Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Σύγκριση περιλήψεων κυματιδίων (wavelet synopses) για διάφορες μετρικές σφάλματος

Διπλωματικη Εργασια

του

ΜΙΧΑΗΛ Γ. ΜΑΘΙΟΥΔΑΚΗ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων

Αθήνα, Ιούλιος 2006

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Σύγκριση περιλήψεων κυματιδίων (wavelet synopses) για διάφορες μετρικές σφάλματος

Διπλωματικη Εργασια

του

ΜΙΧΑΗΛ Γ. ΜΑΘΙΟΥΔΑΚΗ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Ιουλίου 2006.

.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2006

.....

Μιχαηλ Μαθιουδακης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περίληψη

Η χρήση του μετασχηματισμού κυματιδίων αποδεικνύεται ότι είναι ένα αποδοτικό εργαλείο για την κατασκευή περιλήψεων χρονικών σειρών αλλά και πολυδιάστατων δεδομένων. Κύρια χαρακτηριστικά του είναι η απλότητα και η ταχύτητά του καθώς και η υψηλή συμπίεση που προσφέρει. Όπως κάθε απωλεστικός αλγόριθμος συμπίεσης, έτσι και οι περιλήψεις κυματιδίων εισάγουν σφάλματα στην αναπαραγωγή του αρχικού σήματος. Υπάρχουν διάφοροι τρόποι να μετρηθεί το συνολικό σφάλμα και κατά συνέπεια υπάρχουν αντίστοιχοι αλγόριθμοι παραγωγής περιλήψεων που ελαχιστοποιούν τις διάφορες μετρικές σφάλματος. Οι υπάρχοντες αλγόριθμοι που εξετάζουμε θεωρητικά και πειραματικά σε αυτήν την εργασία κατασκευάζουν περιλήψεις μονοδιάστατων συνόλων δεδομένων και ελαχιστοποιούν μετρικές που εκτιμούν το σφάλμα των περιλήψεων για σημειακά σφάλματα και αθροιστικά σφάλματα εύρους. Προτείνουμε, ακόμα, ένα νέο δυναμικό αλγόριθμο που ελαχιστοποιεί τη μετρική weighted-L_p για δυαδική ιεραρχία από αθροιστικά σφάλματα εύρους.

Λέξεις Κλειδιά

Αθροιστικό Σφάλμα Εύρους, Αλγόριθμος Περίληψης, Δένδρο Σφάλματος, Δυαδική Ιεραρχία Αθροιστικών Σφαλμάτων Εύρους, Μετασχηματισμός Haar Wavelet, Μετρική Σφάλματος, Περίληψη Δεδομένων, Περίληψη Κυματιδίων, Σημειακό Σφάλμα

Abstract

The wavelet transformation is a proven tool for constructing synopses of time series and multidimensional data. Its simplicity and the time efficiency it provides are some of its main characteristics. Like every lossy compression algorithm, wavelet synopses introduce some error in the reconstruction of the initial signal. There are several ways to measure the total error and therefore, there are the different synopsis construction algorithms that minimize several error metrics. The existing algorithms we study theoretically and experimentally in this diploma thesis construct wavelet synopses for one-dimensional data sets and minimize error metrics that calculate the synopses error over point and range sum errors. Furthermore, we propose a new dynamic algorithm that minimizes the weighted-L_p error metric for dyadic hierarchies of range sum errors.

Keywords

Range Sum Error, Synopsis Algorithm, Error Tree, Dyadic Hierarchy of Range Sum Errors, Haar Wavelet Transformation, Error Metric, Data Synopsis, Wavelet Synopsis, Point Error

Ευχαριστίες

Επιθυμώ να ευχαριστήσω θερμά τον καθηγητή μου, κ. Τίμο Σελλή, για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα ενδιαφέρον θέμα στη διπλωματική καθώς και για όλη τη βοήθεια που μου έχει προσφέρει. Ευχαριστώ επίσης θερμά το διδακτορικό φοιτητή Δημήτρη Σαχαρίδη για τον πολύτιμο συμβουλευτικό του ρόλο και την αμέριστη συμπαράστασή του σε όλη τη διάρκεια εκπόνησης της διπλωματικής εργασίας. Τέλος, θέλω να ευχαριστήσω τους γονείς μου, οι οποίοι με στηρίζουν σε κάθε μου βήμα ως σήμερα.

Κεφάλαιο 1

Εισαγωγή

Η προσεγγιστική επεξεργασία ερωτημάτων (approximate query processing) πάνω σε περιλήψεις δεδομένων έχει προσελκύσει πρόσφατα μεγάλο ενδιαφέρον ως μια αποτελεσματική μέθοδος χειρισμού μεγάλων συνόλων δεδομένων. Οι περιλήψεις wavelet είναι ένας τύπος περιλήψεων που αποδεινύονται κατάλληλες και αποτελεσματικές για την αναπαράσταση μεγάλου όγκου δεδομένων. Είναι χαρακτηριστικό ότι στο νέο πρότυπο JPEG προβλέπεται η χρήση του μετασχηματισμού wavelet αντί του μετασχηματισμού Fourier.

Η βασική ιδέα είναι ότι το (διακριτό) σήμα εισόδου μετασχηματίζεται με χρήση του M/Σ Haar Wavelet (βλ. Ενότητα 2.2) και παράγονται οι συντελεστές wavelet, που αποτελούν το μετασχηματισμένο μας σήμα. Στη συνέχεια, εφαρμόζοντας κάποιο κριτήριο σφάλματος, κρατάμε ένα υποσύνολο των συντελεστών αυτών — οι οποίοι αποτελούν την περίληψη των δεδομένων μας — και απορρίπτουμε τους υπόλοιπους (δηλαδή τους θεωρούμε μηδενικούς). Το κριτήριο σφάλματος μας το προσφέρει μια μετρική σφάλματος. Οι μετρικές σφάλματος είναι συναρτήσεις που μετράνε με διάφορους τρόπους το σφάλμα που εισάγεται στα δεδομένα μας από την απόρριψη κάποιων συντελεστών wavelet. Δηλαδή, μετράνε το σφάλμα που έχουμε στα δεδομένα μας όταν για την αναπαράστασή τους και την ανακατασκευή τους χρησιμοποιήσουμε μόνο τους συντελεστές που ανήκουν στην περίληψη. Στόχος των αλγορίθμων που μελετάμε είναι η κατασκευή περίληψης που ελαχιστοποιεί μια συγκεκριμένη μετρική σφάλματος, με δεδομένο το μέγιστο πλήθος συντελεστών wavelet που μπορούμε να κρατήσουν στην περίληψη.

Μπορούμε να πούμε κάπως απλουστευτικά, ότι η ιδιότητα που κάνει το μετασχηματισμό Haar Wavelet κατάλληλο και αποτελεσματικό για αυτή τη διαδικασία, είναι ότι όταν κάποια γειτονικά δεδομένα έχουν παραπλήσιες τιμές, εμφανίζονται συντελεστές wavelet με τιμή κοντά στο μηδέν τους οποίους μπορούμε να παραλείψουμε χωρίς σημαντικό σφάλμα.

Οι μετρικές σφάλματος που συναντώνται σε αυτή τη διπλωματική ανήκουν στην κατηγορία των weighted-L_p μετρικών. Οι μετρικές αυτές επιστρέφουν το σφάλμα μιας περίληψης συνεκτιμώντας είτε τα σημειακά σφάλματα — point errors (δηλαδή το σφάλμα στην τιμή κάθε στοιχείου του σήματος — βλ. Κεφάλαιο 3) είτε τα αθροιστικά σφάλματα εύρους — range sum errors (δηλαδή τα σφάλματα στο άθροισμα των τιμών των στοιχείων που ανήκουν σε συγκεκριμένα διαστήματα του σήματος — βλ. Κεφάλαιο 4).

1.1 Αντικείμενο της διπλωματικής

Αντικείμενο της διπλωματικής είναι η θεωρητική και πειραματική μελέτη αλγορίθμων που κατασκευάζουν περιλήψεις δεδομένων με χρήση του Διακριτού Μετασχηματισμού Haar wavelet.

Πιο συγκεκριμένα, η διπλωματική αυτή περιλαμβάνει:

1. **Θεωρητική Μελέτη Υπαρχόντων Αλγορίθμων**. Κατηγοριοποιούνται και παρουσιάζονται υπάρχοντες αλγόριθμοι οι οποίοι κατασκευάζουν περιλήψεις δεδομένων με χρήση του μετασχηματισμού Haar Wavelet, ελαχιστοποιώντας ο καθένας μια συγκεκριμένη μετρική σφάλματος. Οι μετρικές σφάλματος που ελαχιστοποιούνται από τους αλγορίθμους αυτούς ανήκουν στην κατηγορία weighted-L_p. Οι αλγόριθμοι αναλύονται θεωρητικά και μελετώνται ως προς τη χωρική και χρονική τους πολυπλοκότητα — βλ. Κεφάλαιο 3 και Ενότητα 4.3.

2. **Κατασκευή Περιλήψεων για range sum errors**. Μέχρι σήμερα δεν υπάρχει αλγόριθμος wavelet ο οποίος να ελαχιστοποιεί μια μετρική για range sum errors και ο οποίος να τρέχει κατευθείαν στα δεδομένα. Παρουσιάζεται και αναλύεται, λοιπόν, ένας νέος αλγόριθμος (Rangewave) ο οποίος ελαχιστοποιεί τη μετρική weighted-L_p για δυαδική ιεραρχία από range sum errors. Ο αλγόριθμος αυτός χρησιμοποιεί το Διακριτό Μετασχηματισμό Haar Wavelet και τρέχει πάνω από δεδομένα που βρίσκονται στην αρχική μορφή τους — βλ. Ενότητα 4.4.2.

3. **Πειραματική Μελέτη Υπαρχόντων Αλγορίθμων**. Υλοποιούνται, εκτελούνται και συγκρίνονται πειραματικά μερικοί από τους αλγορίθμους που εξετάζονται στη διπλωματική. Η εκτέλεσή τους γίνεται για διαφορετικά μεγέθη περίληψης και πλήθος δεδομένων και η σύγκρισή τους γίνεται ως προς το χρόνο εκτέλεσής τους και την ακρίβεια της περίληψης που κατασκευάζουν. Η ακρίβεια της περίληψης μετράται με χρήση της weighted-L_2 μετρικής. Ένα βασικό ερώτημα που μας απασχολεί είναι κατά πόσο οι πολύπλοκοι αλγόριθμοι που είναι σχεδιασμένοι να ελαχιστοποιούν τη μετρική weighted-L_2, αποδίδουν σημαντικό κέρδος στην ακρίβεια της περίληψης σε σχέση με θεωρητικά λιγότερο ακριβείς αλλά και λιγότερο χρονοβόρους αλγορίθμους. Γι' αυτό το λόγο στην πειραματική σύγκριση συμπεριλαμβάνουμε και αλγορίθμους που δεν ελαχιστοποιούν τη μετρική weighted-L_2, αλλά που ενδεχομένως η απώλεια που έχουν σε ακρίβεια να αντισταθμίζεται από την ταχύτητα που προσφέρουν — βλ. Ενότητα 5.2.

1.2 Η διάρθρωση της διπλωματικής

Στο κεφάλαιο 2 κάνουμε μια ιστορική εισαγωγή στα wavelets (Ενότητα 2.1), ορίζουμε και περιγράφουμε τον κλασσικό μετασχηματισμό Haar Wavelet (Ενότητα 2.2), ο οποίος είναι ο απλούστερος από τους μετασχηματισμούς wavelet και ο ευρύτερα χρησιμοποιούμενος και εισάγουμε την έννοια των μετρικών σφάλματος (Ενότητα 2.3).

Η πρώτη οικογένεια μετρικών σφάλματος είναι εκείνες που ο υπολογισμός του σφάλματος βασίζεται στα σημειακά σφάλματα - point errors που εισάγονται στο ανακατασκευασμένο από

την περίληψη σήμα. Αυτές οι μετρικές επιστρέφουν μια τιμή συνυπολογίζοντας τα N σημειακά σφάλματα που περιλαμβάνει ένα ανακατασκευασμένο σήμα μεγέθους N και χρησιμοποιούνται όταν εφαρμόζουμε point queries πάνω στα δεδομένα μας. Μια κατηγορία μετρικών για point errors είναι οι Weighted-L_p μετρικές, οι οποίες αθροίζουν τα σημειακά σφάλματα όλων των στοιχείων του σήματος υψωμένα σε μια δύναμη p και ίσως πολλαπλασιασμένα με κάποιο βάρος που τους αντιστοιχεί. Στο κεφάλαιο 3, λοιπόν, ορίζουμε τις weighted-L_p μετρικές για point errors. Στη συνέχεια, παρουσιάζουμε αλγορίθμους που, δεδομένου κάποιου ορίου στο χώρο που μπορούμε να διαθέσουμε για τους συντελεστές που μένουν στην περίληψη και δεδομένης κάποιας συνάρτησης που ανήκει στην κατηγορία των Weighted-L_p μετρικών, κατασκευάζουν μια περίληψη που ελαχιστοποιεί το σφάλμα της μετρικής. Οι αλγόριθμοι χωρίζονται σε δύο κατηγορίες με βάση το αν χρησιμοποιούν τον κλασσικό μετασχηματισμό Haar (Ενότητα 3.2) ή εάν μετασχηματίζουν τα δεδομένα ενσωματώνοντας στο μετασχηματισμό τα βάρη των point errors (Ενότητα 3.3).

Η δεύτερη οικογένεια μετρικών σφάλματος βασίζει τον υπολογισμό του σφάλματος στα αθροιστικά σφάλματα εύρους - range sum errors και χρησιμοποιούνται όταν εφαρμόζουμε range sum queries πάνω στα δεδομένα μας. Παρόμοια με την περίπτωση των σημειακών σφαλμάτων, ορίζονται στο κεφάλαιο 4 οι μετρικές weighted-L_p για range sum errors (Ενότητα 4.1). Επιπλέον, εξετάζουμε αν ο κλασσικός μετασχηματισμός Haar είναι ορθοκανονικός ως προς τη μετρική L_2 (Ενότητα 4.2) και παρουσιάζουμε ένα νέο αλγόριθμο που ελαχιστοποιεί τη weighted-L_p μετρική για δυαδικά ιεραρχημένα range sum errors(Ενότητα 4.4).

Στο κεφάλαιο 5 παρουσιάζουμε πειραματικά αποτελέσματα στα οποία συγκρίνεται η απόδοση κάποιων αλγορίθμων για point και dyadic range sum queries (ενότητες 5.2 και 5.3, αντίστοιχα). Στόχος των πειραμάτων είναι να συγκρίνουμε την ακρίβεια των περιλήψεων που επιτυγχάνουν οι αλγόριθμοι αλλά και την ταχύτητα στην κατασκευή τους. Οι μετρήσεις γίνονται σε σχέση αφενός με το πλήθος των δεδομένων και αφετέρου με το μέγεθος της περίληψης. Παράλληλα, μας απασχολεί το ερώτημα κατά πόσο αξίζει να χρησιμοποιούμε πολύπλοκους αλγορίθμους που ελαχιστοποιούν μια μετρική σφάλματος έναντι απλούστερων μη-βέλτιστων αλγορίθμων. Η μετρική σφάλματος που χρησιμοποιείται για να εκτιμήσουμε την ακρίβεια της κάθε περίληψης είναι η weighted-L_2 . Έτσι, στην πειραματική σύγκριση περιλαμβάνονται και αλγόριθμοι που δεν είναι βέλτιστοι για τη μετρική weighted-L_2, ώστε να δούμε εάν κάποια απώλεια στην ακρίβεια της περίληψης αντισταθμίζεται, ενδεχομένως, από την ταχύτητα που προσφέρουν.

Τέλος, στο κεφάλαιο 6 παρουσιάζουμε κάποια γενικά συμπεράσματα από τη μελέτη και τη σύγκριση των αλγορίθμων wavelet και προτείνουμε πιθανές μελλοντικές επεκτάσεις του υπάρχοντος έργου πάνω στις περιλήψεις wavelet.

Κεφάλαιο 2

Wavelets και Περίληψη Δεδομένων

2.1 Εισαγωγή στα wavelets

Η ιδέα της προσεγγιστικής αναπαράστασης σήματος με χρήση υπερτιθέμενων συναρτήσεων υπάρχει από τις αρχές του 19ου αιώνα, όταν ο Joseph Fourier παρατήρησε ότι με την υπέρθεση ημιτονοειδών και συνημιτονοειδών συναρτήσεων διαφόρων συχνοτήτων μπορούσε να παραστήσει άλλες συναρτήσεις. Σταδιακά, η προσοχή των ερευνητών στράφηκε από την ανάλυση συχνότητας στην ανάλυση κλίμακας (scale analysis), καθώς άρχισε να διαφαίνεται ότι η προσέγγιση ενός σήματος δια του υπολογισμού μέσων διακυμάνσεων για τμήματα διαφορετικής κλίμακας, μπορούσε να αντιμετωπίσει καλύτερα την ύπαρξη θορύβου στο σήμα. Η ανάλυση κλίμακας μας επιτρέπει να παρατηρούμε το σήμα από διάφορα επίπεδα εστίασης. Παρατηρώντας το σήμα μέσα από ένα μικρό 'παράθυρο', διακρίνουμε λεπτομερή χαρακτηριστικά του, ενώ αν το παρατηρήσουμε 'μακροσκοπικά', διακρίνουμε γενικότερα και πιο χονδροειδή χαρακτηριστικά.

Οι ημιτονοειδείς και συνημιτονοειδείς συναρτήσεις, που αποτελούν τη βάση του μετασχηματισμού Fourier είναι εξ' ορισμού μη-τοπικές (εκτείνονται στο άπειρο) κι έτσι καθίστανται ακατάλληλες για την αναπαράσταση σημάτων που παρουσιάζουν απότομες διακυμάνσεις (sharp spikes). Με την ανάλυση wavelet, όμως, μπορούμε να χρησιμοποιήσουμε προσεγγιστικές συναρτήσεις που εκτείνονται σε πεπερασμένο χώρο. Έτσι, οι συναρτήσεις wavelet είναι κατάλληλες για την προσέγγιση σημάτων με μεγάλες ασυνέχειες.

Η πρώτη καταγεγραμμένη αναφορά σε αυτό που σήμερα ονομάζουμε wavelet βρίσκεται σε μια εργασία του Alfred Haar από το 1909. Η σύγχρονη θεωρητική τεκμηρίωση των wavelets, ξεκίνησε γύρω στο 1975 με τον Jean Morlet και τη μετέπειτα συνεργασία του με τον Alex Grossmann του Marseille Theoretical Physics Center. Οι μέθοδοι της ανάλυσης wavelet αναπτύχθηκαν κυρίως από το Γάλλο Yves Meyer, ενώ στο συνεργάτη του, Stephane Mallat, αποδίδεται ο κύριος αλγόριθμος του μετασχηματισμού wavelet (1988). Έκτοτε, η ανάλυση wavelet προσέλκυσε διεθνές ενδιαφέρον. Ανάμεσα σ' εκείνους που συνέβαλαν στη σχετική έρευνα είναι οι Ingrid Daubechies, Ronald Coifman, και Victor Wickerhauser.

Τα wavelets σήμερα βρίσκουν εφαρμογή σε μεγάλο εύρος επιστημονικών πεδίων. Μερικά από αυτά είναι: συμπίεση δεδομένων, επεξεργασία σήματος και εικόνας, οπτική, όραση υπολογιστών, ακουστική, μουσική τεχνολογία, επίλυση μερικών διαφορικών εξισώσεων, πυρηνική μηχανική, σεισμολογία, ραντάρ και αλλού.

Μια καλή εισαγωγή στη θεωρία wavelet βρίσκεται στα [2] και [9].

2.2 Ο Διακριτός Μετασχηματισμός Haar Wavelet

Υπάρχουν πολλές διαφορετικές εκδοχές μετασχηματισμού wavelet, τόσο σε συνεχές όσο και σε διακριτό πεδίο τιμών. Ο Διακριτός Μετασχηματισμός Haar Wavelet είναι ο πιο απλός. Στην περίπτωση μονοδιάστατων δεδομένων, εφαρμόζεται πάνω σε ένα σήμα διάνυσμα διακριτών τιμών (διάνυσμα) μεγέθους N, με το N να είναι μια δύναμη του 2 και παράγει ένα διακριτό σήμα ίδιου μεγέθους, διατηρώντας την ενέργεια του σήματος.

Πιο συγκεκριμένα, ορίζουμε τις N συναρτήσεις βάσης του Haar. Η πρώτη συνάρτηση βάσης, η ψ_0, ορίζεται σταθερή και ίση με $+1/\sqrt{N}$ στο διακριτό διάστημα $[0, N-1]$ και μηδενική εκτός αυτού.

$$\psi_0(x) = \frac{1}{\sqrt{N}}, x \in [0, N-1]$$

$$\psi_0(x) = 0, x \notin [0, N-1]$$

Στη συνέχεια, για κάθε ακέραιο j και k, με $0 \leq j < log(N)$ και $0 \leq k < 2^j$, ορίζονται οι υπόλοιπες $N-1$ συναρτήσεις.

$$\phi(x)[j,k] = +\sqrt{\frac{2^j}{N}}, x \in [kN/2^j, kN/2^j + N/2^{j+1} - 1]$$

$$\phi(x)[j,k] = -\sqrt{\frac{2^j}{N}}, x \in [kN/2^j + N/2^{j+1}, (k+1)N/2^j - 1]$$

$$\phi(x)[j,k] = 0, x \notin [kN/2^j, (k+1)N/2^j - 1]$$

Οι συναρτήσεις αυτές δεικτοδοτούνται κατά σειρά αυξανόμενων j και k ως $\psi_1, \ldots, \psi_{N-1}$. Ορίζουμε ως support μιας συνάρτησης βάσης το διάστημα του πεδίου ορισμού της στο οποίο δεν είναι μηδενική. Έτσι, το support μιας συνάρτησης βάσης μπορεί να είναι είτε το διάστημα $[0, N-1]$ είτε, αναδρομικά, το αριστερό ή το δεξιό μισό του support μιας άλλης συνάρτησης.

Ένα σήμα $A[0 \ldots N-1]$ μετασχηματίζεται σε ένα σήμα $C[0 \ldots N-1]$, με το i-οστό του στοιχείο να ισούται με το εσωτερικό γινόμενο του αρχικού σήματος A με την i-οστή συνάρτηση βάσης.

$$C[i] = \langle A, \psi_i \rangle$$

Τα στοιχεία του σήματος C ονομάζονται συντελεστές του μετασχηματισμού Haar. Με δεδομένα τη βάση Haar και το σήμα C είναι δυνατή η ανάκτηση του αρχικού σήματος.

$$A = \sum_{i \in [0, N-1]} C[i]\psi_i$$

Όπως μπορούμε εύκολα να διαπιστώσουμε, ο Μ/Σ Haar είναι ορθοκανονικός. Αυτό, εξαιτίας ενός γνωστού θεωρήματος του Parseval, συνεπάγεται ότι το μετασχηματιμένο σήμα διατηρεί την ενέργεια του αρχικού.

$$\sum_{i\in[0,N-1]} A[i]^2 = \sum_{i\in[0,N-1]} C[i]^2$$

Ακόμα, κοιτώντας το Μ/Σ Haar από αλγοριθμική σκοπιά, βλέπουμε ότι τόσο η διαδικασία μετασχηματισμού όσο και η διαδικασία ανάκτησης του αρχικού σήματος γίνονται σε O(N) χρόνο.

Εξετάζουμε στη συνέχεια ένα παράδειγμα εφαρμογής Μ/Σ Haar. Έστω ότι έχουμε το σήμα A μεγέθους $N = 4$, με $A = (a, b, c, d)$. Με βάση τα όσα προηγήθηκαν, υπολογίζουμε τη βάση του μετασχηματισμού (παραλείπουμε τα διαστήματα που οι συναρτήσεις είναι μηδενικές).

$$\psi_0(x) = +\frac{1}{2}, x \in [0, 3]$$

$$\psi_1(x) = +\frac{1}{2}, x \in [0, 1]$$

$$\psi_1(x) = -\frac{1}{2}, x \in [2, 3]$$

$$\psi_2(x) = +\sqrt{\frac{1}{2}}, x \in [0, 0]$$

$$\psi_2(x) = -\sqrt{\frac{1}{2}}, x \in [1, 1]$$

$$\psi_3(x) = +\sqrt{\frac{1}{2}}, x \in [2, 2]$$

$$\psi_3(x) = -\sqrt{\frac{1}{2}}, x \in [3, 3]$$

Κατόπιν υπολογίζουμε τους συντελεστές του μετασχηματισμένου σήματος.

$$C[0] = \langle A, \psi_0 \rangle = \frac{1}{2}(a + b + c + d)$$

$$C[1] = \langle A, \psi_1 \rangle = \frac{1}{2}(a + b - c - d)$$

$$C[2] = \langle A, \psi_2 \rangle = \sqrt{\frac{1}{2}}(a - b)$$

$$C[3] = \langle A, \psi_3 \rangle = \sqrt{\frac{1}{2}}(c - d)$$

Παρατηρούμε ότι ο μετασχηματισμός διατηρεί την ενέργεια του σήματος.

$$C[0]^2 + C[1]^2 + C[2]^2 + C[3]^2 = a^2 + b^2 + c^2 + d^2$$

Θα δώσουμε τώρα ένα διαφορετικό και λιγότερο τυπικό τρόπο προσέγγισης του μετασχηματισμού Haar DWT. Αν παρατηρήσουμε τις συναρτήσεις βάσης του, θα δούμε ότι οι συντελεστές προκύπτουν ως αθροίσματα ή διαφορές δυαδικού πλήθους αρχικών στοιχείων του σήματος. Η ιδέα στην οποία βασίζεται ο μετασχηματισμός είναι ότι ένα ζεύγος τιμών $p = [a, b]$ μπορεί να παρασταθεί και να ανακτηθεί από το ημιάθροισμα και την ημιδιαφορά των δύο τιμών, δηλ από το ζεύγος $m = [\frac{a+b}{2}, \frac{a-b}{2}]$. Πράγματι, $a = \frac{a-b}{2} + \frac{a+b}{2}$ και $b = \frac{a+b}{2} - \frac{a-b}{2}$.

Έστω, λοιπόν, ότι έχουμε το διάνυσμα τιμών A, μεγέθους N, με το N να είναι μια δύναμη του 2. Στο πρώτο βήμα σχηματίζουμε ένα νέο διάνυσμα $S_1[0 \dots (N/2-1)]$ μεγέθους $N/2$ επιλέγοντας διαδοχικά ζεύγη τιμών από το A και υπολογίζοντας το ημιάθροισμά τους. Με τον ίδιο τρόπο σχηματίζουμε το διάνυσμα ημιδιαφορών $D_1[0 \dots (N/2-1)]$, δηλαδή για διαδοχικά ζεύγη τιμών του A υπολογίζουμε την ημιδιαφορά τους και την τοποθετούμε στο D. Παρατηρούμε ότι ως εδώ δεν έχουμε απώλεια πληροφορίας, αφού κάθε στοιχείο του αρχικού διανύσματος ανακτάται εύκολα από τα στοιχεία των νέων διανυσμάτων. Για παράδειγμα, $A[0] = S[0] + D[0] = (A[0] + A[1])/2 + (A[0] - A[1])/2$, $A[1] = S[0] - D[0] = (A[0] + A[1])/2 - (A[0] - A[1])/2$, κοκ. Η διαδικασία που περιγράψαμε επαναλαμβάνεται αναδρομικά για το διάνυσμα ημιαθροισμάτων, μέχρι να φτάσουμε σε διανύσματα μεγέθους 1 . Το συνολικό πλήθος ημιδιαφορών που υπολογίζουμε είναι $1 + 2 + \ldots + N/2 = N - 1$. Από αυτή τη διαδικασία σχηματίζεται το διάνυσμα W μεγέθους N, το οποίο αποτελείται από τό μέσο όρο των τιμών του αρχικού σήματος (το τελευταίο ημιάθροισμα που υπολογίζουμε) και τις $N-1$ ημιδιαφορές. Το διάνυσμα αυτό είναι ο μη-κανονικοποιημένος μετασχηματισμός Haar του A. Μπορούμε διαισθητικά να καταλάβουμε ότι για να έχουμε κανονικοποιημένο αποτέλεσμα πρέπει να δώσουμε περισσότερο βάρος στα τελευταία από τα $logN$ βήματα, καθώς αυτά αφορούν περισσότερα στοιχεία του αρχικού σήματος. Έτσι, πολλαπλασιάζουμε κάθε συντελεστή που υπολογίστηκε στο k-οστό βήμα με $\sqrt{N/2^{logN-k}}$ και παράγουμε τον κανονικοποιημένο μετασχηματισμό C (συντελεστές wavelet) του A.

Εξετάζουμε και πάλι το παράδειγμα εφαρμογής Μ/Σ Haar για το σήμα A μεγέθους $N = 4$, με $A = (a, b, c, d)$. Με βάση τα όσα προηγήθηκαν, υπολογίζουμε διαδοχικά τα διανύσματα ημιαθροισμάτων και ημιδιαφορών (Σχήμα 2.1).

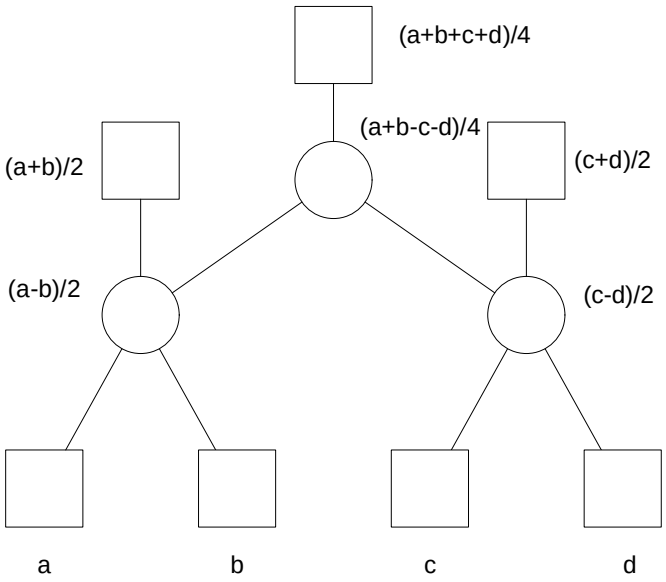
$$S_1 = [\frac{a+b}{2}, \frac{c+d}{2}] \quad D_1 = [\frac{a-b}{2}, \frac{c-d}{2}]$$

$$S_2 = [\frac{a+b+c+d}{4}] \quad D_2 = [\frac{a+b-c-d}{4}]$$

Έτσι, παίρνουμε το διάνυσμα $W = [\frac{a+b+c+d}{4}, \frac{a+b-c-d}{4}, \frac{a-b}{2}, \frac{c-d}{2}]$, ενώ οι κανονικοποιημένοι συντελεστές wavelet προκύπτουν ένας προς έναν ίσοι με τους συντελεστές που υπολογίστηκαν με τον τυπικό ορισμό του Μ/Σ Haar.

$$C[0] = \sqrt{\frac{4}{2^{2-2}}} \frac{a+b+c+d}{4} = \frac{1}{2}(a+b+c+d)$$

$$C[1] = \sqrt{\frac{4}{2^{2-2}}} \frac{a+b-c-d}{4} = \frac{1}{2}(a+b-c-d)$$



Σχήμα 2.1: Παράδειγμα Μ/Σ Haar για το διανύσμα (a, b, c, d).

$$C[2] = \sqrt{\frac{4}{2^{2-1}}}\frac{a-b}{2} = \sqrt{\frac{1}{2}}(a-b)$$

$$C[3] = \sqrt{\frac{4}{2^{2-1}}}\frac{c-d}{2} = \sqrt{\frac{1}{2}}(c-d)$$

Το αρχικό σήμα A μπορεί να ανακατασκευαστεί είτε από τους κανονικοποιημένους συντελεστές wavelet ($C[N]$.

$$A[0] = \frac{1}{2}C[0]+\frac{1}{2}C[1]+\sqrt{\frac{1}{2}}C[2] = \frac{1}{2}\cdot\frac{1}{2}(a+b+c+d)+\frac{1}{2}\cdot\frac{1}{2}(a+b-c-d)+\sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}}(a-b) = a$$

$$A[1] = \frac{1}{2}C[0]+\frac{1}{2}C[1]-\sqrt{\frac{1}{2}}C[2] = \frac{1}{2}\cdot\frac{1}{2}(a+b+c+d)+\frac{1}{2}\cdot\frac{1}{2}(a+b-c-d)-\sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}}(a-b) = b$$

$$A[2] = \frac{1}{2}C[0]-\frac{1}{2}C[1]+\sqrt{\frac{1}{2}}C[3] = \frac{1}{2}\cdot\frac{1}{2}(a+b+c+d)-\frac{1}{2}\cdot\frac{1}{2}(a+b-c-d)+\sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}}(c-d) = c$$

$$A[3] = \frac{1}{2}C[0]-\frac{1}{2}C[1]-\sqrt{\frac{1}{2}}C[3] = \frac{1}{2}\cdot\frac{1}{2}(a+b+c+d)-\frac{1}{2}\cdot\frac{1}{2}(a+b-c-d)-\sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}}(c-d) = d$$

Ή απευθείας από τους μη-κανονικοποιημένους συντελεστές ($W[N]$).

$$A[0] = W[0] + W[1] + W[2] = \frac{a+b+c+d}{4} + \frac{a+b-c-d}{4} + \frac{a-b}{2} = a$$

$$A[1] = W[0] + W[1] - W[2] = \frac{a+b+c+d}{4} + \frac{a+b-c-d}{4} - \frac{a-b}{2} = b$$

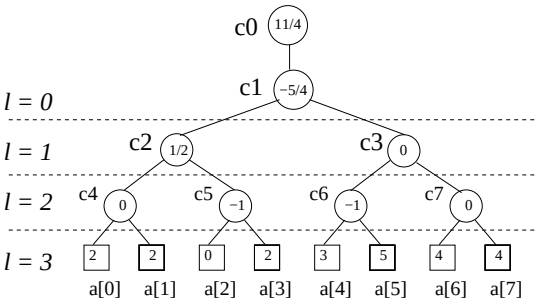
$$A[0] = W[0] - W[1] + W[3] = \frac{a+b+c+d}{4} - \frac{a+b-c-d}{4} + \frac{c-d}{2} = c$$

$$A[0] = W[0] - W[1] - W[3] = \frac{a+b+c+d}{4} - \frac{a+b-c-d}{4} - \frac{c-d}{2} = c$$

Θα προχωρήσουμε με ένα αριθμητικό παράδειγμα. Έστω, λοιπόν, ότι έχουμε το διάνυσμα

Resolution	Averages	Detail Coefficients
3	[2, 2, 0, 2, 3, 5, 4, 4]	-----
2	[2, 1, 4, 4]	[0, −1, −1, 0]
1	[3/2, 4]	[1/2, 0]
0	[11/4]	[−5/4]

(a)



(b)

Σχήμα 2.2: Παράδειγμα δένδρου σφάλματος για το διανύσμα a.

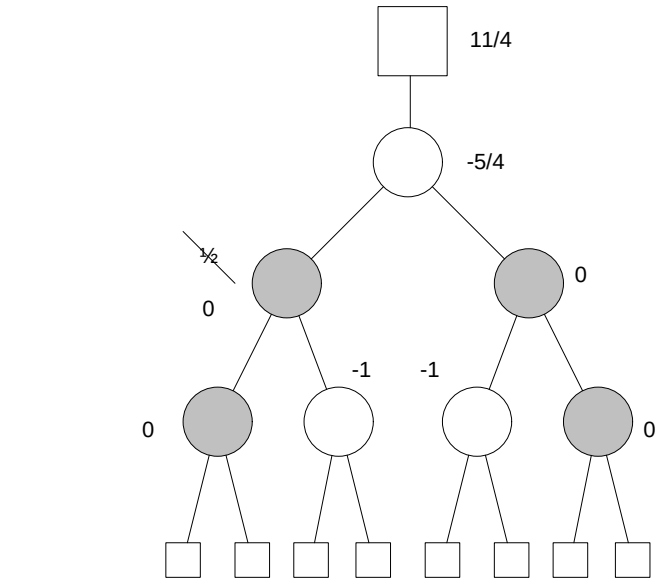
τιμών $a = [2, 2, 0, 2,\ 3, 5, 4, 4]$, μεγέθους $N = 8$ (Σχήμα 2.2). Ο μη-κανονικοποιημένος μετασχηματισμός του a είναι το διάνυσμα $w_a = [11/4, -5/4, 1/2,\ 0, 0, -1, -1, 0]$ μεγέθους $N = 8$. Αποτελείται από το μέσο όρο των τιμών του a και από τις τιμές των διανυσμάτων ημιδιαφορών. Οι τιμές αυτές αποτελούν τους συντελεστές του μετασχηματισμού wavelet ενώ οι ημιδιαφορές ονομάζονται και *λεπτομέρειες* (detail coefficients). Στο παράδειγμά μας, ο μέσος όρος των τιμών του a είναι $11/4$, η λεπτομέρεια επιπέδου $l = 0$ είναι $-5/4$, οι λεπτομέρειες επιπέδου $l = 1$ είναι $1/2$ και 0 και οι λεπτομέρειες επιπέδου $l = 2$ είναι $0, -1, -1$ και 0.

Ένας καλός τρόπος να παραστήσουμε και να κατανοήσουμε την ιεραρχική φύση του μετασχηματισμού Haar είναι το δένδρο σφαλμάτων, όπως αυτό φαίνεται στο σχήμα 2.2(b). Η ρίζα του δένδρου, c_0, είναι ο μέσος όρος των τιμών, οι εσωτερικοί κόμβοι αντιστοιχούν στους υπόλοιπους συντελεστές (λεπτομέρειες) και τα φύλλα αντιστοιχούν στα αρχικά δεδομένα. Παρατηρήστε ότι η τιμή ενός φύλλου μπορεί να ανασκευαστεί από τις τιμές των $logN + 1$ εσωτερικών κόμβων που βρίσκονται στο μονοπάτι από τη ρίζα προς το φύλλο. Για παράδειγμα, $a[5] = c_0 - c_1 + c_3 - c_6 \Leftrightarrow 5 = \frac{11}{4} - \left(-\frac{5}{4}\right) + 0 - (-1)$. Το πρόσημο του όρου στο άθροισμα είναι $+$ ή $-$ όταν το φύλλο βρίσκεται στο αριστερό ή το δεξί φύλλο του όρου, αντίστοιχα.

2.3 Κατασκευή Περιλήψεων

Όπως μπορούμε να παρατηρήσουμε από την περιγραφή του M/Σ Haar, όταν γειτονικά δεδομένα έχουν παρόμοιες τιμές, παράγονται συντελεστές - λεπτομέρειες με μικρό μέτρο (κοντά στο 0). Αν θέσουμε την τιμή αυτών των συντελεστών ίση με 0 περιμένουμε ότι το σφάλμα που θα παρουσιάζεται στα νέα ανακατασκευασμένα δεδομένα, θα είναι ΄μικρό΄. Αυτή η ιδιότητα του M/Σ Haar τον κάνει κατάλληλο στη χρήση του για κατασκευή περιλήψεων δεδομένων.

Μια περίληψη B όρων $\hat{C}$ ορίζεται επιλέγοντας ένα σύνολο $\Lambda \subset C$ συντελεστών, με $B = |\Lambda| \ll N$, ενώ οι υπόλοιποι $N - B$ όροι θεωρούνται ίσοι με μηδέν.Το πόσο ΄μικρό΄ είναι το σφάλμα που εισάγεται στα δεδομένα μας όταν επιλέξουμε να θεωρήσουμε κάποιους συντελεστές ίσους με 0 και να κρατήσουμε την τιμή των υπολοίπων, μπορεί να μετρηθεί με διάφορους τρόπους. Το σφάλμα αυτό υπολογίζεται από συναρτήσεις που ονομάζονται μετρικές σφάλματος. Αν A είναι το αρχικό σήμα και $\hat{A}$ το ανακατασκευασμένο από την περίληψη



Αρχικές Τιμές	2	2	0	2	3	5	4	4
Νέες Τιμές	1.5	1.5	0.5	2.5	3	5	4	4
Απόλυτο Σφάλμα	0.5	0.5	0.5	0.5	0	0	0	0

Σχήμα 2.3: Παράδειγμα περίληψης για τη μετρική L_∞.

σήμα, η μετρική σφάλματος στην ουσία μας παρέχει ένα μέτρο του διανύσματος $A - \hat{A}$.

$$error = f_{metric}(A - \hat{A}) = ||A - \hat{A}||_{f_{metric}}$$

Το πρόβλημα που καλούμαστε να λύσουμε δεδομένων του διαθέσιμου χώρου περίληψης B και μιας μετρικής $f_{metric}()$ είναι να επιλέξουμε τους B όρους από το μετασχηματισμένο σήμα ώστε το σφάλμα που δίνει η μετρική να είναι ελάχιστο.

Μία μετρική είναι η L_∞ ή μετρική μέγιστου απόλυτου σφάλματος.

$$L_\infty = maxAbsErr(A - \hat{A}) = \max_{0 \le i < N} |\hat{A}[i] - A[i]|$$

Αν υποθέσουμε ότι έχουμε το παράδειγμα του σχήματος 2.2 και θέλουμε να φτάξουμε μια περίληψη με $B = 4$ συντελεστές ώστε να ελαχιστοποιείται η μετρική L_∞, επιλέγουμε να αγνοήσουμε τους συντελεστές c_2, c_3, c_4 και c_7 (οι τρεις τελευταίοι είναι ήδη μηδενικοί) με μέγιστο απόλυτο σφάλμα $1/2$.

Στις επόμενες ενότητες θα περιγράψουμε διάφορες μετρικές σφάλματος και κάποιους αλγορίθμους που λύνουν το πρόβλημα της βέλτιστης περίληψης για κάποιες από αυτές.

Κεφάλαιο 3

Weighted-L_p μετρικές για Point Errors

3.1 Εισαγωγή

Η πρώτη μεγάλη οικογένεια μετρικών σφάλματος είναι εκείνες που ο υπολογισμός του σφάλματος βασίζεται στα σημειακά σφάλματα - point errors που εισάγονται στο ανακατασκευασμένο από την περίληψη σήμα. Ως σημειακό σφάλμα, εννοούμε το σφάλμα στην τιμή κάθε στοιχείου του σήματος, πριν και μετά την περίληψη. Δύο συχνά χρησιμοποιούμενοι τύποι σημειακού σφάλματος είναι το απόλυτο και το σχετικό σφάλμα στην τιμή ενός στοιχείου.

$$err_{abs}(i) = |A[i] - \hat{A}[i]|$$

$$err_{rel}(i) = \frac{|A[i] - \hat{A}[i]|}{\max\{s, |A[i]|\}}$$

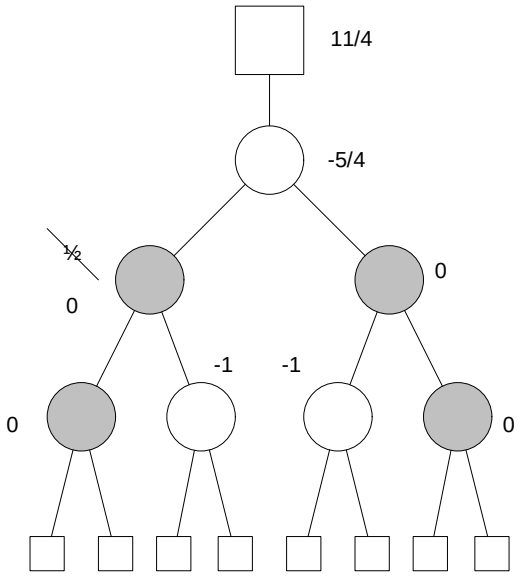
Η σταθερά s χρησιμοποιείται ώστε πολύ μικρές τιμές να μην κυριαρχούν στο σχετικό σφάλμα. Οι μετρικές σφάλματος που χρησιμοποιούν σημειακά σφάλματα, επιστρέφουν μια τιμή συνυπολογίζοντας τα N σημειακά σφάλματα που περιλαμβάνει ένα ανακατασκευασμένο σήμα μεγέθους N. Στη συνέχεια θα χρησιμοποιούμε το απόλυτο σφάλμα ως σημειακό σφάλμα, εκτός κι αν αναφέρεται διαφορετικά.

Αυτή η οικογένεια μετρικών χρησιμοποιείται όταν εφαρμόζουμε point queries πάνω στα δεδομένα μας. Αυτό συμβαίνει διότι μας ενδιαφέρει να έχουμε μια εκτίμηση για το μέσο αναμενόμενο σφάλμα που θα έχουμε για κάθε point query.

Μια κατηγορία μετρικών για point errors είναι οι Weighted-L_p μετρικές.

$$weightedL_p = \sum_i w_i \cdot (err(i))^p$$

Οι μετρικές αυτές αθροίζουν τα σημειακά σφάλματα όλων των στοιχείων του σήματος υψωμένα σε μια δύναμη p και ίσως πολλαπλασιασμένα με κάποιο ΄βάρος΄ που τους αντιστοιχεί. Μερικές μετρικές σφάλματος που ανήκουν στην κατηγορία αυτή είναι οι L_∞ με απόλυτο σημειακό



Αρχικές Τιμές	2	2	0	2	3	5	4	4
Νέες Τιμές	1.5	1.5	0.5	2.5	3	5	4	4
Απόλυτο Σφάλμα	0.5	0.5	0.5	0.5	0	0	0	0
Τετραγωνικό Σφάλμα	0.25	0.25	0.25	0.25	0	0	0	0

Σχήμα 3.1: Παράδειγμα υπολογισμού τετραγωνικών σφαλμάτων για τη μετρική L_2.

σφάλμα, L_∞ με σχετικό σημειακό σφάλμα, L_2 και weighted-L_2.

$$absL_\infty = \max_i \left(err_{abs}(i)\right)$$

$$relL_\infty = \max_i \left(err_{rel}(i)\right)$$

$$L_2 = \sum_i \left(err_{abs}(i)\right)^2$$

$$weightedL_2 = \sum_i w_i \cdot \left(err_{abs}(i)\right)^2$$

Στο σχήμα 3.1 φαίνεται ο υπολογισμός των τετραγωνικών σφαλμάτων για το προηγούμενο παράδειγμα περίληψης. Για τη συγκεκριμένη επιλογή συντελεστών, το σφάλμα που δίνει η μετρική L_2 προκύπτει ίσο με $0.25 + 0.25 + 0.25 + 0.25 = 1$.

Στη συνέχεια θα δούμε αλγορίθμους που δεδομένου κάποιου ορίου στο χώρο που μπορούμε να διαθέσουμε για τους συντελεστές που μένουν στην περίληψη και δεδομένης κάποιας συνάρτησης που ανήκει στην κατηγορία των Weighted-L_p μετρικών, επιλέγουν τους συντελεστές ώστε το σφάλμα αυτό να ελαχιστοποιείται. Όπως είπαμε ήδη, είναι δυνατόν να δίνουμε διαφορετικό βάρος σε κάθε point query και στο αντίστοιχο σφάλμα που εισάγεται στην τιμή του λόγω της περίληψης. Οι αλγόριθμοι που εξετάζουμε χωρίζονται σε δύο κατηγορίες με βάση τον τρόπο που χειρίζονται τα διαφορετικά βάρη στα σημειακά σφάλματα. Η μία προσέγγιση εφαρμόζει τον κλασσικό μετασχηματισμό Haar Wavelet, όπως τον περιγράψαμε στην

ενότητα 2.2 και εντάσσει το χειρισμό των βαρών (αν υπάρχουν) στο αλγόριθμο που υπολογίζει το ελάχιστο σφάλμα και επιλέγει συντελεστές. Η δεύτερη προσέγγιση τροποποιεί τον κλασσικό M/Σ Haar ώστε να ενσωματώνονται σε αυτόν τα διαφορετικά βάρη των σημειακών σφαλμάτων και στη συνέχεια υπολογίζεται το ελάχιστο σφάλμα και επιλέγονται οι συντελεστές της περίληψης.

3.2 Η πρώτη προσέγγιση: τα δεδομένα μετασχηματίζονται με τον κλασσικό M/Σ Haar

3.2.1 Εισαγωγή

Στην ενότητα αυτή εξετάζουμε αλγορίθμους που υπολογίζουν το ελάχιστο σφάλμα που μπορεί να περιέχει μια περίληψη B όρων, αφού πρώτα εφαρμοστεί ο κλασσικός μετασχηματισμός Haar στο αρχικό σήμα. Οι αλγόριθμοι αυτοί, δηλαδή, δέχονται ως είσοδο το αρχικό σήμα μετασχηματισμένο με τον κλασσικό MΣ Haar και το μέγιστο επιτρεπόμενο μέγεθος περίληψης B και δίνουν ως έξοδο το ελάχιστο δυνατό σφάλμα, υπολογισμένο με τη χρήση μιας μετρικής. Η μετρική που χρησιμοποιείται ανήκει στις weighted-L_p μετρικές για point errors. Η επιλογή των συντελεστών με τους οποίους επιτυγχάνεται η βέλτιστη περίληψη, μπορεί να γίνεται προγραμματιστικά είτε σε ένα δεύτερο πέρασμα, αφού πρώτα υπολογιστεί το ελάχιστο σφάλμα, είτε παράλληλα με τον υπολογισμό του ελάχιστου σφάλματος. Για λόγους απλότητας, η επιλογή των συντελεστών δεν περιγράφεται στην περιγραφή των περισσότερων αλγορίθμων που ακολουθούν.

3.2.2 Ο Άπληστος αλγόριθμος επιλογής συντελεστών για τη μετρική L_2

Στην ενότητα αυτή παρουσιάζεται ο απλούστερος αλγόριθμος περίληψης. Ο αλγόριθμος αυτός κρατάει στην περίληψη τους B μεγαλύτερους κατ' απόλυτη τιμή συντελεστές, πρόκειται δηλαδή για έναν άπληστο αλγόριθμο επιλογής συντελεστών. Όταν εφαρμόζεται σε σύνολα συντελεστών που έχουν προκύψει από τον κλασσικό M/Σ Haar ελαχιστοποιεί τη μετρική σφάλματος L_2. Η μετρική σφάλματος L_2 ή SSE - Sum of Squared Errors όπως ονομάζεται ορίζεται ως εξής:

$$SSE(e) = \sum_{i=0}^{N-1} e_i^2$$

με A το αρχικό σήμα-διάνυσμα, C το μετασχηματισμένο και κανονικοποιημένο διάνυσμα, $\hat{C}$ η περίληψη Λ, $\hat{A} = W^{-1}\{\hat{C}_A\}$ τα δεδομένα όπως επανακτώνται μετά την περίληψη και $e = A - \hat{A}$ το διάνυσμα σφάλματος.

Ας αναλύσουμε λίγο το πώς διαπιστώνουμε αν ο άπληστος αλγόριθμος επιλογής συντελεστών μπορεί να χρησιμοποιηθεί για την κατασκευή της βέλτιστης περίληψης Λ ως προς κάποια μετρική. Έστω ο διανυσματικός χώρος $\Re^N$, εφοδιασμένος με ένα εσωτερικό γινόμενο $p(u, v)$. Το εσωτερικό αυτό γινόμενο ορίζει μια νόρμα για κάθε διάνυσμα που ανήκει στο

χώρο, την $n(u) = ||u|| = \sqrt{p^2(u,u)}$. Επίσης, μια βάση S του διανυσματικού χώρου $\Re^N$ είναι ορθοκανονική ως προς το εσωτερικό γινόμενο $p(u,v)$, όταν τα διανύσματα που την αποτελούν είναι ανά δύο κάθετα και επιπλέον η νόρμα του καθενός ισούται με τη μονάδα.

$$p(s_i, s_j) = 0$$

$$p(s_i, s_i) = 1$$

Έστω, λοιπόν, ότι η βάση S του διανυσματικού χώρου είναι ορθοκανονική. Τότε, κάθε διάνυσμα v του $\Re^N$ μπορεί να γραφτεί ως γραμμικός συνδυασμός των στοιχείων της βάσης S, ως $v = \sum a_i \cdot s_i$ και το διάνυσμα σφάλματος $e \in \Re^N$ γράφεται ως $e = A - \hat{A} = \sum_{0 \le i < N} c_i \cdot s_i - \sum_{0 \le i < N} \hat{c}_i \cdot s_i = \sum_{0 \le i < N} c_i \cdot s_i - \sum_{i \in \Lambda} c_i \cdot s_i = \sum_{i \notin \Lambda} c_i \cdot s_i$. Από το θεώρημα Parseval, για κάθε διάνυσμα v ισχύει:

$$||v||^2 = \sum_i a_i^2$$

Ειδικά για το διάνυσμα σφάλματος , έχουμε $||e||^2 = \sum_{i \notin \Lambda} c_i^2$ Τότε, αν η μετρική σφάλματος $f(e)$ ταυτίζεται με το τετράγωνο της νόρμας του διανύσματος σφάλματος, κατά την προσέγγιση το πολύ μιας πολλαπλασιαστικής σταθεράς, μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο για να κατασκευάσουμε τη βέλτιστη περίληψη.

Στην προκειμένη περίπτωση, στο διανυσματικό χώρο $\Re^N$ ανήκουν τα N-διάστατα αρχικά δεδομένα $A[N]$, αλλά και το διάνυσμα σφάλματος. Εφοδιάζουμε το χώρο με το ευκλείδειο εσωτερικό γινόμενο και η νόρμα που ορίζεται από αυτό είναι το γνωστό ευκλείδειο μέτρο. Η βάση του διανυσματικού χώρου είναι η βάση του Haar, όπως περιγράφτηκε στην ενότητα 2.2 και εύκολα βλέπουμε ότι είναι ορθοκανονική ως προς το ευκλείδειο εσωτερικό γινόμενο. Τέλος, η μετρική L_2 ταυτίζεται με το τετράγωνο της ευκλείδειας νόρμας του διανύσματος σφάλματος, δηλαδή ισούται με το άθροισμα των τετραγώνων των (κανονικοποιημένων) συντελεστών που μένουν εκτός της περίληψης. Έτσι, το SSE γίνεται ελάχιστο επιλέγοντας τους B μεγαλύτερους κατ΄ απόλυτη τιμή συντελεστές.

$$SSE = L_2(e) = ||e||^2 = \sum_{i \notin \Lambda} c_i^2$$

Ο αλγόριθμος δουλεύει ως εξής: αρχικά, σε χρόνο $O(N)$, φτιάχνουμε ένα σωρό (max-heap, δηλαδή δυαδικό δένδρο στο οποίο κάθε κόμβος έχει μεγαλύτερη τιμή από τα παιδιά του) των N κανονικοποιημένων συντελεστών του μετασχηματισμού, υπολογίζουμε και αποθηκεύουμε το άθροισμα των τετραγώνων όλων των συντελεστών (έστω Sum). Στη συνέχεια, επιλέγουμε και αφαιρούμε το συντελεστή που βρίσκεται στην κορυφή του σωρού και σε χρόνο $O(\log N)$ ξαναφτιάχνουμε το σωρό. Το βήμα αυτό επαναλαμβάνεται B φορές. Το σφάλμα υπολογίζεται αφαιρώντας κατά την επιλογή των συντελεστών τα τετράγωνά τους από το Sum. Η πολυπλοκότητα του αλγορίθμου είναι $O(N)$ σε χώρο και $O(N + B \log N)$ σε χρόνο (Σχήμα 3.2).

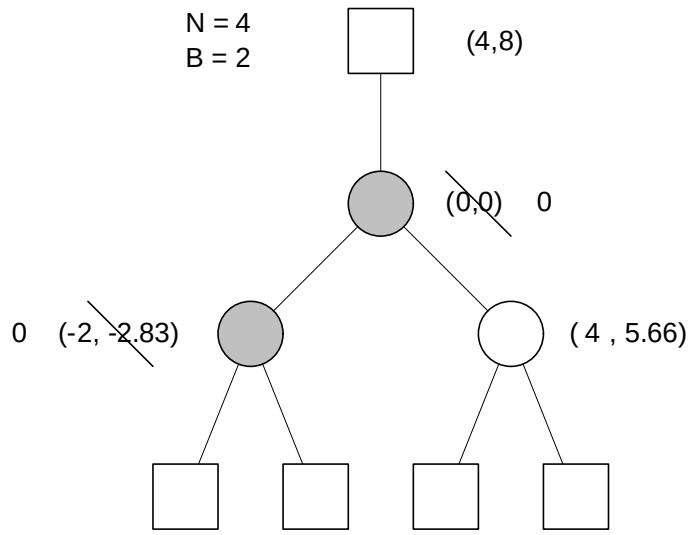
```
procedure greedy (C[N], heap[N])
Input:   C[N] -> οι N συντελεστές Haar σε απόλυτες τιμές
         Heap[N] -> (άδειος) σωρός μεγέθους N
Output: Synopses -> ένα διάνυσμα με τους B μεγαλύτερους (κατ'
                    απόλυτη τιμή) συντελεστές Haar
        SSE -> το άθροισμα τετραγωνικών σφαλμάτων

Begin
    real Sum = 0.0;
    for (i=1 to N) do
        Sum += C[i]*C[i];
    heap = constructHeapCombine (C[N]);        // O(N)
    vector Synopses;
    for (i=1 to B) do
        begin
            Synopses.add(heap.top);
            Sum -= (heap.top.value)*(heap.top.value);
            heap.top.value = -1;
            heap = maxHeapify(heap, heap.top);
        end
    real SSE = Sum;
    return (Synopses, SSE);
end
```

Σχήμα 3.2: Greedy αλγόριθμος για την ελαχιστοποίηση της L_2.

Υπάρχει ένας δεύτερος τρόπος να υλοποιηθεί ο άπληστος αλγόριθμος. Αρχικά, κρατάμε σε ένα min-heap (δηλαδή δυαδικό δένδρο στο οποίο κάθε κόμβος έχει μικρότερη τιμή από τα παιδιά του) μεγέθους $B+1$ τους πρώτους $B+1$ κανονικοποιημένους συντελεστές (κατά σειρά δεικτοδότησης για παράδειγμα). Στη συνέχεια, επιλέγουμε και αφαιρούμε το συντελεστή που βρίσκεται στην κορυφή του σωρού και στη θέση του τοποθετούμε τον επόμενο από τους υπόλοιπους συντελεστές ($N-B-1$ στο πλήθος, για το πρώτο βήμα) και ξαναφτιάχνουμε το σωρό σε χρόνο $O(\log(B+1) = \log B)$. Το βήμα αυτό επαναλαμβάνεται $N-B-1$ φορές, συνολικά. Στο τέλος, οι B συντελεστές του σωρού που βρίσκονται κάτω από τη ρίζα είναι οι B μεγαλύτεροι συντελεστές, οι οποίοι απαρτίζουν την περίληψη. Ο συνολικός χώρος που χρησιμοποιεί ο αλγόριθμος τώρα είναι $O(N)$ και πάλι, το working space του, όμως, είναι $O(B)$ (το μέγεθος του σωρού, δηλαδή) και η χρονική του πολυπλοκότητα είναι $O(B + (N-B-1)\log(B+1)) = O(N\log B + B(1-\log B)) = O(N\log B - B\log B)$. Αυτή η εκδοχή του αλγορίθμου εμφανίζει ελαφρώς μεγαλύτερη χρονική πολυπλοκότητα (στις σταθερές) από την πρώτη για μικρές τιμές του B ($B << N$). Για παράδειγμα, για $B = N/5$, η πολυπλοκότητα της πρώτης εκδοχής προκύπτει ίση με $O(N + \frac{1}{5}N\log N)$ ενώ της δεύτερης ίση με $O(\frac{4}{5}N\log N - \frac{4}{5}Nlog5)$. Όπως φαίνεται, για μεγάλα N κυριαρχεί ο όρος $N\log N$ που εμφανίζεται με καλύτερη σταθερά στην πρώτη περίπτωση.

Στη συνέχεια της εργασίας επιλέγουμε να χρησιμοποιούμε την πρώτη εκδοχή, με χρονική πολυπλοκότητα $O(N + B\log N)$.

Στο σχήμα 3.3 φαίνεται ένα παράδειγμα εφαρμογής του άπληστου αλγορίθμου για την



αρχικές τιμές	2	6	8	0
νέες τιμές	4	4	8	0
απόλυτο σφάλμα	2	2	0	0
τετραγωνικό σφάλμα	4	4	0	0

Σχήμα 3.3: Παράδειγμα περίληψης με τον άπληστο αλγόριθμο για τη μετρική L_2.

επιλογή περίληψης $B = 2$ συντελεστών, ώστε να ελαχιστοποιείται το σφάλμα L_2. Δίπλα σε κάθε συντελεστή, σε παρένθεση, φαίνεται η μη-κανονικοποιημένη και η κανονικοποιημένη του τιμή. Επιλέγονται οι συντελεστές με τη απολύτως μεγαλύτερη κανονικοποιημένη τιμή και το ελάχιστο L_2 σφάλμα που προκύπτει είναι $4 + 4 = 8$.

3.2.3 Ο Αλγόριθμος των Garofalakis και Kumar για τη μετρική L_∞

Η μετρική σφάλματος L_∞ μας δίνει το μέγιστο από τα σφάλματα που περιέχει μια περίληψη για τα στοιχεία του αρχικού σήματος (μέγιστο σημειακό σφάλμα). Το σφάλμα αυτό μπορεί να είναι απόλυτο ή σχετικό. Έχουμε, λοιπόν δύο δυνατούς ορισμούς για αυτή τη μετρική

$$L_\infty = absErr(a - \hat{a}) = \max_{1 \le i \le N} |\hat{a}[i] - a[i]|$$

και

$$L_\infty = relErr(a - \hat{a}) = \max_{1 \le i \le N} \frac{|\hat{a}[i] - a[i]|}{max\{|a[i]|, s\}}$$

όπου s μια σταθερά εκλογίκευσης του σχετικού σφάλματος που δεν επιτρέπει σε πολύ μικρές τιμές του σήματος να κυριαρχούν στη μέτρηση του σφάλματος. Το πρόβλημά που αντιμετωπίζουμε, λοιπόν, είναι η επιλογή B συντελεστών της περίληψης ώστε να ελαχιστοποιείται το μέγιστο σημειακό σφάλμα. Η επιλογή του ενός ή του άλλου ορισμού δεν επηρεάζει σημαντικά τη λύση του προβλήματος.

Η αρχική μορφή του αλγορίθμου

Οι Garofalakis και Kumar στο [1] προτείνουν έναν αλγόριθμο δυναμικού προγραμματισμού (MinMaxErr, Σχήμα 3.4) που λύνει το πρόβλημα για τη μετρική L_∞. Η βασική ιδέα του αλγορίθμου είναι ότι λύνει το (υπο)πρόβλημα της επιλογής συντελεστών για το υποδένδρο με ρίζα τον κόμβο c_i λαμβάνοντας υπόψιν την επιλογή συντελεστών που έχει γίνει για το μονοπάτι από τη ρίζα του δένδρου (κόμβος c_0).

Προτού προχωρήσουμε στην περιγραφή του αλγορίθμου δίνουμε κάποιους ορισμούς και παραδοχές που θα χρησιμοποιήσουμε. Μας δίνονται το αρχικό σήμα μαζί με το μετασχηματισμένο και κανονικοποιημένο σήμα (μέσω του δένδρου σφάλματος) και το πλήθος B των συντελεστών που θα αποτελέσουν την περίληψή μας. Με T_j συμβολίζουμε το υποδένδρο σφάλματος που έχει ρίζα τον κόμβο c_j και ως $coeff(T_j)$ και $data(T_j)$ ορίζουμε τα σύνολα των συντελεστών (εσωτερικοί κόμβοι) και των αρχικών δεδομένων (φύλλα), αντίστοιχα, που ανήκουν στο T_j. Με $path(c_j)$ συμβολίζουμε το σύνολο των συντελεστών που ανήκουν στο μονοπάτι από τη ρίζα του δένδρου σφαλμάτων (c_0) ως τον κόμβο c_j (χωρίς αυτόν). Τέλος, με $M[j, b, S]$ συμβολίζουμε την ελάχιστη τιμή του μέγιστου σημειακού σφάλματος που εισέρχεται στην περίληψη επιλέγοντας b συντελεστές του T_j με την υπόθεση ότι έχουμε ήδη επιλέξει ένα σύνολο $S \subseteq path(c_j)$ μεγέθους το πολύ $min\{B - b, logN + 1\}$. Έτσι, θεωρώντας ότι δουλεύουμε με το σχετικό σφάλμα,

$$M[j, b, S] = \min_{S_j \subseteq coeff(T_j), \|S_j\| \le b} \{ \max_{d_i \in data(T_j)} relErr_i \}$$

με

$$relErr_i = \frac{|d_i - \sum_{c_k \in path(d_i) \cap S_i \cup S} sign_{i,k} \cdot c_k|}{max\{|d_i|, s\}}.$$

Με παρόμοιο τρόπο αντιμετωπίζουμε και το απόλυτο σφάλμα.

Το επιθυμητό αποτέλεσμα δίνεται από την τιμή του $M[0, B, \emptyset]$. Η βάση της αναδρομής βρίσκεται στα φύλλα, δηλαδή τους κόμβους $c_j = d_{j-N+1}$, για $j \geq N$. Τα φύλλα φυσικά δεν ανήκουν στην περίληψη οπότε έχουμε $b = 0$.

$$M[j, 0, S] = \frac{|d_{j-N+1} - \sum_{c_k \in S} sign_{j-N,k} \cdot c_k|}{max\{|d_{j-N+1}, s\}}$$

Για τους εσωτερικούς κόμβους του δένδρου σφάλματος ο αλγόριθμος εξετάζει δύο επιλογές: να κρατήσει τον κόμβο c_j στην περίληψη ή να τον απορρίψει. Αν τον απορρίψει, τότε το ελάχιστο μέγιστο σφάλμα για το T_j είναι το μεγαλύτερο από τα ελάχιστα μέγιστα σφάλματα για τα υποδένδρα T_{2j} και T_{2j+1}. Ο συνολικός χώρος περίληψης και οι προεπιλεγμένοι κόμβοι που μπορούν να εκμεταλλευτούν τα δύο υποδένδρα είναι ίδιοι με του T_j.

$$M_{drop}[j, b, S] = \min_{0 \leq \dot{b} \leq b} max\{M[2j, \dot{b}, S], M[2j+1, b-\dot{b}, S]\}$$

Αν από την άλλη κρατήσει τον κόμβο c_j, τότε κατά τον υπολογισμό του ελάχιστου μέγιστου σφάλματος των T_{2j} και T_{2j+1} ο κόμβος c_j προστίθεται στους προεπιλεγμένους κόμβους ενώ προσαρμόζεται και ο διαθέσιμος χώρος περίληψης.

$$M_{keep}[j, b, S] = \min_{0 \leq \dot{b} \leq b-1} max\{M[2j, \dot{b}, S \cup \{c_j\}], M[2j+1, b-\dot{b}-1, S \cup \{c_j\}]\}$$

Τελικά ο αλγόριθμος επιλέγει την καλύτερη από τις δύο επιλογές.

$$M[j, b, S] = min\{M_{drop}[j, b, S], M_{keep}[j, b, S]\}$$

Για ένα συγκεκριμένο κόμβο c_j επιπέδου l στο δένδρο σφάλματος, ο αλγόριθμος έχει να εξετάσει το πολύ $B+1$ περιπτώσεις όσον αφορά το πλήθος των συντελεστών που θα κρατήσει στο υποδένδρο T_j (συνυπολογίζοντας την περίπτωση να κρατήσει 0 συντελεστές). Ακόμα, για έναν κόμβο επιπέδου l υπάρχουν 2^l υποσύνολα προγόνων να εξεταστούν. Έτσι, στον κόμβο c_j αντιστοιχούν $O(B2^l)$ τιμές του πίνακα $M[]$. Αφού υπάρχουν 2^l κόμβοι επιπέδου l, ο συνολικός χώρος του αλγορίθμου είναι

$$\sum_{l=0}^{l=logN} 2^l B 2^l = O(N^2 B).$$

Ακόμα, για να υπολογίσουμε το κάθε στοιχείο του πίνακα χρειαζόμαστε $O(logB)$ χρόνο: αφού το $M[2j, \dot{b}, S]$ είναι φθίνουσα συνάρτηση του $\dot{b}$ ενώ το $M[2j+1, b-\dot{b}, S]$ είναι αύξουσα συνάρτηση του $\dot{b}$, μπορούμε να εκτελέσουμε δυαδική αναζήτηση για το $\dot{b}$, ώστε να βρούμε το σημείο που τα δύο σφάλματα γίνονται ίσα (και άρα το σφάλμα του κόμβου-γονέα γίνεται

procedure MinMaxErr(W_A, B, root, S, err)
Input: Array W_A=[c0,c1,...,cN-1] of N Haar wavelet coefficients, space budget **B** (number of retained coefficients), error-subtree root-node index **root**, subset of retained ancestors of root node **S**, target maximum error metric **err**
Output: Value of M[root,B,S] according to our optimal dynamic program (M[root,B,S].value), decision made for the root node (M[root,B,S].retained) and space alloted to left child subtree (M[root,B,S].leftAllot). (The last two are used for re-tracing the optimal solution to build the synopsis).

```
BEGIN
1   IF (M[root,B,S].computed = TRUE) THEN
2     RETURN M[root,B,S].value          //optimal value already in M[]
3   IF (N ≤ root < 2N) THEN
4     IF (B=0) THEN {
5       M[root,B,S].value:=|a_{j-N} − Σ_{c_k∈S} sign_{j-N,k} · c_k|
6       IF (err=relErr) THEN
7         M[root,B,S].value:=M[root,B,S].value / max{|a_{root-N}|,S}
8     }
9   ELSE {
10      M[root,B,S].value:=∞
11      FOR b:=0 TO B STEP 1 DO {          //first choice: drop root
12        left:=MinMaxErr(W_A, b, 2*root, S, err)
13        right:=MinMaxErr(W_A, B−b, 2*root+1, S, err)
14        IF (max{left,right} < M[root,B,S].value) THEN {
15          M[root,B,S].value:=max{left,right}
16          M[root,B,S].retained:=FALSE
17          M[root,B,S].leftAllot:=b
18        }
19      }
20      FOR b:=0 TO B-1 STEP 1 do {        //second choice: keep root
21        left:=MinMaxErr(W_A, b, 2*root, S∪{root}, err)
22        right:=MinMaxErr(W_A, B−b, 2*root+1, S∪{root}, err)
23        IF (max{left,right} < M[root,B,S].value) THEN {
24          M[root,B,S].value:=max{left,right}
25          M[root,B,S].retained:=TRUE
26          M[root,B,S].leftAllot:=b
27        }
28      }
29  }
30  M[root,B,S].computed:=TRUE
31  RETURN M[root,B,S].value
END
```

Σχήμα 3.4: Ο αλγόριθμος MinMaxErr

ελάχιστο). Όπως και προηγουμένως, υπολογίζουμε ότι ο συνολικός χρόνος του αλγορίθμου είναι $O(N^2BlogB)$. Έτσι, προέκυψε ότι η πολυπλοκότητα του αλγορίθμου είναι $O(N^2BlogB)$ σε χρόνο και $O(N^2B)$ σε χώρο.

Οι παρατηρήσεις του Guha

Ο Guha στο [3] αποδεικνύει ότι ο αλγόριθμος των Garofalakis και Kumar μπορεί να βελτιωθεί ως προς την πολυπλοκότητα. Η βασική του παρατήρηση είναι ότι δοθέντος ενός εσωτερικού κόμβου c_j επιπέδου l -άρα με $l+1$ προγόνους - και διαθέσιμου χώρου περίληψης B, το μέγιστο πλήθος των συντελεστών του δένδρου T_j που μπορούν να προστεθούν στην περίληψη είναι $min\{B,t\}$, όπου t το πλήθος των κόμβων που ανήκουν στο υποδένδρο με ρίζα το c_j, συμπεριλαμβανομένου του ιδίου. Είναι $t=t_0=0$ για τη ρίζα του δένδρου σφάλματος και $t=t_1=2^{logN-l}-1$ για τους υπόλοιπους κόμβους. Ο κόμβος καλείται $2^l<2N$ φορές, όσα και τα διαφορετικά μονοπάτια απογόνων από τη ρίζα προς τον κόμβο, ανάλογα με το αν

ένας πατρικός κόμβος έχει μείνει στην περίληψη ή όχι. Ο συνολικός χρόνος που απαιτείται για έναν κόμβο είναι, λοιπόν, $2^l min\{B,t\} log min\{B,t\}$ και αφού οι κόμβοι επιπέδου l είναι 2^l στο πλήθος (εκτός από το επίπεδο $l=0$ έχουμε 2 κόμβους), ο συνολικός χρόνος του αλγορίθμου είναι

$$\sum_{l=0}^{logN} (2^l 2^l \min\{B,t_1\} \log \min\{B,t_1\}) + \min\{B,t_0\} \log \min\{B,t_0\}.$$

Στη χειρότερη περίπτωση (worst case) είναι $B = N \geq t$. Ακόμα, ισχύει $2^l(t+1) = 2N \Rightarrow 2^l t = O(2N)$. Έτσι, η πολυπλοκότητα είναι (θέτοντας $r = \log N - l$)

$$O(\sum_{l=0}^{logN} 2^l 2^l t_1 \log t_1 + t_0 \log t_0) = O(\sum_{l=0}^{logN} (2^l 2^l (2^{logN-l} - 1) \log (2^{logN-l} - 1)) + N \log N) =$$

$$= O(\sum_{l=0}^{logN} N 2^l log \frac{N}{2^l} + N \log N) = O(\sum_{r=0}^{logN} \frac{N^2 r}{2^r} + N \log N) = O(N^2 + N \log N) = O(N^2).$$

Όσον αφορά τη χωρική πολυπλοκότητα, για έναν κόμβο απαιτούνται $O(2^l min\{B, 2^{logN-l}\}) = O(min\{B2^l, N\})$ θέσεις στον πίνακα $M[]$. Συνολικά, ο χώρος που καταλαμβάνει ο αλγόριθμος είναι

$$O(\sum_{l=0}^{logN} 2^l min\{B2^l, N\}) \leq O(N \sum_{l=0}^{logN} 2^l) = O(N^2)$$

3.2.4 Ο Αλγόριθμος των Garofalakis και Kumar για τις μετρικές weighted-L_p και οι Κατανεμημένες μετρικές σφάλματος

Ο αλγόριθμος των Garofalakis και Kumar για τη μετρική L_∞ που περιγράψαμε στην προηγούμενη ενότητα, μπορεί να τροποποιηθεί ώστε να εφαρμόζεται για όλες τις μετρικές της κατηγορίας weighted-L_p. Η μετρική L_p και η weighted εκδοχή της ορίζονται ως

$$errL_p(e) = \sum_{i=0}^{N-1} e[i]^p = \sum_{i=0}^{N-1} |a[i] - \hat{a}[i]|^p.$$

$$errWL_p(e) = \sum_{i=0}^{N-1} w_i \cdot (e[i]^p) = \sum_{i=0}^{N-1} w_i \cdot (|a[i] - \hat{a}[i]|^p).$$

Προφανώς η L_p αποτελεί υποπερίπτωση της weighted-L_p. Παρά την εμφανή ομοιότητα με την L_2, ωστόσο, το θεώρημα Parseval δεν ισχύει για τη μετρική L_p για $p \neq 2$ ή την weighted-L_p. Έτσι, ο άπληστος αλγόριθμος που εφαρμόσαμε για την L_2 δεν μπορεί να εφαρμοστεί σε αυτήν την κλάση μετρικών.

Οι μετρικές L_p και weighted-L_p ανήκουν σε μια ευρύτερη κλάση μετρικών, τις *κατανεμημένες μετρικές σφάλματος*. Έστω ένα διάνυσμα-περίληψη δεδομένων A. Με $f(R)$ ορίζουμε το σφάλμα που περιέχει η περίληψη για όλο το εύρος τιμών R του A. Λέμε ότι η μετρική

σφάλματος $f()$ είναι κατανεμημένη αν και μόνο αν, για κάθε συλλογή ξένων διαστημάτων $R_1, R_2, ..., R_k$, υπάρχει κάποια συνδυαστική συνάρτηση $g()$ τέτοια ώστε το σφάλμα όλης της περιοχής $\bigcup_{i=1}^{k} R_i$ μπορεί να εκφραστεί ως

$$f(\bigcup_{i=1}^{k} R_i) = g(f(R_1), f(R_2), ..., f(R_k)).$$

Στο [1] οι Garofalakis και Kumar επεκτείνουν το δυναμικό αλγόριθμο που πρότειναν για τη μετρική L_∞ (βλ. 3.2.3) στις μετρικές κατανεμημένου σφάλματος. Το σκεπτικό είναι το ίδιο. Βάση της αναδρομής είναι και πάλι τα φύλλα. Για τους εσωτερικούς κόμβους έχουμε και πάλι

$$M[j, b, S] = min\{M_{drop}[j, b, S], M_{keep}[j, b, S]\}$$

με

$$M_{keep}[j, b, S] = \min_{0 \le \dot{b} \le b-1} g(M[2j, \dot{b}, S \cup \{c_j\}], M[2j+1, b-\dot{b}-1, S \cup \{c_j\}])$$

$$M_{drop}[j, b, S] = \min_{0 \le \dot{b} \le b} g(M[2j, \dot{b}, S], M[2j+1, b-\dot{b}, S]).$$

Βλέπουμε ότι εδώ τη θέση της συνάρτησης $max()$ έχει πάρει η συνδυαστική συνάρτηση $g()$.

Ας δούμε σε αυτό το σημείο πώς ακριβώς εφαρμόζεται η ιδέα των Garofalakis και Kumar για τη μετρική weighted-L_1 με σχετικό σημειακό σφάλμα sum of weighted relative errors.

$$wL_{1_{rel}} = \sum_i w_i \cdot relErr_i = \sum_i \frac{w_i |\hat{d}_i - d_i|}{\max\{|d_i|, s\}}$$

Η βάση της αναδρομής βρίσκεται και πάλι στα δεδομένα - φύλλα του δένδρου σφάλματος, $c_j = d_{j-N}$ για $j \ge N$. Με γνωστούς τους συντελεστές που ανήκουν στο μονοπάτι από τη ρίζα προς το εκάστοτε φύλλο (S) μπορούμε να υπολογίσουμε το weighted relative error στο κάθε δεδομένο.

$$M[j, 0, S] = \frac{w_{j-N} \cdot |d_{j-N} - \sum_{c_k \in S} sign_{j-N,k} \cdot c_k|}{\max\{|d_{j-N}|, s\}}$$

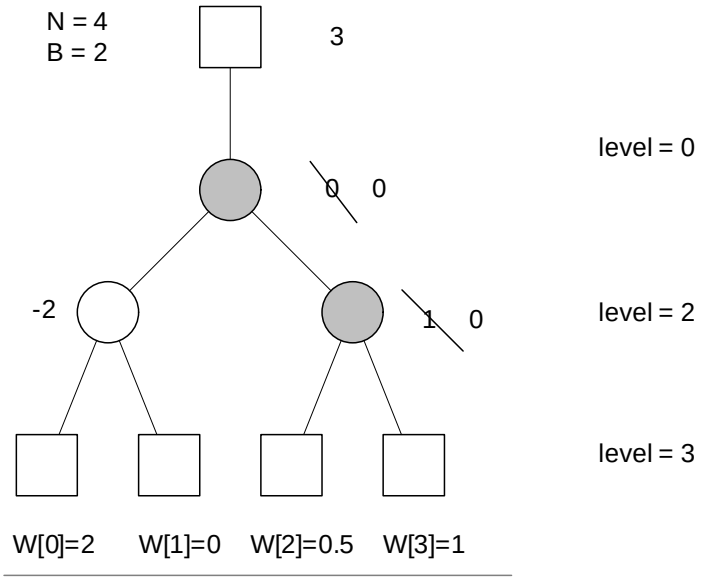
Για τους εσωτερικούς κόμβους, ως συνδυαστική συνάρτηση χρησιμοποιείται η πρόσθεση (αφού τα σχετικά σφάλματα αθροίζονται).

$$M_{drop}[j, b, S] = \min_{0 \le \dot{b} \le b} \{M[2j, \dot{b}, S] + M[2j+1, b-\dot{b}, S]\}$$

$$M_{keep}[j, b, S] = \min_{0 \le \dot{b} \le b-1} \{M[2j, \dot{b}, S] + M[2j+1, b-\dot{b}-1, S \cup \{c_j\}]\}$$

Με παρόμοιο σκεπτικό, ο αλγόριθμος που αντιμετωπίζει τη μετρική weighted-L_2, χρησιμοποιεί το απλό άθροισμα στη θέση της συνδυαστικής συνάρτησης και το μόνο που αλλάζει είναι ο τρόπος υπολογισμού του σφάλματος στα δεδομένα-φύλλα.

$$wL_2 = \sum_i w_i \cdot |\hat{d}_i - d_i|^2.$$



αρχικές τιμές	1	5	4	2
νέες τιμές	1	5	3	3
απόλυτο σφάλμα	0	0	1	1

Σχήμα 3.5: Παράδειγμα εφαρμογής του αλγορίθμου Garofalakis-Kumar για τη μετρική weighted-L_2

$$M[j,0,S] = w_{j-N} \cdot |d_{j-N} - \sum_{c_k \in S} sign_{j-N,k} \cdot c_k|^2$$

$$M_{drop}[j,b,S] = \min_{0 \le \dot{b} \le b} \{M[2j,\dot{b},S] + M[2j+1, b-\dot{b}, S]\}$$

$$M_{keep}[j,b,S] = \min_{0 \le \dot{b} \le b-1} \{M[2j,\dot{b},S] + M[2j+1, b-\dot{b}-1, S \cup \{c_j\}]\}$$

Χρησιμοποιώντας την αρχική μορφή του αλγορίθμου των Garofalakis και Kumar το πρόβλημα λύνεται με χρονική πολυπλοκότητα $O(N^2B^2)$ και χωρική πολυπλοκότητα $O(N^2B)$. Χρησιμοποιώντας τις παρατηρήσεις του Guha οι απαιτήσεις σε χρόνο βελτιώνονται σε $O(N^2 logB)$.

Στο σχήμα 3.5 φαίνεται ένα παράδειγμα εφαρμογής του αλγορίθμου Garofalakis - Kumar για τη μετρική weighted-L_2. Το διάνυσμα $A = (1, 5, 4, 2)$ μετασχηματίζεται κατά Haar και σχηματίζεται μια περίληψη $B = 2$ όρων, των $C[0]$ και $C[2]$ με βέλτιστο σφάλμα 1.5.

$$wL_2err = 0.5 \times 1^2 + 1 \times 1^2 = 1.5$$

3.2.5 Ο Αλγόριθμος του Muthukrishnan για τη μετρική weighted-L_2

Όπως είδαμε στην ενότητα 3.2.4, οι Garofalakis και Kumar προτείνουν ένα γενικό αλγόριθμο για κατανεμημένες μετρικές σφάλματος που λύνει το πρόβλημα της μετρικής weighted-

L_2.

$$errwL_2(e) = \sum_{i=0}^{N-1} (w_i \cdot e[i]^2) = \sum_{i=0}^{N-1} (w_i \cdot |a[i] - \hat{a}[i]|^2).$$

Ο αλγόριθμος αυτός, χρησιμοποιώντας και τις παρατηρήσεις του Guha (Ενότητα 3.2.3), προσαρμόζεται εύκολα ώστε να λύνει το πρόβλημα της μετρικής weighted-L_2 σε χρόνο $O(N^2 \log B)$ (Ενότητα 3.2.4).

Ο Muthukrishnan στο [7] αντιμετωπίζει μια ειδική περίπτωση του προβλήματος της μετρικής weighted-L_2: υποθέτει ότι η κατανομή των βαρών w είναι τμηματικά σταθερή πάνω στα δεδομένα. Δηλαδή, για $p_0 = 0 < p_1 \cdots < p_k = N$ έχουμε $w_j = w_{j+1}$, όπου $j \in [p_i, p_{i+1} - 1]$.

Αρχικά θεωρούμε την περίπτωση που τα διαστήματα $I_i = [p_i, p_{i+1}] = J_u$ αντιστοιχούν στα φύλλα κάποιου υποδένδρου σφάλματος, με ρίζα κάποιο κόμβο u. Συμβολίζουμε το σύνολο αυτών των κόμβων με L με $|L| = k$. Κανένας κόμβος αυτού του συνόλου δεν είναι απόγονος κάποιου άλλου από το ίδιο σύνολο. Αν με t συμβολίσουμε το τμήμα του δένδρου σφάλματος που περιλαμβάνει τους κόμβους του L και όλους τους προγόνους τους, τότε το μέγεθός του είναι $2k - 1$ (k φύλλα και $k - 1$ εσωτερικοί κόμβοι). Η κεντρική ιδέα του αλγορίθμου είναι να εφαρμόσουμε δυναμικό προγραμματισμό στους εσωτερικούς κόμβους και τον άπληστο αλγόριθμο επιλογής συντελεστών στα φύλλα του t.

Στους εσωτερικούς κόμβους του t, λοιπόν, εφαρμόζουμε το δυναμικό αλγόριθμο που περιγράψαμε και στην ενότητα 3.2.3. Τα υποπροβλήματα που έχουμε να λύσουμε είναι $O(kB2^{min\{k,logN\}})$, εφόσον οι κόμβοι του t είναι $2k - 1$, το μέγιστο πλήθος συντελεστών που κρατάμε είναι B και το μέγεθος του S φράσσεται είτε από το ύψος του δένδρου είτε από το πλήθος των εσωτερικών κόμβων του t. Ο χρόνος που απαιτείται για τη λύση του κάθε υποπροβλήματος είναι $O(B)$ - τόσο χρειάζεται για να υπολογίσουμε το σφάλμα για κάθε κατανομή του πλήθους των διαθέσιμων συντελεστών στα δύο υποδένδρα του εξεταζόμενου δένδρου.

Όταν φτάσουμε στα φύλλα του t τότε εφαρμόζουμε τοπικά άπληστους αλγορίθμους για να επιλέξουμε τους συντελεστές. Αποδεικνύεται, δηλαδή, ότι το $M[u, b, S]$, όπου $u \in L$, ελαχιστοποιείται επιλέγοντας τις b μεγαλύτερες λεπτομέρειες του υποδένδρου σφάλματος με ρίζα το u, ανεξάρτητα από το S.

Για κάθε κόμβο του $u \in L$, και για κάθε δυνατή τιμή του b, επιλέγουμε τοπικά b συντελεστές, σε χρόνο $O(|J_u| + b \log |J_u|)$. Έτσι, ο συνολικός χρόνος που αφιερώνεται σε τοπικές αναζητήσεις είναι $O(BN)$.

Από τα παραπάνω προκύπτει ότι ο αλγόριθμος δουλεύει σε χρόνο $O(BN + kB^2 2^{min\{k,logN\}})$, που είναι περίπου γραμμικός εφόσον $k, B << N$.

Στην περίπτωση που τα διαστήματα $I_i = [p_i, p_{i+1}]$ δεν αντιστοιχούν σε φύλλα υποδένδρων σφάλματος, μπορούμε να επεκτείνουμε τον προηγούμενο αλγόριθμο, διαιρώντας κάθε διάστημα σταθερού βάρους σε $O(logN)$ άλλα διαστήματα σταθερού βάρους, ώστε τα νέα διαστήματα να αντιστοιχούν σε φύλλα υποδένδρων σφάλματος. Είναι εύκολο να δούμε ότι η χρονική πολυπλοκότητα του αλγορίθμου σε αυτήν την περίπτωση είναι $O(NkB^2 logN)$.

3.2.6 Ο Αλγόριθμος του Muthukrishnan για τη μετρική weighted-L_∞

Ο Muthukrishnan στο [7] περιγράφει έναν αλγόριθμο που λύνει το δυαδικό του προβλήματος που εξετάζουμε, σε χρόνο $o(N^2)$. Αποδεικνύει, δηλαδή, ότι: *δοθέντος ενός ορίου* δ *με* $\max_i w_i|R[i] - A[i]| \leq \delta$, *μπορούμε να βρούμε το ελάχιστο* B *των συντελεστών wavelet που είναι απαραίτητοι για να αναπαραστήσουν το σήμα, σε χρόνο* $o(N^2)$.

Σε μια πρώτη προσέγγιση θα μπορούσαμε να χρησιμοποιήσουμε ένα σχήμα δυναμικού προγραμματισμού για να λύσουμε το πρόβλημα. Έστω, λοιπόν, ένα διάστημα δεδομένων I και ένα σύνολο S συντελεστών wavelet που υπερκαλύπτουν το διάστημα I (λέγοντας ότι ένας συντελεστής c *καλύπτει* ένα διάστημα I, εννοούμε ότι τα δεδομένα που ανήκουν στο διάστημα αυτό είναι φύλλα στο υποδένδρο σφάλματος με ρίζα το c - και πιθανόν όχι μόνο αυτά, οπότε λέμε ότι *υπερκαλύπτει* το διάστημα). Ορίζουμε ως t_I το υποδένδρο σφάλματος που έχει φύλλα (μόνο) τα δεδομένα που ανήκουν στο διάστημα I, c_I τη ρίζα του, Λ_I τους συνελεστές περίληψης του t_I και ως $B(I, \delta, S)$ τον ελάχιστο αριθμό συντελεστών του δένδρου t_I που πρέπει να επιλεχθούν ώστε συνυπολογίζοντας τους συντελεστές του S να ικανοποιείται η συνθήκη $\max_i w_i|R[i] - A[i]| \leq \delta$, $i \in I$.

Η λύση του προβλήματος δίνεται από την τιμή του $B([1..N], \delta,)$ και υπολογίζεται δυναμικά. Η βάση της αναδρομής βρίσκεται στα φύλλα, με $B([i, i], \delta, S) = 0$, αν η συνθήκη ικανοποιείται ή $B([i, i], \delta, S) = +\infty$, διαφορετικά. Για τους εσωτερικούς κόμβους του δένδρου σφάλματος, αν ο συντελεστής c_I προστεθεί στην περίληψη, τότε ισχύει

$$B_{keep}(I, \delta, S) = B(I_L, \delta, S \cup c_I) + B(I_R, \delta, S \cup c_I) + 1,$$

διαφορετικά ισχύει

$$B_{drop}(I, \delta, S) = B(I_L, \delta, S) + B(I_R, \delta, S).$$

Ο αλγόριθμος κάνει την επιλογή που οδηγεί στη μικρότερη περίληψη.

$$B(I, \delta, S) = \min\{B_{keep}(I, \delta, S), B_{drop}(I, \delta, S)\}$$

Με την προσέγγιση αυτή έχουμε να λύσουμε $O(N^2)$ υποπροβλήματα ($1+2+4+\cdots+\frac{N}{2} = N-1$ διαστήματα I και $2^{logN} = N$ μέγιστο μέγεθος του S, δ σταθερό) σε χρόνο $O(1)$ το καθένα. Άρα η χρονική πολυπλοκότητά της είναι $O(N^2)$.

Μπορούμε να έχουμε έναν αλγόριθμο με μικρότερη πολυπλοκότητα, παρατηρώντας ότι δε χρειάζεται να υπολογίσουμε το $B(I, \delta, S)$ για κάθε πιθανό S, αλλά από κάποιο σημείο και μετά αρκεί να κρατάμε το μέγιστο και το ελάχιστο απόλυτο σφάλμα για κάθε υποσύνολο συντελεστών της περίληψης. Αυτό συμβαίνει επειδή

$$\min_{S, i \in 2^{\Lambda_I}} \max_{j \in I} |\delta_i^j - u_S| = \min_{S, i \in 2^{\Lambda_I}} \max\{|(\max_{j \in I} \delta_i^j) - u_S|, |(\min_{j \in I} \delta_i^j) - u_S|\},$$

όπου δ_i^j το σφάλμα του j-οστού δεδομένου λόγω της i-οστής περίληψης Λ_I (χωρίς να λαμβάνουμε υπόψιν το S) και u_S η συνεισφορά του S στο δεδομένο j. Εφαρμόζουμε, λοιπόν, το δυναμικό αλγόριθμο μέχρι να φτάσουμε διαστήματα κάποιου μήκους 2^k και στη συνέχεια

εφαρμόζουμε τοπικά τις παρατηρήσεις αυτής της παραγράφου. Ο συνδυαστικός αλγόριθμος που προκύπτει έχει χρονική πολυπλοκότητα $O(\frac{N}{2^k}\frac{N}{2^k} + \frac{N}{2^k}2^{2^k}2^k k)$ (ο δυναμικός αλγόριθμος θα εφαρμοστεί σε $O(\frac{N}{2^k})$ διαστήματα, με μέγιστο μέγεθος περίληψης $log(\frac{N}{2^k})$ ενώ ο τοπικός θα εφαρμοστεί σε $\frac{N}{2^k}$ υποδένδρα, εκεί που σταματάει ο δυναμικός αλγόριθμος, για 2^{2^k} δυνατές περιλήψεις που στην καθεμία θα εξετάζονται 2^k δεδομένα και χρειάζονται k πράξεις για να προσδιοριστεί το σφάλμα στο καθένα). Επιλέγοντας κατάλληλα το k, ο αλγόριθμος αποκτά πολυπλοκότητα $O(N^2/logN) = o(N^2)$.

3.3 Η δεύτερη προσέγγιση: ο τροποποιημένος μετασχηματισμός Haar των Matias και Urieli και ο άπληστος αλγόριθμος για τη μετρική weighted-L_2

3.3.1 Εισαγωγή

Οι Matias και Urieli στο [8] λύνουν το πρόβλημα για τη μετρική weighted-L_2, ακολουθώντας μια διαφορετική προσέγγιση από εκείνη στο [1]: ενσωματώνουν τα βάρη w_i της μετρικής στην παραγωγή της περίληψης, έτσι ώστε αν για δύο τιμές a, b έχουμε τα βάρη w_a και w_b, η μέση τιμή που παίρνουμε είναι η $\frac{w_a a + w_b b}{2}$. Με τον τρόπο αυτό, το weighted-L_2 σφάλμα είναι ισοδύναμο με το SSE σφάλμα της τροποποιημένης περίληψης. Με δεδομένη την τροποποιημένη περίληψη, η επιλογή των συντελεστών γίνεται με τον άπληστο αλγόριθμο που εφαρμόσαμε και για την περίπτωση του SSE.

Στην περιγραφή της προσέγγισης που θα ακολουθήσει, θα θεωρούμε το διάνυσμα δεδομένων a μεγέθους $N = 2^j$ ως μια τμηματικά σταθερή συνάρτηση f στο διάστημα $[0, 1)$ και θα συμβολίζουμε με V_j το χώρο των συναρτήσεων με $N = 2^j$ σταθερά τμήματα. Έτσι, ένα μονοδιάστατο διάνυσμα θα αντιστοιχεί σε μια συνάρτηση του χώρου V_0 που είναι σταθερή στο διάστημα $[0, 1)$, ένα δισδιάστατο διάνυσμα θα αντιστοιχεί σε μια συνάρτηση του χώρου V_1 που θα έχει δύο σταθερά τμήματα στα διαστήματα $[0, \frac{1}{2})$ και $[\frac{1}{2}, 1)$, κοκ.

Αρχικά ορίζουμε ένα νέο εσωτερικό γινόμενο που ενσωματώνει τα βάρη της μετρικής σφάλματος.

$$< f, g >= N \cdot (\sum_{i=0}^{N-1} w_i \int_{\frac{1}{N}}^{\frac{i+1}{N}} f(x)g(x)dx)$$

με $0 < w_i \leq 1$, $\sum_{i=0}^{N-1} w_i = 1$. Θεωρούμε,δηλ, τα βάρη είναι κανονικοποιημένα. Όταν όλα τα βάρη είναι ίσα με $\frac{1}{N}$ έχουμε τη μετρική L_2. Βασιζόμενοι σε αυτό το εσωτερικό γινόμενο ορίζουμε το μέτρο

$$\|f\|_{IPB} = \sqrt{< f, f >} = \sqrt{\sum_{i=0}^{N-1} w_i f_i^2},$$

όπου f_i η τιμή της f στο διάστημα i. Έτσι, αν f είναι μια συνάρτηση και f' μια προσέγγισή

της, με $f, f' \in V_j$, τότε για το σφάλμα $e = f - f' \in V_j$ έχουμε

$$\|e\|_{IPB}^2 =< e, e >= \sum_{i=0}^{N-1} w_i e_i^2$$

Έχοντας ορίσει ένα εσωτερικό γινόμενο και τη νόρμα που προκύπτει από αυτό θέλουμε τώρα να έχουμε μια βάση του χώρου V_j που να είναι ορθοκανονική ως προς το εσωτερικό γινόμενο. Αν έχουμε μια τέτοια βάση, τότε από το θεώρημα του Parseval, θα μπορούμε να κατασκευάζουμε τη βέλτιστη περίληψη χρησιμοποιώντας τον άπληστο αλγόριθμο που χρησιμοποιούμε και για τη μετρική L_2. Έστω η υπ' αριθμόν k συνάρτηση βάσης που το θετικό τμήμα της ισούται με x_k και το αρνητικό με y_k. Τα x_k και y_k πρέπει να είναι τέτοια ώστε η βάση μας να είναι ορθοκανονική ως προς το εσωτερικό γινόμενο. Αποδεικνύεται ότι αν l_k (r_k) είναι το άθροισμα των βαρών που αντιστοιχούν στο θετικό (αρνητικό) τμήμα της συνάρτησης βάσης, τότε για να είναι η βάση μας ορθοκανονική πρέπει να επιλέξουμε

$$x_k = \sqrt{\frac{r_k}{l_k r_k + l_k^2}}, y_k = \sqrt{\frac{l_k}{l_k r_k + r_k^2}}.$$

Στην ειδική περίπτωση της συνάρτησης που αντιστοιχεί το σταθμισμένο μέσο όλων των δεδομένων, επιλέγουμε

$$x_0 = y_0 = \sqrt{\frac{1}{l_0 + r_0}} = 1.$$

3.3.2 Ο Αλγόριθμος των Matias και Urieli για τη μετρική weighted-L_2

Τώρα είμαστε έτοιμοι να περιγράψουμε τον αλγόριθμο της σύνοψης. Στο πρώτο μέρος του αλγορίθμου, έχοντας δεδομένα τα βάρη w_i κατασκευάζουμε τη βάση του μετασχηματισμού (σχήματα 3.6, 3.7). Στο δεύτερο μέρος υπολογίζουμε το μετασχηματισμένο σήμα χρησιμοποιώντας τη βάση μετασχηματισμού.

Για να κατασκευάσουμε τη βάση μετασχηματισμού πρέπει να έχουμε τα l_k και r_k. Ο υπολογισμός τους μπορεί να γίνει σε γραμμικό χρόνο. Παρατηρούμε αρχικά ότι για τις συναρτήσεις με δείκτες $\frac{N}{2}, \frac{N}{2}+1, ..., N-1$ τα l_k και r_k είναι δεδομένα και ταυτίζονται με τα βάρη w_i. Οι συναρτήσεις του επόμενου επιπέδου έχουν δείκτες $\frac{N}{4}, ..., \frac{N}{2}-1$, κοκ, ενώ η σταθερή συνάρτηση έχει δείκτη 0. Ισχύει

$$l_k = l_{2k} + r_{2k} \quad r_k = l_{2k+1} + r_{2k+1}.$$

Έτσι, έχοντας τα l_k και r_k ενός επιπέδου, μπορούμε να υπολογίσουμε τα αντίστοιχα του επόμενου, με το συνολικό χρόνο να είναι $\frac{N}{2} + \frac{N}{4} + ... + 1 = N - 1 = O(N)$. Στη συνέχεια υπολογίζουμε τις συναρτήσεις της βάσης με τους τύπους

$$x_k = \sqrt{\frac{r_k}{l_k r_k + l_k^2}}, y_k = \sqrt{\frac{l_k}{l_k r_k + r_k^2}}.$$

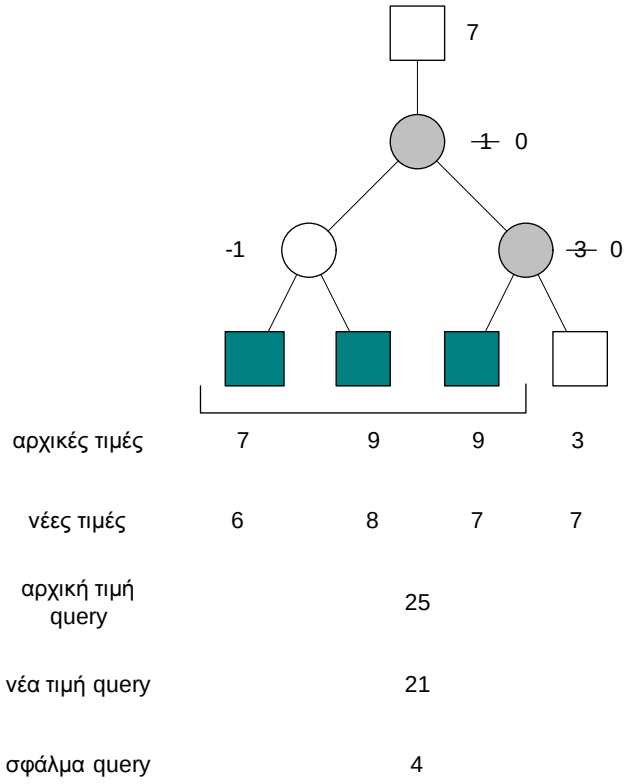
```
Procedure computeWeights ( w[N], N )
BEGIN
        level := logN – 1;
        FOR ( i=N/2; i < N; i++ ) do {
                l[i] = w [ 2i – N ];
                r[i] = w [ 2i + 1 – N];
        }
        WHILE (level > 0)  {
                level = level -1;
                FOR ( i=2^level to 2^level+1 - 1) {
                        l[i] = l[2i] + r[2i];
                        r[i] = l[2i+1] + r[2i+1];
                }
        }
END
```

Σχήμα 3.6: Αλγόριθμος για τον υπολογισμό των l_k και r_k.

procedure computeBasis(l[N],r[N])
BEGIN
x[0]:=1, y[0]:=1
FOR (i:=1; i¡N; i++)
x[i]:= $\sqrt{\frac{r[i]}{l[i]r[i]+l[i]^2}}$
y[i]:= $\sqrt{\frac{l[i]}{l[i]r[i]+r[i]^2}}$

END

Σχήμα 3.7: Αλγόριθμος για τον υπολογισμό των συναρτήσεων βάσης.



Σχήμα 3.8: Υπολογισμός απόλυτου σφάλματος του range query $q(0:2)$.

Ο μετασχηματισμός ενός σήματος (u_0, u_1) μεγέθους $N = 2$ αποτελείται από το σταθμισμένο μέσο a_0 και τη σταθμισμένη λεπτομέρεια a_1. Είναι εύκολο να δούμε ότι

$$a_0 = \frac{yu_0 + xu_1}{x+y} \quad a_1 = \frac{u_0 - u_1}{x+y}$$

Ανακατασκευάσουμε το σήμα, με τις εξισώσεις

$$u_0 = a_0 + xa_1 \quad u_1 = a_0 - ya_1$$

Αν θυμηθούμε τώρα τη διαδικασία που περιγράψαμε στην ενότητα 2.2 βλέπουμε ότι ουσιαστικά ο μετασχηματισμός παράγεται υπολογίζοντας έναν μέσο και μια λεπτομέρεια για ζεύγη τιμών, ξεκινώντας με δεδομένα του αρχικού σήματος και συνεχίζοντας με μέσους του προηγούμενου επιπέδου. Έτσι, εφαρμόζοντας την ίδια διαδικασία για τη νέα βάση και χρησιμοποιώντας τις τελευταίες εξισώσεις, έχουμε έναν αλγόριθμο που υπολογίζει το μετασχηματισμένο σήμα σε $O(N)$ χρόνο. Από το θεώρημα του Parseval έχουμε ότι η περίληψη που ελαχιστοποιεί το σφάλμα $weighted - L_2$ κατασκευάζεται επιλέγοντας τους B μεγαλύτερους συντελεστές, εφόσον στο μετασχηματισμένο σήμα έχουν ενσωματωθεί τα βάρη w_i. Αυτό, όπως περιγράψαμε στην ενότητα 3.2.2 για τον άπληστο αλγόριθμο και τη μετρική L_2, απαιτεί $O(N + BlogN)$ χρόνο και $O(N)$ χώρο.

Κεφάλαιο 4

Weighted-L_p μετρικές για Range Sum Errors

4.1 Εισαγωγή

Η δεύτερη μεγάλη οικογένεια μετρικών σφάλματος είναι εκείνες που ο υπολογισμός του σφάλματος βασίζεται στα αθροιστικά σφάλματα εύρους - range sum errors που εισάγονται στο ανακατασκευασμένο από την περίληψη σήμα. Ως αθροιστικό σφάλμα εύρους, εννοούμε το σφάλμα στο άθροισμα των τιμών των στοιχείων που ανήκουν σε ένα συγκεκριμένο διάστημα (εύρος) του σήματος, πριν και μετά την περίληψη. Αυτή η οικογένεια μετρικών χρησιμοποιείται όταν εφαρμόζουμε range queries πάνω στα δεδομένα μας. Αυτό συμβαίνει διότι μας ενδιαφέρει να έχουμε μια εκτίμηση για το μέσο σφάλμα που θα έχουμε για κάθε range sum query. Έστω ότι το $A[0 \dots N-1]$ είναι ένα σήμα με $N = 2^j$ τιμές και το $q_{l:r}$ είναι το range sum query που επιστρέφει το άθροισμα των τιμών του σήματος από το l-οστό ως το r-οστό στοιχείο. Η τιμή του είναι γενικά διαφορετική πριν και μετά την περίληψη.

$$d_{(l:r)} = \sum_{i=l}^{r} A[i]$$

$$\hat{d}_{(l:r)} = \sum_{i=l}^{r} \hat{A}[i]$$

Αν $d_{l:r}$ είναι η τιμή του $q_{(l:r)}$ με βάση τα αρχικά δεδομένα και $\hat{d}_{l:r}$ είναι η τιμή του με βάση τα ανακατασκευασμένα από την περίληψη δεδομένα, τότε το απόλυτο σφάλμα στο $q_{(l:r)}$ λόγω της περίληψης είναι

$$err_{abs}(l:r) = |d_{l:r} - \hat{d}_{l:r}|$$

Στο σχήμα 4.1 φαίνεται ένα απλό παράδειγμα υπολογισμού του απόλυτου σφάλματος σε ένα query. Όπως και στην περίπτωση των point errors, έτσι κι εδώ μπορεί με ανάλογο τρόπο να οριστεί σχετικό αθροιστικό σφάλμα εύρους, στην περιγραφή που θα ακολουθήσει, όμως, περιοριζόμαστε στο απόλυτο.

Το σύνολο των range queries που αφορούν το σήμα A ορίζει ένα διάνυσμα αθροιστικών σφαλμάτων εύρους, $E = (err_{0:0}, \ldots, err_{0:N-1}, err_{1:1}, \ldots, err_{1:N-1}, \ldots, err_{N-1:N-1})$, μεγέθους $N(N+1)/2$. Ο ρόλος των διαφόρων μετρικών σφάλματος για range errors είναι να συνυπολογίζουν με κάποιο τρόπο τα $N(N+1)/2$ αυτά αθροιστικά σφάλματα και να δίνουν μια εκτίμηση για το αναμενόμενο σφάλμα ενός τυχαίου range sum query.

Παρόμοια με την περίπτωση των point errors, οι Weighted-L_p μετρικές μπορούν να χρησιμοποιηθούν για την εκτίμηση του σφάλματος.

$$weightedL_p = \sum_{l,r|l\leq r} w_i \cdot (err(l:r))^p$$

Οι μετρικές αυτές αθροίζουν τα αθροιστικά σφάλματα όλων των στοιχείων του διανύσματος E υψωμένα σε μια δύναμη p και ίσως πολλαπλασιασμένα με κάποιο ΄βάρος ΄ που τους αντιστοιχεί.

Ας εξετάσουμε ένα παράδειγμα υπολογισμού της μετρικής L_2 για range errors. Όπως φαίνεται στο σχήμα 4.1, έχουμε κατασκευάσει το δένδρο σφάλματος για το διάνυσμα δεδομένων $A = [7, 9, 9, 3]$, έχουμε λάβει το διάνυσμα συντελεστών $C = [7, 1, -1, 3]$, έχουμε κρατήσει τους συντελεστές $C[0]$ και $C[2]$ στην περίληψη και έχουμε υπολογίσει τις νέες τιμές δεδομένων $\hat{A} = [6, 8, 7, 7]$. Πιο κάτω φαίνονται οι τιμές του απόλυτου αθροιστικού σφάλματος εύρους (range sum error) για κάθε range query. Αθροίζοντας τα τετράγωνα των range sum errors, λαμβάνουμε την τιμή σφάλματος που δίνει η μετρική L_2 (ή αλλιώς το SSE των σφαλμάτων εύρους).

$$SSE = err_{0:0}^2 + err_{0:1}^2 + err_{0:2}^2 + err_{0:3}^2 + err_{1:1}^2 + err_{1:2}^2 + err_{1:3}^2 + err_{2:2}^2 + err_{2:3}^2 + err_{3:3}^2 =$$

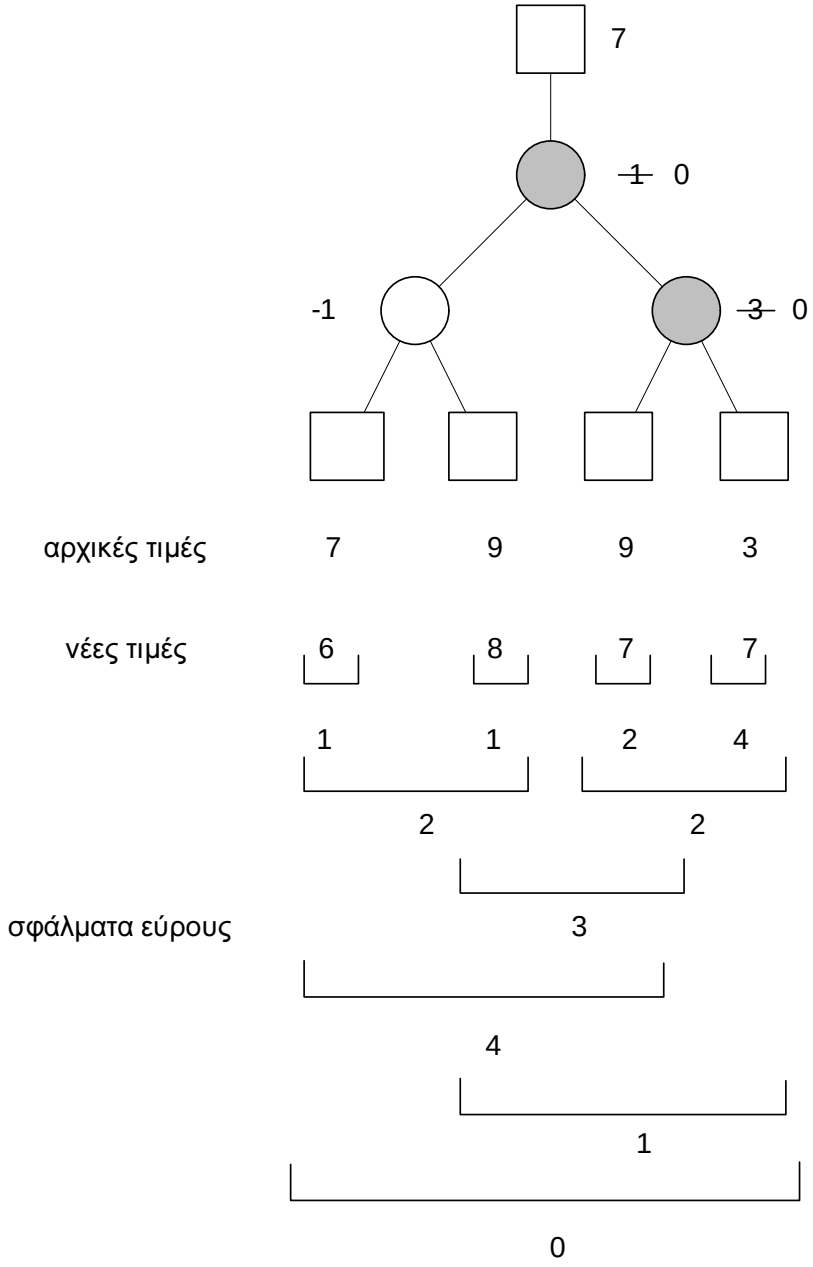
$$= 1^2 + 2^2 + 4^2 + 0^2 + 1^2 + 3^2 + 1^2 + 2^2 + 2^2 + 4^2 = 1 + 4 + 16 + 0 + 1 + 9 + 1 + 4 + 4 + 16 = 56$$

Οι μετρικές σφάλματος που θα εξετάσουμε πιο συγκεκριμένα στη συνέχεια είναι οι μετρικές L_2 και weighted-L_2, οι οποίες εμπίπτουν στην κατηγορία των Weighted-L_p μετρικών, με τη δεύτερη να περιορίζεται στην ειδική περίπτωση των δυαδικά ιεραρχημένων range sum errors .

$$L_2 = \sum_{l,r|l\leq r} (err(l:r))^2$$

$$dyadic - weighted - L_2 = \sum_{l,r|[l,r]\ dyadic} w_i \cdot (err(l:r))^2$$

Για κάθε περίπτωση που θα εξετάσουμε θα δούμε έναν αλγόριθμο που, δεδομένου κάποιου ορίου στο χώρο που μπορούμε να διαθέσουμε για τους συντελεστές που μένουν στην περίληψη, επιλέγει τους συντελεστές ώστε το σφάλμα αυτό να ελαχιστοποιείται. Οι αλγόριθμοι που εξετάζουμε χωρίζονται σε δύο κατηγορίες με βάση τον τρόπο που παριστάνονται τα αρχικά δεδομένα. Ο αλγόριθμος που ελαχιστοποιεί το σφάλμα της μετρικής L_2 έχει ως είσοδο μετασχηματισμένα δεδομένα της μορφής prefix sums, ενώ ο αλγόριθμος που ελαχιστοποιεί το σφάλμα της weighted-L_2 για δυαδικά ιεραρχημένα range sum errors τρέχει πάνω από μετασχηματισμένα raw data.



Σχήμα 4.1: Υπολογισμός απόλυτων σφαλμάτων εύρους για όλα τα range queries

Πριν από αυτό, όμως, θα δούμε αν ο μετασχηματισμός Haar Wavelet είναι ορθοκανονικός ως προς τη μετρική L_2 για range sum errors.

4.2 Ο μετασχηματισμός Haar και τα Range Sum Errors

Έχουν επικρατήσει δύο τρόποι παράστασης των δεδομένων όταν επεξεργαζόμαστε range errors.

Ο πρώτος είναι τα δεδομένα να βρίσκονται σε μορφή raw data. Πρόκειται για τον κλασσικό τρόπο παράστασης, με τον οποίο το σήμα υπό επεξεργασία παριστάνεται από κάποιο διάνυσμα, με κάθε στοιχείο του διανύσματος να αντιστοιχεί σε μια διακριτή τιμή του σήματος.

Ο δεύτερος τρόπος είναι τα δεδομένα να βρίσκονται σε μορφή prefix sums. Αν θεωρήσουμε ότι το σήμα σε μορφή raw data παριστάνεται με ένα διάνυσμα $A[0 \dots N-1]$, τότε το ίδιο σήμα σε μορφή prefix sums παριστάνεται με ένα διάνυσμα PS, κάθε στοιχείο i του οποίου ισούται με το άθροισμα όλων των στοιχείων του A από το $A[0]$, μέχρι το i-οστό στοιχείο του A.

$$PS[i] = \sum_{j=0}^{i} A[i]$$

Ή αναδρομικά, μπορούμε να πούμε πως το i-οστό στοιχείο του PS ισούται με το άθροισμα του $(i-1)$-οστού στοιχείου του PS και του i-οστού στοιχείου του A.

$$PS[0] = A[0]$$

$$PS[i] = A[i] + PS[i-1],\ 0 < i < N$$

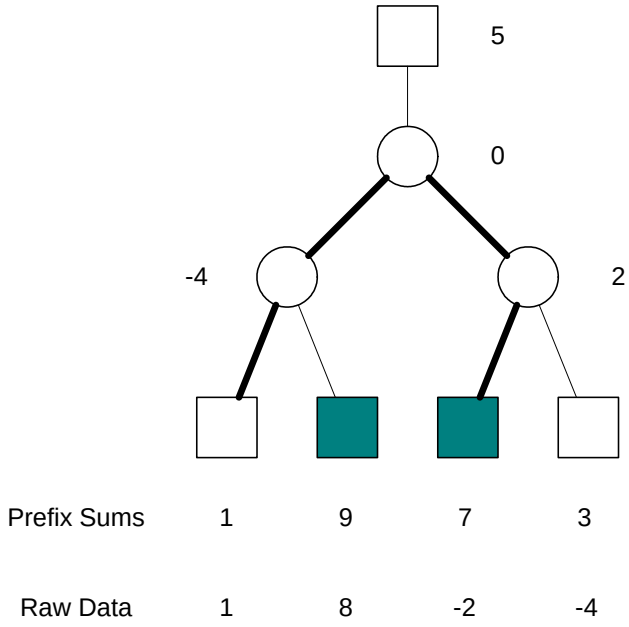
Ο μετασχηματισμός Haar Wavelet μπορεί να εφαρμοστεί τόσο πάνω σε raw data όσο και πάνω από prefix sums. Όπως έχουμε περιγράψει στην ενότητα 2.2 η τιμή ενός στοιχείου του διανύσματος δεδομένων, είτε αυτό παριστάνει raw data είτε πάριστάνει prefix sums, ανακτάται εύκολα προσθαφαιρώντας $(\log N + 1)$ μη-κανονικοποιημένους συντελεστές του μετασχηματισμένου σήματος, οι οποίοι στο δένδρο σφάλματος βρίσκονται πάνω στο μονοπάτι που ενώνει τη ρίζα του δένδρου με το φύλλο που αντιστοιχεί στο εκάστοτε στοιχείο.

$$\hat{A}[i] = \sum_{c_j \in path(A[i])} sign_{i,j} \cdot c_j$$

Το πρόσημο με το οποίο συμμετέχει κάποιος συντελεστής στο αλγεβρικό άθροισμα εξαρτάται από το αν το στοιχείο $A[i]$ βρίσκεται στο αριστερό (+) ή στο δεξιό (-) υποδένδρο με ρίζα το συντελεστή.

Στην περίπτωση των prefix-sums, η ανάκτηση της τιμής ενός range sum query στη γενική περίπτωση γίνεται εύκολα αφαιρώντας την τιμή δύο στοιχείων του διανύσματος prefix-sums PS.

$$q_{(0:r)} = PS[r] = \sum_{c_j \in path(PS[r])} sign_{i,j} \cdot c_j, \quad 0 \le r < N$$



Σχήμα 4.2: Υπολογισμός τιμής range query στην περίπτωση των range-sums

$$q_{(l:r)} = PS[r] - PS[l-1] = \sum_{c_j \in path(PS[r])} sign_{i,j} \cdot c_j - \sum_{c_j \in path(PS[l-1])} sign_{i,j} \cdot c_j, \ 0 < l \leq r < N$$

Για παράδειγμα, όπως φαίνεται και στο σχήμα 4.2, όταν τα δεδομένα μας βρίσκονται σε μορφή range-sums, ο υπολογισμός της τιμής του $q_{1:2}$ γίνεται αφαιρώντας τις τιμές $PS[2]$ και $PS[0]$.

$$q_{1:2} = PS[2] - PS[0] = 7 - 1 = 6 = A[1] + A[2] = 8 + (-2)$$

Όταν χρησιμοποιούνται raw data, ένας συντελεστής c_j συνεισφέρει στην τιμή ενός range sum query μόνο όταν ανήκει στο ένα από τα δύο μονοπάτια που ενώνουν τη ρίζα του δένδρου σφάλματος με τα δύο ακραία φύλλα - στοιχεία του query.

$$q_{(l:r)} = \sum_{c_j \in path(A[l]) \cup path(A[r])} x_j$$

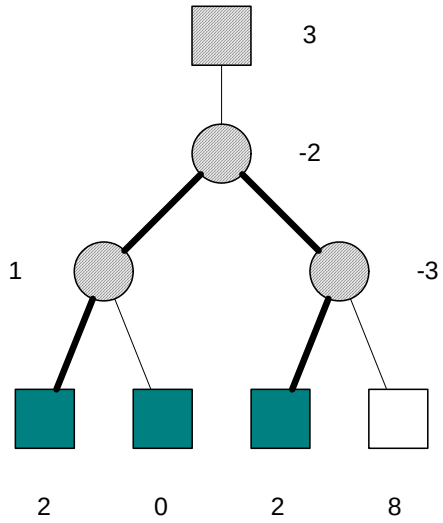
$$x_0 = (r - l + 1) \cdot c_0$$

$$x_j = (leftleaves(c_j, l:h) - rightleaves(c_j, l:h)) \cdot c_j \quad j \neq 0$$

Με $leftleaves(c_j, l:h)$ συμβολίζουμε το πλήθος των στοιχείων του A που ανήκουν στο αριστερό υποδένδρο του c_j και ταυτόχρονα ανήκουν στο εύρος του query - ομοίως για το $rightleaves(c_j, l:h)$. Έτσι, για να επανακτήσουμε την τιμή ενός query από το μετασχηματισμένο σήμα, χρειάζεται να αθροίσουμε $2logN + 1$ όρους.

Για παράδειγμα, αν έχουμε το δένδρο σφάλματος του σχήματος 4.3, όλοι οι συντελεστές wavelet συμμετέχουν στην τιμή του query $q_{0:2}$.

$$q_{0:2} = x_0 + x_1 + x_2 + x_3 = 3 \times 3 + 1 \times (-2) + 0 \times 1 + 1 \times (-3) = 4$$



Σχήμα 4.3: Υπολογισμός τιμής range query στην περίπτωση των raw data

Οι Matias και Urieli στο [6] εξετάζουν αν μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο που χρησιμοποιήσαμε και στην ενότητα 3.2.2 για να επιλέξουμε την περίληψη B όρων που ελαχιστοποιεί το SSE, όταν τα δεδομένα μας, σε μια από τις δύο μορφές που αναφέραμε, μετασχηματίζονται με τον M/Σ Haar Wavelet. Για να συμβαίνει αυτό, ο M/Σ Haar Wavelet πρέπει να είναι τουλάχιστον ορθογώνιος, αν όχι ορθοκανονικός, ως προς τη νόρμα που εκφράζει τη μετρική SSE.

Στην ενότητα3.2.2 περιγράψαμε κάποιες προϋποθέσεις που πρέπει να τηρούνται ώστε ο άπληστος αλγόριθμος να είναι βέλτιστος ως προς κάποια μετρική σφάλματος: ο μετασχηματισμός που χρησιμοποιούμε αρκεί να είναι ορθοκανονικός ως προς μια νόρμα και το τετράγωνο της νόρμας να ταυτίζεται με τη μετρική σφάλματος (κατά την προσέγγιση το πολύ μιας πολλαπλασιαστικής σταθεράς). Η συνθήκη αυτή μπορεί να γίνει λιγότερο αυστηρή, αφού στην πραγματικότητα αρκεί και η ορθογωνιότητα του μετασχηματισμού ως προς τη νόρμα.

Έστω και πάλι ο διανυσματικός χώρος $\Re^N$, εφοδιασμένος με το εσωτερικό γινόμενο $p(u, v)$ και $n(v) = ||v|| = \sqrt{p^2(v, v)}$ η νόρμα που ορίζεται από αυτό για κάθε διάνυσμα u του χώρου. Μια βάση S του διανυσματικού χώρου $\Re^N$ είναι ορθογώνια ως προς το εσωτερικό γινόμενο $p(u, v)$, όταν τα διανύσματα που την αποτελούν είναι ανά δύο κάθετα, δηλαδή ισχύει $p(s_i, s_j) = 0$. Έστω, λοιπόν, ότι η βάση S του διανυσματικού χώρου είναι ορθογώνια. Τότε, κάθε διάνυσμα v του $\Re^N$ μπορεί να γραφτεί ως γραμμικός συνδυασμός των στοιχείων της βάσης S, ως $v = \sum a_i \cdot s_i$ και το διάνυσμα σφάλματος $e \in \Re^N$ γράφεται ως $e = A - \hat{A} = \sum_{0 \le i < N} c_i \cdot s_i - \sum_{0 \le i < N} \hat{c}_i \cdot s_i = \sum_{0 \le i < N} c_i \cdot s_i - \sum_{i \in \Lambda} c_i \cdot s_i = \sum_{i \notin \Lambda} c_i \cdot s_i$. Μπορούμε τότε να εκφράσουμε το ίδιο διάνυσμα ως γραμμικό συνδυασμό στοιχείων μιας ορθοκανονικής βάσης, ως $u = \sum \hat{a}_i \cdot \hat{s}_i$, με $\hat{a}_i = a_i \cdot ||s_i||$ και $\hat{s}_i = \frac{s_i}{||s_i||}$. Για την ορθοκανονική βάση $\hat{S}$ ισχύει το θεώρημα Parseval.

$$||u||^2 = \sum_i \hat{a}_i^2$$

Ειδικά για το διάνυσμα σφάλματος, έχουμε $||e||^2 = \sum_{i \notin L} (c_i \cdot ||s_i||)^2$ Τότε, αν η μετρική σφάλματος $f(e)$ ταυτίζεται με το τετράγωνο της νόρμας του διανύσματος σφάλματος, κατά

την προσέγγιση το πολύ μιας πολλαπλασιαστικής σταθεράς, μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο για να κατασκευάσουμε τη βέλτιστη περίληψη.

Αν, λοιπόν, εφοδιάσουμε το διανυσματικό μας χώρο με ένα εσωτερικό γινόμενο και διαπιστώσουμε ότι η βάση του Haar είναι ορθογώνια ως προς αυτό, αλλά και ότι η νόρμα που ορίζεται από το εσωτερικό μας γινόμενο ταυτίζεται με τη μετρική SSE κατά την προσέγγιση μιας πολλαπλασιαστικής σταθεράς, τότε, αφού μετατρέψουμε τη βάση του Haar σε ορθοκανονική όπως περιγράψαμε στην προηγούμενη παράγραφο, μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο για την κατασκευή της βέλτιστης περίληψης ως προς τη μετρική L_2. Αυτό ακριβώς κάνουν οι Matias και Urieli στο [6] για την περίπτωση των δεδομένων σε μορφή prefix sums. Αποδεικνύουν, δηλαδή, ότι αν τα δεδομένα μας είναι σε μορφή prefix sums και τα μετασχηματίσουμε με την ορθογώνια μορφή του Haar - με υπολογισμό ημιαθροισμάτων και ημιδιαφορών δηλαδή, τότε μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο περίληψης. Για να κανονικοποιηθούν οι συντελεστές, πολλαπλασιάζονται με $\sqrt{\frac{N}{2^i}(N+1)}$, εκτός του συντελεστή που παριστάνει το μέσο όρο των τιμών, που πολλαπλασιάζεται με $\sqrt{N}$.

Στην περίπτωση των raw data, οι Matias και Urieli αποδεικνύουν ότι με τον ΜΣ Haar δεν μπορεί να χρησιμοποιηθεί ο άπληστος αλγόριθμος. Για την απόδειξη του παραπάνω, βρίσκουν ένα εσωτερικό γινόμενο το οποίο ορίζει μια νόρμα που ταυτίζεται με το SSE του διανύσματος σφάλματος. Στη συνέχεια διαπιστώνεται ότι ο ΜΣ Haar δεν είναι ορθογώνιος ως προς το εσωτερικό αυτό γινόμενο. Και αφού δεν υπάρχει άλλο εσωτερικό γινόμενο που να δίνει την ίδια νόρμα, προκύπτει ότι δεν υπάρχει άλλο εσωτερικό γινόμενο ως προς το οποίο ο Haar να είναι ορθογώνιος και η νόρμα που ορίζει αυτό να ταυτίζεται με τη μετρική SSE. Άρα, δεν μπορούμε να χρησιμοποιούμε τον άπληστο αλγόριθμο περίληψης για να ελαχιστοποιούμε το σφάλμα SSE.

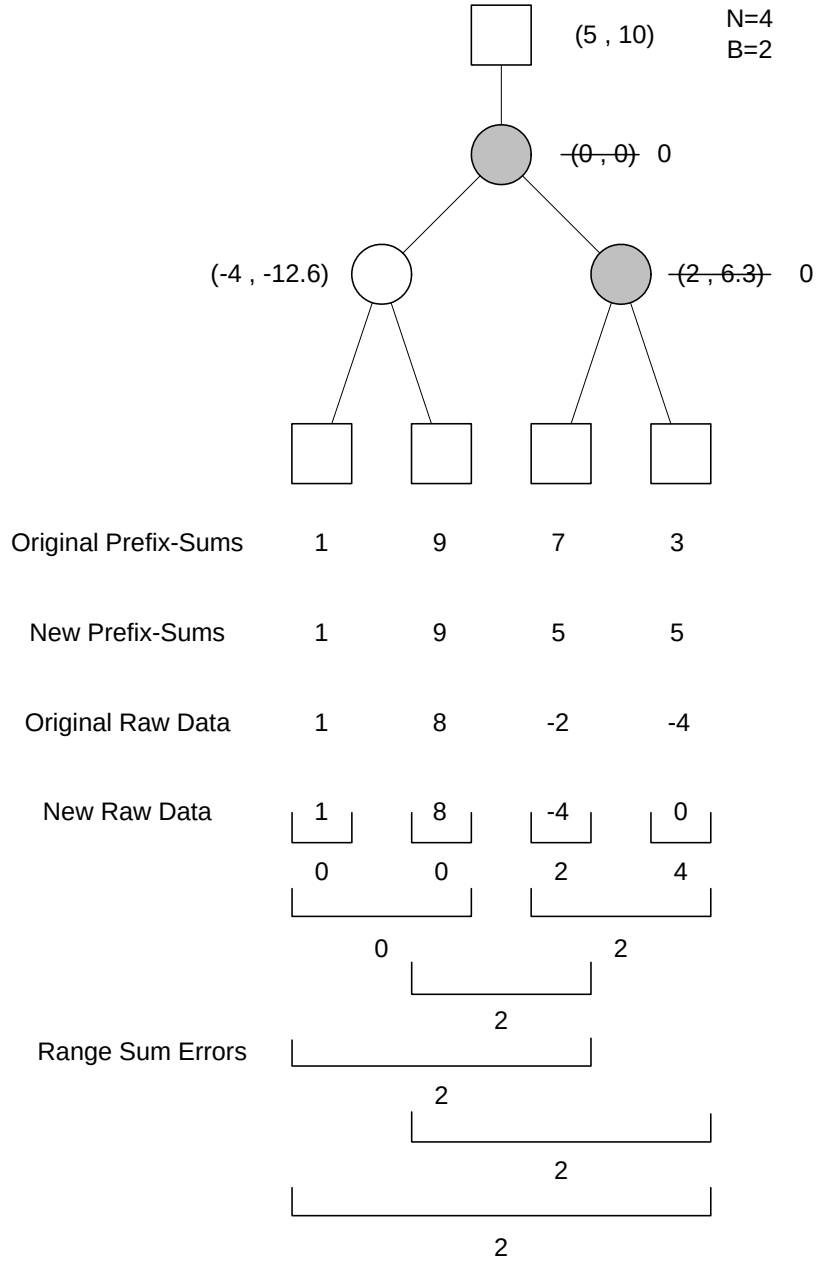
4.3 Prefix-Sums και ο Άπληστος Αλγόριθμος για το SSE

Όπως είδαμε στην προηγούμενη ενότητα, οι Matias και Urieli στο [6] αποδεικνύουν ότι ο μετασχηματισμός Haar Wavelet είναι ορθογώνιος ως προς τη μετρική SSE όταν τα δεδομένα μας βρίσκονται στη μορφή prefix sums. Έτσι, αφού μετασχηματίσουμε τα δεδομένα με τον ορθογώνιο μετασχηματισμό Haar, του οποίου η βάση αποτελείται από διανύσματα με μη μηδενικές συντεταγμένες ± 1 και αφού κανονικοποιήσουμε τους συντελεστές, μπορούμε να εφαρμόσουμε τον άπληστο αλγόριθμο επιλογής συντελεστών, ώστε να κρατήσουμε στην περίληψη τους B μεγαλύτερους από αυτούς. Όπως είπαμε και στην προηγούμενη ενότητα, η κανονικοποίηση των συντελεστών γίνεται πολλαπλασιάζοντας το μέσο όρο του διανύσματος των prefix sums επί $\sqrt{N}$ και τους υπόλοιπους συντελεστές επί $\sqrt{\frac{N}{2^i}(N+1)}$.

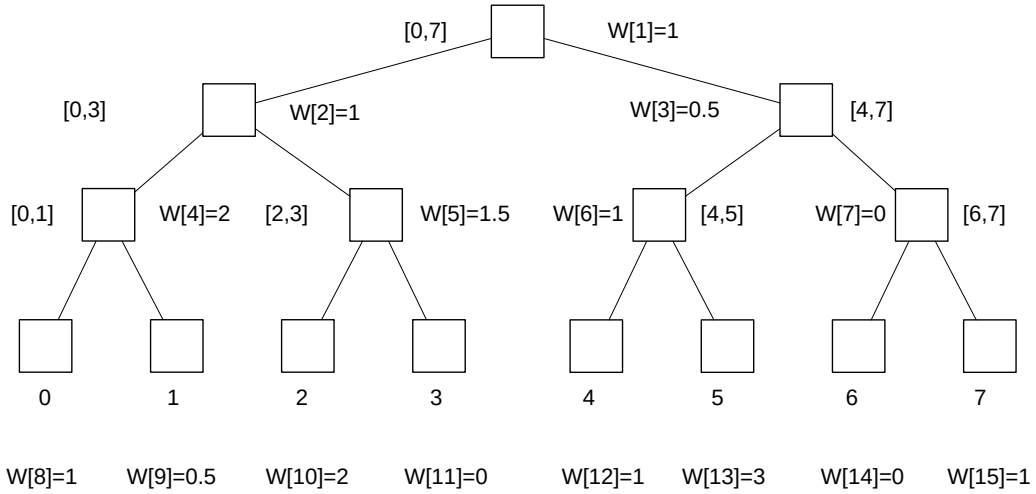
Επομένως, με $C[N]$ το διάνυσμα των κανονικοποιημένων συντελεστών, ισχύει ότι

$$SSE = \sum_{0 \le i \le j < N} err_{i:j}^2 = \sum_{i \notin L} C[i]^2$$

Στο παράδειγμα του σχήματος 4.4 φαίνονται σε παρένθεση οι μη-κανονικοποιημένοι και οι κανονικοποιημένοι συντελεστές Haar που προέκυψαν από το μετασχηματισμό του prefix-



Σχήμα 4.4: Εφαρμογή Άπληστου αλγορίθμου για την L_2 και prefix-sums



Σχήμα 4.5: Παράδειγμα Δυαδικής Ιεραρχίας Range Sum Queries

sums διανύσματος $PS = (1, 9, 7, 3)$. Με εφαρμογή του άπληστου αλγορίθμου μένουν οι συντελεστές $C[0]$ και $C[2]$. Το σφάλμα SSE προκύπτει να είναι ίσο με 40.

$$SSE = \sum_{0 \le i \le j < 4} err_{i:j}^2 = 6 \times 2^2 + 4^2 = 40$$

$$SSE = \sum_{i \notin L} C[i]^2 = C[1]^2 + C[3]^2 = 0^2 + (2\sqrt{10})^2 = 40$$

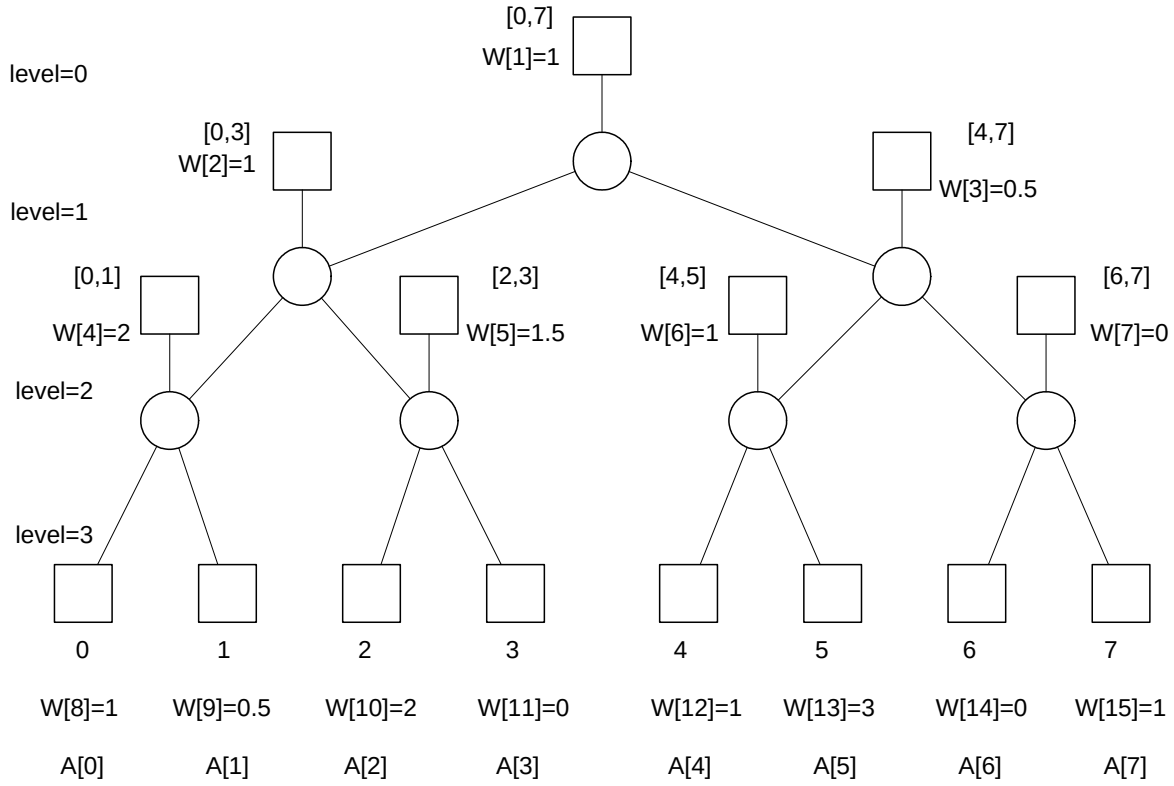
Ο άπληστος αλγόριθμος, έχει χρονική και χωρική πολυπλοκότητα $O(N + B \log N)$ και $O(N)$, αντίστοιχα - (Ενότητα 3.2.2). Το διάνυσμα των prefix sums, αν δεν είναι εκ των προτέρων διαθέσιμο, μπορεί να κατασκευαστεί κατά τη διάρκεια της κατασκευής του μετασχηματισμένου σήματος, οπότε ο μετασχηματισμός παραμένει με πολυπλοκότητα $O(N+B \log N)$.

4.4 Δυαδική ιεραρχία Range Errors πάνω από Raw Data

4.4.1 Εισαγωγή

Στην ενότητα αυτή παρουσιάζουμε έναν αλγόριθμο που επιστρέφει την περίληψη με το ελάχιστο σφάλμα ως προς τη μετρική weighted-L_p, για ένα ειδικό υποσύνολο από range sum errors. Ως είσοδο δέχεται το μετασχηματισμό Haar ενός σήματος δεδομένων, με τα δεδομένα αυτά να βρίσκονται στη μορφή raw data.

Βασική παραδοχή στο πρόβλημα που εξετάζουμε είναι ότι τα range queries τα οποία μας ενδιαφέρουν και πάνω στα οποία μετράμε το σφάλμα, ακολουθούν δυαδική ιεραρχία. Αυτό σημαίνει ότι υπάρχει ένα range query το οποίο βρίσκεται στην κορυφή της δυαδικής ιεραρχίας που καλύπτει ολόκληρο το εύρος των $N = 2^j$ δεδομένων και ότι κάθε άλλο range query της ιεραρχίας καλύπτει το αριστερό ή το δεξιό μισό εύρος του query-γονέα του. Υποθέτουμε ακόμα, χωρίς βλάβη της γενικότητας, ότι το δένδρο που παριστάνει την ιεραρχία είναι πλήρες δυαδικό, δηλαδή δεν υπάρχει query-κόμβος που να έχει μόνο ένα query-παιδί. Για ένα raw data διάνυσμα δεδομένων μεγέθους $N = 2^j$, το πλήθος των range sum queries που μας



Σχήμα 4.6: Παράδειγμα δένδρου σφάλματος για Δυαδική Ιεραρχία Range Sum Queries

ενδιαφέρουν είναι $2N-1$, συμπεριλαμβανομένων και εκείνων που έχουν μοναδιαίο εύρος, που είναι δηλαδή range queries "εκφυλισμένα" σε point queries. Σε κάθε κόμβο της ιεραρχίας αντιστοιχεί κάποιο βάρος w_i, $i \in [1 \dots (2N-1)]$ με το οποίο το σφάλμα rse_i στο αντίστοιχο range sum query συμμετέχει στη μετρική weighted-L_p. Το βάρος κάθε άλλου range query θεωρείται ίσο με 0.

$$dyadic - wL_p = \sum_{i=1}^{2N-1} w_i \cdot rse_i^p$$

Το βάρος w_i μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός, ακόμα και 0. Δεν είναι ανάγκη, δηλαδή, να ακολουθεί κάποια κατανομή ή να πληροί κάποια ειδική συνθήκη (πχ. $\sum w_i = 1$).

Στο σχήμα 4.5 φαίνεται ένα παράδειγμα δυαδικής ιεραρχίας από range sum queries. Δίπλα σε κάθε κόμβο φαίνεται το εύρος του query και το βάρος του αντίστοιχου range sum error για τη μετρική weighted-L_p.

4.4.2 Ο Νέος Δυναμικός Αλγόριθμος για τη weighted-L_p

Θα περιγράψουμε τώρα ένα δυναμικό αλγόριθμο που επιστρέφει την περίληψη B συντελεστών που ελαχιστοποιεί το weighted-L_p σφάλμα.

Ο αλγόριθμος τρέχει πάνω στο δυαδικό δένδρο σφάλματος, οι κόμβοι του οποίου αντιστοιχούν στους μη - κανονικοποιημένους συντελεστές του απλού μετασχηματισμού Haar, όπως τον περιγράψαμε στην ενότητα 2.2 με τη βοήθεια των διανυσμάτων ημιαθροισμάτων και ημιδια-

φορών. Παρατηρούμε επίσης ότι υπάρχει μια αντιστοιχία ανάμεσα στους κόμβους του δένδρου σφάλματος και τους κόμβους της δυαδικής ιεραρχίας: η τιμή του range query που παριστάνεται από έναν κόμβο i της δυαδικής ιεραρχίας μπορεί να υπολογιστεί αθροίζοντας με το κατάλληλο πρόσημο τους συντελεστές του δένδρου σφάλματος που ανήκουν στο $path(i)$, δηλαδή τους πατρικούς κόμβους του i που βρίσκονται στο μονοπάτι από τη ρίζα προς τον κόμβο i - βλ. σχήμα 4.6. Για την αποθήκευση των ενδιάμεσων αποτελεσμάτων χρησιμοποιούμε πίνακες E_i δύο (2) διαστάσεων, έναν για κάθε κόμβο c_i του δένδρου σφάλματος. Η πρώτη διάσταση αναφέρεται στο χώρο της περίληψης που είναι διαθέσιμος, για το υποδένδρο σφάλματος με ρίζα το συγκεκριμένο κόμβου, όπως αυτός προσδιορίζεται από την πρώτη συντεταγμένη. Το μέγεθός της είναι $min\{B, 2^{logN-level(i)} - 1\}$, όπου $level(i)$ το επίπεδο του κόμβου. Και η δεύτερη διάσταση αναφέρεται στους συντελεστές - προγόνους που έχουν κρατηθεί στην περίληψη και οι οποίοι ανήκουν στο μονοπάτι από τη ρίζα του δένδρου σφάλματος ως τον κόμβο. Το μέγεθός της είναι $2^{level(i)+1}$.

Θα περιγράψουμε τώρα τον τρόπο που λειτουργεί ο δυναμικός αλγόριθμος και θα δώσουμε τις εξισώσεις που χρησιμοποιεί για να επιλέγει τους κατάλληλους συντελεστές. Έστω ότι οι συντελεστές είναι αποθηκευμένοι στον πίνακα $C[0 \dots N-1]$ και τα δεδομένα σε μορφή raw data είναι αποθηκευμένα στον πίνακα $A[0 \dots N-1]$. Τα βάρη αποθηκεύονται στον πίνακα $W[1 \dots 2N-1]$. Ειδικότερα, το τμήμα του πίνακα $W[1 \dots N-1]$ περιέχει τα βάρη των range queries με εύρος τουλάχιστον 2, ενώ το τμήμα $W[N \dots 2N-1]$ περιέχει το βάρος των μοναδιαίων range queries, που αντιστοιχούν ουσιαστικά στα δεδομένα. Η βάση της αναδρομής του αλγορίθμου είναι τα φύλλα-δεδομένα. Το σφάλμα ενός range query μοναδιαίου εύρους ισούται με το απόλυτο σφάλμα στην τιμή του, υψωμένο στη δύναμη p, επί το βάρος του query.

$$E_i[b,S] = W[i] \cdot (\sum_{j \in path(i)} sign(i,j) \cdot C[j] - \sum_{j \in S} sign(i,j) \cdot C[j] \ )^p$$

$$= W[i] \cdot (\sum_{j \in path(i)-S} sign(i,j) \cdot C[j] \ )^p$$

Για τους εσωτερικούς κόμβους - συντελεστές του δυαδικού δένδρου σφάλματος, πλην της ρίζας, ο αλγόριθμος μπορεί να υπολογίσει το σφάλμα που αναφέρεται στο αντίστοιχο query, γνωρίζοντας τους συντελεστές - προγόνους του που έχουν ήδη κρατηθεί στην περίληψη, παρόμοια όπως κάνει και για τα φύλλα-δεδομένα. Στη συνέχεια καλείται να επιλέξει εάν θα κρατήσει το συντελεστή ή αν θα τον απορρίψει και για κάθε περίπτωση, υπολογίζει και προσθέτει στο συνολικό σφάλμα το σφάλμα που εισάγεται στα queries-παιδιά του.

$$E_i[b,S] = \min\{E_i^{keep}[b,S], E_i^{drop}[b,S]\}$$

$$E_i^{keep}[b,S] = W[i] \cdot (\sum_{j \in path(i)-S} sign(i,j) \cdot C[j] \ )^p + \min_{0 \le \hat{b} \le b-1} (E_{2i}[\hat{b}, S \cup \{i\}] + E_{2i+1}[b-\hat{b}-1, S \cup \{i\}])$$

$$E_i^{drop}[b,S] = W[i] \cdot (\sum_{j \in path(i)-S} sign(i,j) \cdot C[j] \ )^p + \min_{0 \le \hat{b} \le b} (E_{2i}[\hat{b}, S] + E_{2i+1}[b-\hat{b}-1, S])$$

Η κορυφή της αναδρομής βρίσκεται στη ρίζα c_0 του δένδρου σφάλματος. Εδώ ο αλγόριθμος δεν υπολογίζει το σφάλμα κάποιου query καθώς η ρίζα της ιεραρχίας των queries αντιστοιχεί στον κόμβο c_1 του δένδρου σφάλματος. Ο αλγόριθμος απλά εξετάζει και πάλι τις περιπτώσεις να κρατήσει ή να απορρίψει τον κόμβο c_0. Είναι προφανές ότι στην αρχή του αλγορίθμου κανένας κόμβος δεν έχει ήδη κρατηθεί ($S = \emptyset$, $b = B$).

$$E_0[B, \emptyset] = \min\{E_0^{keep}[B, \emptyset], E_0^{drop}[B, \emptyset]\}$$

$$E_0^{keep}[B, \emptyset] = \min_{0 \le \hat{b} \le B-1} E_1[\hat{b}, \{0\}]$$

$$E_0^{drop}[B, \emptyset] = \min_{0 \le \hat{b} \le B} E_1[\hat{b}, \emptyset]$$

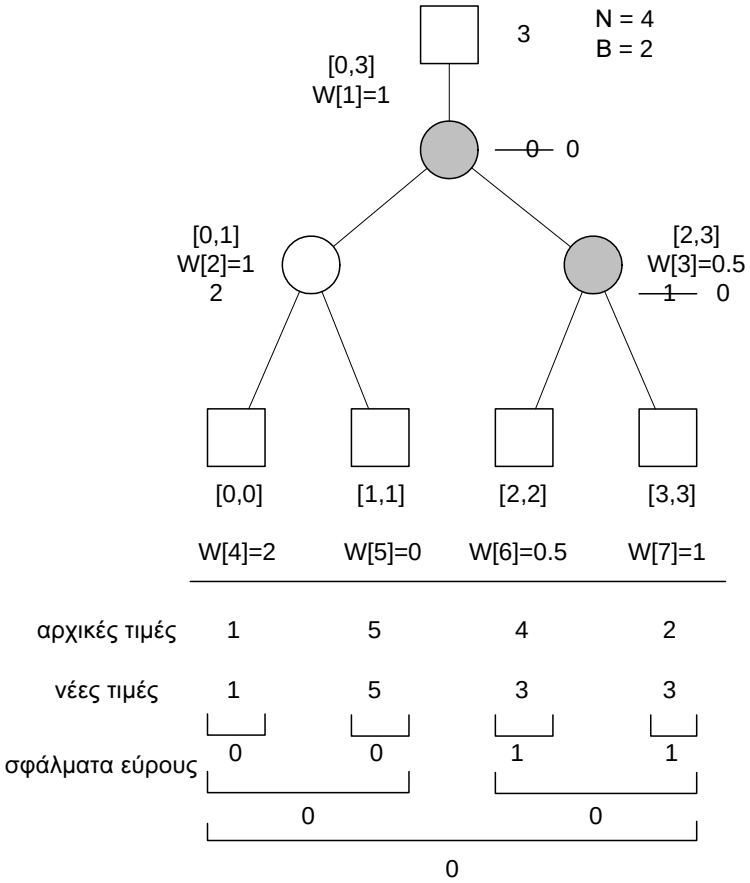
Ο αλγόριθμος, λοιπόν, ξεκινάει από τη ρίζα του δένδρου σφάλματος και αναδρομικά φτάνει στα φύλλα του. Εκεί υπολογίζονται οι πίνακες E_i των φύλλων. Οι πίνακες αυτοί, αφού συμπληρωθούν, επιστρέφονται στα προηγούμενα στάδια της αναδρομής, στους κόμβους - γονείς, δηλαδή, των φύλλων και χρησιμοποιούνται για τον υπολογισμό των αντίστοιχων πινάκων E_i. Αυτό το βήμα επαναλαμβάνεται μέχρι να επιστρέψει ο αλγόριθμος στην κορυφή της αναδρομής: όταν βρισκόμαστε στον κόμβο του συντελεστή c_i, διαθέτουμε τους πίνακες E_{2i} και E_{2i+1} και υπολογίζουμε τον πίνακα E_i για όλες τις δυνατές τιμές των b και S. Στη συνέχεια μπορούμε να αποδεσμεύσουμε τους πίνακες E_{2i} και E_{2i+1} και να επιστρέψουμε τον πίνακα E_i στον πατρικό κόμβο του συντελεστή c_i, τον $c_{i/2}$.

Μπορούμε τώρα να υπολογίσουμε το συνολικό χώρο που καταλαμβάνει ο αλγόριθμος. Για κάθε κόμβο c_i επιπέδου $level(i) = l_i$ του δένδρου σφάλματος δεσμεύουμε έναν πίνακα E_i, ο οποίος έχει διαστάσεις $\min\{B, 2^{logN-l_i} - 1\} \times 2^{l_i+1}$. Για τους κόμβους επιπέδου $l_i = level(i) > logN - logB$, ο κάθε πίνακας έχει $(2^{logN-l_i} - 1) \times 2^{l_i+1} = O(2N) = O(N)$ στοιχεία. Συνολικά δηλαδή έχουμε $O(\sum_{l=logN-logB+1}^{logN} 2^l \cdot N) = O(N^2)$ στοιχεία. Για τους κόμβους επιπέδου $l_i = level(i) \le logN - logB$, ο κάθε πίνακας έχει $B \cdot 2^{l+1}$ στοιχεία. Άρα, συνολικά για τους κόμβους αυτών των επιπέδων έχουμε $\sum_{l=0}^{logN-logB} 2^l \cdot B \cdot 2^{l+1} = B \cdot \sum_{l=0}^{logN-logB} 2^{2l+1} = O(B \cdot \frac{N^2}{B^2}) = O(\frac{N^2}{B})$. Επομένως η χωρική πολυπλοκότητα του αλγορίθμου μας είναι $O(N^2) + O(\frac{N^2}{B}) = O(N^2)$.

Ο χρόνος που χρειάζεται ο αλγόριθμος υπολογίζεται με παρόμοιο τρόπο. Για κάθε κόμβο επιπέδου $level(i) = l$ και για κάθε στοιχείο του πίνακα που υπολογίζουμε, χρειαζόμαστε χρόνο $O(\min\{B,\, 2^{logN-l+1} - 1\})$. Έτσι, ο συνολικός χρόνος που διαρκεί ο αλγόριθμος είναι

$$\sum_{l=0}^{logN-logB} (2^l \cdot B^2 \cdot 2^l) + \sum_{l=logN-logB}^{logN} (2^l \cdot (2^{logN-l})^2 \cdot 2^l) = O(N^2) + O(N^2 logB) = O(N^2 logB)$$

Στο παράδειγμα του σχήματος 4.7 έχει μετασχηματιστεί το raw data διάνυσμα $A = (1, 5, 4, 2)$ και έχει εφαρμοστεί ο δυναμικός αλγόριθμος για τη δυαδική ιεραρχία του σχήματος. Σχηματίζεται περίληψη $B = 2$ συντελεστών, των $C[0]$ και $C[2]$ με βέλτιστο weighted-L_2 σφάλμα 1.5. Στο σχήμα φαίνονται οι νέες τιμές μετά την περίληψη, $\hat{A} = (1, 5, 3, 3)$ και τα σφάλματα εύρους των δυαδικών range sum queries της ιεραρχίας. Υπολογίζουμε, τέλος το



Σχήμα 4.7: Παράδειγμα εφαρμογής του Δυναμικού Αλγορίθμου για Δυαδική Ιεραρχία Range Sum Queries

weighted-L_2 σφάλμα της περίληψης.

$$wL_2err = 0.5 \times 1^2 + 1 \times 1^2 = 1.5$$

Κεφάλαιο 5

Πειραματική σύγκριση αλγορίθμων

5.1 Δεδομένα και Μεθοδολογία

Για τη διεξαγωγή των πειραμάτων υλοποιήσαμε και εκτελέσαμε κάποιους αλγορίθμους κατασκευής περιλήψεων. Στόχος των πειραμάτων είναι να συγκρίνουμε την ακρίβεια των περιλήψεων που επιτυγχάνουν οι αλγόριθμοι αλλά και την ταχύτητα στην κατασκευή τους. Οι μετρήσεις γίνονται σε σχέση αφενός με το πλήθος των δεδομένων και αφετέρου με το μέγεθος της περίληψης. Παράλληλα, μας απασχολεί το ερώτημα κατά πόσο αξίζει να χρησιμοποιούμε πολύπλοκους αλγορίθμους που ελαχιστοποιούν μια μετρική σφάλματος έναντι απλούστερων μη-βέλτιστων αλγορίθμων. Η μετρική σφάλματος που χρησιμοποιείται για να εκτιμήσουμε την ακρίβεια της κάθε περίληψης είναι η weighted-L_2 . Έτσι, στην πειραματική σύγκριση περιλαμβάνονται και αλγόριθμοι που δεν είναι βέλτιστοι για τη μετρική weighted-L_2, ώστε να δούμε εάν κάποια απώλεια στην ακρίβεια της περίληψης αντισταθμίζεται, ενδεχομένως, από την ταχύτητα που προσφέρουν.

Τα πειράματα που εκτελέσαμε χωρίζονται σε δύο ομάδες. Στην πρώτη, μας ενδιαφέρει η ακρίβεια της περίληψης σε σχέση με όλα τα weighted point errors, ενώ στη δεύτερη, μετράμε την ακρίβεια της περίληψης για τη δυαδική ιεραρχία των weighted range sum errors που κτίζεται πάνω από τα δεδομένα μας.

Η είσοδος των αλγορίθμων είναι ένα διάνυσμα N δεδομένων, το μέγεθος B του χώρου περίληψης και ένα διάνυσμα βαρών. Στην πρώτη περίπτωση, το πλήθος των βαρών είναι N, όσο και των δεδομένων, ενώ στη δεύτερη είναι $2N-1$, όσο και το πλήθος των κόμβων της δυαδικής ιεραρχίας. Η έξοδός τους είναι το weighted-L_2 σφάλμα της περίληψης που κατασκευάζουν (για point errors και dyadic range sum errors, αντίστοιχα) καθώς και ο χρόνος $Time$ που χρειάστηκαν για την κατασκευή της περίληψης. Οι μετρήσεις γίνονται αρχικά για σταθερό πλήθος δεδομένων και μεταβλητό μέγεθος περίληψης. Στη συνέχεια, κάνουμε τις μετρήσεις μεταβάλλοντας το πλήθος των δεδομένων και αφήνοντας για χώρο περίληψης ένα σταθερό ποσοστό αυτού.

Για την εκτέλεση των πειραμάτων χρησιμοποιήσαμε συνθετικά data sets μεγέθους μέχρι

$N = 2^{10}$ που ακολουθούν την κατανομή Zipf με παράμετρο 0.6 και 1.2. Πιο συγκεκριμένα, τα δεδομένα εισόδου ακολουθούν την κατανομή Zipf με παράμετρο 0.6. Στην πρώτη ομάδα πειραμάτων, τα N βάρη που αντιστοιχούν στα N point errors ακολουθούν την κατανομή Zipf με παράμετρο 1.2. Στη δεύτερη ομάδα πειραμάτων τα βάρη χωρίζονται σε δύο ανεξάρτητα σύνολα που ακολουθούν το καθένα την κατανομή Zipf με παράμετρο 1.2. Η πρώτη, μεγέθους $N-1$, αντιστοιχεί στα μη-μοναδιαία range sum errors της δυαδικής ιεραρχίας και η δεύτερη, μεγέθους N αντιστοιχεί στα μοναδιαία range sum errors της δυαδικής ιεραρχίας (που ουσιαστικά ταυτίζονται με τα point errors των δεδομένων).

5.2 Point Errors και η μετρική weighted-L_2

5.2.1 Εισαγωγή

Στην ενότητα αυτή παρουσιάζουμε τα αποτελέσματα της πειραματικής σύγκρισης που κάναμε μεταξύ αλγορίθμων περίληψης που επιχειρούν να ελαχιστοποιήσουν τη μετρική weighted-L_2 για point errors.

Οι αλγόριθμοι που συγκρίναμε είναι οι εξής:

- **Matias - Urieli**. Όπως περιγράψαμε στην ενότητα 3.3, ο αλγόριθμος αυτός εφαρμόζει τον άπληστο αλγόριθμο περίληψης πάνω σε ένα wavelet μετασχηματισμό του αρχικού σήματος, παραλλαγή του κλασσικού μετασχηματισμού Haar (Ενότητα 2.2), που ενσωματώνει τα βάρη των point queries. Η χρονική πολυπλοκότητα του μετασχηματισμού είναι $O(N)$ ενώ η πολυπλοκότητα του άπληστου αλγορίθμου είναι $O(N + B \log N)$

- **Garofalakis-Kumar**. Αρχικά περιγράψαμε αναλυτικά την εκδοχή του αλγορίθμου αυτού που ελαχιστοποιεί τη μετρική L_∞ (Ενότητα 3.2.3). Στη συνέχεια περιγράψαμε πώς επεκτείνεται στις weighted-L_p μετρικές και συγκεκριμένα στη μετρική weighted-L_2 (Ενότητα 3.2.4), για την οποία τον υλοποιήσαμε. Η χρονική πολυπλοκότητά του είναι $O(N^2 \log B)$.

- **Classic Wavelets**. Πρόκειται για τον άπληστο αλγόριθμο που εφαρμόζεται πάνω στον κλασσικό μετασχηματισμό Haar (Ενότητα 3.2.2). Όπως έχουμε ήδη εξηγήσει προσφέρει τον απλούστερο τρόπο να κατασκευάσουμε μια περίληψη wavelet ενός σήματος. Δεν ελαχιστοποιεί όμως τη μετρική weighted-L_2 αλλά τη μετρική L_2. Η χρονική πολυπλοκότητα του κλασσικού μετασχηματισμού Haar είναι $O(N)$ ενώ η πολυπλοκότητα του άπληστου αλγορίθμου είναι $O(N + B \log N)$.

- **Classic Histograms**. Ο τελευταίος αλγόριθμος που χρησιμοποιούμε στην πειραματική σύγκριση δεν κατασκευάζει περίληψη wavelet αλλά βέλτιστο ιστόγραμμα για τη μετρική weighted-L_2. Παρουσιάζεται στο [5]. Η χρονική του πολυπλοκότητα είναι $O(N^2 B)$.

Οι κλασσικός αλγόριθμος wavelet (Classic Wavelets), δεν ελαχιστοποιεί τη μετρική weighted-$L2$ για το μετασχηματισμό που εκτελεί (κλασσικός Haar), αλλά μπαίνει στη σύγκριση για να δούμε εάν κάποια απώλεια στην ακρίβεια της περίληψης μπορεί ενδεχομένως να αντισταθμίζεται από κάποιο κέρδος στο χρόνο κατασκευής της.

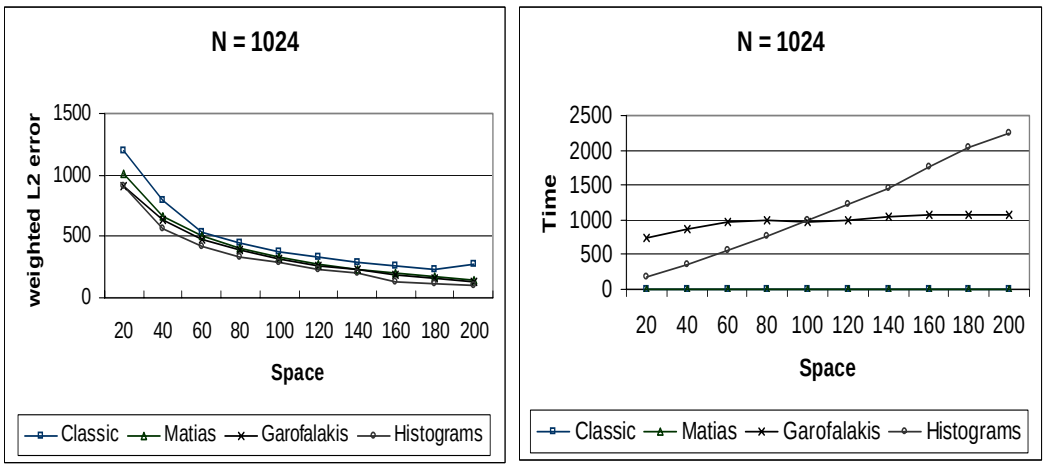
Αλγόριθμος	Χρόνος	Χώρος	Βέλτιστος
Matias-Urieli	$N + B \log N$	N	ΝΑΙ
Garofalakis-Kumar	$N^2 \log B$	N^2	ΝΑΙ
Classic Wavelets	$N + B \log N$	N	ΟΧΙ
Classic Histograms	$N^2 B^2$	BN	ΝΑΙ

Πίνακας 5.1: Αλγόριθμοι προς σύγκριση για point errors

5.2.2 Τα πειραματικά αποτελέσματα

Επίδοση σε σχέση με το μέγεθος της περίληψης

Στο σχήμα 5.1 φαίνεται η ακρίβεια της περίληψης που κατασκευάζουν οι 4 αλγόριθμοι, για ένα σήμα μεγέθους $N = 2^{10}$ και ένα σχετικά μεγάλο εύρος τιμών του διαθέσιμου χώρου περίληψης. Όπως αναμενόταν για τους βέλτιστους αλγορίθμους, το σφάλμα φθίνει όσο μεγαλώνει η περίληψη, κάτι που δεν ισχύει πάντα για τον κλασσικό αλγόριθμο που δεν είναι βέλτιστος για τη μετρική weighted-L_2. Ο κλασσικός αλγόριθμος δίνει μεγαλύτερο σφάλμα από τους αλγορίθμους Garofalakis - Kumar και Matias - Urieli, οι οποίοι είναι βέλτιστοι για τη μετρική weighted-L_2, για το ίδιο μέγεθος περίληψης B. Παρατηρούμε ακόμα, ότι ο κλασσικός αλγόριθμος κατασκευής ιστογραμμάτων δίνει γενικά ακριβέστερη ή το ίδιο ακριβή περίληψη με τους αλγορίθμους wavelet Garofalakis - Kumar και Matias - Urieli.

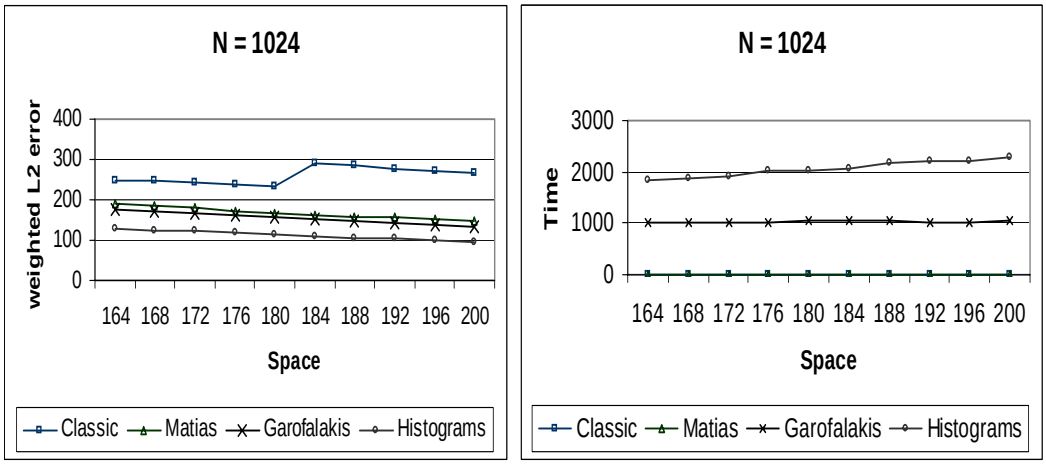


Σχήμα 5.1: Point Errors - Αποτελέσματα για δεδομένα μήκους 2^{10}

Στο πεδίο του χρόνου (Σχήμα 5.1) βλέπουμε ότι ο κλασσικός αλγόριθμος wavelet και ο αλγόριθμος των Matias - Urieli είναι σημαντικά πιο γρήγοροι από εκείνους των Garofalakis - Kumar και του κλασσικού αλγορίθμου ιστογραμμάτων. Επιπλέον, όπως αναμενόταν από τη θεωρητική ανάλυση της χρονικής πολυπλοκότητας που προηγήθηκε, ο χρόνος των δύο πρώτων αλγορίθμων είναι μικρός και παρουσιάζεται πρακτικά ανεξάρτητος του χώρου περίληψης, ο χρόνος του Garofalakis - Kumar αυξάνεται λογαριθμικά, ενώ ο χρόνος του κλασσικού αλγορίθμου ιστογραμμάτων αυξάνεται γραμμικά με την αύξηση του χώρου περίληψης.

Στο σχήμα 5.2 φαίνονται τα αποτελέσματα στην ακρίβεια και την ταχύτητα των αλγορίθμων

για ένα στενότερο εύρος τιμών του χώρου περίληψης.

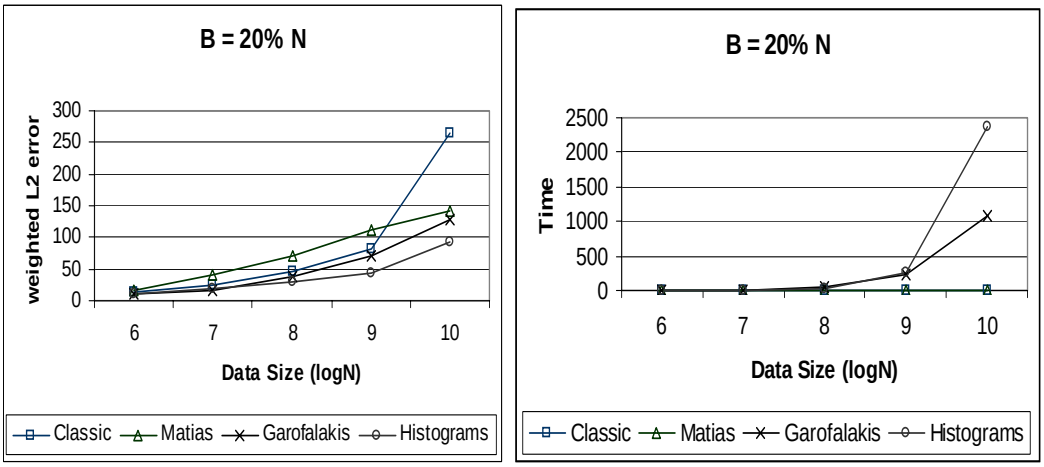


Σχήμα 5.2: Point Errors - Αποτελέσματα για δεδομένα μήκους 2^{10}

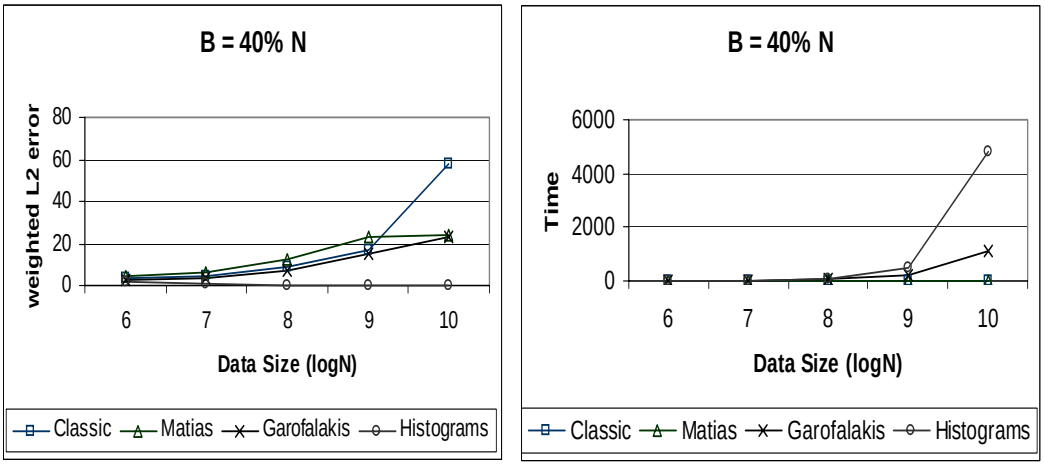
Επίδοση σε σχέση με το πλήθος των δεδομένων

Στη συνέχεια, εκτελούμε τους αλγορίθμους για μεταβαλλόμενο πλήθος δεδομένων και με το χώρο περίληψης να είναι ένα σταθερό ποσοστό του πλήθους των δεδομένων. Στο σχήμα 5.3 φαίνεται η ακρίβεια και ο χρόνος της περίληψης για $B = 20\%N$ και μεταβλητό μήκος δεδομένων. Βλέπουμε ότι γενικά η ακρίβεια της περίληψης ελαττώνεται για αυξανόμενο πλήθος δεδομένων, παρόλο που το μέγεθος της περίληψης μεγαλώνει κατά τον ίδιο παράγοντα. Βλέπουμε ότι ο κλασσικός αλγόριθμος περίληψης μπορεί να πετυχαίνει καλή ακρίβεια σε αρκετές περιπτώσεις, αλλά μπορεί και το σφάλμα που δίνει να είναι ΄απρόβλεπτα΄ μεγάλο, όπως φαίνεται για $\log N = 10$, αφού δε συνυπολογίζει τα βάρη των point errors. Όσον αφορά τους βέλτιστους αλγορίθμους, δεν εμφανίζεται μεγάλη απόκλιση μεταξύ τους, αλλά για το διάστημα που εξετάζουμε φαίνεται ο αλγόριθμος Garofalakis-Kumar να δίνει μεγαλύτερη ακρίβεια από τον Matias - Urieli και ο κλασσικός αλγόριθμος ιστογραμμάτων παρουσιάζεται σταθερά πιο ακριβής από όλους. Ωστόσο, για τις ίδιες εκτελέσεις, ο κλασσικός αλγόριθμος ιστογραμμάτων είναι με αρκετή διαφορά ο πλέον χρονοβόρος (χρονική πολυπλοκότητα $O(N^2B^2)$), ειδικά για τα μεγαλύτερα datasets, ο Garofalakis-Kumar λιγότερο χρονοβόρος (πολυπλοκότητα $O(N^2 \log B)$),ενώ ο κλασσικός αλγόριθμος περίληψης και ο Matias - Urieli είναι εξαιρετικά ταχείς (πολυπλοκότητα $O(N + B \log N)$).

Παρόμοια συμπεράσματα προκύπτουν και από την εκτέλεση των αλγορίθμων για μεγαλύτερη περίληψη, μεγέθους $B = 40\%N$ (Σχήμα 5.4). Ας προσέξουμε, όμως, ότι η ακρίβεια που δίνει ο κλασσικός αλγόριθμος ιστογραμμάτων είναι απόλυτη για μεγάλα datasets.



Σχήμα 5.3: Point Errors - Αποτελέσματα για χώρο περίληψης $20\%N$



Σχήμα 5.4: Point Errors - Αποτελέσματα για χώρο περίληψης $B = 40\%N$

5.3 Δυαδικά Range Sum Errors και η μετρική Weighted-L_2

5.3.1 Εισαγωγή

Στην ενότητα αυτή παρουσιάζουμε τα αποτελέσματα της πειραματικής σύγκρισης που κάναμε μεταξύ αλγορίθμων περίληψης που επιχειρούν να ελαχιστοποιήσουν τη μετρική weighted-L_2 για dyadic range sum errors.

Οι αλγόριθμοι που συγκρίναμε είναι οι εξής:

- **Rangewave**. Είναι ο νέος αλγόριθμος που παρουσιάσαμε στην ενότητα 4.4.2. Χρησιμοποιεί τον κλασσικό μετασχηματισμό Haar και ελαχιστοποιεί τη μετρική weighted-L_2 για δυαδική ιεραρχία range sum errors. Η χωρική και χρονική πολυπλοκότητά του είναι $O(N^2)$ και $O(N^2 \log B)$ αντίστοιχα.

- **Koudas-Guha**. Ο αλγόριθμος αυτός δεν κατασκευάζει περίληψη wavelet αλλά βέλτιστο ιστόγραμμα για τη μετρική weighted - L_2 πάνω από δυαδική ιεραρχία range sum errors και παρουσιάζεται στο [4]. Η χρονική του πολυπλοκότητα είναι $O(N^7B^2)$ και οι

απαιτήσεις του σε χώρο είναι $O(N^5B)$.

- **Matias - Urieli**. Όπως περιγράψαμε στην ενότητα 4.3, όταν ο άπληστος αλγόριθμος περίληψης εφαρμοστεί πάνω από δεδομένα σε μορφή prefix-sums, ελαχιστοποιεί τη μετρική L_2 για το σύνολο των range sum errors (δε συνυπολογίζει τα βάρη, δηλαδή). Η χρονική πολυπλοκότητα του μετασχηματισμού και της κατασκευής των prefix-sums είναι $O(N)$ ενώ η πολυπλοκότητα του άπληστου αλγορίθμου είναι $O(N + B \log N)$

- **Classic**. Χρησιμοποιούμε και πάλι τον κλασσικό αλγόριθμο περίληψης, λόγω της ταχύτητάς του, παρόλο που ελαχιστοποιεί τη μετρική L_2 για point errors. Η χρονική πολυπλοκότητα του κλασσικού μετασχηματισμού Haar είναι $O(N)$ ενώ η πολυπλοκότητα του άπληστου αλγορίθμου είναι $O(N + B \log N)$.

Αλγόριθμος	**Χρόνος**	**Χώρος**	**Βέλτιστος**
Rangewave	$N^2 \log B$	N^2	ΝΑΙ
Koudas-Guha	N^7B^2	N^5B	ΝΑΙ
Matias-Urieli	$N + B \log N$	N	Μόνο για ίσα βάρη
Classic	$N + B \log N$	N	ΟΧΙ

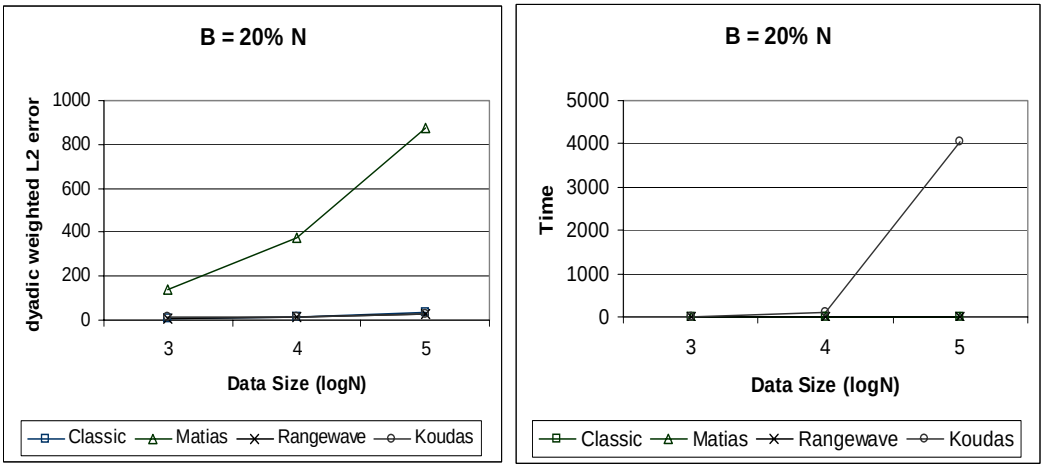
Πίνακας 5.2: Αλγόριθμοι προς σύγκριση για dyadic range sum errors

5.3.2 Τα πειραματικά αποτελέσματα

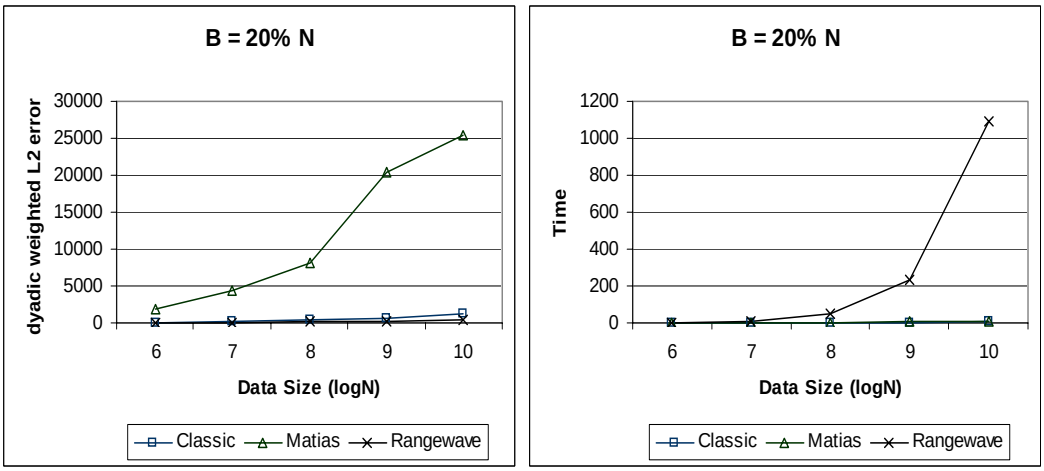
Επίδοση σε σχέση με το πλήθος των δεδομένων

Αρχικά εκτελούμε τους αλγορίθμους για χώρο περίληψης $B = 20\%N$ και datasets μικρού μεγέθους. Στο σχήμα 5.5 φαίνεται η απόκρισή τους. Οι αλγόριθμοι, με εξαίρεση εκείνον των Matias-Urieli, εμφανίζουν παρόμοια συμπεριφορά και μικρό σφάλμα για τα μικρά datasets (θυμίζουμε ότι ο Matias-Urieli ελαχιστοποιεί τη weighted-L_2 για όλα τα range sum errors). Παρατηρούμε ότι ο αλγόριθμος των Koudas και Guha αποκλίνει πολύ γρήγορα σε χρόνο από τους υπόλοιπους ενώ οι απαιτήσεις του σε χώρο καθιστούν πρακτικά αδύνατη την εκτέλεσή του σε υπολογιστές με συμβατικές δυνατότητες μνήμης (για δεδομένα μεγέθους $N = 2^7 = 128$ και μέγεθος περίληψης $B = 4$, απαιτείται μνήμη πολλαπλάσια των 128GB). Γι΄ αυτό δεν τον εξετάζουμε στη συνέχεια.

Στη συνέχεια εκτελούμε και πάλι τους αλγορίθμους (πλην εκείνον των Koudas-Guha) για χώρο περίληψης $B = 20\%N$ αλλά για μεγαλύτερα datasets. Στο σχήμα 5.6 φαίνεται ότι ο αλγόριθμος Rangewave πετυχαίνει το μικρότερο σφάλμα, ενώ ο κλασσικός αλγόριθμος περίληψης, παρότι ελαχιστοποιεί τη μετρική L_2 για point errors εμφανίζει πολύ μικρότερο σφάλμα από τον αλγόριθμο των Matias - Urieli, που ελαχιστοποιεί τη μετρική L_2 για όλα τα range sum errors. Από την άλλη πλευρά, το κόστος του Rangewave σε χρόνο είναι αρκετά μεγαλύτερο από εκείνο των άπληστων αλγορίθμων.



Σχήμα 5.5: Dyadic Range Sum Errors - Αποτελέσματα για χώρο περίληψης $B = 20\%N$



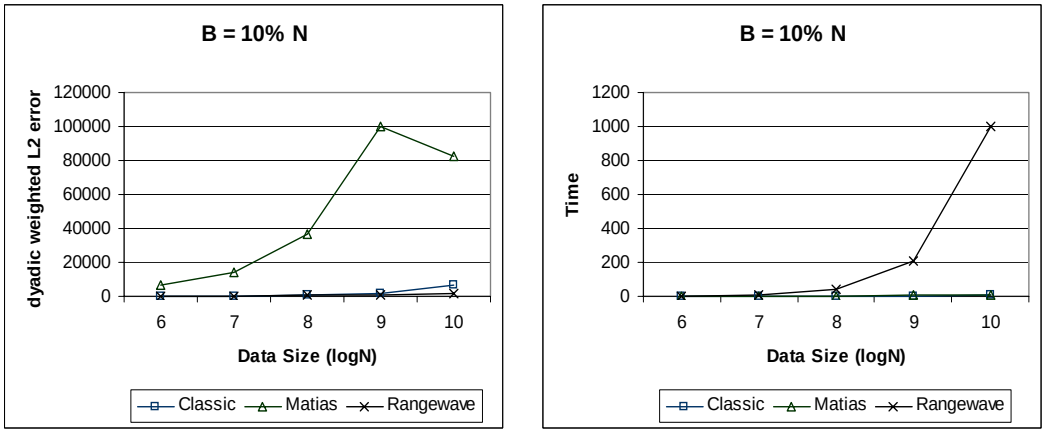
Σχήμα 5.6: Dyadic Range Sum Errors - Αποτελέσματα για χώρο περίληψης $B = 20\%N$

Παρόμοια συμπεράσματα προκύπτουν από την εκτέλεση των αλγορίθμων για μικρότερο χώρο περίληψης. Στο σχήμα 5.7 φαίνεται η απόκριση των αλγορίθμων για χώρο περίληψης $B = 10\%N$ και μεταβλητό πλήθος δεδομένων.

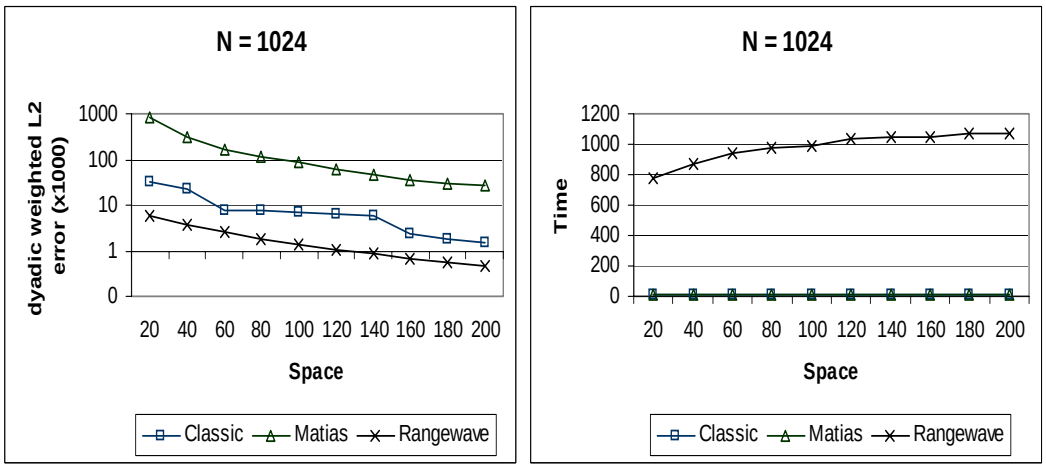
Επίδοση σε σχέση με το μέγεθος της περίληψης

Τέλος, εκτελέσαμε τους αλγορίθμους για σταθερό πλήθος δεδομένων $N = 2^{10} = 1024$ και μεταβλητό χώρο περίληψης B που δεν ξεπερνάει το $0.2 \times N$. Αρχικά παρουσιάζουμε την απόκριση των αλγορίθμων για μεγάλο εύρος στην τιμή του B (Σχήμα 5.8). Φαίνεται ότι ο αλγόριθμος Rangewave παρουσιάζει το μικρότερο σφάλμα σε σύγκριση με τους άλλους δύο, ενώ η κλασσική περίληψη wavelet παρουσιάζει και πάλι σημαντικά καλύτερη ακρίβεια από την περίληψη των Matias - Urieli. Όσον αφορά την ταχύτητα, ο Rangewave αποδεικνύεται αρκετά πιο χρονοβόρος σε σχέση με τους δύο άπληστους αλγορίθμους και η επίδοσή του σε ταχύτητα μειώνεται λογαριθμικά με την αύξηση του μεγέθους της περίληψης. Από την άλλη, οι άπληστες περιλήψεις wavelet (κλασσική και Matias - Urieli) εμφανίζονται και πάλι πρακτικά ανεξάρτητες του χώρου περίληψης για τέτοιο εύρος τιμών του.

Στο σχήμα 5.9 φαίνεται η απόκριση των αλγορίθμων για σύνολο δεδομένων μήκους $N =$



Σχήμα 5.7: Dyadic Range Sum Errors - Αποτελέσματα για χώρο περίληψης $B = 10\%N$



Σχήμα 5.8: Dyadic Range Sum Errors - Αποτελέσματα για δεδομένα μήκους 2^{10}

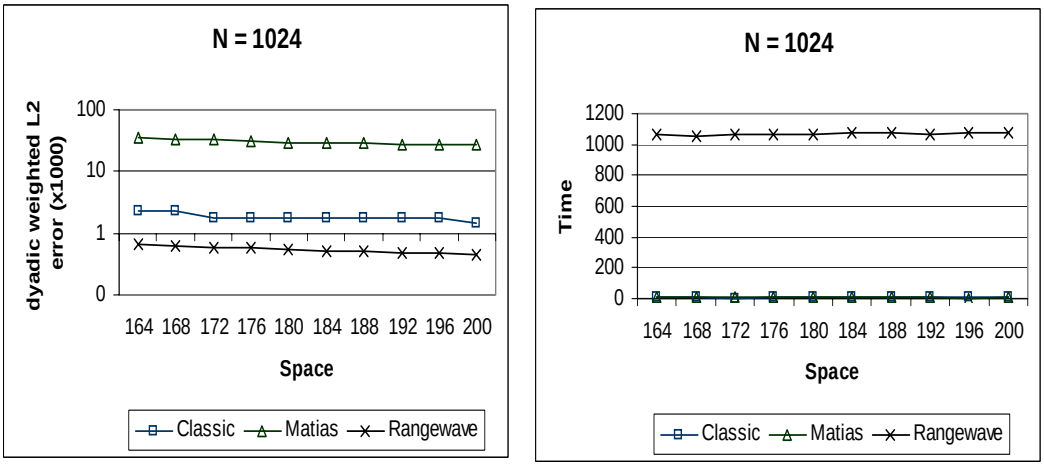
1024 και για τιμές του χώρου περίληψης εστιασμένες γύρω στο $B = 15\%N - 20\%N$.

5.4 Συμπεράσματα

Για τα πειράματα που εκτελέσαμε και τα αποτελέσματα που πήραμε μπορούμε να κάνουμε συνοπτικά κάποιες παρατηρήσεις.

Όσον αφορά τα πειράματα για τα (weighted) point errors, είδαμε ότι ο κλασσικός αλγόριθμος ιστογραμμάτων πετυχαίνει την καλύτερη ακρίβεια για τα datasets που χρησιμοποιήσαμε, χωρίς όμως οι βέλτιστοι αλγόριθμοι ιστογραμμάτων (Matias-Urieli και Garofalakis-Kumar) να υστερούν σημαντικά. Από την άλλη, τα ιστογράμματα υστερούν σημαντικά σε ταχύτητα σε σχέση με τον αλγόριθμο Matias - Urieli, ο οποίος εμφανίζεται εξαιρετικά ταχύς σε όλα τα αποτελέσματα, και σε σχέση με τον Garofalakis-Kumar. Ο κλασσικός αλγόριθμος περίληψης wavelet, μη-βέλτιστος για τη μετρική weighted-L_2, αν και είναι το ίδιο γρήγορος με τον Matias-Urieli, δεν αποδεικνύεται αρκετά αξιόπιστος στην ακρίβεια που δίνει, ώστε να χρησιμοποιηθεί στη θέση του.

Φαίνεται, λοιπόν, ότι ο αλγόριθμος Matias-Urieli πετυχαίνει τον καλύτερο συνδυασμό ακρίβειας και ταχύτητας. Αν, ωστόσο μας ενδιαφέρει η καλύτερη δυνατή ακρίβεια και όχι η



Σχήμα 5.9: Dyadic Range Sum Errors - Αποτελέσματα για δεδομένα μήκους 2^{10}

ταχύτητα, τα κλασσικά ιστογράμματα προσφέρουν την ενδεδειγμένη λύση.

Όσον αφορά τα (weighted) dyadic range sum errors, δύο ήταν οι βέλτιστοι αλγόριθμοι που συγκρίναμε: ο (νέος) αλγόριθμος Rangewave και ο αλγόριθμος των Koudas-Guha. Για τα μικρά datasets που χρησιμοποιήσαμε, η ακρίβεια που παρουσίασαν ήταν παραπλήσια. Όπως αναμενόταν, όμως, και από τη θεωρητική πολυπλοκότητά του, ο Koudas-Guha αποδείχθηκε εξαιρετικά δαπανηρός σε χώρο και χρόνο, τόσο ώστε η χρησιμοποίησή του να φαίνεται αδύνατη σε συμβατικά μηχανήματα. Οι δύο μη-βέλτιστοι αλγόριθμοι Matias - Urieli για prefix-sums και ο κλασσικός, είχαν πολύ καλή απόδοση σε ταχύτητα, με τον πρώτο, όμως, να είναι πολύ λιγότερο ακριβής σε σχέση με το δεύτερο. Θυμίζουμε ότι ο Matias - Urieli για prefix-sums βελτιστοποιεί τη μετρική L_2 για όλα τα range sum errors, ενώ ο κλασσικός βελτιστοποιεί την L_2 για όλα τα point errors.

Συγκρίνοντας τώρα τον Rangewave με τον κλασσικό αλγόριθμο, βλέπουμε ότι ο Rangewave είναι πιο ακριβής και ο κλασσικός είναι αρκετά πιο γρήγορος. Αν, λοιπόν, μας ενδιαφέρει περισσότερο η ακρίβεια στην περίληψη και μια μέτρια ταχύτητα, τότε ο Rangewave είναι ο πιο κατάλληλος να χρησιμοποιήσουμε. Ωστόσο, αν μας ενδιαφέρει πρωταρχικά η ταχύτητα και μπορούμε να θυσιάσουμε ένα μέρος της ακρίβειας, ο κλασσικός αλγόριθμος μπορεί να προσφέρει μια καλή λύση.

Κεφάλαιο 6

Επίλογος

6.1 Συνοπτικές Παρατηρήσεις

Στη διπλωματική αυτή παρουσιάσαμε αλγορίθμους που κατασκευάζουν περιλήψεις wavelet ελαχιστοποιώντας κάποια weighted-L_p μετρική σφάλματος για point ή range sum errors.

Στην περίπτωση των point errors συναντήσαμε αρκετούς βέλτιστους και αποδοτικούς αλγορίθμους, τους οποίους κατηγοριοποιήσαμε με κριτήριο τον τρόπο που αντιμετωπίζουν τα βάρη για τις weighted-L_p μετρικές. Τους αναλύσαμε θεωρητικά και υπολογίσαμε τη χρονική και χωρική τους πολυπλοκότητα.

Στην περίπτωση των range sum errors, είδαμε αρχικά ότι ο κλασσικός μετασχηματισμός Haar είναι ορθοκανονικός για τη μετρική L_2 μόνο για δεδομένα σε μορφή prefix-sums και ότι γι΄ αυτά μπορούμε να χρησιμοποιήσουμε τον άπληστο αλγόριθμο επιλογής συντελεστών ώστε να ελαχιστοποιήσουμε τη μετρική.

Προτείναμε, επίσης, ένα νέο αποδοτικό δυναμικό αλγόριθμο που ελαχιστοποιεί τη μετρική weighted-L_p για range sum errors που ακολουθούν δυαδική ιεραρχία. Ο νέος αλγόριθμος, ο Rangewave, είναι ο πρώτος αλγόριθμος για range sum errors και δεδομένα που βρίσκονται σε μορφή raw data.

Μετά τη θεωρητική ανάλυση των αλγορίθμων, παρουσιάσαμε τα αποτελέσματα της πειραματικής σύγκρισης που έγινε για κάποιους από αυτούς. Για τη μετρική weighted-L_2 με την οποία μετρήσαμε την ακρίβεια των περιλήψεων που κατασκευάστηκαν, συμπεριλάβαμε τόσο σχετικά πολύπλοκους αλγορίθμους που ελαχιστοποιούν το σφάλμα, όσο και γρήγορους αλγορίθμους που δεν είναι βέλτιστοι ως προς αυτήν. Συμπεριλάβαμε ακόμα, βέλτιστους αλγορίθμους κατασκευής ιστογραμμάτων, ώστε να έχουμε μια σύγκριση μεταξύ των δύο μεθόδων (wavelets και ιστογράμματα). Στην περίπτωση των point errors είδαμε ότι ο αλγόριθμος wavelet των Matias - Urieli αποδείχθηκε μια καλή λύση για κατασκευή περιλήψεων που συνδυάζουν την ταχύτητα με την ακρίβεια και ότι ο κλασσικός αλγόριθμος κατασκευής ιστογραμμάτων, αν και πιο αργός, πέτυχε ελαφρώς μικρότερο σφάλμα. Στην περίπτωση των range sum errors είδαμε ότι ο νέος αλγόριθμος αποτελεί μια αποδοτική λύση περιλήψεων με μικρό σφάλμα, αλλά και ότι ο κλασσικός άπληστος αλγόριθμος wavelet μπορεί να επιτύχει αρκετά καλή ακρίβεια σε μικρό χρόνο.

6.2 Μελλοντική Εργασία

Ενώ για την οικογένεια των point error μετρικών έχει διεξαχθεί αρκετή έρευνα και έχουν προταθεί αποδοτικοί αλγόριθμοι wavelet, για την περιοχή των range sum error μετρικών δεν έχουν μελετηθεί σε σημαντικό βαθμό τρόποι ελαχιστοποίησης του σφάλματος για μετρικές weighted-L_p. Έτσι, αποτελεί άμεσο ενδιαφέρον να αναζητηθούν αποδοτικοί αλγόριθμοι που θα ελαχιστοποιούν το σφάλμα των περιλήψεων wavelet για κάποιες ή όλες τις μετρικές weighted-L_p. Όπως είδαμε, ο άπληστος αλγόριθμος που χρησιμοποιεί τον κλασσικό ΜΣ Haar πάνω από prefix-sums ελαχιστοποιεί τη μετρική L_2 για όλα τα range sum errors. Αποδοτικοί δυναμικοί αλγόριθμοι που θα τρέχουν πάνω στο (δυαδικό) δένδρο σφάλματος και που θα ελαχιστοποιούν μια μετρική για όλα τα range sum errors είναι φαίνεται μάλλον απίθανο να υπάρξουν, καθώς το σύνολο των range sum errors δεν ακολουθεί κάποια ιεραρχία. Σαν πρώτο βήμα, ωστόσο, μπορεί να διερευνηθεί αν υπάρχει κάποια παραλλαγή του ΜΣ Haar Wavelet για prefix sums που ενσωματώνει τα βάρη των αθροιστικών σφαλμάτων και για την οποία ο άπληστος αλγόριθμος επιλογής συντελεστών θα ελαχιστοποιεί τη μετρική weighted-L_2 (παρόμοια με τον αλγόριθμο των Matias-Urieli που ελαχιστοποιεί τη weighted-L_2 για όλα τα point errors).

Για την περίπτωση που τα range sum errors ακολουθούν δυαδική ιεραρχία, ο αλγόριθμος Rangewave αποτελεί μια καλή λύση. Φαίνεται ότι μια επέκτασή του σε γενικότερη ιεραρχία από range sum errors είναι προσιτή. Μια ιδέα που ίσως αξίζει να εξεταστεί είναι η μετατροπή μιας γενικής ιεραρχίας σε δυαδική και υπό ποιες προϋποθέσεις μπορεί να γίνει αυτή. Μια άλλη ιδέα είναι η εύρεση ενός γενικού σχήματος μετασχηματισμού wavelet, ο οποίος θα ακολουθεί τη γενική ιεραρχία, όπως ο Haar ακολουθεί τη δυαδική ιεραρχία των range sum errors. Ελλείψη ενός αλγορίθμου που ελαχιστοποιεί μια μετρική (πλην της L_2) για όλα τα range sum errors, μια τέτοια επέκταση θα μπορεί να αποδειχθεί χρήσιμη για μια ευρεία κατηγορία περιλήψεων.

Bibliography

[1] M. Garofalakis και A. Kumar. Deterministic wavelet thresholding for maximum-error metrics. Στο *Proceedings ACM Principles of Database Systems (PODS)*, σελίδες 166–176, 2004.

[2] Amara Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2, 1995.

[3] Sudipto Guha. Space efficiency in synopsis construction algorithms. Στο *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, σελίδες 409–420. VLDB Endowment, 2005.

[4] Sudipto Guha, Nick Koudas και Divesh Srivastava. Fast algorithms for hierarchical range histogram construction. Στο *PODS*, σελίδες 180–187, 2002.

[5] H. V. Jagadish, Nick Koudas, S. Muthukrishnan, Viswanath Poosala, Kenneth C. Sevcik και Torsten Suel. Optimal histograms with quality guarantees. Στο *VLDB*, σελίδες 275–286, 1998.

[6] Y. Matias και D. Urieli. On the optimality of the greedy heuristic in wavelet synopses for range queries. Τεχνική Αναφορά υπ. αρίθμ. TR-TAU, 2005.

[7] S. Muthukrishnan. Subquadratic algorithms for workload-aware haar wavelet synopses. Στο *FSTTCS*, 2005.

[8] D. Urieli και Y. Matias. Optimal workload-based weighted wavelet synopses. Στο *Proceedings of International Conference on Database Theory (ICDT)*, 2005.

[9] Clemens Valens. A really friendly guide to wavelets. *http://perso.orange.fr/polyvalens/clemens/wavelets/wavelets.html*, 2004.