Streaming Data Mining

PRESENTED BY Edo Liberty | April 11, 2014

Copyright © 2014 Yahoo! All rights reserved. No reproduction or distribution allowed without express written permission.

Parts of this presentation were given with Jelani Nelson (Harvard) as a KDD tutorial on streaming data mining.
Single machine data mining

The World

Data

Computation

Result
Distributed storage

The World

Data Data Data Data

Computation

Result
Distributed model (map/reduce, message passing, …)
Distributed model (indexes, tables, databases, …)
207 big-data infographics (meta infographic)
The streaming model

The World

Computation

Sketch

Query

Query Algorithm

Result

Result
The parallel streaming model

The World

Compute + Sketch
Compute + Sketch
Compute + Sketch
Compute + Sketch

Aggregate+ Sketch

Query Algorithm

Query

Result

Result
The streaming model (more accurately)

\[ O(n) \text{ items} \]

\[ O(\text{polylog}(n)) \text{ computation per item} \]

\[ O(\text{polylog}(n)) \text{ space} \]
Communication complexity
Frequent items

Demaine, Lopez-Ortiz, Munro. Frequency estimation of internet packet streams with limited space, 2002
The name "Lossy Counting" was used for a different algorithm by Manku and Motwani, 2002
Metwally, Agrawal, Abbadi, Efficient Computation of Frequent and Top-k Elements in Data Streams, 2006
\[ d \begin{bmatrix} \text{blue} & \text{orange} \\ \text{purple} & \text{green} \\ \text{red} & \text{green} \end{bmatrix} \Rightarrow f(\text{purple}) = 5 \]
$f(\color{purple}\square) = 5$
\[ f'(\text{light green}) = 2 \]

\[ f'(\text{dark green}) = 0 \]
The proof (very short)

First fact: \( f'(x) \leq f(x) \)

Assume we do this \( t \) times

Second fact: \( f'(x) \geq f(x) - t \)
The proof (very short)

Third (not so obvious) fact:

\[ 0 \geq \sum f'(x) = \sum f(x) - t \cdot \ell = n - t \cdot \ell \]

Which gives \( t \leq n/\ell \). In words:

**We can only delete \( \ell \) items \( n/\ell \) times!**

\[ |f'(x) - f(x)| \leq n/\ell \]
Useful form...

Define \( p(x) = \frac{f(x)}{n} \)
And \( p'(x) = \frac{f'(x)}{n} \)

We get that
\[
|p'(x) - p(x)| \leq \frac{1}{\ell}
\]

This is very useful for keeping approx’ distributions!
Threading Machine Generated Email
A simple email thread (that’s not very hard to do…)

- **Emma Brunskill**: Hi Edo, It was very interes

- **Me**: Hi Emma, Thanks for reaching out, I ha

- **Emma Brunskill**: To Me
Threading Machine Generated Email

Ailon, Karnin, Maarek, Liberty, Threading Machine Generated Email, WSDM 2013
Threading Machine Generated Email

- Order Confirmation (retail) → Shipping Notification: 64%
- 19% back to Order Confirmation (retail)

- Utility bill payment due → Payment received: 44%
- 35% back to Insurance payment due
- 53% to Service cancellation

- Insurance payment due → Service cancellation: 15%
Threads Machine Generated Email

PayPal.com: “You submitted an order in the amount of * usd to overstock.com.”

overstock.com: “Overstock.com password reset request.”

payless.com “Order confirmation”

C=632
w=1,221

C=769
w=1,490

overstock.com: “Your overstock.com order has shipped.”

payless.com “Your order is shipped”

C=193
w=12,098

C=652
w=1,300

C=753
w=1,395

C=153
w=704

C=1,742
w=6,446
What else can we do in the streaming model...

Items (words, IP-addresses, events, clicks,...):
- Item frequencies
- Counting distinct elements
- Moment and entropy estimation
- Approximate set operations

Vectors (text documents, images, example features,...)
- Dimensionality reduction
- Clustering (k-means, k-median,...)
- Linear Regression
- Machine learning (some of it at least)

Matrices (text corpora, user preferences, graphs...)
- Covariance estimation matrix
- Low rank approximation
- Sparsification
Thanks!

Yahoo does big data algorithms, software and systems!

Speak to our Talent Team or visit Careers.Yahoo.com and explore our career opportunities in NYC or Sunnyvale, CA

Seth Tropper
satropper@yahoo-inc.com

Doug DeSimone
desimone@yahoo-inc.com

Keith Daniels
kdnl@yahoo-inc.com

Yahoo is an equal opportunity employer.