Семинар 10

Хеш-функции. Фильтры Блума

Используем открытую адресацию для борьбы с коллизиями

Каждый индекс в Х-Т вычисляется с помощью хеш функции, которая зависит от ключа и номера i в зондажной последовательности $(i = 0, 1, \dots, n-1)$. Например,

- линейное зондирование:

$$h(k, i) = (h'(k) + i) mod\ n, h'(k)$$ - вспомогательная хеш-функция

- Двойное зондирование:

$$h(k, i) = \big(h_1(k) + ih_2(k)\big) mod\ n$$

Вставить ключ в Х-Т

```
Hash-Insert(A, k)
 for i:=0 to n-1 do
     j:=h(k, i)
     if A[j]=nil  then   // ячейка с номером  j пуста
        A[j]:=k      // вставим ключ в эту ячейку
     return j
 Error «переполнение таблицы»   // i=n
```

Коэффициент заполнения Х-Т $\alpha = \frac{|S|}{n}$

Пусть $\alpha < 1$. Можно доказать, что матожидание количества исследований неудачного поиска и вставки при условии равномерного хеширования $\leq \frac{1}{1-\alpha}$

Какие хеш-функции можно считать хорошими, а какие – плохими?

Пример

ТЕСТОВОЕ ЗАДАНИЕ

Рассмотрим хеш-таблицу длиной $n \geq 1$, и пусть h равно хеш-функции, где $h(k) = 0$, для каждого ключа $k \in U$. Предположим, что набор данных S вставлен в хеш-таблицу, где $|S| \leq n$. Каково типичное время выполнения последующих операций Просмотреть?

а) $\Theta(1)$ со сцеплением, $\Theta(1)$ с открытой адресацией.

б) $\Theta(1)$ со сцеплением, $\Theta(|S|)$ с открытой адресацией.

в) $\Theta(|S|)$ со сцеплением, $\Theta(1)$ с открытой адресацией.

г) $\Theta(|S|)$ со сцеплением, $\Theta(|S|)$ с открытой адресацией.

Плохая!

Коллизии неизбежны. При паталогическом наборе данных O(1) →O(n)

Хеш-функция должна быть построена так, чтобы выбор позиции происходил независимо от других и равномерно.

Случайная хеш-функция

- Возьмем h(k) – случайную хеш-функцию с равномерным распределением по позициям.

Почему нецелесообразно использовать совершенно случайный выбор хеш-функции? (Выберите все подходящие варианты.)

а) На самом деле это практично.

б) Выбор не является детерминированным.

в) Ее хранение займет слишком много места.

г) Ее оценивание займет слишком много времени.

Фильтр Блума

- Подобно Х-Т позволяет выполнять операции «*Просмотреть*» и «*Вставить*», причем за О(1).
- Имеем битовый массив длины n и m хеш-функций $h_1, h_2, \dots, h_m$.

«*Просмотреть*» : по ключу k вернуть «да», если k был ранее вставлен в фильтр Блума, и «нет» - в противном случае.

«*Вставить*» : добавить новый ключ k в фильтр Блума.

Пример:

- Пример: фильтр Блума хранит множество из 3 объектов $\{x, y, z\}$, количество хеш-функций $m = 3, n = 18$.

x

0	1	0	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*

y

z

$h_1(x) = 1$ $h_2(y) = 4$ $h_3(z) = 3$

$h_2(x) = 5$ $h_2(y) = 11$ $h_3(z) = 5$

$h_3(x) = 13$ $h_2(y) = 16$ $h_3(z) = 11$

Пусть требуется найти w, у которого $h_1(w) = 4$, $h_2(w) = 13$, $h_3(w) = 15$

Т.к. один из битов B(15)=0, то w <u>не входит</u> в множество (Верно)

Пусть требуется найти t, у которого $h_1(t) = 4$, $h_2(t) = 13$, $h_3(t) = 16$. Все три бита установлены в 1, но объект не входит в множество$\{x, y, z\}$. <u>Ложное срабатывание</u>

Вероятность ложноположительного срабатывания

Эвристический анализ *— это подход к открытию и разрешению проблем, основанный на правилах, анализе, оценках и обоснованных предположениях.*

Примем следующие допущения:

1. Для каждого ключа $k \epsilon U$ в совокупности данных и хеш-функции h_i фильтра Блума значения $h_i(k)$ равномерно распределены, причем каждая из n позиций в массиве является равновероятной.
2. Все $h_i(k)$, охватывающие все ключи $k \epsilon U$ и хеш-функции $h_1, h_2, \dots, h_m$, являются независимыми случайными величинами.

Пусть $q, r \in \{0, 1, \dots, n-1\}$.

Тогда (1) $\Rightarrow P(h_i(k) = q) = \frac{1}{n}$, из (2) $\Rightarrow P(h_i\ (k_1) = q, h_j(k_2) = r)) = \frac{1}{n^2}$

Тестовое задание

Предположим, что набор данных S вставляется в фильтр Блума, который использует m хеш-функций и битовый массив длины n. В рамках наших эвристических допущений какова вероятность того, что первый бит массива равен 1?

а) $(1/n)^{|S|}$

б) $(1 - 1/n)^{|S|}$

в) $(1 - 1/n)^{m|S|}$

г) $1 - (1 - 1/n)^{m|S|}$

Хотя бы одна из m функций установит 1-ый бит в 1.

Рассмотрим $\left(1 - \frac{1}{n}\right)^{m|S|}$

$e^x - 1 \backsim x$ (x мало) $\Rightarrow e^x \backsim 1 + x$

Возьмем $x = -\frac{1}{n} \Rightarrow 1 - \frac{1}{n} \backsim e^{-\frac{1}{n}}$

$$\left(1 - \frac{1}{n}\right)^{m|S|} \backsim \left(e^{-\frac{1}{n}}\right)^{m|S|};\ 1 - \left(1 - \frac{1}{n}\right)^{m|S|} \backsim 1 - e^{-\frac{m|S|}{n}}$$

Итак,

Заменим $1-\left(1-\frac{1}{n}\right)^{m|S|}$ (вероятность того, что данный бит будет установлен в 1) на $1-e^{-\frac{m|S|}{n}}$. Обозначим $\frac{n}{|S|}=b$ – количество бит в расчете на 1 вставку.

$$1-e^{-\frac{m|S|}{n}}=1-e^{-\frac{m}{b}}$$

Оценим частоту ложных срабатываний $P_{Л}$. Это случается, когда $k\notin S$, но все m бит $h_1(k),h_2(k),\ldots,h_m(k)$ установлены в 1.

$$P_{Л}=\left(1-e^{-\frac{m|S|}{n}}\right)^m.$$

Найдем, при каком числе хеш-функций m $P_{Л}$ будет минимальным (b – фиксировано).

$$y=\left(1-e^{-\frac{m}{b}}\right)^m.$$

Вычислим производную с помощью логарифмического дифференцирования:

$$y = \left(1 - e^{-\frac{m}{b}}\right)^{m}$$

1) $\ln y = m \ln\left(1 - e^{-\frac{m}{b}}\right)$

2) $(\ln y)' = \ln\left(1 - e^{-\frac{m}{b}}\right) + \frac{m\, e^{-\frac{m}{b}}}{b\left(1 - e^{-\frac{m}{b}}\right)}$

3) $y' = y \cdot \left(\ln\left(1 - e^{-\frac{m}{b}}\right) + \frac{m\, e^{-\frac{m}{b}}}{b\left(1 - e^{-\frac{m}{b}}\right)}\right) = 0$

$$\ln\left(1 - e^{-\frac{m}{b}}\right) = -\frac{m\, e^{-\frac{m}{b}}}{b\left(1 - e^{-\frac{m}{b}}\right)}, \qquad t = e^{-\frac{m}{b}}, -\frac{m}{b} = \ln t$$

$$\ln(1 - t) = \frac{\ln t \cdot t}{1 - t}, \qquad (1 - t)\ln(1 - t) = t \ln t$$

$$\ln(1 - t)^{1-t} = \ln t^{t}$$

Подбором $t = \frac{1}{2} \Rightarrow \ln\frac{1}{2} = -\frac{m}{b} \Rightarrow m_{\text{опт}} = b \ln 2$

$$P_{\text{л}}(\text{опт}) = \left(1 - e^{\ln\frac{1}{2}}\right)^{m_{\text{опт}}} = \left(\frac{1}{2}\right)^{b \ln 2} = 0{,}6185^{b}$$

Примеры

- Вычислить $P_{Л}$ уже существующего фильтра Блума, который начал работать некорректно. Емкость фильтра $n = 3$ МгБ, со временем он хранит $|S| = 10^7$ элементов и использует 2 хеш-функции. (*m*=2)

Решение: $b = \frac{3 \cdot 10^6 \cdot 2^3}{10^7} = \frac{24}{10}$ ($b \approx 2.5$ бита на элемент)

$P_{Л}$= $\left(1 - e^{-\frac{m}{b}}\right)^m = \left(1 - e^{-\frac{2 \cdot 10}{24}}\right)^2 = \left(1 - \left(\frac{1}{e}\right)^{\frac{5}{6}}\right)^2$ =32%

При $m = \frac{24}{10} \ln 2 = 1.66 \approx 2$ Т.е. вероятность ложного срабатывания в самом благоприятном случае очень высока.

При известной $P_{Л}$ и размере набора данных S легко вычислить оптимальный размер фильтра.

$$P_{Л} = \left(\frac{1}{2}\right)^{\frac{n}{|S|}\ln 2} \Rightarrow n = -\frac{\ln P_{Л} \cdot |S|}{(\ln 2)^2}$$

- Каков оптимальный размер фильтра Блума, если в *S* содержится миллиард элементов при вероятности ложноположительного результата в 1%?

Получим около 10 млн битов (1,25 гБ)