Maximal repetition & “Runs” conjecture

Riccardo Lo Iacono
Docente: Gabriele Fici

13 novembre 2024

Notazione

1. Σ è un alfabeto ordinato e finito di simboli
2. Un elemento $S \in \Sigma^*$ è detto stringa, la cui lunghezza è denotata da $|S|$
3. Per ogni i, $1 \leq i \leq n = |S|$, S[i] indica l'i-esimo carattere in S.
4. Per ogni i, j, $1 \leq i \leq j \leq n = |S|$, S[i, j] indica la sottostringa compresa tra le posizioni i e j,

Periodo ed esponente

Definizione: Data S una stringa, si definisce *periodo* un intero $p \geq 1$ tale che S[i] = S[i + p], $\forall i = 1, \dots, |S| - p$

Definizione: Si definisce *esponente* exp il rapporto tra la lunghezza di S e del suo più piccolo periodo.

Definizione: Una coppia (i, j) è detta ripetizione massimale (o run) di una stringa S, se $\exp(S[i, j]) \geq 2p$ e la periodicità non può essere estesa ne a destra ne a sinistra.

Un esempio

Sia considerata la stringa $S = babbabbababbabbabc$; questa contiene nove run, mostrati in *Figura 1*.

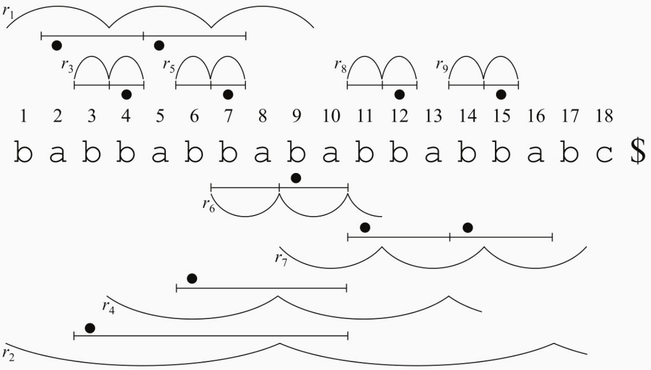

Figura 1: Illustrazione dei run nella stringa S = babbabbababbabbabc.

Background storico

Sia $\rho(S)$ il numero dei run in una stringa S.

Kolpakov e Kucherov[1] dimostrano che $\rho(S) = \mathcal{O}(|S|)$.

Congettura (Runs conjecture): $\rho(S) < |S|$ per ogni stringa S.

[1] R. M. KUCHEROV AND G. KOLPAKOV, *Finding maximal repetitions in a word in linear time. FOCS, 1999*

Punti chiave della discussione

1. Dimostrazione della runs conjecture.
2. Applicazione della soluzione algoritmica.

Sia $\prec$ l'ordine totale definito su Σ, e si indichi con lo stesso l'ordine indotto su Σ^*.

Assunto $\prec$ su Σ, si può definire $\tilde{\prec}$ l'ordine rovesciato. Ossia, un'ordine tale per cui presi $a, b \in \Sigma$, se

$$a \prec b \implies b \tilde{\prec} a$$

Siano $\prec_0, \prec_1$ rispettivamente l'ordine indotto su Σ^* e l'ordine rovesciato.

Data S una stringa e un indice i, $1 \leq i \leq n$, la stringa S[i, n]S[1, i] è detta *shift ciclico* di S (se i > 1 *shift ciclico non banale*).

Definizione (Lyndon word): Una stringa $S \in \Sigma^*, S \neq \varepsilon$, è detta Lyndon word se S è minore di ogni suo shift ciclico non banale.

Congettura: $\rho(S) < |S|$, per ogni stringa S.

Dimostrazione: Sia S una stringa di lunghezza n, e sia (i, j), $0 \leq i < j < n$, un suo run con periodo p minore.

Se j + 1 < n, ed inoltre S[j - p + 1] $\prec_0$ S[j + 1], al run si assegna un indice k tale per cui S[k, j] è il *longest proper suffisso* di S[i, j]. Viceversa k definisce il più lungo suffisso per $\prec_1$.

Si noti che se k > i allora k > 0, e S[k, j] contiene un intero periodo del run. Inoltre, S[k, k + p - 1] è il più lungo coniugato di S[i ,i + p - 1] rispetto i due ordini. Ciò implica l’essere *border-free* che è una proprietà nota delle Lyndon word.

Si intende dunque dimostrare che ogni k > 0 è posizione iniziale di al più un solo *longest proper suffix* in un run.

Siano quindi considerati due run: (i, j) e (k, l) rispettivamente un p-run e un q-run. Inoltre, sia assunto per assurdo che i loro longest proper suffix condividano la stessa posizione iniziale k.

Poichè run distinti non possono avere uno stesso periodo si può assumere $p \neq q$.

1. $j = l$, ma allora $S[k, j] = S[k, l]$. Assumendo per esempio $p < q$, $S[k, k + q - 1]$ ha periodo p, ma allora questa non è border-free, che è una contraddizione.
2. sia $j < l$, ragionamento analogo vale nel caso $j > l$, inoltre siano i suffissi i più lunghi rispetto un stesso ordine. Posto $d = S[j + 1]$, carattere successivo il p-run, per definizione si ha $S[j - p + 1] \prec d$. Inoltre, $S[k, k - p + 1] \prec S[k, j - p]d$. Ma da ciò segue che $S[k, l - 1]$ non è massimale come supposto, poichè $S[i + p, j]$ è fattore del q-run.

3. $j \neq l$ e i suffissi risultano essere massimali secondo ordini diversi. Senza perdita di generalita sia p < q e il suffisso del p-run sia massimale rispetto $\prec_0$. Poichè q > 1, si hanno S[k + q - 1] $\prec_0$ S[k] e S[k + q - 1] $\prec_1$ S[k - 1], da cui S[k - 1] $\prec_0$ S[k]. Si ha però che p $\not\geq$ 1 (S[k] $\prec_0$ S[k - 1]) e p $\neq$ 1 (S[k - 1] = S[k]). Da cui un'altra contraddizione.

Ciò conclude la dimostrazione, mostrando come il numero di run non possa essere superiore agli n - 1 potenziali valori di k.