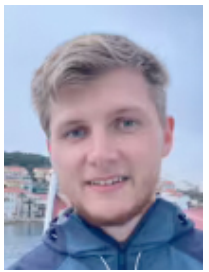
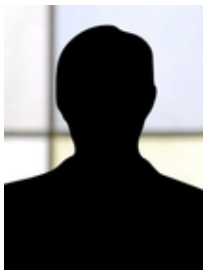
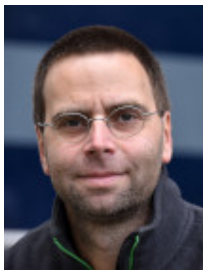
A Mastodon Corpus to Evaluate Federated Microblog Search




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.

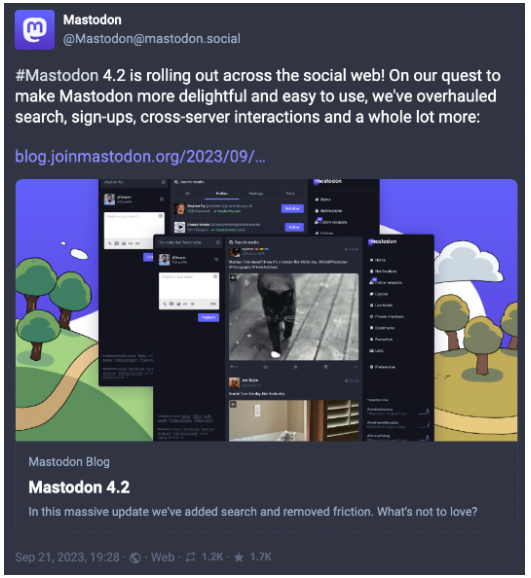
https://mastodon-analytics.com/

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/



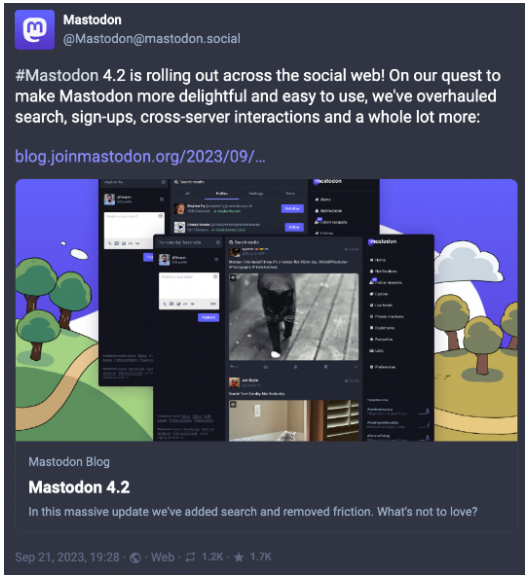
Motivation

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/

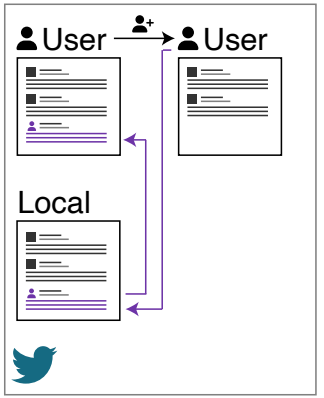


Motivation

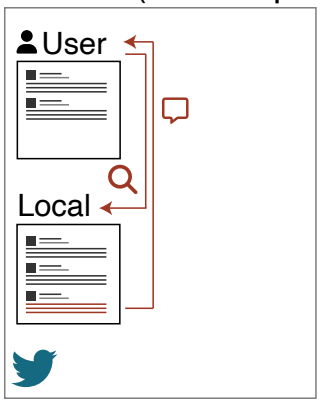
Challenges:

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

Post



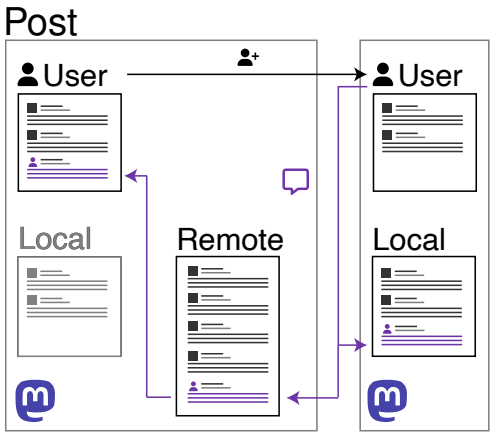
Search (status quo)

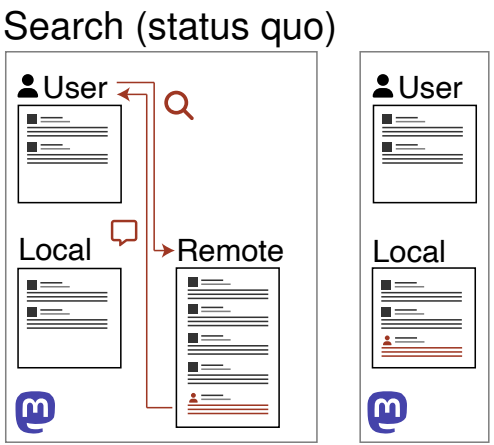


Motivation

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.

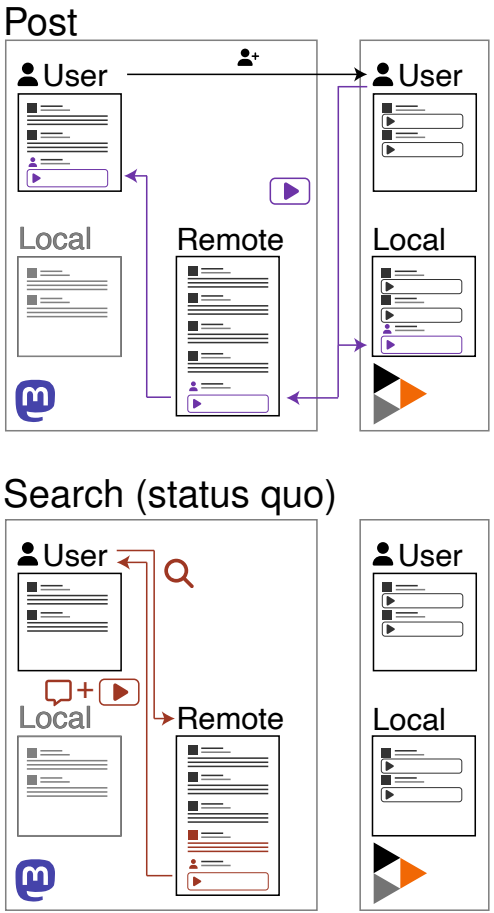




Motivation

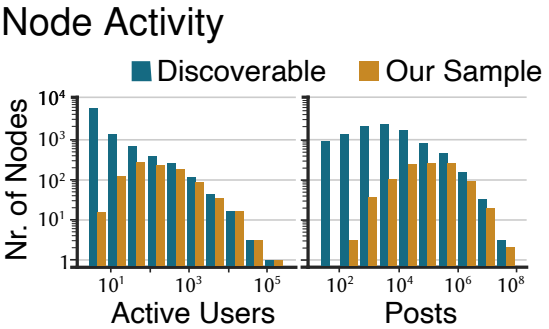
Challenges:

- 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.



Corpus Construction

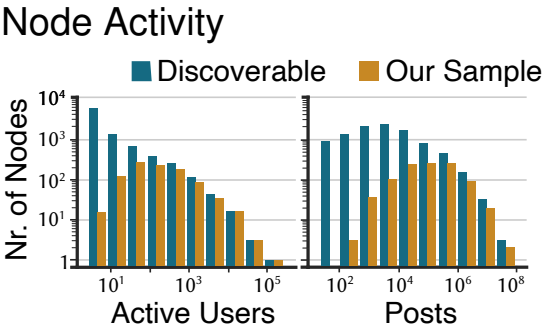
1. **Node sampling**
   - Select 1,000 nodes (ca. 10%) based on 6 activity statistics.
   - Replace nodes that went dark (15).



Corpus Construction

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

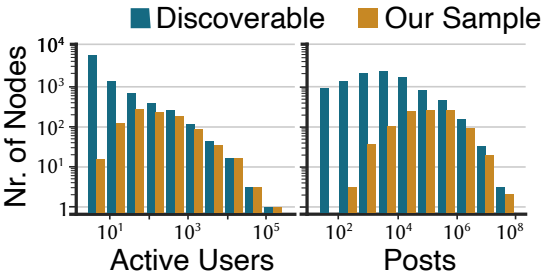
2. **Crawling** of the local + remote timeline
   - For 61 days (12. Dec to 21. Feb)
   - Discard posts with `noindex`.
   - Via streaming API with search API as backup.



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

Node Activity



Stored Fields

Fields	Total	Req.
Post	38	12
Author	29	12
Emb. Content	15	0
Media Attach.	18	0
Total	100	24

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.

Node		Posts	% of unique	% in remote	% in local
1	`mastodon.social`	16M	44%	80%	20%
2	`mastodon.online`	10M	28%	97%	3%
...					
10	`toot.community`	7M	19%	100%	0%
	total posts	733M	—	99%	1%
	unique posts	35M	100%	72%	28%

Insights

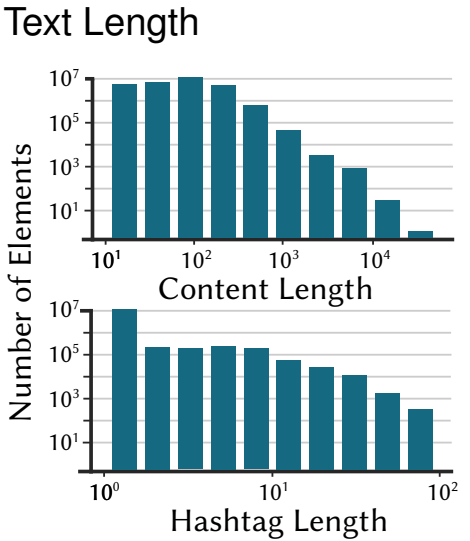
- **Number of Posts and Centrality**
  - Most posts on Mastodon are duplicates.
    → Querying many nodes is inefficient. duplicates, traffic, resources
    → Querying fewer nodes is more efficient but less effective.
  - Big nodes collect a large % of all unique posts.
    → Can we not just search `mastodon.social`?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

    Expert communities? Principles of Decentralization?

Node		Posts	% of unique	% in remote	% in local
1	`mastodon.social`	16M	44%	80%	20%
2	`mastodon.online`	10M	28%	97%	3%
...					
10	`toot.community`	7M	19%	100%	0%
	total posts	733M	—	99%	1%
	unique posts	35M	100%	72%	28%

Insights

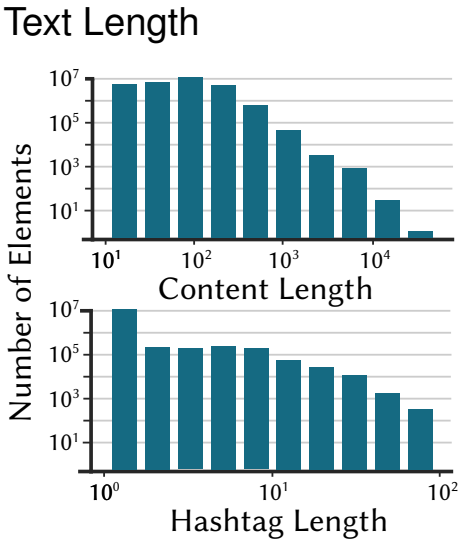
- **Number of Posts and Centrality**
- **Text Length**



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

Insights

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

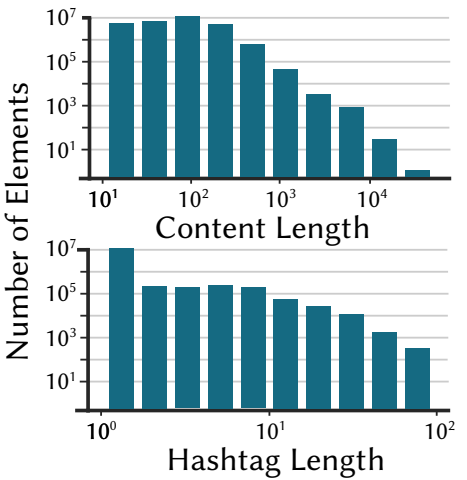


Node		Posts	% of unique	% in remote	% in local
1	`mastodon.social`	16M	44%	80%	20%
2	`mastodon.online`	10M	28%	97%	3%
...					
10	`toot.community`	7M	19%	100%	0%
	total posts	733M	—	99%	1%
	unique posts	35M	100%	72%	28%

Insights

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

Text Length



Node		Posts	% of unique	% in remote	% in local
1	`mastodon.social`	16M	44%	80%	20%
2	`mastodon.online`	10M	28%	97%	3%
...					
10	`toot.community`	7M	19%	100%	0%
	total posts	733M	—	99%	1%
	unique posts	35M	100%	72%	28%

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.