MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Anders Skovsgaard
Aarhus University


Darius Šidlauskas
Aarhus University


Christian S. Jensen
Aarhus University

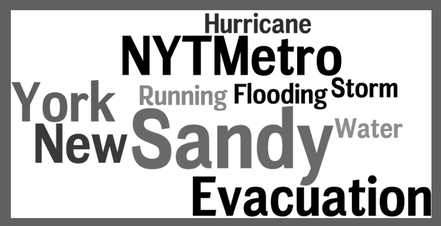
Danmarks Grundforskningsfond
Danish National Research Foundation

Scalable Top-*k* Spatio-Temporal Term Querying

Introduction

Motivation
What's the "talk of the town" at a specific time interval?

The top-k most popular terms:



Scalability
We seek a solution that is capable of supporting the entire world, a long history of content, and a stream with much higher rates than what Twitter currently sees (5,000 tweets/sec).

Problem Definition

- D – a set of spatio-temporal objects o.
- $o = (\lambda, \varphi, ts)$
  - λ – a point location (latitude and longitude),
  - φ – a text document (set of terms t),
  - ts – a timestamp.
- $score(t, D) = |\{o \in D \mid t \in o.\varphi\}|$ – a score of a term t for a set D of objects.
- Input: $q = (k, R, I)$ – a top-k scored terms query
  - k – number of top-k terms,
  - R – a rectangular range,
  - I – a time interval.
- Output:
  - k top scored terms from objects $\{o \in D \mid o.\lambda \in R \cap o.ts \in I\}$,
  - k_g – an integer ($\leq k$) guaranteeing that the first k_g terms to have the highest scores (the rest $k - kg$ terms are approximate).

Adaptive Frequent Item Aggregator (*AFIA*)

- Dynamic summaries: extend SpaceSaving [1] to *dynamically adjusts* to incoming stream.
- Multiple spatio-temporal granularities (Figures 2 and 3).
- The top-k spatio-temporal query processor including:
  - Support for *ad-hoc* spatio-temporal ranges.
  - Computing k_g that captures which part of the result is *exact* rather than approximate.

Fig. 1: Dynamic Summaries

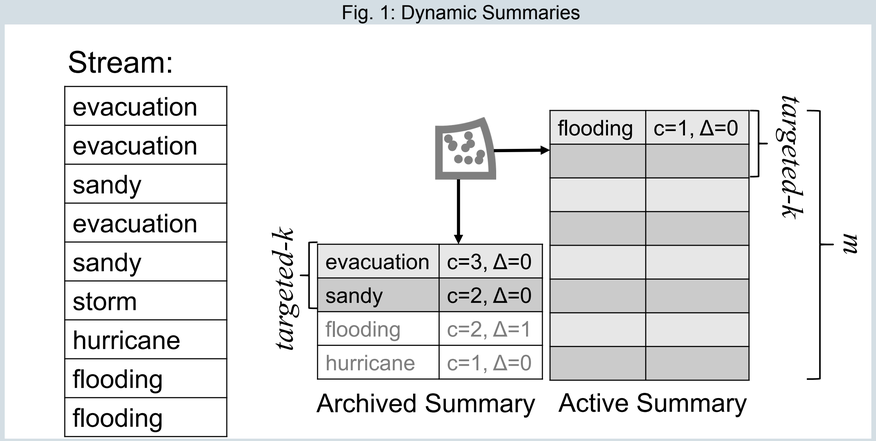


Fig. 2: Multiple Spatial Granularities

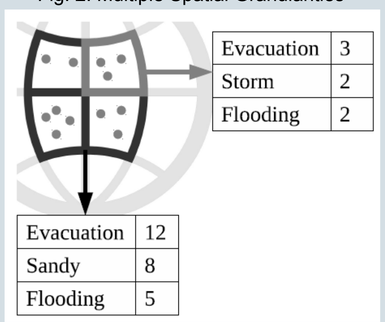


Fig. 3: Multiple Temporal Granularities

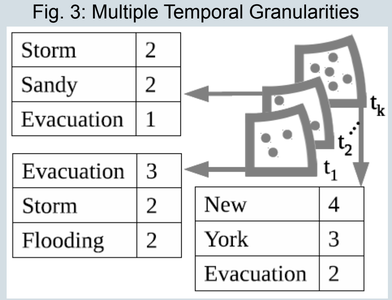
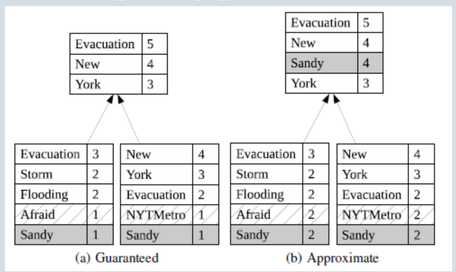


Fig. 4: Merging of Summaries



Empirical Study

Data

- All geo-tagged posts from Twitter's Streaming API during May, 2013.
- The total number of tweets is 110,426,053 (41 tweets/second).

Baselines

- SS: (approximate) frequent item aggregation using SpaceSaving [1].
- HT: (exact) frequent item counting using a hash table.

Fig. 5: Accuracy at Different Spatio-Temporal Granularities

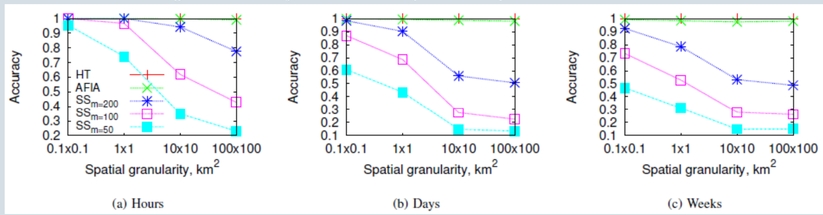


Fig. 6: Average Number of Counters Maintained at Different Spatio-Temporal Granularities

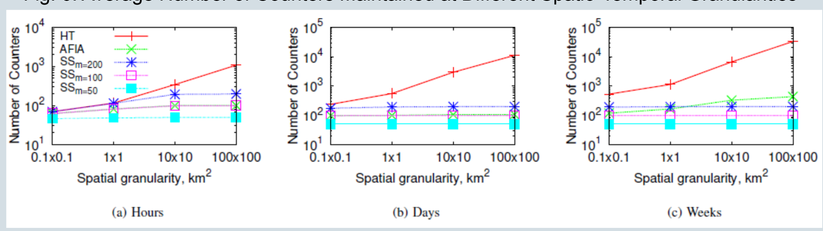


Fig. 7: Large Scale Stream Processing

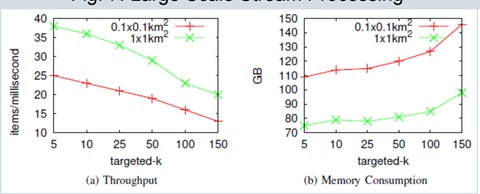
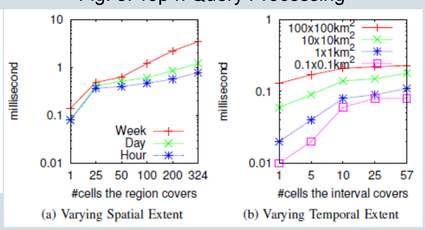


Fig. 8: Top-*k* Query Processing



Conclusion

- AFIA's throughput exceeds Twitter's current average rate by a factor of 4–10.
- One month of dynamic summaries require 120 GB of memory.
- The lowest observed accuracy was 97%.

References

[1] A. Metwally, D. Agrawal, and A. El Abbadi. *Efficient computation of frequent and top-k elements in data streams*. ICDT, 2005.