Лекция 11. Основы теории вероятностей. Хеширование. Хеш-таблицы.

#вшпи #аисд #теория #структуры_данных

Автор конспекта: Гридчин Михаил

Вероятностное пространство

Def. *Вероятностное пространство событий* Ω - некоторое конечное множество.
Def. $\mathcal{P}$ - *вероятностная мера* — некоторая функция из 2^Ω в $[0, 1]$
такая, что:

1. $\mathcal{P}(\varnothing) = 0, \mathcal{P}(\Omega) = 1$
2. Пусть $A, B \in 2^\Omega$ и $A \cap B = \varnothing$, тогда $\mathcal{P}(A \cap B) = \mathcal{P}(a) \cdot \mathcal{P}(b)$ и $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(b)$.
   Def. *Вероятностное пространство* - тройка $(\Omega, \mathcal{F}, \mathcal{P})$, где Ω - пространство событий, $F = 2^\Omega$, $\mathcal{P}$ - вероятностная мера на Ω.

Пример. Рассмотрим классический шестигранный кубик. Пусть ω_i — событие, что на кубике выпало i. Соответственно $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$. Пусть кубик справедливый, то есть вероятность выпадения всех граней одинакова. Тогда $\mathcal{P}(\omega_i) = \frac{1}{6}$. Таким образом, корректно определены вероятности, например,
$\mathcal{P}(\text{выпало простое число}) = \mathcal{P}(\omega_2) + \mathcal{P}(\omega_3) + \mathcal{P}(\omega_5) = \frac{1}{2}$.

Определение случайное величины

Def. *Случайная величина* - функция Ω в $\mathbb{R}$.

Пример. Рассмотрим все тот же кубик и пусть $\xi(\omega_i) = i^2$. То есть функция, которая берёт количество точек на грани кубика и возводит в квадрат. Надо посчитать $\mathcal{P}(\xi < 30)$.
$\mathcal{P}(\xi < 30) = 1 - \mathcal{P}(\xi \geq 30) = 1 - \mathcal{P}(\omega_i \geq \sqrt{30}) = 1 - \mathcal{P}(\omega_6) = \frac{5}{6}$. Действительно, не подходит только 6, поэтому ответ $\frac{5}{6}$.

Независимость

Def. *Две случайные величины* ξ, η *независимы*, если
$\forall x, y \in \mathbb{R} \Rightarrow P(\xi = x, \eta = y) = P(\xi = x) \cdot P(\eta = y)$.

Пример. Независимо друг от друга бросаются два кубика. Надо посчитать вероятность того, что сумма квадратов выпавших значений больше 9. Пусть ξ - бросок первого, а η - бросок второго. Тогда надо посчитать вероятность того, что $\mathcal{P}(\xi^2 + \eta^2 > 9)$. Будем считать вероятность дополнения. $\mathcal{P}(\xi^2 + \eta^2 > 9) = 1 - \mathcal{P}(\xi^2 + \eta^2 \leq 9)$. Вспомним, что значения

квадратов: $1, 4, 9, 16, 25, 36$, то есть
устраивают только пары $(1, 1), (1, 4), (4, 1), (4, 4)$.

$$\mathcal{P}(\xi^2 + \eta^2 > 9) = 1 - \mathcal{P}(\xi^2 + \eta^2 \leq 9) = 1 - 4 \cdot \frac{1}{36} = 1 - \frac{1}{9} = \frac{8}{9}$$

Можно рассмотреть другую вероятностную модель

$$\Omega_1 = \{(i, j) \mid i, j \in \{1, 2, 3, 4, 5, 6\}\}$$

$$\Omega_2 = \{(i, j) \mid 1 \leq i \leq j \leq 6\}$$

В зависимости от вероятностного пространства получатся разные результаты, поэтому договариваться о выборе вероятностной модели нужно заранее.

Матожидание

Def. *Матожиданием* случайной величины ξ называют величину

$$\mathbb{E}\xi = \sum_{\omega \in \Omega} \xi(\omega) P(\omega)$$

Пример. Все ещё честный шестигранный кубик. Надо посчитать матожидание квадрата выпавших точек.

$$\mathbb{E}\xi = \sum_{\omega \in \Omega} \xi(\omega) P(\omega) = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{91}{6}$$

Свойства матожидания:

1. Пусть c - константа, тогда $\mathbb{E}c = c$.
2. Пусть $\alpha, \beta \in \mathbb{R}$, ξ, η - случайные величины. Тогда
3. $E(\alpha \cdot \xi + \beta \cdot \eta) = \alpha \cdot \mathbb{E}\xi + \beta \cdot \mathbb{E}\eta$.
4. Пусть ξ, η независимы. Тогда $\mathbb{E}(\xi\eta) = \mathbb{E}\xi \cdot \mathbb{E}\eta$.
5. $|\mathbb{E}\xi| \leq \mathbb{E}|\xi|$.
6. $\mathbb{E}^2\xi \leq \mathbb{E}\xi^2$

Дисперсия

Def. *Дисперсия* случайной величины $\mathbb{D}\xi = \mathbb{E}\left[(\xi - \mathbb{E}\xi)^2\right]$.
Утверждение. $\mathbb{D}\xi = \mathbb{E}\xi^2 - \mathbb{E}^2\xi$.
□

$$\mathbb{D}\xi = \mathbb{E}[(\xi - \mathbb{E}\xi)^2] = \mathbb{E}[\xi^2 - 2\xi \cdot \mathbb{E}\xi + \mathbb{E}^2\xi] = \mathbb{E}\xi^2 - 2\mathbb{E}^2\xi + \mathbb{E}^2\xi = \mathbb{E}\xi^2 - \mathbb{E}^2\xi$$

■

Direct addressing

Def. Пусть U - множество рассматриваемых объектов (например, все строки), тогда $h : U \to \{0, 1, \dots, k-1\}$ называется *хеш-функцией*.

Def. Элементы $x, y \in U$, где $x \neq y$ образуют *коллизию*, если $h(x) = h(y)$. Введем массив размера $|(h(U)|$ и будем класть булев флаг в ячейку $h(x)$, если элемент там есть.

Проблема: огромная память и коллизии.

Заведем массив определенного небольшого размера m и кладем элемент по индексу $h(x)\%m$.

Проблема: всё ещё коллизии.

Давайте хранить по индексу не элемент, а цепочку из данных.

Проблема: поиск в цепочке работает за линейное время.

Def. *Бакетом (bucket)* называют нечто, хранящее все элементы, образующие коллизию.

Simple Uniform hashing

Хочется как-то минимизировать максимальную длину цепочки, то есть добиться того, чтобы хеш-функция примерно равномерно раскидывала ключи по бакетам.
Пусть $\mathcal{H} = \{h : U \to \{0, 1, \dots, k-1\}\}$ - наше множество хеш-функций, и мы хотим уметь выбирать случайную хеш-функцию из него.

Note. Далее для анализа нам важно, что $U \subset \mathbb{N}$ и $|U| < \infty$, то есть $U = \{0, 1, \dots, n-1\}$.

В таких предположениях на самом деле можно сказать, что хеш-функция - отображение $\{0, 1, \dots, n-1\} \to \{0, 1, \dots, k-1\}$. Тогда мы можем определить случайную хеш-функцию как то, что образ для каждого ключа это случайный элемент из множества $\{0, 1, \dots, k-1\}$.

Проблема: Требуется $O(|U| \log |h(U)|)$ памяти на хранение такого отображения, поэтому это неприменимо на практике.

Def. Модель, описанная выше, называется *простое равномерное хеширование (simple uniform hashing)*.

Рассмотрим величину L_q - длина цепочки, отвечающая ключу q. Пусть таблица построена для различных ключей $k_1, \dots, k_n$, тогда:

$$L_q = \sum_i l(h(q) = k(k_i))$$

Посчитаем матожидание длины цепочки.

$$\mathbb{E}L_q = \mathbb{E}\sum_i l(h(q) = h(k_i)) = \sum_i P(h(q) = h(k_i))$$

Посчитаем вероятность выше.

$$P(h(x) = h(y)) = \begin{cases} 1, & x = y \\ \frac{1}{k}, & x \neq y \end{cases}$$

С учетом того, что все ключи ki выше различны, получаем, что

$$\mathbb{E}L_q = \mathbb{E}\sum_i l(h(q) = h(k_i)) = \sum_i P(h(q) = h(k_1)) \leq 1 + \frac{n-1}{k} \leq 1 + \frac{n}{k}$$

Def. *Коэффициентом загруженности (load factor)* называют величину $\alpha = \frac{n}{k}$.

Note. Из оценок выше мы для хеш-таблицы заведомо задаем какую-то константу C такую, что всегда $\alpha < C$. Тогда в среднем сложность операций выше составит $O(1)$.

Заметим, что в ходе обсуждения мы уже получили хеш-таблицу. А именно, это будет массив цепочек, где все операции работают за длину цепочки. Если соблюдать ограничение на load factor, то сложность всех операций составит O(1) в среднем. То есть получаем контейнер, способный делать вставку, удаление и поиск в среднем за O(1).

Теорема(б/д). Семейство хеш-функций $\mathcal{H}_{\alpha,\beta}$:

$$H = (\alpha x + \beta) \pmod p, \alpha \in \{1, \dots, n-1\}, \beta \in \{0, \dots, n-1\}, p \text{ - простое}, p > n$$

Гарантирует матожидание длины цепочки в хеш-таблице $O(1)$.

□
Пусть хеш-таблица имеет размер $m = n$. Тогда хеш-функция для ключа x определена как $h(x) = ((\alpha x + \beta) \mod p) \mod m$, где α, β, p - из условия теоремы.
Для двух различных ключей $x \neq y$ коллизия происходит, если

$$(\alpha x + \beta) \mod p) \mod m = ((\alpha y + \beta) \mod p) \mod m$$
$$(\alpha(x - y)) \mod p \equiv 0 \mod m$$

Обозначим $d = x - y \mod p$, где $d \neq 0$ (так как $x \neq y$ и $p > n$). Тогда коллизия возникает, если

$$\alpha d \mod p = km \quad \text{для некоторого целого } k$$

Поскольку $(m, p) = 1$, то $ad \equiv km \mod p$ имеет ровно одно решение $\alpha \in \mathbb{Z}_p$ для каждого k. Количество подходящих k, для которых $km < p$ равно $t = \lfloor \frac{p-1}{m} \rfloor$. Таким образом, существует t возможных значений k, каждому из которых соответствует ровно одно $\alpha \in \mathbb{Z}_p$. Но $\alpha \in \{1, \dots, n-1\}$, а не $\mathbb{Z}_p$. Обозначим через s количество таких α, которые удовлетворяют условию коллизии. Вероятность коллизии равна:

$$\mathcal{P} = \frac{s}{|\mathcal{H}|} = \frac{s}{n \cdot (n-1)}$$

Но поскольку коллизия не зависит от β, количество пар (α, β), приводящих к коллизии, ровно $s \cdot n$. Тогда вероятность коллизии:

$$P = \frac{s}{n-1}$$

Оценим s. Так как $p > m$ и $p < 2m$, то $t = \lfloor \frac{p-1}{m} \rfloor < 2$. Это означает, что $s \leq 2$, так как каждое k даёт ровно одно α, и интервал $[1, n-1]$ содержит не более двух подходящих значений. Поэтому:

$$P \leq \frac{2}{n-1} \leq \frac{2}{n}$$

Матожидание длины цепочки равно

$$\mathbb{E}L = 1 + (n-1) \cdot P < 1 + (n-1) \cdot \frac{2}{n} < 3$$

Таким образом, математическое ожидание длины цепочки ограничено константой 3, то есть $O(1)$.
■