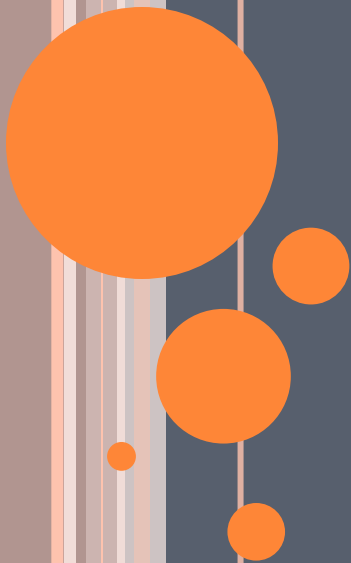


Advanced Data Indexing

(Προηγμένη ευρετηρίαση δεδομένων)

Ροές Δεδομένων
(3^ο Μέρος)

Άλλες Μέθοδοι για Ροές Αποτύπωμα



Αποτύπωμα

- Είναι μία μέθοδος τοποθέτησης διακριτών αναγνωριστικών ετικετών σε σύνολα **μεγάλων** αντικειμένων.
- **Στόχοι**
 - Αποφυγή αποθήκευσης ολόκληρων των αντικειμένων.
 - Γρήγορος και αποτελεσματικό έλεγχος ισότητας αντικειμένων.
 - Γρήγορη αναζήτηση αντικειμένων με βάση το αποτύπωμά τους.
- **Ιδιότητες**
 - Οι ετικέτες/αποτυπώματα πρέπει να είναι μικρές.
 - Διαφορετικά αποτυπώματα **πρέπει** να αντιστοιχούν σε διαφορετικά αντικείμενα.
 - Διαφορετικά αντικείμενα **πιθανώς να έχουν** διαφορετικά αποτυπώματα.



Η Δομή ενός Αποτυπώματος

- Ένα αποτύπωμα αποτελείται από δυαδικά ψηφία (bits).
- Ένα αποτύπωμα μεγέθους k αντιστοιχεί σε αντικείμενα με μέγεθος χώρου $N=2^k$.
- Για την αντιστοιχία ορίζεται μία οικογένεια συναρτήσεων του αποτυπώματος από τον χώρο των αντικειμένων Ω σε λέξη των k bits:
$$F = \{f : \Omega \rightarrow \{0,1\}^k\}$$
- Κάθε τυχαία συνάρτηση της οικογένειας $f \in F$ πρέπει να έχει τις εξής ιδιότητες:
 - $f(A) \neq f(B) \Rightarrow A \neq B$ (διαφορετικά αποτυπώματα αντιστοιχούν σε διαφορετικά αντικείμενα).
 - $P[f(A) = f(B) \mid A \neq B] \rightarrow 0$ (η πιθανότητα δύο διαφορετικά αντικείμενα να έχουν το ίδιο αποτύπωμα πρέπει να είναι κοντά στο 0 [ιδεατά να είναι $2^{O(-k)}$]).
- **Τυπική Περίπτωση:**
 - Αν έχουμε ένα σύνολο αντικειμένων S με $|S| = n \ll 2^k$
 - Για να πετύχουμε στην 2^η ιδιότητα αρκεί: $|f(S)| = |S|$ με μεγάλη πιθανότητα.
 - Έχουμε n^2 πιθανές συγκρούσεις ζευγαριών, άρα απαιτείται: $2^k > n^2$.

Παράδειγμα – URL Αποτυπώματα

- Εφαρμογή στις Μηχανές Αναζήτησης
 - Διαχείριση μεγάλου πλήθους URLs λέξεων.
 - Μεγάλες, μεταβλητού μεγέθους λέξεις.
- Τι μέγεθος αποτυπώματος απαιτείται;
 - 4 δις ιστοσελίδες $\rightarrow n=2^{32}$
 - $N \approx n^2 \rightarrow k=64$
 - Αποτύπωμα $\rightarrow$ αναπαράσταση με 8 bytes
- Πως ορίζονται όμως οι συναρτήσεις;



Ορισμός Συναρτήσεων Αποτυπώματος

- Δοθείσης μίας λέξης των m -bits: $A = a_m a_{m-1} \cdots a_1$ όπου $a_i \in \{0,1\}$, ένα αποτύπωμα των k -bits υπολογίζεται ως εξής:

1. Ορίζουμε το πολυώνυμο $A(t)$:

$$A(t) = a_m t^{m-1} + a_{m-1} t^{m-2} + \cdots + a_2 t + a_1$$

2. Επιλέγουμε ένα **ανάγωγο** πολυώνυμο (που δεν μπορεί να εκφρασθεί σαν γινόμενο παραγόντων) $P(t)$:

$$P(t) = p_k t^k + p_{k-1} t^{k-1} + \cdots + p_0$$

3. Υπολογίζουμε το αποτύπωμα $f(A)$ ως εξής:

$$f(A) = A(t) \bmod P(t)$$



Ιδιότητες

- Για λέξεις από **bit**, οι μόνες πράξεις που απαιτούνται για το $A(t) \bmod P(t)$ είναι η *ολίσθηση* και το **XOR**:

$$f(b_1, \dots, b_l) = r_k t^{k-1} + \dots + r_2 t + r_1$$

$$f(b_1, \dots, b_l, b_{l+1}) = r_{k-1} t^{k-1} + \dots + r_1 t + r_k (t^k \bmod P(t))$$

Αριστερή Ολίσθηση $f(b_1, \dots, b_l)$ $\oplus$ $P(t)$ χωρίς 1^ο bit

The diagram illustrates the calculation of the polynomial remainder. The expression $f(b_1, \dots, b_l, b_{l+1}) = r_{k-1} t^{k-1} + \dots + r_1 t + r_k (t^k \bmod P(t))$ is shown. Below it, the components are identified: $f(b_1, \dots, b_l)$ is the left-shifted polynomial, $\oplus$ is the XOR operation, and $P(t)$ without its first bit is the polynomial being subtracted. Arrows indicate the flow of information from the terms in the equation to their respective labels.

- Απαιτείται λοιπόν γραμμικός χρόνος υπολογισμού.



Παράδειγμα

- Δοθείσης της λέξης **A=101010**, θέλουμε να υπολογίσουμε ένα **4-bit** αποτύπωμα.
- Επιλέγουμε **P=11111** (ώστε το **P(t)** να είναι ανάγωγο).

$$\begin{array}{r} 1\ 0\ 1\ 0\ 1\ 0 \\ 1\ 1\ 1\ 1\ 1\ \oplus \\ \hline 1\ 0\ 1\ 0\ 1 \\ 1\ 1\ 1\ 1\ 1\ \oplus \\ \hline 1\ 0\ 1\ 0 \longrightarrow f(A) \end{array}$$



Παράδειγμα Εφαρμογής

- Ταίριασμα λέξεων: μπορούμε να βρούμε τη λέξη $y_1y_2\ldots y_m$ στην λέξη $x_1x_2\ldots x_n$ ($m < n$);

Αλγόριθμος Σύγκρισης ένα προς ένα:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\dots$	x_n
?	?	?	X						
y_1	y_2	y_3	y_4	$\dots$	y_m				

Πολυπλοκότητα Χρόνου: $O(m \cdot n)$



Παράδειγμα Εφαρμογής

- Ταίριασμα λέξεων: μπορούμε να βρούμε τη λέξη $y_1y_2\cdots y_m$ στην λέξη $x_1x_2\cdots x_n$;

Με χρήση αποτυπωμάτων

$$f(y_1y_2y_3y_4\cdots y_m)$$

$$f(x_1x_2\cdots x_m)_1)$$

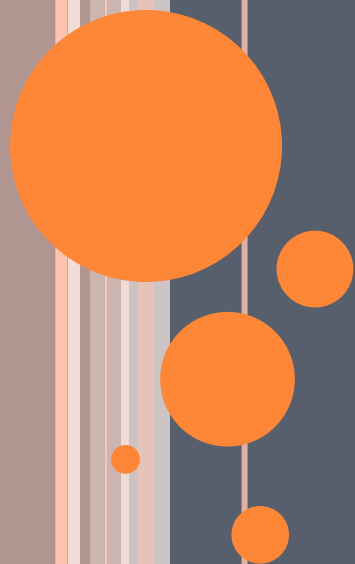
|?

Η πολυπλοκότητα χρόνου για τον υπολογισμό του $f(x_2x_3\cdots x_{m+1})$ από το $f(x_1x_2\cdots x_m)$ είναι $O(1)$

Πολυπλοκότητα Χρόνου: $O(m+n)$



Άλλες Μέθοδοι για Ροές Wavelets



Wavelets

- Στην επεξεργασία σήματος, τα **wavelets** χρησιμοποιούνται για να διασπάσουν το σύνθετο σήμα σε πιο απλά κομμάτια. Παρόμοια, τα **wavelets** χρησιμοποιούνται για να διασπάσουν μεγάλα σύνολα δεδομένων σε πιο απλές μορφές.
- Χρησιμοποιούνται σε διάφορες εφαρμογές ροών δεδομένων (χρονοσειρές, σήματα, ήχος, κλπ). Δίνουν μία σύνοψη των δεδομένων.
- Εφαρμόζονται σε ροές του μοντέλου των χρονοσειρών.
- Ο μετασχηματισμός **wavelet** δείχνει τις **κυρίαρχες τάσεις** στο σήμα (στα δεδομένα της ροής).
- Τα ανακατασκευασμένα δεδομένα από λίγους σημαντικούς **wavelet συντελεστές**, προσεγγίζουν με τον καλύτερο τρόπο τα αρχικά δεδομένα.



Wavelets

- Οι wavelet συντελεστές είναι **προβολές** του σήματος/των δεδομένων σε ένα **(ορθογώνιο) σύνολο διανυσμάτων βάσης**.
- Μία κατηγορία με μεγάλη εφαρμογή είναι τα **Haar wavelets** που χρησιμοποιούνται στις βάσεις δεδομένων λόγω ευκολίας υπολογισμού τους.
 - Οι υπολογισμοί γίνονται ιεραρχικά κατά επίπεδα (levels).
 - Στην πιο απλή τους μορφή υπολογίζεται ο μέσος όρος κατά ζεύγη για το επόμενο επίπεδο ($\frac{a+b}{2}$) [για ορθογώνια βάση: $\frac{a+b}{\sqrt{2}}$].
 - Οι συντελεστές προκύπτουν με αφαίρεση του πρώτου αριθμού του ζεύγους και του μέσου όρου που προέκυψε.
- **Έξοδος wavelet**: ένα διάνυσμα βάσης που αποτελείται από **έναν μέσο όρο** και από τους τελικούς **συντελεστές**.



Παράδειγμα Haar Wavelet

Έστω ότι έχουμε τα δεδομένα: **2, 2, 0, 2, 3, 5, 4, 4**.

Ανάλυση	Μέσοι Όροι	Αναλυτικοί Συντελεστές
Level-3	$[2, 2, 0, 2, 3, 5, 4, 4]$	----
Level-2	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
Level-1	$[1.5, 4]$	$[0.5, 0]$
Level-0	$[2.75]$	$[-1.25]$

Διάνυσμα wavelet μετασχηματισμού:

$[2.75, -1.25, 0.5, 0, 0, -1, -1, 0]$

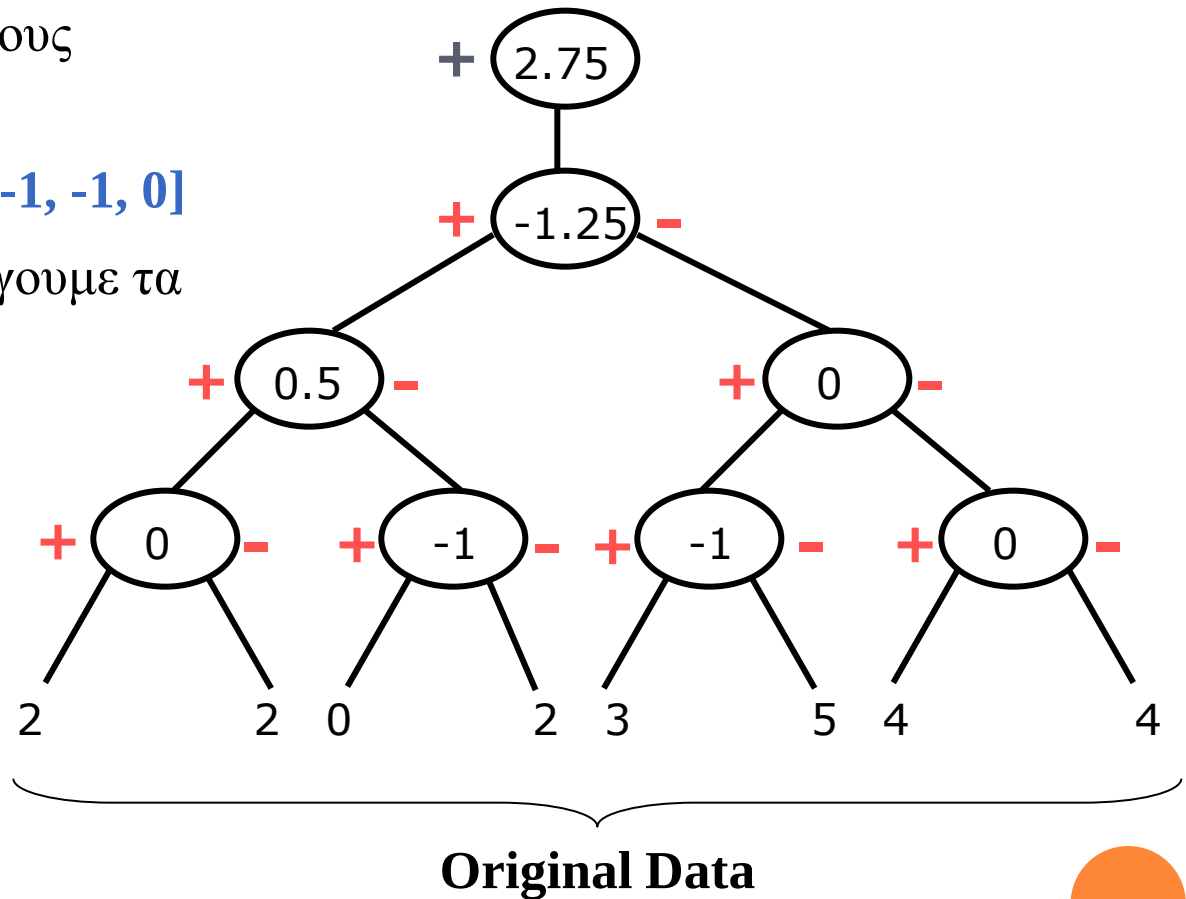


Παράδειγμα Haar Wavelet

- Χρησιμοποιώντας το διάνυσμα του μετασχηματισμού με τους συντελεστές:

$[2.75, -1.25, 0.5, 0, 0, -1, -1, 0]$

μπορούμε να αναπαράγουμε τα δεδομένα ως εξής:



Παράδειγμα για ορθογώνια βάση

Haar Wavelet Decomposition of Array $A = [1, 3, 5, 11, 12, 13, 0, 1]$

Level 3	1	3	5	11	12	13	0	1
Level 2	2.8284	11.3137	17.6777	0.7071	1.4142	4.2426	0.7071	0.7071
Level 1	10.0000	13.0000	6.0000	-12.0000				
Level 0	16.2635	2.1213						

Level=3	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
2	$\frac{a_1+a_2}{\sqrt{2}}$	$\frac{a_3+a_4}{\sqrt{2}}$	$\frac{a_5+a_6}{\sqrt{2}}$	$\frac{a_7+a_8}{\sqrt{2}}$	$\frac{a_2-a_1}{\sqrt{2}}$	$\frac{a_4-a_3}{\sqrt{2}}$	$\frac{a_6-a_5}{\sqrt{2}}$	$\frac{a_8-a_7}{\sqrt{2}}$
1	$\frac{a_1+a_2+a_3+a_4}{2}$		$\frac{a_5+a_6+a_7+a_8}{2}$		$\frac{(a_3+a_4)-(a_1+a_2)}{2}$		$\frac{(a_7+a_8)-(a_5+a_6)}{2}$	
0	$\frac{a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8}{2\sqrt{2}}$				$\frac{(a_5+a_6+a_7+a_8)-(a_1+a_2+a_3+a_4)}{2\sqrt{2}}$			



Haar Wavelet Συντελεστές

- Κρατάμε και **αποθηκεύουμε** μόνο **B ($B < N$) συντελεστές** (συνήθως τους μεγαλύτερους ή αυτούς που προσεγγίζουν καλά την συνολική «**ενέργεια**» του «**σήματος**»). Όλοι οι υπόλοιποι γίνονται μηδέν στην ανακατασκευή. Π.χ.

[2.75, -1.25, 0.5, 0, 0, -1, -1, 0]

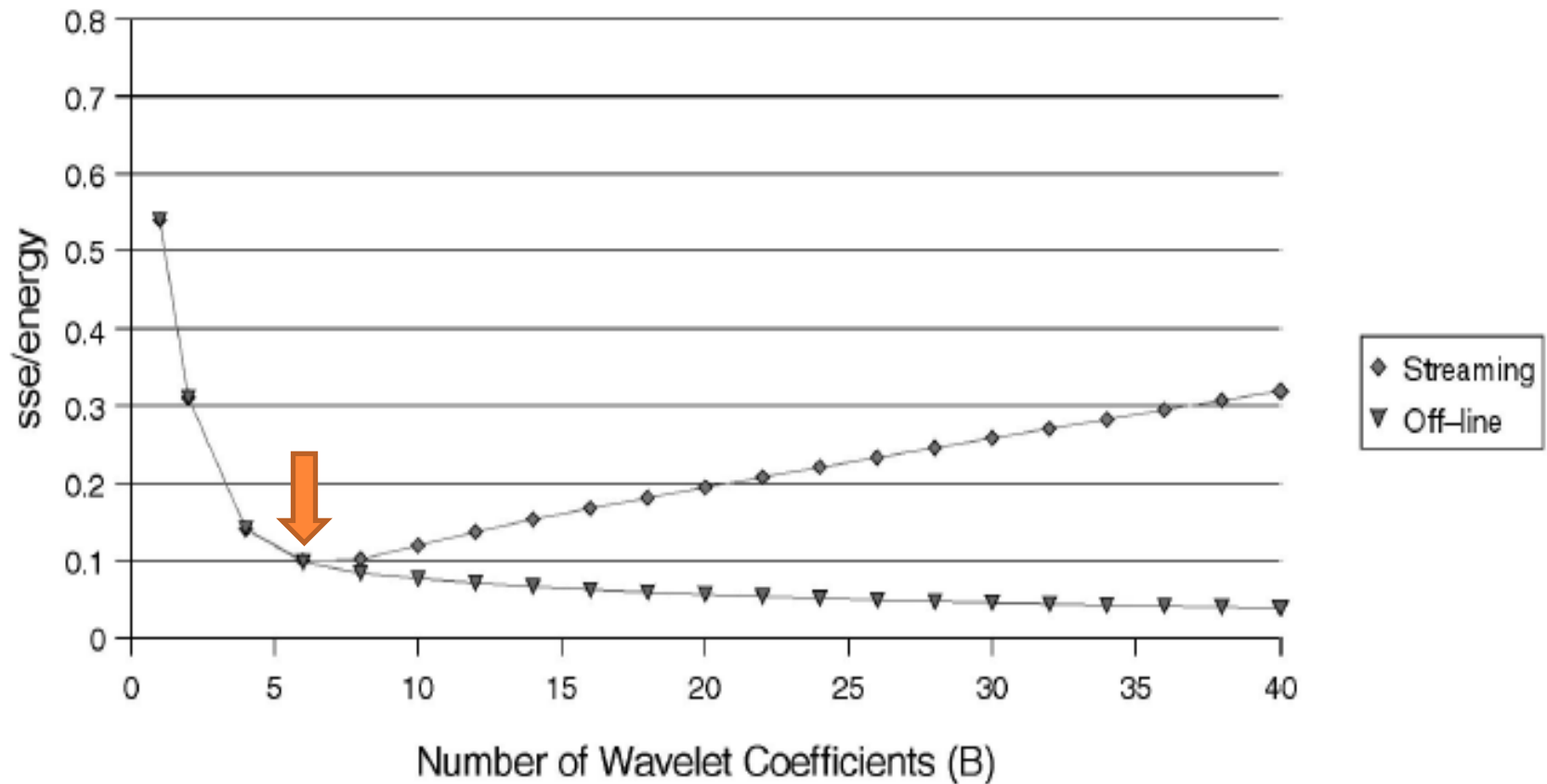
[2.75, -1.25, 0.5, 0, 0, 0, 0, 0] → σύνοψη: [2.75, -1.25, 0.5]

- Η διαγραφή μικρών συντελεστών δίνει μικρό λάθος στην επανακατασκευή των αρχικών δεδομένων.
- Οποσδήποτε όμως κρατάμε **έναν τουλάχιστον συντελεστή από το κάθε επίπεδο του δέντρου** του wavelet (δηλαδή **$\log N$** συντελεστές που μπορεί να περιέχονται ή όχι στους **B**).



Haar Wavelet Συντελεστές

- Η καλύτερη εκλογή των **B** συντελεστών (που προσεγγίζουν την συνολική ενέργεια) γίνεται με βάση την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (**SSE**) ως προς την ενέργεια. Η εκλογή περισσότερων συντελεστών δεν οδηγεί πάντα σε περαιτέρω ελαχιστοποίηση:



Haar Wavelet Απόδοση

- Ο Haar wavelet μετασχηματισμός απαιτεί χώρο $O(B+\log N)$.
- Ο χρόνος υπολογισμού/προσέγγισης μίας τιμής είναι επίσης $O(B+\log N)$.
- Αν έχουμε ενημερώσεις στοιχείων από την ροή, τότε πρέπει να ενημερωθούν και τα αντίστοιχα wavelets.
- Η ενημέρωση ενός wavelet γίνεται εύκολα αν αυτό αναπαρασταθεί με sketches μεγέθους $\log N$. [A.Gilbert, Y.Kotidis, S.Muthukrishnan, M.Strauss, 2002].
- Το πρόβλημα των συχνών στοιχείων λύνεται αποδοτικά ακόμα και με μεθόδους που χρησιμοποιούν wavelets [A.Gilbert et al.].



ΤΕΛΟΣ

