Analysis of Hashing

COMPSCI 230 — Discrete Math

April 20, 2017
Outline

1 Probabilistic Analysis of Hashing
Clicker Test

I know that the final exam is on Saturday, May 6th, starting at 9am in Physics 128, that it lasts 90 minutes, and that topics coverage and other rules are spelled out in the sample exam.

Pick one  (Not Graded)

A: Yes
B: That short? I want it longer, with more questions
C: Where is Physics 128?
   My chauffeur just drives me places
D: Sample? What sample?
E: Exam? What exam?
Python

A list comprehension was used to make a list of 10 empty lists:

\[ T4 = [[] \text{ for } b \text{ in } \text{range}(10)] \]

Here is an alternative:

\[ T4 = [[]] * 10 \]

The alternative ...

(Not Graded)

A: is exactly equivalent to a list comprehension
B: works like a list comprehension, but is slower
C: works like a list comprehension, and is faster
D: would effectively return a single bucket
E: would return a single empty list, rather than a list of empty lists
A Hash Table with 10 Buckets

grade = [[] for b in range(10)]
insert('Smith', [87, 80, 93], grade)
insert('Brown', [82, 90, 78], grade)
insert('Jones', [67, 74, 77], grade)
# collision (different keys with same hash value)
Tradeoffs on \( b \)

- Designing \( h \) is nontrivial (needs to lead to low collision: look like “uniform at random”)
- With enough buckets (\( b \approx n \)), collision has low probability (much more about this)
- Can still have \( b \approx n \) and \( b \ll \) possible number of names
- Number of names \( \approx 26^{\text{name length}} \), astronomical
- Empty buckets take up some space but not much (can also make buckets only when needed)
- **Insertion and lookup take constant time almost always**, and just a bit longer sometimes
- **How few buckets can we get away with and still have low collision probability?**
- This is a probabilistic question
Probabilistic Analysis of Hashing

• Assume \( n \) items and \( b \) buckets

• Modeling and warmup questions:
  • How to model hashing? (Sample space, probability)
  • Probability that some items collide?
  • How big must \( n \) be for given \( b \) so that \( \text{prob}(\text{at least one collision}) \geq 1/2 \)?
  • Probability that some buckets are empty?

• Expected...
  • ... number of items per bucket
  • ... lookup time
  • ... number of collisions
  • ... number of empty buckets
  • ... number of items until all buckets have some items
  • ... maximum number of items per bucket (hard)
Checking Our Intuition

With \( n = 1000 \) items and \( b = 100 \) buckets, how many items per bucket do we expect, on average?

Pick one (Not Graded)

A: 1
B: 5
C: 10
D: 50
E: 100
Checking Our Intuition

With $b = 1000$ buckets, how many items do we expect to have to insert, more or less, before it is more likely than not that some bucket has more than one item?

Pick one (Not Graded)

A: 10  
B: 40  
C: 100  
D: 400  
E: 500
Checking Our Intuition

When the table has one item per bucket on average, what fraction of the buckets is empty, on average?

<table>
<thead>
<tr>
<th>Pick one</th>
<th>(Not Graded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 0%</td>
<td></td>
</tr>
<tr>
<td>B: 3%</td>
<td></td>
</tr>
<tr>
<td>C: 7%</td>
<td></td>
</tr>
<tr>
<td>D: 16%</td>
<td></td>
</tr>
<tr>
<td>E: 37%</td>
<td></td>
</tr>
</tbody>
</table>
Modeling Hashing

• Each bucket index $B_k = h(key_k)$ is a simple experiment with sample space $𝕊_1 = \{0, \ldots, b - 1\}$
• $|𝕊_1| = b$ (number of buckets)
• To answer questions about the hash table, we consider a repeated experiment, because one table has multiple items in it
• Fixed $n$: each $n$-sequence of indices is an outcome
  • $B = (B_1, \ldots, B_n)$ with $B_k \in 𝕊_1$
  • $B \in 𝕊 = 𝕊_1^n$
  • There are $|𝕊| = b^n$ $n$-sequences on $𝕊_1$
• Unbounded $n$: each sequence is an outcome
  • $B = (B_1, B_2, \ldots)$ with $B_k \in 𝕊_1$
  • $B \in 𝕊 = 𝕊_1^\infty$ and $|𝕊| = 𝕀$
  • There are infinitely many sequences
• Probabilities computed as limits from the fixed-$n$ case
Probability Model

• Outcomes of the simple experiment $B_k$ are equally likely

$$\text{prob}(\{B_j = \beta\}) = \frac{1}{b} \quad \text{for} \quad \beta = 0, \ldots, b - 1$$

• Distinct simple experiments are mutually independent

$$j \neq k \rightarrow \text{prob}(\{B_j = \beta_j\} \cap \{B_k = \beta_k\}) = \text{prob}(\{B_j = \beta_j\}) \cdot \text{prob}(\{B_k = \beta_k\})$$

• Leads to a uniform distribution over $\mathbb{S}$ when $\mathbb{S}$ is finite:

$$\mathbb{S} = \mathbb{S}_1^n \rightarrow P(B) = \frac{1}{b^n}$$

• $b$-outcome Bernoulli trial when $\mathbb{S} = \mathbb{S}_1^\infty$
Summary of Hashing Model

- Simple experiment $B_k = h(\text{key}_k) \in S_1 = \{0, \ldots, b - 1\}$
- $|S_1| = b$

- Repeated experiment
  - Either $B = (B_1, \ldots, B_n) \in S = S_1^n$
  - $|S| = b^n$
  - or $B = (B_1, B_2, ...) \in S = S_1^\infty$
  - $|S| = c$

- Probability model
  - Uniform: $P(B_k = \beta_k) = \frac{1}{b}$
  - Independent: If $j \neq k$ then
    $$P(B_j = \beta_j \text{ and } B_k = \beta_k) = P(B_j = \beta_j) P(B_k = \beta_k)$$
  - $P(B) = \frac{1}{b^n}$ for any $n$-sequence or $n$-subsequence
Probability of Some Collision

- No collision, $D(n)$: All $n$ items hash to different buckets
- $P(D(n)) = 0$ for $n > b$
- $n \leq b$: Sequences are equally likely → compute $P(D(n))$ as
  $\#$(collision free sequences) / $\#$(all sequences)
- For $n \leq b$ and no collision, there are $b$ buckets for first item, $b - 1$
  for second ... $b - n + 1$ for $n$-th
- $b(b-1) \cdots (b-n+1) = (b)_n = \frac{b!}{(b-n)!}$
- $P(D(n)) = \frac{(b)_n}{b^n}$ for $n \leq b$ and zero otherwise
- $S(n)$: Some collision
- $P(S(n)) = 1 - P(D(n))$ for $n \leq b$ and 1 otherwise
- $b = 10$:

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S(n))$</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.28</td>
<td>0.496</td>
<td><strong>0.70</strong></td>
<td>0.85</td>
<td>0.94</td>
<td>0.98</td>
<td>0.996</td>
<td>0.9996</td>
<td>1</td>
</tr>
</tbody>
</table>

- With 10 buckets and at least 5 items, it is more likely than not
  that there is at least one collision
Checking Our Intuition

With $b = 10$ buckets and 5 item insertions, it is more likely than not that at least one collision occurs. How many insertions are needed for some collision to become more likely than no collision when there are $b = 1000$ buckets?

Pick one (Not Graded)

<table>
<thead>
<tr>
<th>Pick</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>40</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
</tr>
<tr>
<td>D</td>
<td>500</td>
</tr>
<tr>
<td>E</td>
<td>750</td>
</tr>
</tbody>
</table>
$P(\text{collision}) \geq 1/2$

- With $b = 10$ buckets, it takes $n = 5$ items to bring the probability that some collision occurs to at least 1/2
- What about other values of $b$? (Table computed numerically)

<table>
<thead>
<tr>
<th>$b$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>2</td>
</tr>
<tr>
<td>3-5</td>
<td>3</td>
</tr>
<tr>
<td>6-9</td>
<td>4</td>
</tr>
<tr>
<td>10-16</td>
<td>5</td>
</tr>
<tr>
<td>17-23</td>
<td>6</td>
</tr>
<tr>
<td>24-32</td>
<td>7</td>
</tr>
<tr>
<td>33-42</td>
<td>8</td>
</tr>
<tr>
<td>43-54</td>
<td>9</td>
</tr>
<tr>
<td>55-68</td>
<td>10</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>13</td>
</tr>
<tr>
<td>1000</td>
<td>38</td>
</tr>
<tr>
<td>10000</td>
<td>119</td>
</tr>
</tbody>
</table>

- Surprisingly few items are enough to make collisions likely in very large hash tables