**Algoritmi e Strutture Dati**

# Hashing Universale

Docente: *Nicolò Cesa-Bianchi* versione 4 gennaio 2013

Sia $\mathcal{U}$ un insieme con $|\mathcal{U}| = m \gg 1$. Consideriamo una tabella hash $T$ di dimensione $n > 1$ ed una funzione di hash $h : \mathcal{U} \to \{0, \ldots, n-1\}$. Supponiamo di memorizzare $S \subseteq \mathcal{U}$ nella tabella hash usando la funzione $h$. Ovvero, memorizziamo ogni $u \in S$ nella posizione $h(u)$ di $T$. In un'analisi di caso peggiore, ogni ulteriore operazione (ricerca, inserimento o cancellazione) sulla tabella richiederà tempo $\ell_{\max}$, dove $\ell_{\max}$ è il massimo numero di elementi in $S$ che sono stati mappati da $h$ nella stessa posizione di $T$. Più precisamente,

$$\ell_{\max} = \max_{i=0,\ldots,n-1} \big|\{u \in S \,:\, h(u) = i\}\big| \,.$$

Idealmente, vorremmo avere $\ell_{\max} \simeq |S|/n$, in modo da minimizzare il tempo di esecuzione di un'operazione nel caso peggiore. Vediamo ora come ottenere questo risultato, in valore atteso, tramite la tecnica dello hashing universale.

Una collezione $\mathcal{H}$ di funzioni $h : \mathcal{U} \to \{0, \ldots, n-1\}$ è detta una *famiglia universale di funzioni hash* quando

1. Per ogni $u, v \in \mathcal{U}$ vale $\mathbb{P}\big(H(u) = H(v)\big) \leq \frac{1}{n}$ dove la probabilità è calcolata rispetto all'estrazione casuale della funzione hash $H$ da $\mathcal{H}$.
2. Ogni $h \in \mathcal{H}$ può essere rappresentata con al più $\Theta(\log m)$ bit e calcolata in modo efficiente.

Si noti che la condizione 1 è sufficiente a garantire $\ell_{\max} \leq |S|/n$ in valore atteso rispetto all'estrazione di $H$ da $\mathcal{H}$. Infatti, supponiamo di aver estratto $H$ a caso da $\mathcal{H}$ e di averla usata per inserire $S = \{u_1, \ldots, u_s\}$ nella tabella. Sia $u \in \mathcal{U}$ un elemento arbitrario che vogliamo aggiungere alla tabella. Definiamo le variabili casuali $X_1, \ldots, X_s$ dove

$$X_i = \begin{cases} 1 & \text{se } H(u) = H(u_i) \\ 0 & \text{altrimenti.} \end{cases}$$

Allora, il numero di elementi di $S$ che sono mappati nella stessa posizione di $u$ è dato da

$$\sum_{i=1}^{s} X_i \,.$$

Il valore atteso di questo numero è calcolato come

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{s} X_i\right] &= \sum_{i=1}^{s} \mathbb{E}\, X_i \quad \text{per linearità del valore atteso} \\ &= \sum_{i=1}^{s} \mathbb{P}\big(H(u) = H(u_i)\big) \quad \text{per definizione di } X_i \\ &\leq \sum_{i=1}^{s} \frac{1}{n} \quad \text{per ipotesi su } H \\ &= \frac{|S|}{n}\ .\end{aligned}$$

Ma è possibile trovare classi $\mathcal{H}$ che soddisfano la condizione 1? Per verificarlo, assumiamo per semplicità che $\mathcal{U} = \{1, \ldots, m\}$ e consideriamo la classe $\mathcal{H}$ di tutte le funzioni $h : \{1, \ldots, m\} \to \{0, \ldots, n-1\}$. Ognuna di queste funzioni hash corrisponde a un vettore $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m) \in \{0, \ldots, n-1\}^m$ in modo che $h(u) = \boldsymbol{h}_u$. Allora, il numero di vettori $\boldsymbol{h} \in \{0, \ldots, n-1\}^m$ tali che $\boldsymbol{h}_u = \boldsymbol{h}_v$ è $n^{m-1}$. Quindi, se estraggo $H$ a caso da $\mathcal{H}$,

$$\mathbb{P}\big(H(u) = H(v)\big) = \frac{n^{m-1}}{n^m} = \frac{1}{n}$$

e la condizione 1 è soddisfatta. D'altra parte, la condizione 2 non è soddisfatta in quanto mi servono $\log |\mathcal{U}| = m \log n$ bit per descrivere ciascuna $h \in \mathcal{H}$. Vediamo ora un modo per costruire $\mathcal{H}$ che soddisfi 1 e 2.

Per semplicità, assumiamo $\mathcal{U} = \{1, \ldots, m\}$ e anche $n = p$ per un qualche $p$ primo. Rappresentiamo ciascun elemento $u \in \{1, \ldots, m\}$ come un numero $[u]_p$ in base $p$. Ovvero, $[u]_p = x = (x_1, \ldots, x_r)$ dove $x_i \in [0, \ldots, p-1]$ e $r$ è il più piccolo intero tale che $p^r \geq m$. Ovvero,

$$r = \left\lceil \frac{\log m}{\log p} \right\rceil\ .$$

Sia $\mathcal{A} = \{0, \ldots, p-1\}^r$ l'insieme usato per rappresentare gli elementi di $\mathcal{U}$. Introduciamo ora la famiglia di funzioni hash $\mathcal{H} = \{h_a : a \in \mathcal{A}\}$ di tipo $h_a : \mathcal{A} \to \{0, \ldots, p-1\}$ e definite come

$$h_a(x) = \left(\sum_{i=1}^{r} a_i x_i\right) \mod p\ .$$

Si noti che posso rappresentare ogni $h_a$ con $\Theta(\log m)$ bit, quindi la condizione 2 è soddisfatta. Per verificare la condizione 1 utilizziamo il lemma seguente.

**Lemma 1** *Per ogni primo $p$, per ogni $\alpha, \beta$ e $z$ interi,*

$$z \not\equiv 0 \mod p \quad e \quad \alpha z \equiv \beta z \mod p \quad \textit{implicano} \quad \alpha \equiv \beta \mod p\ .$$

Dimostrazione. Chiaramente, $\alpha z \equiv \beta z \mod p$ se e solo se $z(\alpha - \beta) \equiv 0 \mod p$. Inoltre, $z \not\equiv 0 \mod p$ e $z(\alpha - \beta) \equiv 0 \mod p$ implicano $(\alpha - \beta) \equiv 0 \mod p$. Infatti, se $p$ è primo e $z$ non contiene $p$

come fattore, allora $\alpha - \beta$ lo deve contenere (se invece $p$ non fosse primo, allora $z$ e $\alpha - \beta$ potrebbero spartirsi i fattori primi di $p$). □

Siamo ora pronti a dimostrare che per $\mathcal{H}$ vale la proprietà 1. Dato che, come abbiamo già osservato, per la stessa classe valeva anche la proprietà 2, concludiamo che $\mathcal{H}$ è una famiglia universale di funzioni hash.

**Teorema 2** *$\mathcal{H}$ è tale che per ogni $x, y \in \mathcal{A}$ distinti, la frazione di elementi $a \in \mathcal{A}$ tali che $h_a(x) = h_a(y)$ è al più $\frac{1}{p}$. Quindi, se $H$ è estratta a caso da $\mathcal{H}$, allora*

$$\mathbb{P}\big(H(x) = H(y)\big) \leq \frac{1}{p} \ .$$

Dimostrazione. Se $x \neq y$ allora esiste almeno una coordinata $j \in \{0, \ldots, r\}$ tale che $x_j \neq y_j$. Per estrarre $H$ a caso in $\mathcal{H}$ prendiamo $a \in \mathcal{A}$ estraendo a caso ciascuna coordinata $a_i \in \{0, \ldots, p-1\}$. Per qualunque estrazione dei valori $a_i$ sulle coordinate $i \neq j$, abbiamo che

$$\begin{aligned} h_a(x) = h_a(y) &\iff \left( \sum_{i=1}^{r} a_i x_i \equiv \sum_{i=1}^{r} a_i y_i \right) \mod p \\ &\iff a_j \underbrace{(x_j - y_j)}_{z} \equiv \underbrace{\sum_{i \,:\, i \neq j} a_i (x_i - y_i)}_{k} \mod p \\ &\iff a_j z \equiv k \mod p \ . \end{aligned}$$

Dimostriamo ora che esiste un unico $a_j \in \{0, \ldots, p-1\}$ tale che la congruenza $a_j z \equiv k \mod p$ sia soddisfatta. Per assurdo, supponiamo che esistano $a_j, a'_j$ distinti che soddisfano entrambi la congruenza. Allora avremmo anche che $a_j z \equiv a'_j z \mod p$. Dato che $z \not\equiv 0 \mod p$ per ipotesi, il lemma implica che $a_j \equiv a'_j \mod p$. Siccome $0 \leq a_j, a'_j < p$, ciò implica che $a_j = a'_j$ e si ha una contraddizione. Quindi c'è un solo valore della coordinata $j$ che rende $h_a(x) = h_a(y)$. La probabilità di estrarre questo valore è al più $\frac{1}{p}$. Se $x$ e $y$ differiscono su altre coordinate, allora la probabilità che $h_a(x) = h_a(y)$ diventa ancora più piccola. □