

Μ' ένα Σμπάρο Δυο Τρυγώνια: Εισάπαξ Κυματιδικές Συνόψεις για Μέτρα Μεγίστου Σφάλματος

Παναγιώτης Καρράς

Αθήνα, 26 Αυγούστου 2005

Έρευνα στο HKU με τον Νίκο Μαμουλή



Περίληψη

- Προκαταρκτικά & Κίνητρα
 - Χρησιμότητα Συνόψεων Δεδομένων
 - Μετασχηματισμός κυματιδίων Haar, συμβατικές κυματιδικές συνόψεις
 - Το πρόβλημα με την εγγύηση σφάλματος
- Προηγούμενη Προσέγγιση: Κυματιδικές Συνόψεις με Βέλτιστη Εγγύηση Σφάλματος
 - Μη πρακτικότητα αυτής της προσέγγισης
- Λύση: *Πρακτικές* Κυματιδικές Συνόψεις με Εγγυήσεις Σφάλματος
 - Αλγόριθμοι *Χαμηλής Πολυπλοκότητας* που παρέχουν σχεδόν βέλτιστη εγγύηση σφάλματος
- Επέκταση σε Ρέοντα Δεδομένα
 - Εισάπαξ Προσαρμογές των προταθέντων αλγορίθμων
- Συμπεράσματα & Μελλοντικές Κατευθύνσεις



Συμπαγείς Συνόψεις Δεδομένων χρήσιμες για:

- Προσεγγιστική Επεξεργασία Επερωτήσεων (ακριβείς απαντήσεις δεν απαιτούνται όσο ταχύτητα)
- Μάθηση, Ταξινόμηση, Εντοπισμό Γεγονότων
- Εξόρυξη Δεδομένων, Εκτίμηση Επιλεκτικότητας
- Συνθήκες όπου μαζικά δεδομένα αφίκνυνται εν ροή

Κυματιδιακός Μετασχηματισμός Haar

- **Κυματίδια:** μαθηματικό εργαλείο για την ιεραρχική αποσύνθεση συναρτήσεων/σημάτων
- **Κυματίδια Haar:** απλούστερη κυματιδιακή βάση, εύκολη στην σύλληψη και υλοποίηση
 - Αναδρομική κατά ζεύγος εξαγωγή διαφοράς από μέσο σε διαφορετικές αναλύσεις

Ανάλυση	Μέσοι	Συντελεστές Διαφοράς
3	$D = [2, 2, 0, 2, 3, 5, 4, 4]$	----
2	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
1	$[1.5, 4]$	$[0.5, 0]$
0	$[2.75]$	$[-1.25]$

Κυματιδιακός μετ/σμός Haar: $[2.75, -1.25, 0.5, 0, 0, -1, -1, 0]$

- Κατ' αναλογία επέκταση σε πολλαπλές διαστάσεις

Συντελεστές Κυματιδίων Haar

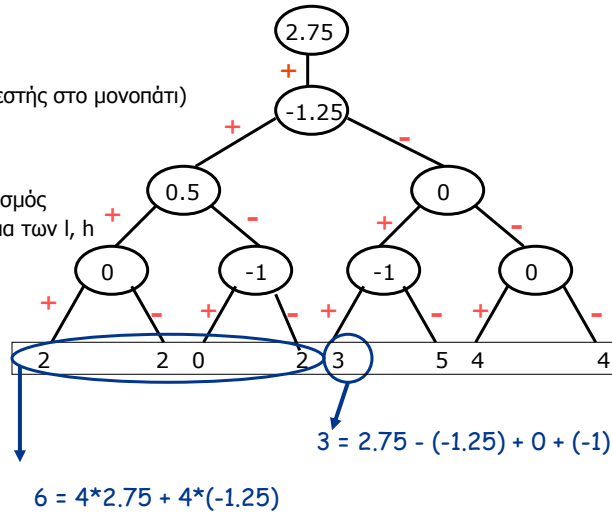
- **Δένδρο Σφάλματος**: Ιεραρχική δομή αποσύνθεσης
 - Εργαλείο για την αναπαράσταση του μετασχηματισμού και της ανακατασκευής

- Ανακατασκευή τιμής $d(i)$
 - $d(i) = \sum (+/-1) * (\text{συντελεστής στο μονοπάτι})$

- Άθροισμα εύρους $d(l:h)$
 - $d(l:h) = \text{γραμμικός συνδυασμός συντελεστών στα μονοπάτια των } l, h$

- Μόνο $O(\log N)$ όροι

Αρχικά δεδομένα



Κυματιδιακές Συνόψεις

- Υπολογίζουμε μετασχηματισμό Haar του D
- **Περικοπή συντελεστών**: κρατάμε $B \ll |D|$ συντελεστές
 - Το B καθορίζεται από τον διαθέσιμο χώρο
- Μηχανή προσεγγιστικών ερωτήσεων μπορεί να λειτουργήσει με τέτοιες συμπαγείς συνόψεις
 - [MWW, SIGMOD'98]; [VW, SIGMOD'99]; [CGRS, VLDB'00]
- Συμβατική περικοπή: Επιλογή B μεγαλύτερων συντελεστών κατ' **απόλυτη κανονικοποιημένη τιμή**
 - Κανονικοποίηση: διαίρεση συντελεστών ανάλυσης j με $\sqrt{2^j}$
 - *Αποδεδειγμένα βέλτιστη* για το Ολικό Τετραγωνικό Σφάλμα (L_2)
- **Δυστυχώς**, αυτή η μέθοδος δεν παρέχει εγγυήσεις ποιότητας της προσέγγισης για:
 - Ατομικές ανακατασκευές τιμών
 - Ατομικά αθροίσματα εύρους

Το Πρόβλημα με τις Συμβατικές Συνόψεις

- Παράδειγμα δεδομένων και συνόψεως ($|D|=16$, $B=8$)

Πάνω από 2,000% σχετικό σφάλμα! Ακριβές!

Αρχικές Τιμές	127 71 87 31 59 3 43 99	100 42 0 58 30 88 72 130
Προσεγγίσεις	65 65 65 65 65 65 65 65	100 42 0 58 30 88 72 130

Προσέγγιση = 195, αληθείς τιμές: $d(0:2)=285$, $d(3:5)=93$

- Μεγάλη διακύμανση στην ποιότητα των προσεγγίσεων
- Αιτία:
 - Ελαχιστοποίηση Ολικού Μέτρου Σφάλματος L_2
 - Απουσία εγγυήσεων για ατομικά σφάλματα

Λύση: Περικοπή για Μέτρα Μεγίστου Σφάλματος

- Μέτρα Σφάλματος παρέχοντα εγγυήσεις για όλες τις ανακατασκευασμένες τιμές:
 - Μέγιστο Απόλυτο Σφάλμα

$$\max_i \{ |\hat{d}_i - d_i| \}$$

- Μέγιστο Σχετικό Σφάλμα με όριο (για την αποφυγή κυριαρχήσεως των μικρών τιμών)

$$\max_i \left\{ \frac{|\hat{d}_i - d_i|}{\max\{|d_i|, s\}} \right\}$$

- Στόχευση στην ελαχιστοποίηση αυτών των μέτρων

Προηγούμενη Πρόταση: Βέλτιστη Περικοπή για Μέτρα Μεγίστου Σφάλματος [GK, PODS'04]

- Βασισμένη σε Δυναμικό Προγραμματισμό
- Αναδρομική συνάρτηση που υπολογίζει ελάχιστο δυνατό μέγιστο σφάλμα για τα υποδένδρα ενός συντελεστή δεδομένου του διαθέσιμου χώρου
- Βέλτιστη κατανομή διαθέσιμου χώρου b μεταξύ των υποδένδρων ενός κόμβου και του κόμβου του ίδιου.

Όμως:

- Απαγορευτική Πολυπλοκότητα:
- $O(BN^2 \log N)$ στον χρόνο
- $O(BN^2)$ στον χώρο
- Μη εφαρμόσιμη για τον σκοπό της
- Αδύνατη σε Περιβάλλον Ροής Δεδομένων
- *Πρόκληση:*
 - Σχεδίαση αποδοτικών σχημάτων περικοπής, χαμηλής πολυπλοκότητας, που επιτυγχάνουν ανταγωνιστικά αποτελέσματα σε σχέση με την βέλτιστη λύση.

Λύση:

Άπληστη Περικοπή για Μέτρα Μεγίστου Σφάλματος

- Ιδέα: Άπληστη λύση που διαλέγει τον *καλύτερο επόμενο* συντελεστή που θα διαγράψει βήμα προς βήμα
- Κάθε κόμβος του δένδρου σφάλματος αποθηκεύει το *Μέγιστο Δυνάμει Σφάλμα* που θα δημιουργηθεί αν ο συντελεστής διαγραφεί:

$$\text{-- Για Απόλυτο Σφάλμα: } MA_k = \max_{d_j \in \text{leaves}_k} \left\{ \left| \text{err}_j - \delta_{jk} \cdot c_k \right| \right\}$$

$$\text{-- Για Σχετικό Σφάλμα: } MR_k = \max_{d_j \in \text{leaves}_k} \left\{ \left| \text{err}_j - \delta_{jk} \cdot c_k \right| / \max \left(\left| d_j \right|, S \right) \right\}$$

- Σωρός επιστρέφει κόμβο *Ελάχιστου Μεγίστου Δυνάμει Σφάλματος*
- **Για Απόλυτο Σφάλμα:**
 - 4 τιμές *Max* και *Min* του *Συσσωρευμένου Σφάλματος* στους κόμβους
- **Για Σχετικό Σφάλμα:**
 - *Συσσωρευμένο Σφάλμα* στα φύλλα
 - Σωροί επιστρέφοντες φύλλο *Μεγίστου Δυνάμει Σφάλματος* για κάθε κόμβο

Λύση:

Άπληστη Περικοπή για Μέτρα Μεγίστου Σφάλματος

Μετά από κάθε διαγραφή:

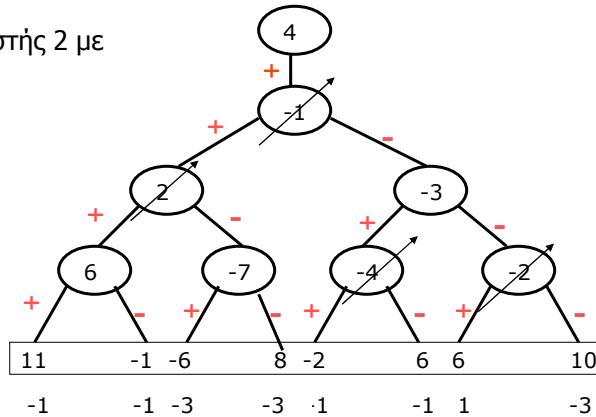
- Αλλαγές στο *Συσσωρευμένο Σφάλμα* διαδίδονται στο δένδρο
- Για κάθε επηρεαζόμενο κόμβο:
- **Για Απόλυτο Σφάλμα:**
 - Ενημέρωση 4 τιμών *Max*, *Min* *Συσσωρευμένου Σφάλματος*
 - Υπολογισμός Νέου Μεγίστου Δυνάμει Απολύτου Σφάλματος ως:

$$MA_k = \max \left\{ \begin{array}{ll} \left| \max_k^l - c_k \right|, & \left| \min_k^l - c_k \right| \\ \left| \max_k^r + c_k \right|, & \left| \min_k^r + c_k \right| \end{array} \right\}$$

- **Για Σχετικό Σφάλμα:**
 - Ενημέρωση Σωρού Απογόνων
 - Επιστροφή Νέου Μεγίστου Δυνάμει Σχετικού Σφάλματος από Σωρό
- Ενημέρωση Ολικού Σωρού συντελεστών

Παράδειγμα (απόλυτο σφάλμα)

- Πρώτα διώχνουμε τον συντελεστή -1
- Διάδοση σφάλματος
- Έπειτα φεύγει το συντελεστής 2 με δύναμη μέγιστο σφάλμα 3
- Κ.ο.κ....



Ανάλυση Πολυπλοκότητας

- Αλγόριθμος για *Απόλυτο Σφάλμα* :

Χρόνος: $O(N \log^2 N)$

Χώρος: $O(N)$

- Αλγόριθμος για *Σχετικό Σφάλμα* :

Χρόνος: $O(N \log^3 N)$

Χώρος: $O(N \log N)$

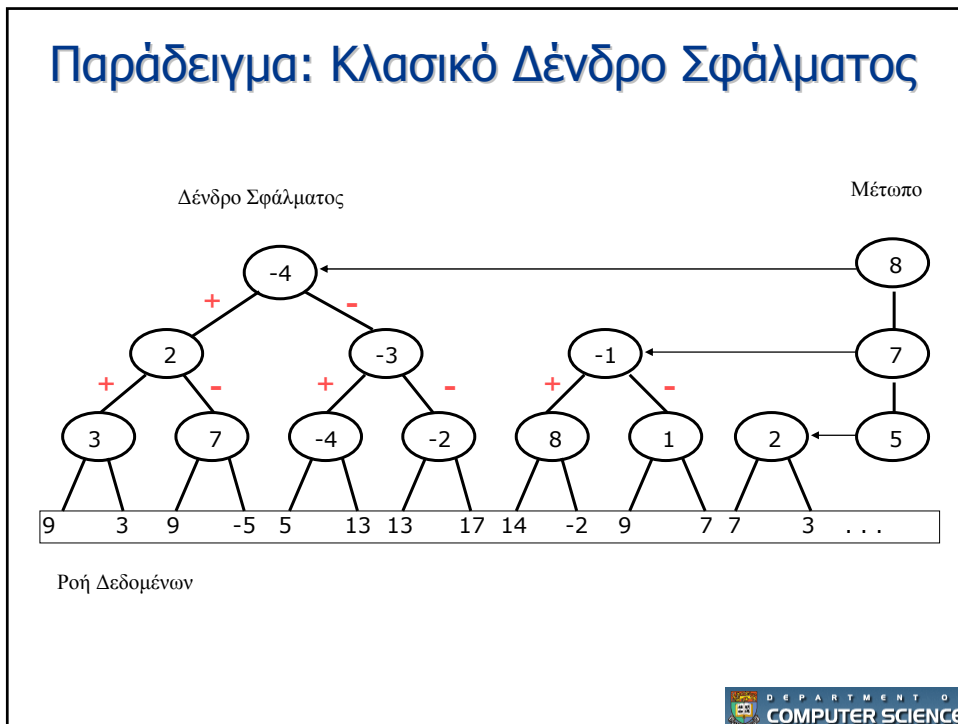
Επέκταση σε Ρέοντα Δεδομένα

- Κύρια Περιοχή Εφαρμογών
- Προταθείσες μέθοδοι μη εφαρμόσιμες
- Παραδοχή: $O(B)$ διαθέσιμη μνήμη
- *Επιπλέον Πρόβλημα:*
 - Επέκταση προταθεισών μεθόδων
 - Εισάπαξ ολική διαδικασία
 - Κατασκευή και Περικοπή του δένδρου ταυτόχρονα

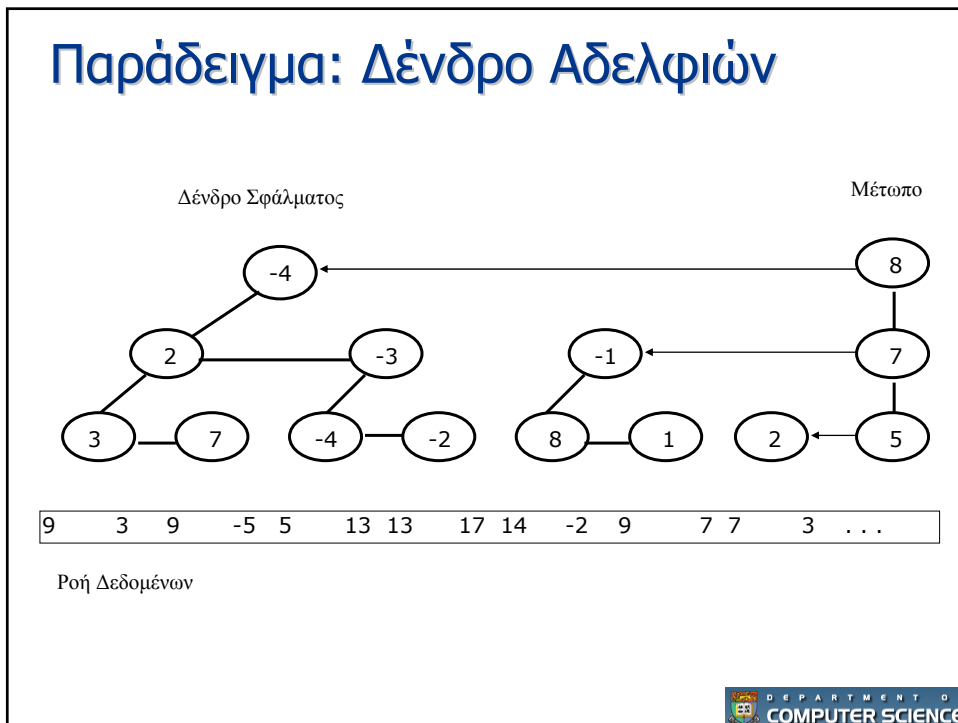
Λύση για το Απόλυτο Σφάλμα

- Μετά τα πρώτα B τιμές, ένα ζεύγος συντελεστών αποβάλλεται για κάθε νέο ζεύγος δεδομένων
- Σκοπιά στο μέχρι τούδε κατασκευασμένο δένδρο
- Ανώτερο επίπεδο δένδρου για κάθε δύναμη του 2 #δεδομένων
- Δομή *Μέτωπο* αποθηκεύει:
 - *Κρεμόμενους* συντελεστές
 - Μέση τιμή για το κρεμόμενο υποδένδρο
 - Πληροφορία Σφάλματος από διαγεγραμμένα *ορφανά*
- Διάδοση Σφάλματος όπως στην στατική περίπτωση, με παραπάνω λεπτομέρειες για την προς τα πάνω διάδοση λόγω αραιότητας του δένδρου

Παράδειγμα: Κλασικό Δένδρο Σφάλματος



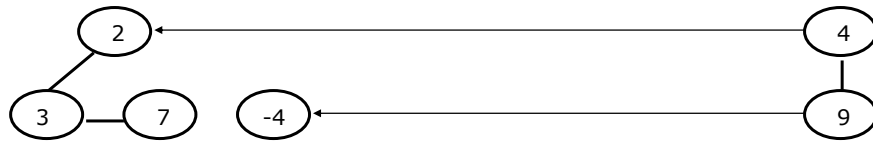
Παράδειγμα: Δένδρο Αδελφιών



Παράδειγμα : $B = 6$, μετά από 6 τιμές

Δένδρο Σφάλματος

Μέτωπο



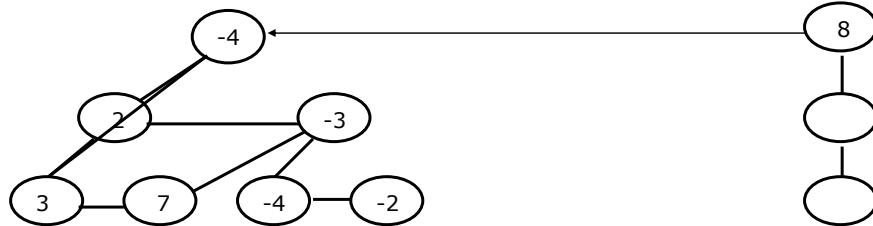
9 3 9 -5 5 13 ...

Ροή Δεδομένων

Παράδειγμα : $B = 6$, μετά από 8 τιμές

Δένδρο Σφάλματος

Μέτωπο



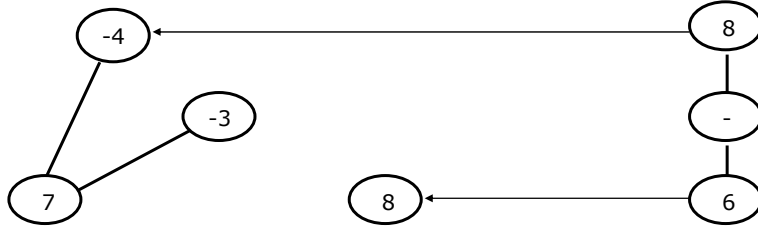
9 3 9 -5 5 13 13 17 ...

Ροή Δεδομένων

Παράδειγμα : $B = 6$, μετά από 10 τιμές

Δένδρο Σφάλματος

Μέτωπο



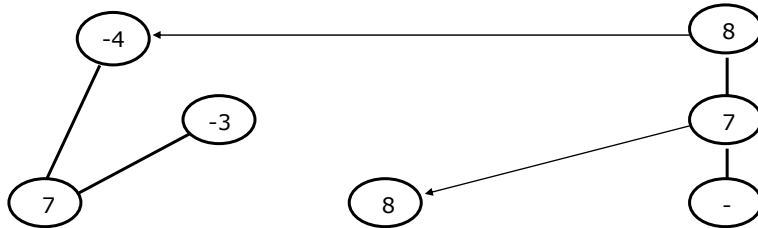
9 3 9 -5 5 13 13 17 14 -2 ...

Ροή Δεδομένων

Παράδειγμα : $B = 6$, μετά από 12 τιμές

Δένδρο Σφάλματος

Μέτωπο



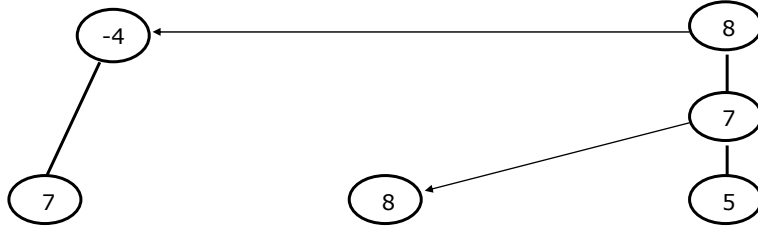
9 3 9 -5 5 13 13 17 14 -2 9 7 ...

Ροή Δεδομένων

Παράδειγμα : $B = 6$, μετά από 14 τιμές

Δένδρο Σφάλματος

Μέτωπο

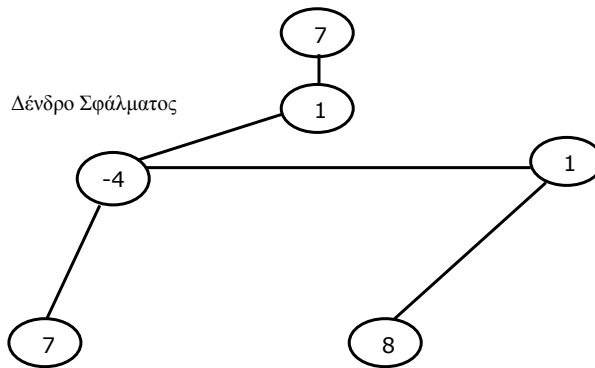


9 3 9 -5 5 13 13 17 14 -2 9 7 7 3

Ροή Δεδομένων

Παράδειγμα: $B = 6$, μετά το συμπλήρωμα

Δένδρο Σφάλματος



4 4 11 -3 12 12 12 12 15 -1 7 7 5 5

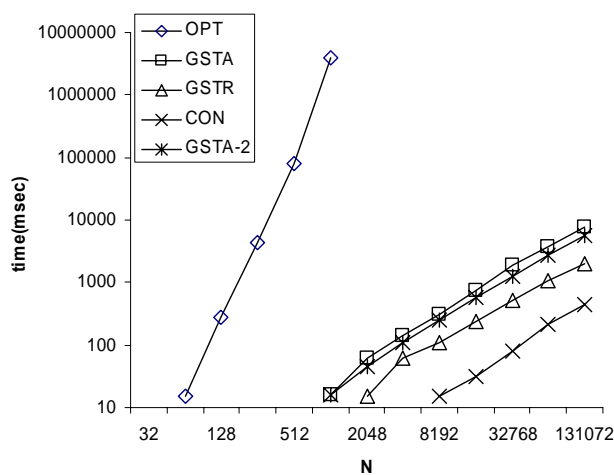
Ανακατασκευή

Λύση για το Σχετικό Σφάλμα

- Ανάλογη Επέκταση αδύνατη
- Λύση: Ευρετικές Τεχνικές
- Εκτιμήτρια του MR_k βασιμμένη σε:
 - 4 ποσότητες όπως για Απόλυτο Σφάλμα (με παρονομαστές)
 - Ελάχιστες Απόλυτες Τιμές σε κάθε υποδένδρο (με σφάλματα)
 - Μια τιμή δείγμα (με σφάλμα) για κάθε υποδένδρο, αρχικοποιούμενη ως Ελάχιστη Απόλυτη Τιμή, μεταβαλλόμενη από της διαδικασίες διάδοσης σφάλματος όταν άλλη τιμή δώσει μεγαλύτερο σχετικό σφάλμα
- Ευρετική Εκτίμηση ως το Μέγιστο Σχετικό Σφάλμα μεταξύ των 8 παραπάνω θέσεων

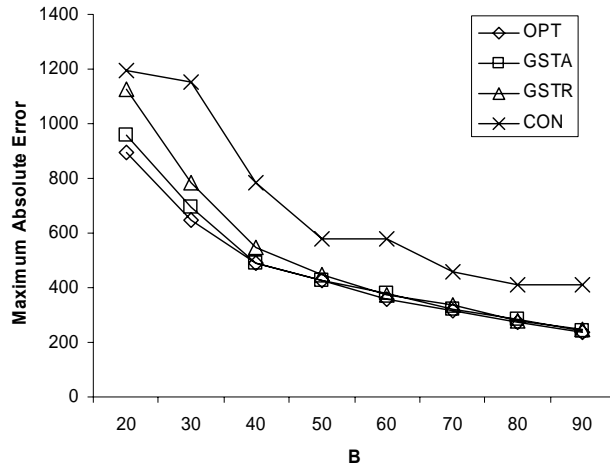
Πειραματικά Αποτελέσματα

- Χρόνος, $B = N / 16$, Σχετικό Σφάλμα



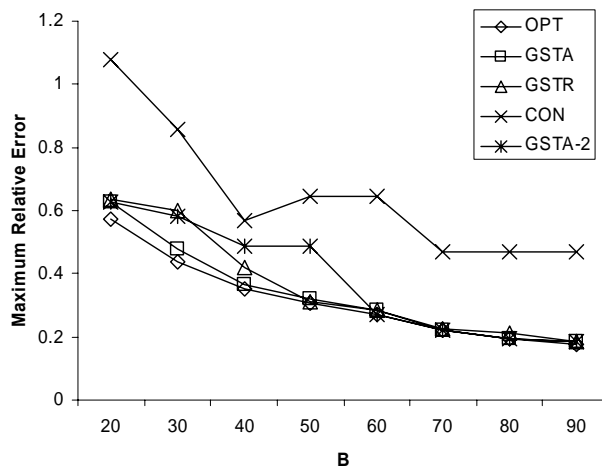
Πειραματικά Αποτελέσματα

- Ποιότητα, Απόλυτο Σφάλμα (μετρητές συχνότητας), $N = 360$



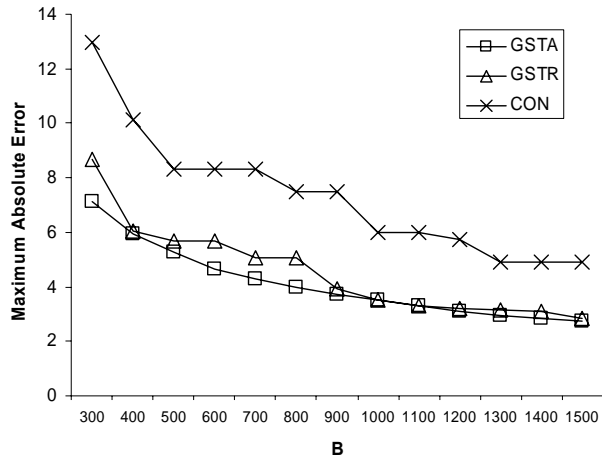
Πειραματικά Αποτελέσματα

- Ποιότητα, Σχετικό Σφάλμα (μετρητές συχνότητας), $N = 360$



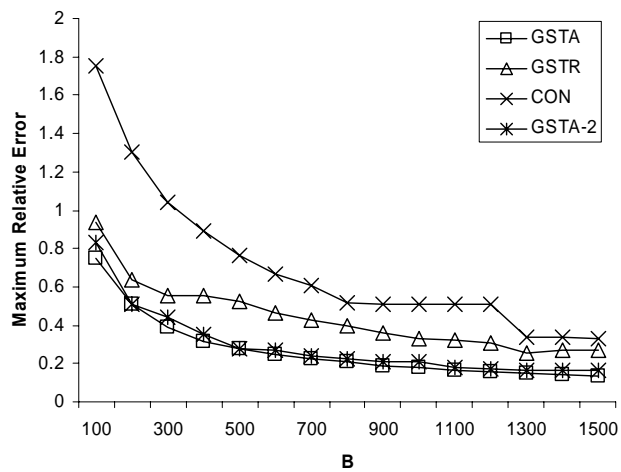
Πειραματικά Αποτελέσματα

- Κλιμακωσιμότητα, Απόλυτο Σφάλμα (μετρητές φωτονίων), $N = 16K$



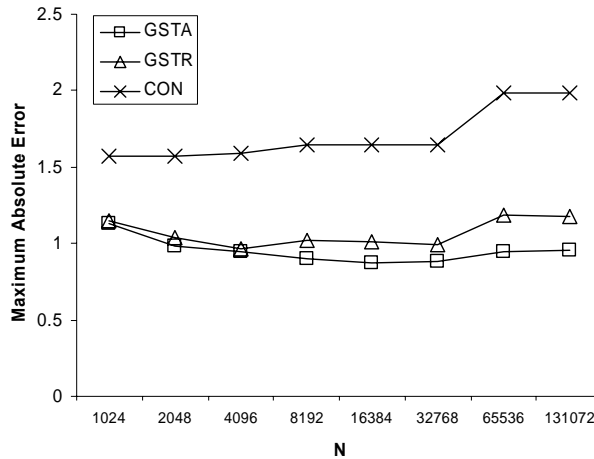
Πειραματικά Αποτελέσματα

- Κλιμακωσιμότητα, Σχετικό Σφάλμα (μετρητές φωτονίων), $N = 16K$



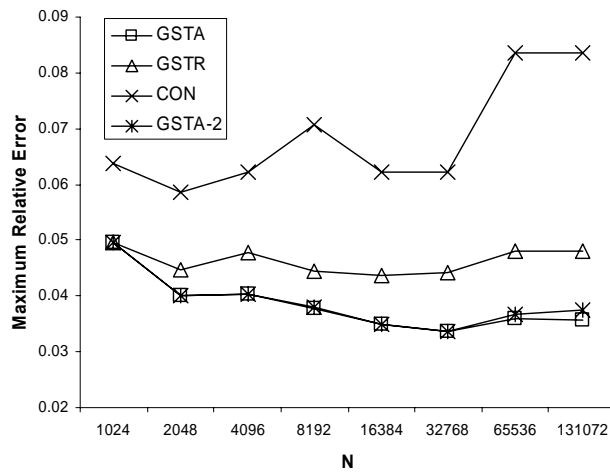
Πειραματικά Αποτελέσματα

- Κλιμακωσιμότητα, Απόλυτο Σφάλμα (θερμοκρασίες), $B = N / 16$



Πειραματικά Αποτελέσματα

- Κλιμακωσιμότητα, Σχετικό Σφάλμα (θερμοκρασίες), $B = N / 16$



Συμπεράσματα & Μελλοντικές Κατευθύνσεις

- Επιτευξιμότητα Κυματιδιακών Συνόψεων με σχεδόν βέλτιστες Εγγυήσεις Μεγίστου Σφάλματος και σχεδόν γραμμικό κόστος για Στατικά και Ρέοντα Δεδομένα
- Επέκταση σε Πολυδιάστατα Κυματίδια;
- Εναλλακτικές Ευρετικές για το Σχετικό Σφάλμα;
- Μεταβλητές Τιμές Συντελεστών;
- Θεωρητική Τεκμηρίωση;

Αναφορές

- Y. Matias, J. S. Vitter, and M. Wang. Wavelet-based histograms for selectivity estimation. *SIGMOD 1998*
- J. S. Vitter and M. Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. *SIGMOD 1999*
- K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. Approximate query processing using wavelets. *VLDB Journal 2001*
- M. Garofalakis and A. Kumar. Deterministic wavelet thresholding for maximum-error metrics. *PODS 2004*
- A. Gilbert, Y. Kotidis, S. Muthukrishnan and Martin Strauss. Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries. *VLDB 2001*

www.sailingissues.com

Ευχαριστώ! Ερωτήσεις;

