A Few Sketches of Sketches

Adam Marcus
B12
@marcua
A Sketching data structure summarizes a dataset, trading off accuracy for space.
Examples

- Sum
- Count
- Is item x in dataset D?
- How many x's are in D?
- How many distinct items are in D?
- What is a histogram of D?
- ...

Examples

- Sum
- Count
- Is item x in dataset D?  \(\rightarrow\) Bloom Filter
- How many x's are in D?  \(\rightarrow\) Count-min Sketch
- How many distinct items are in D?
- What is a histogram of D?
- ...

Bloom Filters
[Bloom, 1970]

Is item x in dataset D?

is
http://badurl.com → Google
Malicious?
↓
Gigabytes
Bloom Filters

[Bloom, 1970]

Is item x in dataset D?

is
http://badurl.com → Chrome
Malicious? → Bloom
Filter
↓
Megabytes → Google
↓
Gigabytes
How it Works

\[ h_1 \quad h_2 \quad h_3 \]
How it Works

\[ h_1 \rightarrow h_2 \rightarrow h_3 \]

Insert http://badurl.com
Nope!

Yes! Yes! No!

0 0 0 1 1 1 0 1 1 0

h₁ h₂ h₃

lookup http://goodurl.com
False Positive!

Yes! Yes! Yes!

0 0 0 1 1 1 0 1 1 0

\[ h_1 \quad h_2 \quad h_3 \]

lookup http://nice.url.com
Bloom filters guarantee an item is not present...

...and have a small chance of telling you an item is present when it's not 😞
How Small a Chance?
\[ \text{# of bits} = - \left( \frac{\text{# of items}}{\ln(2)^2} \right) \ln \left( \frac{\text{false positive rate}}{\ln(2)^2} \right) \]

<table>
<thead>
<tr>
<th>false positive rate</th>
<th>#bits per item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/100</td>
<td>~10</td>
</tr>
<tr>
<td>1/1000</td>
<td>~15</td>
</tr>
<tr>
<td>1/10,000</td>
<td>~20</td>
</tr>
</tbody>
</table>
false positive rate = \( \frac{1}{10,000} \)

# bits per item = \( \sim 20 \)

10 million URLs

30 characters each

Size of strings alone = \( \sim 1,144 \) MB

Bloom Filter = \( \sim 24 \) MB

...
Optimal # Hash Functions?

\[
\text{# hash functions} = \left(\frac{\text{# bits}}{\text{# items}}\right) \ln 2 \approx 0.69 \left(\frac{\text{# bits}}{\text{# items}}\right)
\]

20 bits/item \Rightarrow \sim 14 \text{ hash functions}
Count-Min Sketch
[Cormode & Muthukrishnan, 2003]

How many X's are in dataset D?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cats</td>
<td>35,000,000</td>
</tr>
<tr>
<td>Bats</td>
<td>4,000,000</td>
</tr>
<tr>
<td>Flats</td>
<td>2,000,000</td>
</tr>
<tr>
<td>Mats</td>
<td>100,000</td>
</tr>
<tr>
<td>Drats</td>
<td>50,000</td>
</tr>
</tbody>
</table>

Count the number of words on the web
How it Works

<table>
<thead>
<tr>
<th></th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>3</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>$h_2$</td>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>$h_3$</td>
<td>12</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>31</td>
<td>22</td>
</tr>
</tbody>
</table>
Count "flats"
<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>7</th>
<th>13</th>
<th>2</th>
<th>5</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>7</td>
<td>80</td>
<td>0</td>
<td>1</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>12</td>
<td>20</td>
<td>7</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Count "flats"
\[ \min(5, 7, 12) = 5 \]

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>7</th>
<th>13</th>
<th>2</th>
<th>5</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>7</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>12</td>
<td>20</td>
<td>7</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Count "flats":
- \( h_1 \)
- \( h_2 \)
- \( h_3 \)
overestimate \leq \frac{2 \left( \text{\# total increments} \right)}{\text{\# Counters per hash}}

With probability \(1 - \left(\frac{1}{2}\right)^\text{\# hashes}\)
1 Million Increments

overestimate \leq 100 \rightarrow 20,000 \text{ counters per hash}

probability .999 \rightarrow 10 \text{ hashes}

20,000 \text{ counters} \times 10 \text{ hashes} \times 4 \text{ bytes} \leq 800 \text{ KB} !!!
A Count-Min Sketch Provides

... an overestimate of the count

... that is more meaningful for the frequent items ("heavy hitters") in your data
Final Thoughts
Bloom Filter
Count-Min Sketch
Hyper Log Log
T-Digest

...
Use Sketches When

Large data $\rightarrow$ Small summary

Unbounded stream $\rightarrow$ Bounded memory
Embrace Randomness
Sketch Away!

Adam Marcus
B12
@marcua