Notas sobre o "hash" perfeito

• "Hash" perfeito em espaço $O(n^2)$

• Definição

Uma função $h: U \to A$ diz-se uma função de "hash" perfeita para um conjunto dado $N \subseteq U$, se h restrita a N for injectiva, isto é, se não houver colisões entre os elementos de N.

Exemplos de aplicação

- Dicionários para palavras chave de uma linguagem de programação.
- Thesaurus

• Construção com espaço $O(n^2)$

Tabela de "hash" de tamanho n^2: é muito fácil definir uma função de "hash" perfeita usando um algoritmo aleatorizado de tempo polinomial:

> Com probabilidade pelo menos 1/2 esse algoritmo produz uma função de "hash" perfeita. Repetindo o algoritmo um determinado número de vezes pode conseguir-se, em tempo da mesma ordem de grandeza, "hash" perfeito com probabilidade arbitrariamente próxima de 1.

Como a tabela é grande, a probabilidade não existir qualquer colisão é grande.

Exercício 1 São gerados k inteiros aleatórios, de forma unforme, entre 1 e n. Até que valor de k, a probabilidade de não existirem 2 inteiros gerados iguais é superior a 1/2? □

Exercício 2 Numa sala estão 25 pessoas. Qual a probabilidade de pelo menos 2 dessas pessoas fazerem anos no mesmo dia? Calcule essa probabilidade de forma explícita (e não através de uma expressão). Use os seguintes axiomas: as pessoas nascem com igual probabilidade em todos os dias do ano; nenhum ano é bissexto. □

Teorema

Seja $h(\cdot)$ uma função de "hash" definida através de uma matriz aleatória H correspondente a uma tabela de "hash" com n^2 elementos (ver "hash" universal). A probabilidade de existir pelo menos uma colisão é inferior a 1/2.

Dem.

Há $\binom{n}{2}$ pares (x, y) de elementos do conjunto a representar. A probabilidade de colisão de um qualquer par específico é não superior a $1/a$. Portanto, a probabilidade p de existir pelo menos uma colisão satisfaz (note-se que $a = n^2$)

$$p \leq \binom{n}{2}/a = \frac{n(n-1)}{2n^2} < \frac{1}{2}$$

□

• "Hash" perfeito em espaço $O(n)$

Método relativamente simples, elegante e recente: "hash" a 2 níveis. o "hash" do primeiro nível não é (excepto se tivermos uma sorte excepcional) perfeito.

Construção do "hash" perfeito

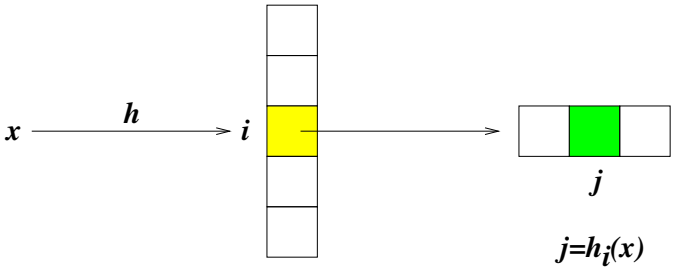
1. "Hash" universal: define-se uma função h de "hash" (do primeiro nível)[1] com uma tabela de tamanho n, $a = n$.
2. Inserem-se todos os n elementos na tabela.
3. Segundo nível de "hash", é construída para cada cadeia uma função de "hash" universal e perfeito com espaço $O(n_i^2)$. Sejam h_1, h_2,..., h_n as respectivas funções de "hash" e A_1, A_2,..., A_n as respectivas tabelas. Note-se que a tabela A_i tem tamanho n_i^2 em que n_i é o número de elementos da cadeia i, isto é, elementos x com $h(x) = i$.

[1] Que em geral não é perfeito.

Algoritmo de pesquisa

Dado x, retorna `V` of `F`.

1. Calcula-se $i = h(x)$
2. Calcula-se $j = h_i(x)$
3. Se $A_i[j]$ está ocupado com x, retorna-se `V`, senão (se estiver vazio ou contiver um valor diferente de x) retorna-se `F`



Notas

- Vamos mostrar que este algoritmo é $O(1)$.
- As cadeias – elementos com o mesmo valor $h(x)$ – têm muitas vezes 1 ou nenhum elemento; assim, é preferível com vista a poupar espaço e tempo, não construir tabelas de "hash" para esses casos, tratando-os de modo especial.

Correcção do algoritmo de construção do "hash" perfeito

Já verificamos que as funções de "hash" perfeito do segundo nível podem ser eficientemente construídas em espaço quadrático. Falta mostrar que o espaço gasto pelas tabelas A_i do segundo nível é $O(n)$.

Teorema

Se h é uma função uniformemente escolhida de um conjunto universal,

$$\text{prob}\left\{\sum_{1\le i\le a} n_i^2 > 4n\right\} < \frac{1}{2}$$

Dem.

Basta mostrar que $E(\sum_{1\le i\le a} n_i^2) < 2n$; na verdade seja a variável aleatória $Y = \sum_{1\le i\le a} n_i^2$. Pela desigualdade de Markov se $E(y) < 2n$, então prob $(Y > 4n) < 1/2$.

Se contarmos o número total de pares que colidem no "hash" do primeiro nível, incluindo as colisões dos elementos consigo próprios, vamos ter $\sum_{1\le i\le a} n_i^2$. Por exemplo, se a, b e c são os elementos de uma das cadeias, temos $3^2 = 9$ colisões que correspondem a todos os pares (x, y) com x e y pertencentes a $\{a, b, c\}$. Assim, e definindo-se c_{xy} como sendo 1 x e y colidem e 0 caso contrário, temos

$$\begin{array}{lcll} E(\sum_i n_i^2) & = & \sum_x \sum_y E(c_{xy}) & \\ & = & n + \sum_x \sum_{y\neq x} E(c_{xy}) & \\ & \le & n + n(n-1)/a & \text{porque } h \text{ é universal} \\ & \le & n + n(n-1)/n & \text{porque } a = n \\ & < & 2n & \end{array}$$

□