35M</td>
<td>100%</td>
<td>72%</td>
<td>28%</td>
</tr>
</tbody>
</table>

@Wiegmann, 2024
Insights

- **Number of Posts and Centrality**
- **Text Length**
- **Languages** EN (35%), JA (23%), DE (5%), ZH (3%)
- Interactions, Spoiler tags, hashtags, accounts, ...

<table>
<thead>
<tr>
<th>Node</th>
<th>Posts</th>
<th>% of unique</th>
<th>% in remote</th>
<th>% in local</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mastodon.social</td>
<td>16M</td>
<td>44%</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>2 mastodon.online</td>
<td>10M</td>
<td>28%</td>
<td>97%</td>
<td>3%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 toot.community</td>
<td>7M</td>
<td>19%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>total posts</td>
<td>733M</td>
<td>—</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>unique posts</td>
<td>35M</td>
<td>100%</td>
<td>72%</td>
<td>28%</td>
</tr>
</tbody>
</table>

@Wiegmann, 2024
Overview

*Mastodon is a federated and open-source microblogging service. We want to improve the full-text search.*

Our Contribution:

- Webis Mastodon Corpus 2024:
  - 35 million unique public posts, 733 million total
  - Timelines of 1,015 Mastodon nodes, 10% of discoverable nodes
  - Across 61 days (Dec-Feb).

- To be hosted on TIREx:
  - Study microblog search.
  - Develop a shared task?
Appendix

Visibility and Consent

- noindex flag: ca. 20% of mastodon posts are opted-in to search.
- Implemented by few non-mastodon apps, but this will likely change.
  → For private (research) index: remove noindex posts, leave the rest.