Κεφάλαιο 11

ΑΝΑΚΤΗΣΗ ΜΕ ΔΕΥΤΕΡΕΥΟΝ ΚΛΕΙΔΙ

Κεφάλαιο 11

ΑΝΑΚΤΗΣΗ ΜΕ ΔΕΥΤΕΡΕΥΟΝ ΚΛΕΙΔΙ

11.1 Εισαγωγή

Όλες οι οργανώσεις αρχείων που αναπτύχθηκαν στα προηγουμένα κεφάλαια είναι αποτελεσματικές, όταν οι υποβαλλόμενες ερωτήσεις γίνονται με βάση τιμές του πρωτεύοντος κλειδιού. Όμως, πολύ συχνά στην πράξη τίθενται ερωτήσεις με βάση την τιμή κάποιου δευτερεύοντος κλειδιού. Στην περίπτωση αυτή λέγεται ότι γίνεται **ανάκτηση με βάση δευτερεύον κλειδί** (secondary key retrieval). Ο δευτερεύων κατάλογος, με οποιαδήποτε δομή και αν είναι οργανωμένος (πχ. B^+-δένδρο κλπ.), έχει ως βασικό χαρακτηριστικό ότι κάθε τιμή ενός δευτερεύοντος κλειδιού συνοδεύεται από μία λίστα δεικτών προς τις αντίστοιχες εγγραφές. Έτσι επιτυγχάνεται ταχύτατη προσπέλαση, βέβαια με τίμημα αυξημένες απαιτήσεις σε μνήμη και επιπρόσθετο χρονικό κόστος για την ενημέρωση των καταλόγων κάθε φορά που ενημερώνται οι εγγραφές. Όσο περισσότεροι είναι οι δευτερεύοντες κατάλογοι για ένα αρχείο, τόσο μεγαλώνουν τα προβλήματα του αυξημένου χώρου μνήμης και των χρονοβόρων ενημερώσεων. Ένα μέτρο για τον προσδιορισμό του πόσοι και ποιοί δευτερεύοντες κατάλογοι πρέπει να δημιουργηθούν είναι ο λόγος του αριθμού των ενημερώσεων προς τον αριθμό των προσπελάσεων.

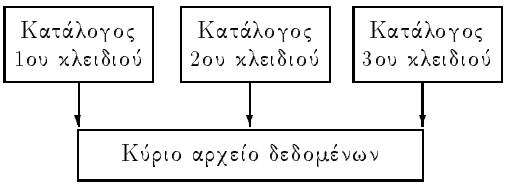
Στο παρόν κεφάλαιο αρχικά πρόκειται να μελετηθούν δύο δομές καταλόγων που στηρίζονται στις συνδεδεμένες λίστες, τα **αντεστραμμένα αρχεία** (inverted files) και οι **πολλαπλές λίστες** (multilists). Οι δομές αυτές μπορούν να χρησιμοποιηθούν για την εξυπηρέτηση λογικών ερωτήσεων που

σχηματίζονται από διαζεύξεις ή/και συζεύξεις απλών ερωτήσεων. Μία άλλη ενδιαφέρουσα οργάνωση καταλόγων είναι η δομή των **συνδυασμένων καταλόγων** (combined indexes), που αποσκοπεί στην αποτελεσματική απάντηση λογικών ερωτήσεων με βάση συζευκτικούς μόνο συνδυασμούς δευτερευόντων κλειδιών. Πιο συγκεκριμένα, σύμφωνα με τη μέθοδο αυτή δημιουργούνται κατάλογοι, όπου κάθε δείκτης προς το κύριο αρχείο αντιστοιχεί σε ζεύγη, τριάδες κλπ. υπαρκτών τιμών των δευτερευόντων κλειδιών από το αρχείο αυτό. Ωστόσο, πρέπει να τονισθεί ότι αυτές οι οργανώσεις είναι επιπρόσθετες δομές σε σχέση με το κυρίως αρχείο, που μπορεί να είναι οποιουδήποτε τύπου από όσους αναφέρθηκαν στα προηγουμένα κεφάλαια. Όλες αυτές οι δομές εξυπηρετούν αποτελεσματικά απλές ερωτήσεις με βάση την τιμή κάποιου δευτερεύοντος κλειδιού.

Μία δενδρική δομή ειδική για ανάκτηση με δευτερεύον κλειδί είναι το **πολυδιάστατο δένδρο** (k-dimensional tree) και, η γενίκευσή του, το **πολυδιάστατο B-δένδρο** (k-dimensional B-tree). Ιδιαίτερο χαρακτηριστικό των δομών αυτών είναι ότι δεν γίνεται διάκριση μεταξύ πρωτεύοντος και δευτερευόντων κλειδιών, αλλά όλα τα κλειδιά αντιμετωπίζονται ισότιμα. Επίσης συχνά στην πράξη μπορεί να τεθούν από το χρήστη ερωτήσεις διαστήματος. Τα πολυδιάστατα δένδρα και τα πολυδιάστατα B-δένδρα μπορούν να απαντήσουν τέτοιου είδους ερωτήσεις. Το παρόν κεφάλαιο κλείνει την παρουσίαση των κυριότερων δομών αρχείων για ανάκτηση με δευτερεύοντα κλειδιά με το **δικτυωτό αρχείο** (grid file), που εξυπηρετεί το σκοπό αυτό αποτελεσματικότερα από κάθε άλλη οργάνωση.

11.2 Αντεστραμμένα αρχεία

Ένα αντεστραμμένο αρχείο για κάποιο δευτερεύον κλειδί περιέχει όλες τις διακριτές τιμές που λαμβάνει το δευτερεύον κλειδί από το πεδίο ορισμού του μαζί με δείκτες προς τις εγγραφές του κύριου αρχείου, οι οποίες περιέχουν αυτήν την τιμή του δευτερεύοντος κλειδιού. Τότε το αρχείο θεωρείται ότι είναι αντεστραμμένο ως προς το συγκεκριμένο κλειδί. **Πλήρως αντεστραμμένο** (fully inverted) ονομάζεται ένα αρχείο αν διατηρεί ένα δευτερεύοντα κατάλογο για κάθε δευτερεύον κλειδί, ενώ στην αντίθετη περίπτωση ονομάζεται **μερικά αντεστραμμένο** (partially inverted). Το Σχήμα 11.1 παριστά τη δομή ενός κύριου αρχείου με τρία αντεστραμμένα αρχεία με σκοπό την ανάκτηση με βάση τα αντίστοιχα 3 δευτερεύοντα κλειδιά.



Σχήμα 11.1: Κύριο αρχείο με τρία αντεστραμμένα αρχεία.

Αν η εγγραφή δεν πρόκειται να μετακινηθεί φυσικά, τότε ως δείκτης χρησιμοποιείται η διεύθυνση της εγγραφής στο δίσκο. Ωστόσο, είναι δυνατόν οι θέσεις των εγγραφών να αλλάζουν εξαιτίας εισαγωγών και διαγραφών. Σε μία τέτοια περίπτωση, οι διακριτές τιμές του δευτερεύοντος κλειδιού πρέπει να συνοδεύονται από τις αντίστοιχες τιμές του πρωτεύοντος κλειδιού, ώστε να συνεχισθεί η αναζήτηση προς τις κατάλληλες εγγραφές. Η μέθοδος αυτή μειονεκτεί επειδή είναι πιο αργή, όμως ταυτόχρονα είναι πιο αξιόπιστη.

Κράτος	Πόλη
Βοσνία-Ερζεγοβίνη	Σαράγιεβο
Βουλγαρία	Σόφια, Φιλιπούπολη
Γιουγκοσλαβία	Βελιγράδι
Ελλάδα	Αθήνα, Θεσσαλονίκη
Κροατία	Ζάγκρεμπ
ΠΓΔΜ	Σκόπια
Ρουμανία	Βουκουρέστι, Κωνστάντζα

Λιμάνι	Πόλη
Ναι	Αθήνα, Θεσσαλονίκη, Κωνστάντζα
`Οχι	Βελιγράδι, Βουκουρέστι, Σαράγιεβο, Σκόπια, Σόφια, Φιλιπούπολη, Ζάγκρεμπ

Σχήμα 11.2: Αντεστραμμένα αρχεία.

Προφανώς, οι λίστες των δεικτών που προσαρτώνται στις διάφορες διακριτές τιμές των δευτερευόντων κλειδιών δεν είναι ίδιου μήκους, αλλά μεταβλητού. Για παράδειγμα, στο Σχήμα 11.2 δίνονται δύο κατάλογοι ενός αρχείου που αφορά σε πόλεις των βαλκανικών κρατών με πληθυσμό περισσότερο από 340.000 κατοίκους. Ο ένας κατάλογος δίνει τις πόλεις για κάθε διακριτή τιμή του δευτερεύοντος πεδίου `Κράτος΄, ενώ ο άλλος κατάλογος

επιμερίζει τις πόλεις σε λιμάνια και μη λιμάνια. Έτσι δημιουργείται πρόβλημα από την ύπαρξη εγγραφών μεταβλητού μήκους στο αντεστραμμένο αρχείο. Στην πράξη ο κατάλογος δεν περιέχει εγγραφές μεταβλητού μήκους, αλλά για όλες τις λίστες δεσμεύεται σταθερός χώρος με μέγεθος που μπορεί να ποικίλει και αποτελεί μία παράμετρο κατά το σχεδιασμό των αρχείων. Αν ο δεσμευμένος χώρος εξαντληθεί, τότε οι υπερχειλίζοντες δείκτες αποθηκεύονται σε νέο σταθερό χώρο που αποδίδεται από το σύστημα. Κατόπιν, η αλληλουχία των σταθερών χώρων για κάθε διακριτή τιμή του δευτερεύοντος κλειδιού συνδέεται, ώστε να σχηματισθεί ένα σειριακό αρχείο ή ένα σειριακό αρχείο με δείκτη. Η μέθοδος αυτή ονομάζεται **κυτταρικό αντεστραμμένο αρχείο** (cellular inverted file).

Οι κατάλογοι δημιουργούνται κατά τη φόρτωση του αρχείου. Με την εισαγωγή κάθε εγγραφής ενημερώνεται και ο αντίστοιχος κατάλογος για κάθε δευτερεύον πεδίο. Δηλαδή, με την εισαγωγή μίας εγγραφής ελέγχεται αν η τιμή ενός συγκεκριμένου πεδίου υπάρχει στον αντίστοιχο κατάλογο. Αν όχι, τότε δημιουργείται η σχετική λίστα με ένα μόνο δείκτη, αλλιώς η υπάρχουσα λίστα ενημερώνεται και επιμηκύνεται κατά άλλον ένα δείκτη. Για τη διαγραφή μίας εγγραφής πρέπει να προσπελασθούν όλοι οι σχετικοί κατάλογοι και να διαγραφούν οι δείκτες (ή οι διευθύνσεις). Αν η λίστα των δεικτών αποτελείται από ένα μόνο δείκτη, τότε διαγράφεται ολόκληρη η λίστα. Ανάλογη είναι και η διαδικασία σε περίπτωση αλλαγής της τιμής κάποιου δευτερεύοντος πεδίου.

Η αναζήτηση με βάση την τιμή ενός δευτερεύοντος κλειδιού είναι πλέον εύκολη. Προσπελάζεται ο σχετικός κατάλογος και επιλέγονται οι αντίστοιχοι δείκτες. Αν η υποβαλλόμενη ερώτηση είναι μία λογική σύζευξη (με and, or και not) πολλών απλών ερωτήσεων με βάση δευτερεύοντα κλειδιά, τότε πρέπει να προσπελασθούν όλοι οι αντίστοιχοι κατάλογοι και να ληφθεί η τομή των σχετικών λιστών πριν γίνει η προσπέλαση στο κύριο αρχείο.

Η μέθοδος των αντεστραμμένων καταλόγων χρησιμοποιείται από όλα σχεδόν τα εμπορικά πακέτα Ανάκτησης Πληροφοριών (Information Retrieval), που κυρίως αφορούν βιβλιοθηκονομικά συστήματα. Το γνωστότερο εμπορικό σύστημα είναι της IBM, το λεγόμενο STAIRS (STorage And Information Retrieval System), και η επεξεργασία του στηρίζεται στην άλγεβρα Boole. Άλλα εμπορικά συστήματα είναι τα BRS, DIALOG, MEDLARS και ORBIT.

11.3 Πολλαπλές λίστες

Σε ένα δευτερεύοντα κατάλογο κάθε διακριτή τιμή του δευτερεύοντος κλειδιού συνοδεύεται από μία σειρά δεικτών, επειδή η ίδια τιμή του κλειδιού μπορεί να εμφανισθεί σε πολλές εγγραφές. Είναι δυνατό να μειωθεί το μέγεθος του δευτερεύοντος καταλόγου, αν δεν αποθηκεύονται όλοι οι δείκτες προς τις αντίστοιχες εγγραφές, αλλά μόνο ένας προς την πρώτη εγγραφή. Βέβαια, για να μην υπάρξει απώλεια δεδομένων, πρέπει όλες οι εγγραφές του κύριου αρχείου με ίδιες τιμές στο δευτερεύον κλειδί να είναι συνδεδεμένες σε μία λίστα. Η συνδεδεμένη λίστα μπορεί να είναι **απλή, διπλή ή κυκλική** (simple, double, circular linked list). Το ίδιο μπορεί να γίνει και για τα υπόλοιπα δευτερεύοντα κλειδιά με αντίστοιχες λίστες. Όταν ένα αρχείο συνοδεύεται από τέτοιες οργανώσεις λέγεται αρχείο πολλαπλών λιστών.

Πόλη	Κράτος	Δείκτης	Λιμάνι	Δείκτης
Αθήνα	Ελλάδα	Θεσσαλονίκη	Ναι	Θεσσαλονίκη
Βελιγράδι	Γιουγκοσλαβία	–	Όχι	Βουκουρέστι
Βουκουρέστι	Ρουμανία	Κωνστάντζα	Όχι	Ζάγκρεμπ
Ζάγκρεμπ	Κροατία	–	Όχι	Σαράγιεβο
Θεσσαλονίκη	Ελλάδα	–	Ναι	Κωνστάντζα
Κωνστάντζα	Ρουμανία	–	Ναι	–
Σαράγιεβο	Βοσνία-Ερζεγοβίνη	–	Όχι	Σκόπια
Σκόπια	ΠΓΔΜ	–	Όχι	Σόφια
Σόφια	Βουλγαρία	Φιλιπούπολη	Όχι	Φιλιπούπολη
Φιλιπούπολη	Βουλγαρία	–	Όχι	–

Λιμάνι	Πόλη	Μήκος
Ναι	Αθήνα	3
Όχι	Βελιγράδι	7

Κράτος	Πόλη	Μήκος
Βοσνία-Ερζεγοβίνη	Σαράγιεβο	1
Βουλγαρία	Σόφια	2
Γιουγκοσλαβία	Βελιγράδι	1
Ελλάδα	Αθήνα	2
Κροατία	Ζάγκρεμπ	1
ΠΓΔΜ	Σκόπια	1
Ρουμανία	Βουκουρέστι	2

Σχήμα 11.3: Αρχείο πολλαπλών λιστών με απλή λίστα.

Τα δεδομένα του Σχήματος 11.2 παρουσιάζονται στο Σχήμα 11.3 με υλοποίηση απλής συνδεδεμένης λίστας. Το κύριο αρχείο έχει ως πρωτεύον κλειδί το όνομα της πόλης και ως δευτερεύοντα κλειδιά τα χαρακτηριστικά

`Κράτος΄ και `Λιμάνι΄. Το αρχείο συνοδεύεται από δύο δευτερεύοντες καταλόγους, ένα για κάθε δευτερεύον κλειδί. Σε κάθε κατάλογο οι είσοδοι μίας διακριτής τιμής του δευτερεύοντος κλειδιού έχουν άλλα δύο πεδία: το δείκτη (πρωτεύον κλειδί) της πρώτης εγγραφής του κυρίως αρχείου και το μήκος της αλυσίδας. Επίσης, το κύριο αρχείο έχει δύο επιπλέον πεδία, που είναι οι σύνδεσμοι των λιστών.

Πόλη	Κράτος	Δ.εμπρός	Δ.πίσω	Λιμάνι	Δ.εμπρός	Δ.πίσω
Αθήνα	Ελλάδα	Θεσσαλ.	–	Ναι	Θεσσαλ.	–
Βελιγράδι	Γιουγκοσλ.	–	–	`Οχι	Βουκ.	–
Βουκουρέστι	Ρουμανία	Κωνστ.	–	`Οχι	Ζάγκρ.	Βελιγρ.
Ζάγκρεμπ	Κροατία	–	–	`Οχι	Σαράγ.	Βουκ.
Θεσσαλονίκη	Ελλάδα	–	Αθήνα	Ναι	Κωνστ.	Αθήνα
Κωνστάντζα	Ρουμανία	–	Βουκ.	Ναι	–	Θεσσαλ.
Σαράγιεβο	Βοσνία	–	–	`Οχι	Σκόπια	Ζάγκρ.
Σκόπια	ΠΓΔΜ	–	–	`Οχι	Σόφια	Σαράγ.
Σόφια	Βουλγαρία	Φιλιπ.	–	`Οχι	Φιλιπ.	Σκόπια
Φιλιπούπολη	Βουλγαρία	–	Σόφια	`Οχι	–	Σόφια

Σχήμα 11.4: Αρχείο πολλαπλών λιστών (διπλή λίστα).

Το Σχήμα 11.4 είναι ένα παράδειγμα υλοποίησης με διπλά συνδεδεμένη λίστα, όπου για κάθε δευτερεύον πεδίο υπάρχουν δύο δείκτες, ένας προς τα εμπρός και ένας προς τα πίσω. Η υλοποίηση αυτή πλεονεκτεί σε σχέση με την προηγούμενη, γιατί οι εισαγωγές και οι διαγραφές εγγραφών γίνονται αποτελεσματικότερα, όμως απαιτείται επιπλέον χώρος. Σε ένα αρχείο πολλαπλών λιστών υπάρχουν πολλοί τρόποι για να απαντηθεί μία ερώτηση γιατί προσφέρεται η δυνατότητα πλεύσης (navigation) μέσα στο αρχείο. Για παράδειγμα, τίθεται η ερώτηση: `Ποιά είναι τα λιμάνια της Βουλγαρίας;΄. Η ερώτηση αυτή μπορεί να απαντηθεί κατά τρεις τρόπους (ποιούς;).

`Ενα αρχείο πολλαπλών λιστών έχει μία σειρά πλεονεκτημάτων σε σχέση με τη μέθοδο των αντεστραμμένων αρχείων:

- εύκολος προγραμματισμός του λογισμικού σε υλοποιήσεις με απλά συνδεδεμένες λίστες,
- καλή επίδοση κατά την αναζήτηση για μικρές λίστες και για μικρό αριθμό δευτερεύοντων κλειδιών,
- πολύ καλή επίδοση κατά την ανανέωση, ιδιαίτερα για διπλά συνδεδεμένες λίστες,

- δεν απαιτείται χώρος στην κύρια μνήμη για την επεξεργασία των λογικών ερωτήσεων,
- οι κατάλογοι απαιτούν λιγότερο χώρο και μπορεί κατά την επεξεργασία να αποθηκευθούν στην κύρια μνήμη. Αυτό έχει ως συνέπεια την ταχύτατη επεξεργασία τους, γιατί δεν απαιτείται μεταφορά δεδομένων από/προς το δίσκο.

Όμως η οργάνωση των πολλαπλών λιστών σε σχέση με τη μέθοδο των αντεστραμμένων αρχείων έχει και μία σειρά μειονεκτημάτων, όπως:

- σχετικά πολύπλοκος προγραμματισμός του λογισμικού εισαγωγών και διαγραφών σε υλοποιήσεις με διπλά συνδεδεμένες λίστες,
- μέτρια επίδοση κατά την αναζήτηση για μεγάλες λίστες,
- απαιτείται περισσότερος χώρος μέσα στο κύριο αρχείο για την αποθήκευση των δεικτών. Σε υλοποιήσεις με διπλά συνδεδεμένες λίστες για πολλά δευτερεύοντα κλειδιά ο επιπρόσθετος χώρος μπορεί να υπερβαίνει το μέγεθος του χώρου, που εξοικονομείται στον κατάλογο από τη μη αποθήκευση όλων των δεικτών.

Έχει προταθεί μία παραλλαγή του αρχείου πολλαπλών λιστών που υπερβαίνει το πρόβλημα της μέτριας επίδοσης κατά την αναζήτηση σε μεγάλες λίστες. Η οργάνωση αυτή ονομάζεται **κυτταρικό αρχείο πολλαπλών λιστών** (cellular multilist). Το μήκος κάθε λίστας περιορίζεται έτσι ώστε οι αντίστοιχες εγγραφές να χωρούν σε ένα `κύτταρο', που προσδιορίζεται από το υλικό, όπως για παράδειγμα σε μία σελίδα ή μία άτρακτο του δίσκου. Έτσι κατά την ανάκτηση των εγγραφών μίας λίστας μειώνεται ο χρόνος αναζήτησης λόγω της παλινδρομικής κίνησης του βραχίονα του δίσκου. Το τίμημα για την καλύτερη απόκριση κατά την αναζήτηση είναι ο καταλαμβανόμενος χώρος. Κάθε είσοδος του αντίστοιχου καταλόγου δεν περιέχει μόνο ένα ζεύγος (πρώτη διεύθυνση, μήκος λίστας), αλλά τόσες τριάδες όσες είναι τα κύτταρα από όπου περνά μία λίστα. Επομένως απαιτείται επιπρόσθετος χώρος στον κατάλογο σε σχέση με μία απλή υλοποίηση πολλαπλών λιστών. Σύμφωνα με μία υλοποίηση δεν επιτρέπεται από κάθε κύτταρο να περνούν πάρα πολλές λίστες, αλλά ο αριθμός των λιστών ανά κύτταρο είναι μία παράμετρος που ορίζεται κατά το σχεδιασμό των αρχείων. Έτσι επιτυγχάνεται καλύτερη συγκέντρωση και τοπικότητα των δεδομένων, αλλά απαιτείται επιπλέον χώρος στο κύριο αρχείο, γιατί διατηρείται κενός χώρος

σε κάθε κύτταρο για μελλοντική εισαγωγή εγγραφών από τις ήδη υπάρχουσες λίστες στο κύτταρο αυτό. Στο Σχήμα 11.5 παρουσιάζεται ο δευτερεύων κατάλογος για τα δεδομένα του Σχήματος 11.2 όταν είναι οργανωμένα με τη μέθοδο αυτή. Υποτίθεται ότι το κύριο αρχείο αποτελείται από τρεις σελίδες με χωρητικότητα τεσσάρων εγγραφών.

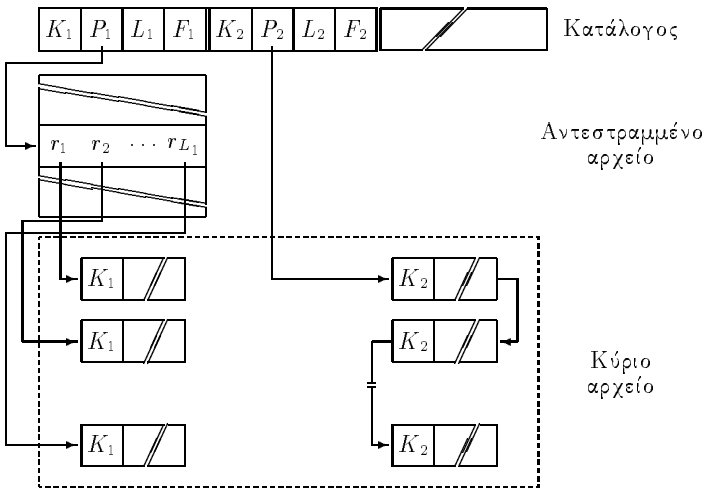
Πόλη	Κράτος	Δείκτης	Λιμάνι	Δείκτης
Αθήνα	Ελλάδα	Θεσσαλονίκη	Ναι	Θεσσαλονίκη
Βελιγράδι	Γιουγκοσλαβία	–	Όχι	Βουκουρέστι
Βουκουρέστι	Ρουμανία	Κωνστάντζα	Όχι	Ζάγκρεμπ
Ζάγκρεμπ	Κροατία	–	Όχι	Σαράγιεβο
Θεσσαλονίκη	Ελλάδα	–	Ναι	Κωνστάντζα
Κωνστάντζα	Ρουμανία	–	Ναι	–
Σαράγιεβο	Βοσνία-Ερζεγοβίνη	–	Όχι	Σκόπια
Σκόπια	ΠΓΔΜ	–	Όχι	Σόφια
Σόφια	Βουλγαρία	Φιλιπούπολη	Όχι	Φιλιπούπολη
Φιλιπούπολη	Βουλγαρία	–	Όχι	–

Λιμάνι	Πόλη	Κύτταρο	Μήκος
Ναι	Αθήνα	1	1
	Θεσσαλονίκη	2	2
Όχι	Βελιγράδι	1	3
	Σαράγιεβο	2	2
	Σόφια	3	2

Σχήμα 11.5: Κυτταρικό αρχείο πολλαπλών λιστών.

Μία άλλη οργάνωση που ανήκει σε αυτήν την κατηγορία είναι το **υβριδικό αρχείο λιστών** (hybrid list file), που προτάθηκε από τον Yang (1978) και αποτελεί μία λύση που συνδυάζει τα πλεονεκτήματα των δομών των αντεστραμμένων αρχείων και των πολλαπλών λιστών. Αν ο αριθμός των εγγραφών που έχουν μία συγκεκριμένη τιμή σε ένα δευτερεύον κλειδί είναι μεγαλύτερος από μία παράμετρο L, που ορίζεται κατά το σχεδιασμό του αρχείου, τότε δημιουργείται μία αντίστοιχη είσοδος σε ένα αντεστραμμένο αρχείο. Στην αντίθετη περίπτωση η σύνδεση των εγγραφών αυτών γίνεται με τη μέθοδο των πολλαπλών λιστών. Αν L=0 τότε η υβριδική αυτή οργάνωση μετατρέπεται σε καθαρή οργάνωση αντεστραμμένων αρχείων, ενώ για πολύ μεγάλες τιμές του L μετατρέπεται σε καθαρή οργάνωση πολλαπλών λιστών. Η επίδοση της μεθόδου αυτής κατά την αναζήτηση υπερτερεί των

επιδόσεων των δύο απλούστερων μεθόδων με αντίστοιχα αυξημένη πολυπλοκότητα λογισμικού.



Σχήμα 11.6: Υβριδικό αρχείο λιστών.

Στο Σχήμα 11.6 παρουσιάζεται ένα υβριδικό αρχείο λιστών. Η εγγραφή του καταλόγου αποτελείται από τέσσερα πεδία: την τιμή του δευτερεύοντος κλειδιού (K_i), το δείκτη προς το αντεστραμμένο αρχείο ή το κύριο αρχείο (P_i), το σύνολο των εγγραφών που έχουν τη συγκεκριμένη τιμή (L_i), και μία σημαία (F_i) που δηλώνει ποιά μέθοδος χρησιμοποιείται για κάθε τιμή του δευτερεύοντος κλειδιού.

Οργανώσεις πολλαπλών λιστών χρησιμοποιούνται στο πακέτο IMS (Information Management System) της IBM που είναι μία ιεραρχική (hierarchical) βάση δεδομένων, καθώς και παλαιότερα στο πακέτο IDMS της Cullinet που είναι μία δικτυωτή (network ή Codasyl) βάση δεδομένων. Οι σχεσιακές (relational) βάσεις δεδομένων δεν χρησιμοποιούν δομές με δείκτες. Σε αντίθεση οι αντικειμενοστραφείς (object-oriented) βάσεις δεδομένων χρησιμοποιούν δομές με δείκτες.

11.4 Συνδυασμένοι κατάλογοι

Οι συνδυασμένοι κατάλογοι είναι μία ομάδα δομών που εξυπηρετούν αποτελεσματικά **ερωτήσεις μερικής ταύτισης** (partial match queries). Ο

τύπος αυτός ερωτήσεων αντιδιαστέλλεται προς τον τύπο των **ερωτήσεων επακριβούς ταύτισης** (exact match queries). Με τον όρο `μερική ταύτιση΄ εννοούνται οι λογικές ερωτήσεις που αποτελούνται από τη σύζευξη (με and) μερικών χαρακτηριστικών μίας εγγραφής, ενώ με τον όρο `επακριβής ταύτιση΄ προσδιορίζονται τιμές για όλα τα χαρακτηριστικά της εγγραφής. Για παράδειγμα, έστω ότι τίθεται μία ερώτηση μερικής ταύτισης, που προσδιορίζει την τιμή a χαρακτηριστικών και ότι σε κάθε χαρακτηριστικό αντιστοιχεί ένας κατάλογος με δομή B^+-δένδρου. Το κόστος ανάκτησης r εγγραφών που ικανοποιούν την ερώτηση είναι:

$$a \times 2 \times (s + r + btt) \ + \ c \ + \ r \times (s + r + btt)$$

όπου ο πρώτος όρος δίνει το κόστος προσπέλασης των δύο τελευταίων επιπέδων των B^+-δένδρων, ο τρίτος όρος δίνει το κόστος προσπέλασης του κύριου αρχείου για την ανάκτηση των εγγραφών, ενώ c είναι το συνήθως αμελητέο κόστος συγχώνευσης στην κύρια μνήμη των λιστών των δεικτών προς το κύριο αρχείο.

Ο χρόνος απόκρισης για την ικανοποίηση μίας τέτοιας ερώτησης θα ήταν μικρότερος αν για κάθε πιθανό συζευκτικό συνδυασμό των a χαρακτηριστικών σε μία ερώτηση μερικής ταύτισης υπήρχε ένας αντίστοιχος κατάλογος. Στον κατάλογο αυτό για κάθε υπαρκτό συνδυασμό των τιμών των υπ΄ όψη χαρακτηριστικών θα έπρεπε να δίνεται μία λίστα από δείκτες προς τις σχετικές εγγραφές του κύριου αρχείου. `Ετσι ο πρώτος όρος του κόστους θα ελαττώνονταν σημαντικά.

Το τίμημα στην περίπτωση αυτή θα ήταν διπλό:

- αντί ενός κλειδιού θα πρέπει να αποθηκεύονται περισσότερα. `Ετσι κάθε κατάλογος θα ήταν ογκωδέστερος και επομένως σε σχέση με έναν απλό κατάλογο θα μειώνονταν ο λόγος διακλάδωσης (y), και το σχετικό κόστος θα ήταν μικρότερο κατά λιγότερο από a φορές.
- ο αριθμός των απαιτούμενων καταλόγων θα ήταν υπερβολικός.

Ας υποτεθεί ότι μία εγγραφή έχει τρία χαρακτηριστικά που ονομάζονται Α, Β και Γ. Εύκολα προκύπτει ότι οι δυνατές ερωτήσεις που μπορούν να απαντηθούν είναι πάρα πολλές, για παράδειγμα μπορεί να είναι ερωτήσεις με βάση μόνο ένα χαρακτηριστικό (Α, Β, ή Γ), ερωτήσεις με βάση δύο χαρακτηριστικά (ΑΒ, ΒΑ, ΑΓ, ΓΑ, ΒΓ, ΓΒ), και ερωτήσεις με βάση και τα τρία χαρακτηριστικά (ΑΒΓ, ΑΓΒ, ΒΑΓ, ΒΓΑ, ΓΑΒ, ΓΒΑ). Δηλαδή, συνολικά

μπορεί να τεθούν 15 διαφορετικές ερωτήσεις μερικής ταύτισης. Εύκολα αποδεικνύεται ότι ο συνολικός αριθμός δυνατών ερωτήσεων που μπορεί να προκύψουν είναι:

$$\sum_{i=1}^{a} i! \begin{pmatrix} a \\ i \end{pmatrix}$$

Αν για κάθε πιθανή ερώτηση υπήρχε και αντίστοιχος κατάλογος, τότε ευνόητο είναι ότι, ακόμη και για μικρές τιμές του α, το κόστος διατήρησης και επεξεργασίας των καταλόγων θα ήταν απαγορευτικό.

Ο Lum (1970) παρατήρησε ότι δεν έχει σημασία η διάταξη των χαρακτηριστικών στους συνδυασμένους καταλόγους και επομένως ο αριθμός των συνδυασμένων καταλόγων μπορεί να ελαττωθεί δραστικά. Πιο συγκεκριμένα απαιτούνται μόνο κατάλογοι με εγγραφές που προκύπτουν από το συνδυασμό όλων των χαρακτηριστικών. Έτσι αν a=3 τότε αρκούν τρεις κατάλογοι, που ονομάζονται ΑΒΓ, ΒΓΑ, και ΓΑΒ. Όσες ερωτήσεις αφορούν στο χαρακτηριστικό Α απαντώνται από τον πρώτο κατάλογο (ΑΒΓ), όσες αφορούν στο Β από το δεύτερο κατάλογο (ΒΓΑ), όσες αφορούν το Γ από τον τρίτο κατάλογο (ΓΒΑ), όσες αφορούν στα Α και Β ή Β και Α από τον πρώτο, όσες αφορούν στα Β και Γ ή Γ και Β από το δεύτερο, ερωτήσεις που αφορούν στα Α και Γ ή Γ και Α από τον τρίτο, ενώ οι ερωτήσεις που αφορούν και στα τρία χαρακτηριστικά μπορούν να απαντηθούν από οποιοδήποτε κατάλογο. Ο συνολικός αριθμός συνδυασμένων καταλόγων για a χαρακτηριστικά είναι:

$$\begin{pmatrix} a \\ \lfloor a/2 \rfloor \end{pmatrix} = \frac{a!}{\lceil a/2 \rceil ! \lfloor a/2 \rfloor !}$$

Έτσι αν a=4 τότε αρκούν έξι κατάλογοι (οι ΑΒΓΔ, ΒΓΔΑ, ΒΔΑΓ, ΓΑΔΒ, ΓΔΑΒ και ΔΑΒΓ), αν και ο αριθμός των πιθανών ερωτήσεων μερικής ταύτισης που μπορεί να τεθούν είναι εξήντα ένα.

Από τον Shneiderman (1977) προτάθηκε η μέθοδος των **μειωμένων συνδυασμένων καταλόγων** (reduced combined indices), που στηρίζεται στην παρατήρηση ότι εκτός από έναν κατάλογο όλοι οι υπόλοιποι κατάλογοι που προκύπτουν σύμφωνα με τη μέθοδο του Lum μπορούν να απλοποιηθούν. Δηλαδή αν a=3, τότε ο αριθμός των απαραίτητων καταλόγων είναι και πάλι τρεις, αλλά απλουστεύοντας τους προκύπτουν οι κατάλογοι ΑΒΓ, ΒΓ και ΓΑ. Αν a=4 τότε ο αριθμός των απαραίτητων καταλόγων είναι και πάλι έξι, αλλά απλουστεύοντας τους προκύπτουν οι κατάλογοι ΑΒΓΔ, ΒΔΑ, ΓΑΔ,

ΔΓΒ, ΑΔ και ΒΓ. Για παράδειγμα, χρησιμοποιώντας τον κατάλογο ΑΒΓΔ απαντώνται οι ερωτήσεις μερικής ταύτισης που αφορούν στους εξής τέσσερις συνδυασμούς χαρακτηριστικών: ΑΒΓΔ, ΑΒΓ, ΑΒ και Α. Κατά τον ίδιο τρόπο και οι υπόλοιποι κατάλογοι εξυπηρετούν περισσότερες από μία ερωτήσεις μερικής ταύτισης, ενώ διασφαλίζεται ότι κάθε πιθανή ερώτηση θα ικανοποιηθεί από αυτήν την ομάδα καταλόγων. Είναι προφανές ότι η μέθοδος των μειωμένων συνδυασμένων καταλόγων υπερτερεί σε σχέση με την αρχική μέθοδο τόσο από πλευράς χώρου όσο και πλευράς χρόνου. Πιο συγκεκριμένα προκύπτει ότι ο αριθμός των καταλόγων που δεικτοδοτούν i χαρακτηριστικά είναι:

$$\binom{a}{i} - \binom{a}{i+1} \qquad \text{για } \lceil a/2 \rceil \leq i < a$$

ενώ απαιτείται και ένας κατάλογος που δεικτοδοτεί όλα τα χαρακτηριστικά.

Από τον Shneiderman επίσης προτάθηκε η εναλλακτική μέθοδος των **τροποποιημένων συνδυασμένων καταλόγων** (modified combined indices), που σε σχέση με τις προηγούμενες μεθόδους διατηρεί περισσότερους καταλόγους, αλλά μικροτέρου μεγέθους. Ο συνολικός αριθμός τροποποιημένων συνδυασμένων καταλόγων για a χαρακτηριστικά είναι:

$$2^{a-1}$$

Αν a=3 τότε ο αριθμός των απαραίτητων καταλόγων είναι τέσσερις (ΑΒΓ, ΑΓ, ΒΓ και Γ), ενώ αν a=4 τότε ο αριθμός των απαραίτητων καταλόγων είναι οκτώ (ΑΒΓΔ, ΑΒΔ, ΑΓΔ, ΑΔ, ΒΓΔ, ΒΔ, ΓΔ και Δ). Πιο συγκεκριμένα προκύπτει ότι ο αριθμός των καταλόγων που δεικτοδοτούν i χαρακτηριστικά είναι:

$$\binom{a-1}{i-1}$$

Επομένως όταν ο χρόνος είναι πιο κρίσιμος παράγοντας από τον πρώτο η μέθοδος των τροποποιημένων καταλόγων πρέπει να προτιμηθεί σε σχέση με τη μέθοδο των μειωμένων, ενώ αν ο χώρος είναι πιο κρίσιμος παράγοντας η προτίμηση θα πρέπει να αντιστραφεί. Βέβαια, εκτός από το χρόνο ανάκτησης, πρέπει να θεωρήσουμε και το χρόνο εισαγωγής και διαγραφής. Ως προς τις παραμέτρους αυτές, η μέθοδος των τροποποιημένων καταλόγων υστερεί επειδή για κάθε εισαγωγή ή διαγραφή απαιτείται η ενημέρωση περισσοτέρων καταλόγων.

Παρά το σημαντικό περιορισμό του αριθμού των καταλόγων όλες οι προηγούμενες παραλλαγές είναι απαγορευτικές για λίγο μεγαλύτερο αριθμό χαρακτηριστικών. Από τον Stonebraker (1974), το θεμελιωτή των Συστημάτων Διαχείρισης Βάσεων Δεδομένων Ingres, Postgres και Illustra μελετήθηκαν τρόποι εύρεσης του βέλτιστου αριθμού συνδυασμένων καταλόγων και του βέλτιστου τρόπου ομαδοποίησης των χαρακτηριστικών σε κάθε ένα συνδυασμένο κατάλογο. Ας θεωρηθεί μία απλή περίπτωση: για παράδειγμα, αν a=9 τότε προκύπτουν 126 συνδυασμένοι κατάλογοι. Αν μερικά χαρακτηριστικά δεν συνδυάζονται ποτέ μεταξύ τους σε μία ερώτηση μερικής ταύτισης, τότε τα πράγματα διευκολύνονται σημαντικά. Ας υποτεθεί ότι a=9 και ότι τα χαρακτηριστικά αυτά υποδιαιρούνται σε τρεις διακριτές τριάδες, ενώ στοιχεία από διαφορετικές τριάδες δεν συνδυάζονται μεταξύ τους. Στην περίπτωση αυτή προκύπτουν εννέα κατάλογοι, δηλαδή αριθμός που είναι ίσος με τον αριθμό των χαρακτηριστικών και συνεπώς ίσος με τον αριθμό των απλών καταλόγων. Όμως, σύμφωνα με τη μέθοδο αυτή, οι κατάλογοι έχουν μεγαλύτερο ειδικό βάρος.

11.5 Πολυδιάστατα δένδρα

Το πολυδιάστατο δένδρο (που απλούστερα λέγεται και k-d δένδρο) είναι μία πολύ αποτελεσματική δομή για ανάκτηση με δευτερεύον κλειδί, καθώς και για ερωτήσεις μερικής ταύτισης ή ερωτήσεις διαστήματος. Η δομή αυτή δεν κάνει διάκριση μεταξύ των k κλειδιών σε πρωτεύον και δευτερεύοντα, αλλά τα αντιμετωπίζει όλα ισότιμα. Έτσι η δομή παρουσιάζει ιδιαίτερα πλεονεκτήματα.

Τα πολυδιάστατα δένδρα προτάθηκαν από τον Bentley (1975). Έκτοτε έχουν μελετηθεί από πολλούς ερευνητές και έχουν προκύψει πολλές βελτιωμένες παραλλαγές. Στην αρχική εκδοχή η δομή χρησιμοποιούνταν για την αποθήκευση των δεδομένων στην κύρια μνήμη αλλά την ίδια στιγμή εξυπηρετεί και ως κατάλογος (όπως άλλωστε συμβαίνει και στα B-δένδρα, που εξετάσθηκαν στο βιβλίο των Δομών Δεδομένων).

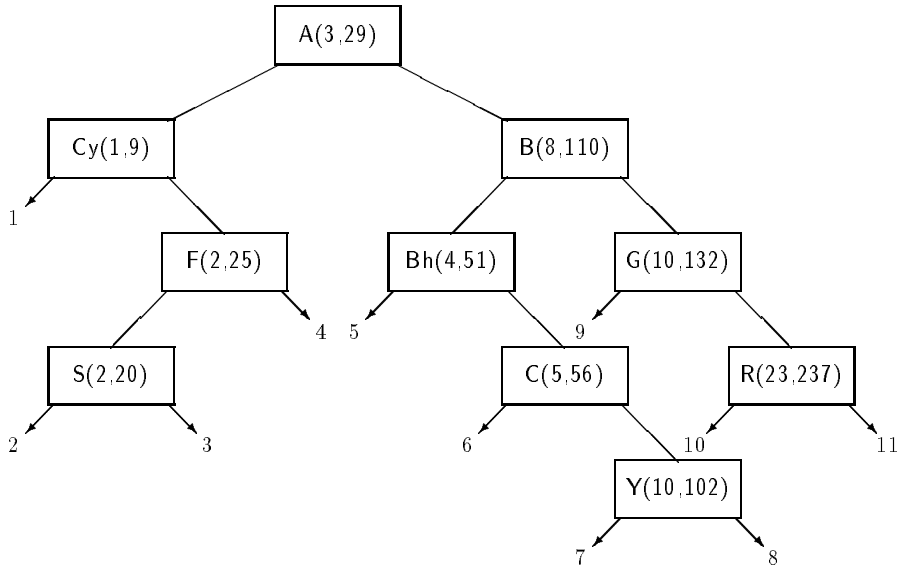
Τα απλά δυαδικά δένδρα αναζήτησης μπορούν να θεωρηθούν ως μονοδιάστατα (1-d) δένδρα. Η ομοιότητα των δύο τύπων δένδρων έγκειται στο γεγονός ότι ο βαθμός τους είναι δύο, οπότε σε κάθε κόμβο αποθηκεύονται δύο μόνο δείκτες. Ωστόσο, η βασική διαφορά των k-d δένδρων από τα απλά δυαδικά δένδρα αναζήτησης είναι ότι σε κάθε επίπεδο του δένδρου χρησιμοποιείται ένα διαφορετικό χαρακτηριστικό της εγγραφής, ώστε οι εγγραφές

να υποδιαιρεθούν σε δύο ανεξάρτητα μεταξύ τους υποσύνολα σύμφωνα με τη στρατηγική `διαίρει και βασίλευε´. Για παράδειγμα, στο επίπεδο της ρίζας οι εγγραφές υποδιαιρούνται σε δύο υποσύνολα με βάση την τιμή του πρώτου χαρακτηριστικού, οπότε οι εγγραφές με τις μικρότερες και τις μεγαλύτερες τιμές στο πρώτο χαρακτηριστικό κατευθύνονται στο αριστερό και στο δεξιό υποδένδρο, αντίστοιχα. Δηλαδή, στο δεύτερο, τρίτο κοκ. επίπεδο η διάκριση των εγγραφών γίνεται με βάση τις τιμές του δεύτερου, του τρίτου κοκ. χαρακτηριστικού, αντίστοιχα. Στο $(k+1)$-οστό επίπεδο η διάκριση γίνεται και πάλι με βάση το πρώτο χαρακτηριστικό. `Ετσι, η διαδικασία προχωρεί κατά κυκλικό τρόπο.

`Εχει αποδειχθεί αναλυτικά ότι για ένα τυχαίο πολυδιάστατο δένδρο η πολυπλοκότητα της εισαγωγής και της τυχαίας διαγραφής είναι $O(\log n)$, ενώ η πολυπλοκότητα μίας λογικής ερώτησης που αφορά στα t από τα k χαρακτηριστικά είναι $O(n^{(k-t)/k})$. Οι επιδόσεις αυτές είναι πολύ καλές, αλλά πρακτικά συνιστάται να χρησιμοποιείται αυτή η δομή μόνο αν ο αριθμός των εγγραφών είναι μεγαλύτερος από 2^{2k}. Ο όρος αυτός τίθεται ώστε να είναι βέβαιο ότι το δένδρο θα έχει περισσότερο από $2k$ επίπεδα, οπότε κάθε χαρακτηριστικό θα χρησιμοποιείται τουλάχιστον δύο φορές για τη διάκριση των εγγραφών σε δύο υποσύνολα. Νεότερη παραλλαγή είναι τα **ισοζυγισμένα πολυδιάστατα δένδρα** (multidimensional balanced binary trees) κατά αναλογία προς τα δένδρα AVL, που εξετάσθηκαν στο βιβλίο των Δομών Δεδομένων. `Εχει αποδειχθεί από τον Vaishnavi (1989) ότι η πολυπλοκότητα για την αναζήτηση, εισαγωγή και διαγραφή μίας εγγραφής είναι της τάξης $O(\log n+k)$, ενώ σε κάθε εισαγωγή ή διαγραφή απαιτούνται περιστροφές της τάξης $O(k)$.

Ωστόσο, όλες οι προηγούμενες παραλλαγές είναι ομογενείς δομές και σχεδιάσθηκαν για αποθήκευση στην κύρια μνήμη. Βέβαια, μπορούν να αποθηκευθούν και στη δευτερεύουσα μνήμη αλλά η επίδοσή τους θα εκφυλισθεί (όπως το απλό δυαδικό δένδρο αναζήτησης). Κάτω από αυτό το πρίσμα μία ενδιαφέρουσα παραλλαγή της οργάνωσης αυτής ονομάζεται **εκτεταμένο πολυδιάστατο δένδρο** (extended k-d tree) και προτάθηκε από την Chang (1981). Η ετερογενής αυτή δομή μπορεί να θεωρηθεί ότι αποτελείται από δύο μέρη: ένα απλό k-d δένδρο που αποθηκεύεται στην κύρια μνήμη και λειτουργεί ως μηχανισμός καταλόγου για τη δεικτοδότηση των δεδομένων του κυρίου αρχείου, που προφανώς είναι αποθηκευμένο στη δευτερεύουσα μνήμη. Στη συνέχεια εξετάζεται ένα παράδειγμα οργάνωσης δεδομένων σύμφωνα με τη μέθοδο αυτή.

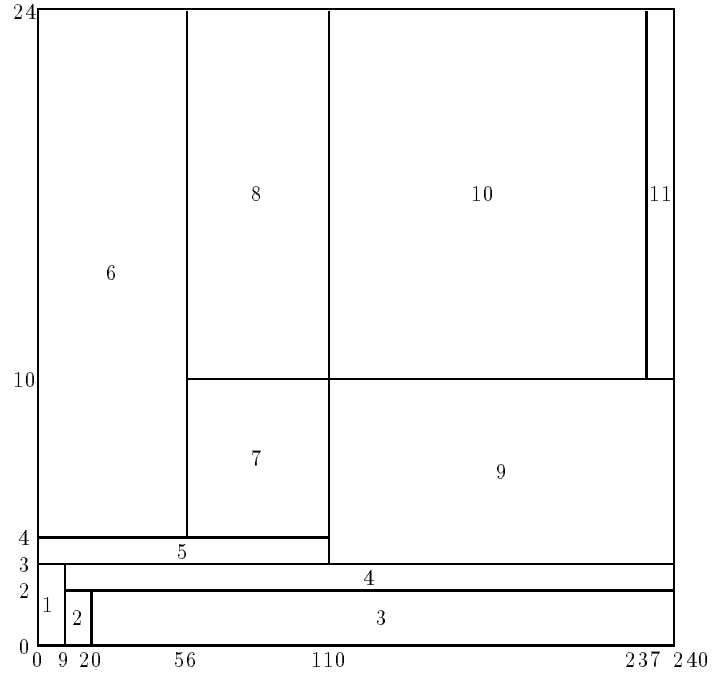
Έστωσαν τα δεδομένα του Πίνακα 7.2 που αφορούν σε δεδομένα (πληθυσμό και έκταση) των Βαλκανικών κρατών. Δηλαδή, τα δευτερεύοντα χαρακτηριστικά είναι δύο (k=2) και επομένως η αποθήκευση των δεδομένων θα γίνει σε ένα δισδιάστατο δένδρο. Ας υποτεθεί ότι τα δεδομένα εισάγονται στο δένδρο με αλφαβητική σειρά: A, B, Bh, C, Cy, F, G, R, S, Y. Η ρίζα του 2-d δένδρου θα είναι η εγγραφή A(3,29). Η επόμενη εισαγωγή της εγγραφής B(8,110) θα γίνει η ρίζα του αριστερού δένδρου γιατί 8>3, ενώ η εγγραφή Bh(4,51) θα κατευθυνθεί στα αριστερά της εγγραφής B(8,110) γιατί 51<110. Στο Σχήμα 11.7 παρουσιάζεται το δένδρο στην τελική του μορφή. Ας σημειωθεί ότι αν σε περίπτωση σύγκρισης προκύψει ισότητα, τότε τα εισαγόμενα δεδομένα κατευθύνονται στο αριστερό υποδένδρο.



Σχήμα 11.7: Δισδιάστατο δένδρο με τα δεδομένα του Πίνακα 7.2.

Στους κόμβους του δένδρου του Σχήματος 11.7 αποθηκεύονται και οι δύο τιμές των σχετικών εγγραφών. Όμως σε μία πραγματική υλοποίηση δεν είναι απαραίτητο να αποθηκεύονται και οι δύο τιμές, αλλά μόνον εκείνη που στο συγκεκριμένο επίπεδο χρησιμεύει για τη σύγκριση. Έτσι ολόκληρες οι εγγραφές αποθηκεύονται στο αρχείο, που είναι αποθηκευμένο στο δίσκο. Από τα φύλλα του δένδρου δεικτοδοτούνται 11 κάδοι του αρχείου, με διευθύνσεις που φαίνονται στο Σχήμα 11.7. Είναι ευνόητο ότι αν n είναι οι

εγγραφές του καταλόγου, τότε οι δεικτοδοτούμενοι κάδοι του αρχείου είναι $n+1$.



Σχήμα 11.8: Υποδιαίρεση χώρου σύμφωνα με τα δεδομένα του Πίνακα 7.2.

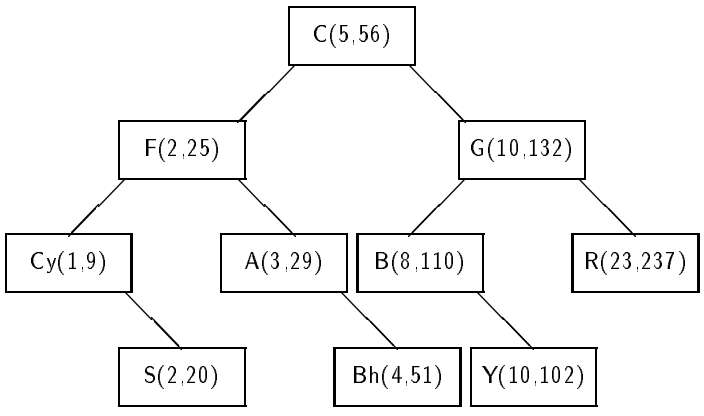
Κάθε κόμβος του 2-d δένδρου διαιρεί το χώρο σε δύο υποχώρους ανάλογα με τις συνθήκες $\leq$ και $>$. Σε κάθε κατώτερο επίπεδο γίνονται διαδοχικές υποδιαιρέσεις των υποχώρων. Στο Σχήμα 11.8 φαίνεται η διαίρεση του χώρου σε υποχώρους, ενώ σε κάθε υποχώρο παρουσιάζεται η διεύθυνση του αντίστοιχου κάδου, όπως προκύπτει από το Σχήμα 11.7. Στον οριζόντιο άξονα φαίνονται οι τιμές του χαρακτηριστικού έκταση (9,20,56,110,237) που χρησιμοποιούνται για σύγκριση στα επίπεδα άρτιας τάξης. Αντίστοιχα, στον κατακόρυφο άξονα φαίνονται οι τιμές του χαρακτηριστικού πληθυσμός (2,3,4,10), που χρησιμοποιούνται για σύγκριση στα επίπεδα περιττής τάξης.

Το κύριο αρχείο, λοιπόν, έχει 11 κάδους. Η αποθήκευση των εγγραφών φαίνεται στον Πίνακα 11.1. Παρατηρείται ότι ένας κάδος περιέχει δύο εγγραφές, ενώ δύο κάδοι παραμένουν κενοί. Δηλαδή, σε μία τέτοια υλοποίηση ο κατάλογος διαμορφώνεται πριν τη φόρτωση του αρχείου και είναι

Χώρα	Κάδος
Αλβανία (A)	4
Βοσνία-Ερζεγοβίνη (Bh)	5
Βουλγαρία (B)	7
Γιουγκοσλαβία (Y)	7
Ελλάδα (G)	9
Κροατία (C)	6
Κύπρος (Cy)	1
ΠΓΔΜ (F)	3
Ρουμανία (R)	10
Σλοβενία (S)	2

Πίνακας 11.1: Αποθήκευση εγγραφών στους κάδους του αρχείου.

στατικός. Σωστά σχεδιασμένος είναι ο κατάλογος που είναι περίπου ισοζυγισμένος. Για το σκοπό αυτό πρέπει κατά τη δημιουργία του καταλόγου σε κάθε διαδοχικό επίπεδο του δένδρου ως ρίζα να επιλέγεται η εγγραφή εκείνη, που στο κατάλληλο πεδίο έχει τη μεσαία τιμή σε σχέση με τις τιμές του ίδιου πεδίου όλων των εγγραφών, που ανήκουν στο συγκεκριμένο υποδένδρο. Στο Σχήμα 11.9 παρουσιάζεται ένα ισοζυγισμένο 2-d δένδρο που προκύπτει εφαρμόζοντας αυτή τη διαδικασία. Αν είναι γνωστό το πλήθος των εγγραφών και η χωρητικότητα των κάδων, τότε εύκολα προκύπτει ο αριθμός των κλειδιών, που χρειάζονται για τη δημιουργία του καταλόγου.



Σχήμα 11.9: Ισοζυγισμένο 2-d δένδρο με τα δεδομένα του Πίνακα 7.2.

Αν οι κάδοι του αρχείου αυτού γεμίσουν, τότε δημιουργούνται κάδοι υπερχείλισης. Αν το φαινόμενο της υπερχείλισης γίνει έντονο και η επίδοση κατά την αναζήτηση εκφυλισθεί, τότε η αναδιοργάνωση του αρχείου και η δημιουργία ενός νέου καταλόγου είναι αναγκαία.

Στη συνέχεια δίνονται παραδείγματα αναζήτησης για ερώτηση μερικής ταύτισης και ερώτηση διαστήματος με βάση τα Σχήματα 11.7 και 11.8. Ας υποτεθεί ότι τίθεται η ερώτηση: `Ποιά βαλκανική χώρα έχει πληθυσμό 10.000.000 κατοίκους και έκταση 102.000 τετραγωνικά χιλιόμετρα;΄. Η αναζήτηση ξεκινά από τη ρίζα, όπου η σύγκριση γίνεται με την τιμή 3, και συνεχίζει στο δεξιό υποδένδρο. Στο νέο κόμβο η σύγκριση γίνεται με βάση την τιμή 110, οπότε η αναζήτηση συνεχίζει στο αριστερό υποδένδρο. Στο τρίτο πλέον επίπεδο μετά από σύγκριση με την τιμή 4, ακολουθείται ο δεξιός δενδρικός δείκτης που παραπέμπει στο τέταρτο επίπεδο. Εκεί η σύγκριση γίνεται με βάση το κλειδί 56, οπότε η αναζήτηση η αναζήτηση συνεχίζει στο δεξιό παιδί (που είναι φύλλο). Από τον κόμβο αυτό λαμβάνεται ο αριστερός δείκτης προς τον κάδο υπ΄ αριθμόν 7. `Ολη αυτή η διαδικασία γίνεται στην κύρια μνήμη και θεωρείται αμελητέου κόστους. Από το δίσκο προσπελάζεται ο κάδος υπ΄ αριθμόν 7, που περιέχει δύο εγγραφές: της Βουλγαρίας και της Γιουγκοσλαβίας. Οι εγγραφές ελέγχονται στην κύρια μνήμη και τελικώς επιστρέφεται στο χρήστη η σωστή απάντηση. Αυτός ο έλεγχος των εγγραφών του προσπελασθέντος κάδου είναι απαραίτητος, αφού έτσι άλλωστε διαπιστώνεται και η ανεπιτυχής αναζήτηση.

Ας θεωρηθεί τώρα η ερώτηση διαστήματος: `Ποιές βαλκανικές χώρες έχουν πληθυσμό περισσότερο από 6.000.000;΄. Και πάλι η αναζήτηση ξεκινά από τη ρίζα και προφανώς συνεχίζει στο δεξιό υποδένδρο. Στο νέο κόμβο δεν μπορεί να γίνει σύγκριση, οπότε η αναζήτηση συνεχίζει και στο αριστερό και στο δεξιό υποδένδρο. Στο τρίτο πλέον επίπεδο μετά από σύγκριση με τις τιμές 4 και 10, η αναζήτηση συνεχίζει σε τρεις κατευθύνσεις. Η τελική κατάληξη είναι να προσπελασθούν οι κάδοι υπ΄ αριθμόν 6, 7, 8, 9, 10, και 11. Οι εγγραφές ελέγχονται στην κύρια μνήμη και τελικώς από το περιεχόμενο των κάδων αυτών απορρίπτεται μόνο η εγγραφή της Κροατίας. Αν και η επεξεργασία του καταλόγου θεωρείται αμελητέου κόστους, στην περίπτωση αυτή πρέπει να τονισθεί ότι η αναζήτηση είναι σχετικά πολύπλοκη διαδικασία, γιατί πρέπει να διατηρούνται ενεργά πολλά μονοπάτια.

Τα **πολυδιάστατα Β-δένδρα** (k-d B-trees), που προτάθηκαν από τον Robinson (1981), είναι μία γενίκευση των πολυδιάστατων δένδρων (όπως τα Β-δένδρα είναι μία γενίκευση των απλών δυαδικών δένδρων). Η δομή

αυτή είναι σχεδιασμένη για υλοποίηση στο δίσκο, οπότε σε κάθε κόμβο των δένδρων αυτών δεν περιέχεται μόνο μία εγγραφή αλλά περισσότερες. Οι αλγόριθμοι εισαγωγής και διαγραφής είναι ιδιαίτερα πολύπλοκοι στον προγραμματισμό και την κατανόηση τους, γιατί η δομή πρέπει να παραμένει ισοζυγισμένη. Έτσι για να απλοποιηθεί η δομή δεν ισχύει το ελάχιστο ποσοστό χρήσης που ισχύει στα B-δένδρα (δηλαδή, το U_{min}=50%), οπότε η μέση χρήση του χώρου είναι μόνο 60% περίπου, όταν τα δεδομένα υπακούουν σε ομοιόμορφη κατανομή.

11.6 Δικτυωτό αρχείο

Έστω ότι από δορυφορικές φωτογραφίες έχει δημιουργηθεί ένα αρχείο-χάρτης με τις πόλεις της Ρωσίας. Κάθε πόλη διακρίνεται από το γεωγραφικό πλάτος και το γεωγραφικό μήκος, που αποτελούν δύο ισότιμα κριτήρια. Αν το αρχείο οργανωθεί με βάση το γεωγραφικό πλάτος, τότε πόλεις με ίδιο γεωγραφικό πλάτος αλλά διαφορετικό γεωγραφικό μήκος θα αποθηκευθούν μαζί. Έτσι, για παράδειγμα, το Μαγκαντάν (59^0 $57'$ βόρεια και 150^0 $43'$ ανατολικά, κοντά στα σύνορα με την Ιαπωνία) θα αποθηκευθεί μαζί με το Λένινγκραντ (59^0 $57'$ βόρεια και 30^0 $20'$ ανατολικά, κοντά στα σύνορα με τη Φιλλανδία). Αυτές οι πόλεις απέχουν μεταξύ τους 4000 μίλια και όμως πρέπει σύμφωνα με αυτήν την οργάνωση να αποθηκευθούν μαζί. Συνεπώς, χρειάζεται μία δομή με χαρακτηριστικά διαφορετικά από όσες εξετάσθηκαν μέχρις εδώ.

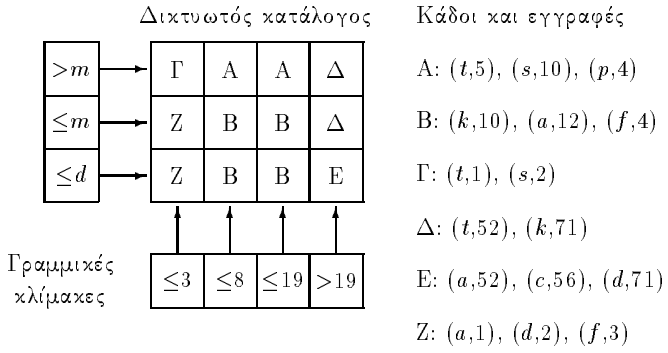
Έστω ότι μία εγγραφή έχει τέσσερα κλειδιά. Θεωρείται ο τετραδιάστατος χώρος που ορίζεται από τα πεδία ορισμού των χαρακτηριστικών. Αυτός ο χώρος είναι δυνατόν να θεωρηθεί ως ένας δυαδικός πίνακας. Στον πίνακα αυτόν η διάσταση i θα έχει τόσα στοιχεία όσες είναι οι διακριτές τιμές του χαρακτηριστικού i. Ένα στοιχείο του τετραδιάστατου πίνακα θα ισούται με 1, αν υπάρχει εγγραφή με τις αντίστοιχες τιμές για τα τέσσερα χαρακτηριστικά. Ένα στοιχείο του πίνακα θα είναι μηδέν, αν δεν υφίσταται η σχετική εγγραφή. Έτσι ο πίνακας των bits θα είναι μία πλήρης αναπαράσταση του συνόλου των εγγραφών.

Αρχικά φαίνεται ότι η δομή αυτή μπορεί να ικανοποιήσει όλες τις απαιτήσεις από ένα σύστημα αρχείων. Επεξεργασία των σημειακών ερωτήσεων γίνεται με εξέταση ενός μόνο στοιχείου, ενώ οι ερωτήσεις διαστήματος απαντώνται με επεξεργασία όλων των στοιχείων του j-διάστατου χώρου, όπου

$j \leq k$. Οι εισαγωγές και οι διαγραφές επιτυγχάνονται θέτοντας τα αντίστοιχα bits ίσα με 1 ή με 0. Τέλος, όλα τα κλειδιά αντιμετωπίζονται κατά τον ίδιο τρόπο. Ωστόσο, αυτή η οργάνωση δεν μπορεί να ανταπεξέλθει σε πρακτικές καταστάσεις. Για παράδειγμα, αν υπήρχαν τέσσερα πεδία με 100 διακριτές τιμές το καθένα, τότε ο πίνακας θα αποτελούνταν από 1.000.000 στοιχεία. Η προηγούμενη οργάνωση, λοιπόν, είναι μία ουτοπική λύση.

Το δικτυωτό αρχείο είναι μία νέα οργάνωση που προτάθηκε από το Nievergelt (1984). Η δομή αυτή διαχειρίζεται το ίδιο αποτελεσματικά τόσο στατικά όσο και δυναμικά δεδομένα, επειδή σχεδιάσθηκε με βάση τις εξής απαιτήσεις:

- αποτελεσματική ικανοποίηση **σημειακών ερωτήσεων** (point queries), που γίνονται με βάση συγκεκριμένες τιμές για όλα τα κλειδιά,
- αποτελεσματική ικανοποίηση των ερωτήσεων διαστήματος και ερωτήσεων μερικής ταύτισης,
- δυναμική προσαρμοστικότητα στις εισαγωγές και διαγραφές εγγραφών, και τέλος
- ισότιμη αντιμετώπιση όλων των κλειδιών, πρωτεύοντος και δευτερεύοντων.



Σχήμα 11.10: Δομή δικτυωτού αρχείου.

Στο Σχήμα 11.10 παρουσιάζεται η δομή του δικτυωτού αρχείου που αποτελείται από τους κάδους A,B,...,Z, που περιέχουν τα δεδομένα, και τον κατάλογο που απαρτίζεται από δύο μέρη: το **δικτυωτό κατάλογο ή πίνακα**

(grid directory, array) και τις **γραμμικές κλίμακες** (linear scales). Για κάθε χαρακτηριστικό υπάρχει και μία αντίστοιχη γραμμική κλίμακα, που περιέχει ζεύγη τιμών κλειδιών και αντίστοιχων διευθύνσεων προς τον κατάλογο. Οι κλίμακες αποθηκεύονται στην κύρια μνήμη και δείχνουν ποιό $(k–1)$-διάστατο τμήμα του καταλόγου περιέχει τους δείκτες προς τους κάδους των εγγραφών.

Ο δικτυωτός κατάλογος είναι ένας πίνακας με διάσταση k (όπου k είναι ο αριθμός των χαρακτηριστικών), που αν είναι μικρός, τότε αποθηκεύεται στην κύρια μνήμη. Το μέγεθος του καταλόγου περιορίζεται τεμαχίζοντας το σύνολο ορισμού κάθε χαρακτηριστικού σε ανάλογα υποσύνολα. Κάθε είσοδος του καταλόγου περιέχει τη διεύθυνση του αντίστοιχου κάδου των εγγραφών. Είναι δυνατόν πολλές είσοδοι να περιέχουν την ίδια διεύθυνση. Συνήθως, ο κατάλογος είναι αρκετά μεγάλος και δεν χωρά στην κύρια μνήμη. Υπάρχει ένας περιορισμός όσον αφορά στις εισόδους του καταλόγου, που μπορεί να περιέχουν την ίδια διεύθυνση. Αν οι είσοδοι $(a_1, a_2, ..., a_k)$ περιέχουν τις ίδιες διευθύνσεις με τις εισόδους $(b_1, b_2, ..., b_k)$, τότε τις ίδιες διευθύνσεις πρέπει να περιέχουν και οι είσοδοι $(x_1, x_2, ..., x_k)$, για κάθε x_i όπου $1 \leq i \leq k$ και $\mathsf{min}(a_i, b_i) \leq x_i \leq \mathsf{max}(a_i, b_i)$. Αυτός ο περιορισμός σημαίνει ότι ένας κάδος περιέχει εγγραφές που ανήκουν σε μία **τετράγωνη ή κοίλη** (rectangular, convex) περιοχή. Στο Σχήμα 11.11 παρουσιάζεται μία περίπτωση που παραβιάζει τον περιορισμό.

A	A	A
B	B	A

Σχήμα 11.11: Λανθασμένη οργάνωση καταλόγου δικτυωτού αρχείου.

Αν ο κατάλογος είναι σχετικά μικρός και μπορεί να αποθηκευθεί στην κύρια μνήμη, τότε η ικανοποίηση μίας απλής ερώτησης είναι εύκολη υπόθεση. Πρώτα, χρησιμοποιείται η γραμμική κλίμακα για να εντοπισθεί η είσοδος του καταλόγου που περιέχει το δείκτη προς το δίσκο. Η είσοδος του καταλόγου μπορεί να βρεθεί εύκολα χρησιμοποιώντας τις γνωστές τεχνικές από το βιβλίο των Δομών Δεδομένων για την επεξεργασία πινάκων. Με την εύρεση του δείκτη και την επεξεργασία του κάδου στην κύρια μνήμη ελέγχεται αν πράγματι η ζητούμενη εγγραφή είναι αποθηκευμένη στο αρχείο.

Αν η υποβαλλόμενη ερώτηση είναι ερώτηση διαστήματος, τότε και πάλι η επεξεργασία δεν είναι δύσκολη. Έστω, για παράδειγμα, ότι ζητείται να

βρεθούν όλες οι πόλεις που απέχουν από τη Μόσχα ± 2 μοίρες ως προς το γεωγραφικό πλάτος και το γεωγραφικό μήκος. Άρα η ερώτηση μπορεί να τεθεί ως:

$$53^0\ 45'\ \text{βόρεια} \leq \text{μήκος} \leq 57^0\ 45'\ \text{βόρεια}$$

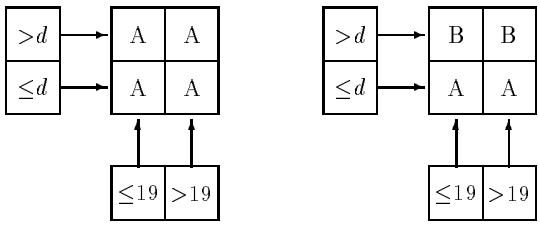
και

$$35^0\ 37'\ \text{ανατολικά} \leq \text{πλάτος} \leq 39^0\ 37'\ \text{ανατολικά}$$

Η επεξεργασία αρχίζει από τις γραμμικές κλίμακες για να εντοπισθεί ποιές είσοδοι του καταλόγου περιέχουν αυτές τις τιμές. Το αποτέλεσμα της επεξεργασίας είναι ένα σύνολο εισόδων. Από τον κατάλογο βρίσκονται οι δείκτες προς τη δευτερεύουσα μνήμη. Μάλιστα, είναι πολύ πιθανό ότι μερικοί δείκτες θα περιέχονται περισσότερο από μία φορά στον κατάλογο, αλλά προφανώς θα γίνει μόνο μία προσπέλαση στο δίσκο.

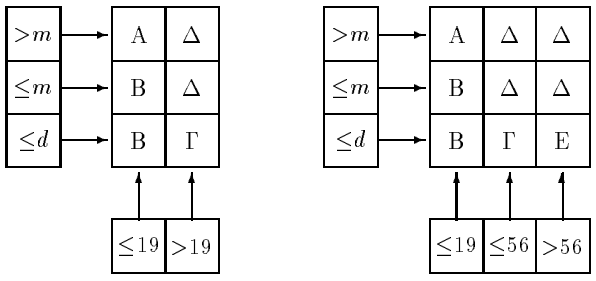
Η διαδικασία της εισαγωγής χρειάζεται ιδιαίτερη προσοχή. Βέβαια, αν ο κάδος όπου θα γίνει η εισαγωγή διαθέτει ακόμη κενό χώρο δεν είναι δύσκολη. Όμως αν ο κάδος υπερχειλίσει, τότε διακρίνονται δύο περιπτώσεις, μία απλή και μία σύνθετη. Και στις δύο περιπτώσεις στο αρχείο παραχωρείται ένας ακόμη κάδος για να στεγάσει μερικές από τις εγγραφές του κάδου που υπερχείλισε.

Αν υπάρχουν περισσότερες είσοδοι του καταλόγου που αναφέρονται στον κάδο που υπερχείλισε, τότε οι εγγραφές διανέμονται εκατέρωθεν μίας οριακής τιμής μεταξύ των τιμών της γραμμικής κλίμακας. Επίσης, πρέπει να ενημερωθούν σχετικά και οι δείκτες του καταλόγου, επειδή μερικοί θα αναφέρονται προς το νέο κάδο. Για παράδειγμα, στο Σχήμα 11.12 η χωρητικότητα των κάδων είναι τρεις εγγραφές. Με την εισαγωγή της εγγραφής $(b,5)$ προκύπτει η δεξιά δομή του Σχήματος 11.12.



Σχήμα 11.12: Εισαγωγή με απλή ενημέρωση καταλόγου.

Η πιο σύνθετη περίπτωση παρουσιάζεται όταν ο κάδος που υπερχείλισε αναφέρεται από μία μόνο είσοδο του καταλόγου ή όταν οι εγγραφές του κάδου που υπερχείλισε ανήκουν σε ένα μόνο διάστημα τιμών της γραμμικής κλίμακας. Στην περίπτωση αυτή εκτός από την επέκταση του αρχείου κατά ένα νέο κάδο επεκτείνεται και ο κατάλογος. Αυτή η διαδικασία φαίνεται στο Σχήμα 11.13, όπου εισάγεται η εγγραφή $(b,90)$ και ο κατάλογος επεκτείνεται κατά ένα μονοδιάστατο πίνακα. Στη γενική περίπτωση, αν ο κατάλογος είναι k-διάστατος, τότε η επέκταση γίνεται με ένα $(k-1)$-διάστατο πίνακα. Αξίζει να σημειωθεί ότι στο παράδειγμα αυτό η διάσπαση του καταλόγου έγινε κατά ένα νοερό κατακόρυφο άξονα. Έπεται ότι σε περίπτωση μελλοντικής διάσπασης του καταλόγου ο νοερός αυτός άξονας θα είναι οριζόντιος. Γενικά, αν υπάρχουν k χαρακτηριστικά, τότε οι διασπάσεις θα γίνονται κατά κυκλικό τρόπο, δηλαδή κάθε φορά ως προς ένα διαφορετικό χαρακτηριστικό της γραμμικής κλίμακας.



Δ: $(t,52)$, $(k,71)$
Γ: $(a,52)$, $(c,56)$, $(d,71)$

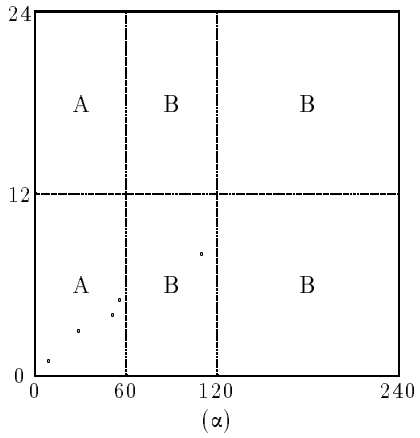
Δ: $(t,52)$, $(k,71)$
Γ: $(a,52)$, $(c,56)$
E: $(d,71)$, $(b,90)$

Σχήμα 11.13: Εισαγωγή με επέκταση καταλόγου.

Στο σημείο αυτό προκύπτει και το κυριότερο πρόβλημα του δικτυωτού αρχείου. Έστω ότι η εγγραφή αποτελείται από τρία κλειδιά, ενώ κάθε γραμμική κλίμακα αποτελείται από 100 τιμές. Όταν ο κατάλογος επεκταθεί κατά ένα δισδιάστατο πίνακα, στην ουσία προστίθενται 10.000 νέες είσοδοι στον κατάλογο. Αν κάθε είσοδος είναι τέσσερις χαρακτήρες, τότε φαίνεται ότι για μία εισαγωγή εγγραφής απαιτείται επέκταση κατά 40 Kb περίπου που ισοδυναμεί με μία ατράκτο σε ένα σύστημα δίσκων IBM 3380. Αν τα δεδομένα δεν υπακούουν σε μία ομοιόμορφη κατανομή, τότε αυτή η διαδικασία μπορεί να συμβαίνει σχετικά συχνά. Έτσι ο κατάλογος μπορεί να χρειάζεται σημαντικά περισσότερο χώρο από ότι καταλαμβάνουν οι εγγραφές.

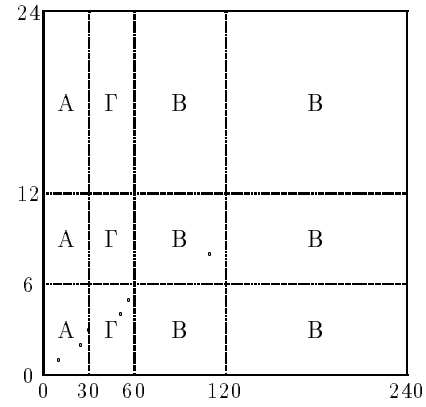
Ακόμη, αν υποβληθεί μία ερώτηση διαστήματος μπορεί να χρειασθεί μία χρονοβόρα επεξεργασία του καταλόγου για να προσπελασθούν τελικά λίγοι μόνο κάδοι του αρχείου.

Πάντως, πρέπει να σημειωθεί ότι (πέρα από το πρόβλημα του καταλόγου) για μία απλή σημειακή ερώτηση που βασίζεται σε οποιοδήποτε κλειδί, απαιτούνται το μέγιστο δύο προσπελάσεις στο δίσκο. Επιπλέον μετά την προσπέλαση των κάδων απαιτείται και επεξεργασία των εγγραφών στην κύρια μνήμη για να διευκρινισθεί πόσες και ποιές εγγραφές ανταποκρίνονται πράγματι την ερώτηση. Από τα προηγούμενα είναι προφανές ότι η ανεπιτυχής αναζήτηση μίας απλής ερώτησης έχει το ίδιο κόστος με την επιτυχή αναζήτηση, δηλαδή δύο προσπελάσεις στο δίσκο. Επίσης είναι ευνόητο ότι οι διαδικασίες ικανοποίησης μίας ερώτησης διαστήματος ή μίας ερώτησης μερικής ταύτισης δεν διαφέρουν σημαντικά από την αντίστοιχη διαδικασία μίας απλής σημειακής ερώτησης.

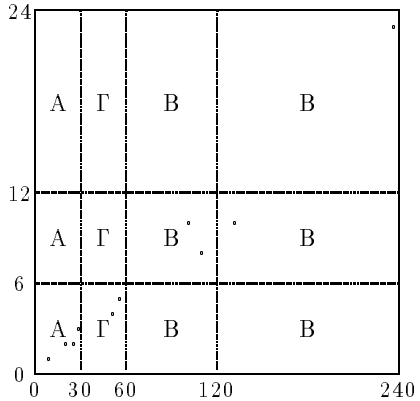


Σχήμα 11.14: Εισαγωγή δεδομένων Πίνακα 7.2 σε δικτυωτό αρχείο.

Στη συνέχεια τα δεδομένα του Πίνακα 7.2 εισάγονται σε ένα δικτυωτό αρχείο με κάδους χωρητικότητας τεσσάρων εγγραφών. Έστω ότι και πάλι τα δεδομένα εισάγονται στη δομή με αλφαβητική σειρά: A, B, Bh, C, Cy, F, G, R, S, Y. Στο Σχήμα 11.14α παρουσιάζεται η μορφή του δικτυωτού καταλόγου και των αντίστοιχων γραμμικών κλιμάκων μετά την εισαγωγή της πέμπτης εγγραφής (Cy) και τη σχετική επεξεργασία, δηλαδή διάσπαση γραμμικών κλιμάκων, δικτυωτού καταλόγου και αρχικού κάδου σε δύο κάδους A και B, ενώ με σημεία παρουσιάζονται οι πέντε εγγραφές. Στο Σχήμα



(β)



(γ)

Σχήμα 11.14: Εισαγωγή δεδομένων Πίνακα 7.2 σε δικτυωτό αρχείο (συνέχεια).

11.14β παρουσιάζεται η κατάσταση της δομής μετά την εισαγωγή της έκτης εγγραφής (F) και τη σχετική επεξεργασία που καταλήγει στη διάσπαση του κάδου Α και τη δημιουργία του κάδου Γ. Στο Σχήμα 11.14γ παρουσιάζεται η τελική κατάσταση, όπου κάθε κάδος περιέχει εγγραφές που ανήκουν σε μία τετράγωνη περιοχή.

Η διαδικασία μετάβασης από την αρχικά κενή δομή του Σχήματος 11.14α στο 11.14β είναι ανάγλυφα παραδείγματα της χρονοβόρας διαδικασίας, που αναφέρθηκε στις προηγούμενες παραγράφους. Αξίζει στο σημείο αυτό να αναφερθεί η συγγένεια αυτής της διαδικασίας των διαδοχικών διαιρέσεων του

χώρου του δικτυωτού καταλόγου με την παθολογική περίπτωση του διπλασιασμού του επεκτατού κατακερματισμού. Σύμφωνα με μία άλλη υλοποίηση ο δικτυωτός κατάλογος μπορεί να μοιάζει με τη μέθοδο του δυναμικού κατακερματισμού. Από τη Regnier (1985) μάλιστα αποδείχθηκε αναλυτικά ότι αν η χωρητικότητα των κάδων είναι λιγότερο από είκοσι εγγραφές ή αν οι τιμές των χαρακτηριστικών δεν υπακούουν σε ομοιόμορφες κατανομές, τότε πρέπει να προτιμηθεί η τελευταία μέθοδος.

Ας υποτεθεί, λοιπόν, ότι με βάση το Σχήμα 11.14 τίθεται η ερώτηση: `Ποιά βαλκανική χώρα έχει πληθυσμό 10.000.000 κατοίκους και έκταση 132.000 τετραγωνικά χιλιόμετρα;΄. Η αναζήτηση ξεκινά από τη γραμμική κλίμακα, όπου ο δείκτης κατευθύνει στην αντίστοιχη είσοδο του δικτυωτού καταλόγου. Η είσοδος αυτή περιέχει ένα δείκτη προς τον κάδο Β, που προσπελάζεται από το δίσκο. Ο κάδος Β περιέχει τέσσερις εγγραφές, που ελέγχονται στην κύρια μνήμη και τελικώς επιστρέφεται στο χρήστη η σωστή απάντηση. Ας θεωρηθεί τώρα η ερώτηση διαστήματος: `Ποιές βαλκανικές χώρες έχουν έκταση λιγότερο από 50.000 τετραγωνικά χιλιόμετρα;΄. Και πάλι η αναζήτηση ξεκινά από τη γραμμική κλίμακα και το δικτυωτό κατάλογο και στη συνέχεια προσπελάζονται οι κάδοι Α και Γ, που περιέχουν τις κατάλληλες εγγραφές.

Η διαγραφή μίας εγγραφής είναι μία πολύπλοκη διαδικασία, αντίστροφη της εισαγωγής, και μπορεί να συνοδεύεται με ελάττωση του αριθμού των κάδων. `Οπως και στις οργανώσεις του δυναμικού κατακερματισμού τίθεται από το σχεδιαστή των αρχείων ένα ελάχιστο όριο για τον παράγοντα χρησιμοποίησης του χώρου των κάδων. Αν η τιμή του παράγοντα χρησιμοποίησης γίνει μικρότερη από την προκαθορισμένη τιμή, τότε εξετάζεται αν μπορεί να γίνει συγχώνευση. Συγχώνευση μπορεί να γίνει με κάποιον κάδο που δημιουργήθηκε από την ίδια διάσπαση με τον υπ΄ όψη κάδο (buddy bucket) σε οποιαδήποτε από τις k διαστάσεις και με την προϋπόθεση ότι το περιεχόμενο και των δύο κάδων μπορεί να αποθηκευθεί σε ένα μόνο κάδο. Σύμφωνα με μία άλλη πιο πολύπλοκη τεχνική, συγχώνευση μπορεί να γίνει με οποιοδήποτε από τους δύο γειτονικούς κάδους στις k διαστάσεις (neighbor bucket) αρκεί το αποτέλεσμα να μην παραβιάζει την απαίτηση που υπαγορεύει κάθε κάδος να αντιστοιχεί σε μία τετράγωνη περιοχή του δικτυωτού καταλόγου. Βέβαια, εκτός από τη συγχώνευση των κάδων πρέπει να γίνει και συρρίκνωση του καταλόγου. `Ομως σε πρακτικές περιπτώσεις η συρρίκνωση αποφεύγεται, ώστε να αποφευχθεί το διπλό κόστος σε μία μελλοντική επέκταση του καταλόγου.

Το δικτυωτό αρχείο σε σύγκριση με το πολυδιάστατο δένδρο έχει μία σειρά πλεονεκτημάτων:

- είναι περισσότερο ισοζυγισμένη δομή,
- γενικότερα σε δυναμικά δεδομένα έχει καλύτερες επιδόσεις σε κάθε είδους αναζήτηση,
- επιτυγχάνει κατά μέσο όρο χρήση του χώρου 69% περίπου,
- έχει καλύτερη επίδοση σε περιβάλλον πολλών χρηστών (για λόγους που θα γίνουν αντιληπτοί στο τελευταίο κεφάλαιο του τόμου αυτού).

Όμως υπάρχουν και πλεονεκτήματα του k-d δένδρου όπως:

- η απλότητα του λογισμικού του,
- έχει καλύτερη επίδοση κατά την εισαγωγή και τη διαγραφή,
- έχει την ίδια πολυπλοκότητα και επίδοση ανεξάρτητα από το πλήθος k των χαρακτηριστικών,
- αν το αρχείο είναι σωστά σχεδιασμένο για στατικά δεδομένα, τότε έχει πολύ καλές επιδόσεις σε όλες τις λειτουργίες, καθώς και πολύ καλή χρήση του χώρου.

Το κεφάλαιο αυτό ήταν αφιερωμένο σε οργανώσεις κατάλληλες για μη απλές ερωτήσεις. Από το πλήθος των οργανώσεων αυτών που είναι μεγάλο εξετάσθηκαν τα αντεστραμμένα αρχεία, οι πολλαπλές λίστες, οι συνδυασμένοι κατάλογοι, τα πολυδιάστατα δένδρα και το δικτυωτό αρχείο. Οι προηγούμενες δομές είναι σύνθετες και απαιτούν ιδιαίτερη προσοχή για την υλοποίηση τους. Εξ άλλου, και η επακριβής μαθηματική έκφραση της επίδοσης τους είναι πολύπλοκη.

11.7 Ασκήσεις

<1> Δανειστική βιβλιοθήκη διατηρεί πληροφορίες για βιβλία και δανειζόμενους, ώστε να απαντώνται πολύ γρήγορα ερωτήσεις του τύπου:

- Ποιά βιβλία που έχει χρεωθεί ο κ.Τάδε΄,

- `Είναι το βιβλίο τάδε δανεισμένο, και αν ναι σε ποιόν;΄.

Να σχεδιασθούν τα αρχεία που θα υποστήριζαν αποτελεσματικά αυτές τις ερωτήσεις.

<**2**> Δίνεται αρχείο με 100.000 εγγραφές ασθενών των 400 bytes, 200 εγγραφές γιατρών των 400 bytes και 500.000 εγγραφές εργαστηριακών εξετάσεων των 100 bytes. Υπάρχουν 100 είδη εργαστηριακών εξετάσεων. Να σχεδιασθεί μία δομή πολλαπλών λιστών και να κοστολογηθούν οι απαντήσεις των εξής ερωτήσεων:

- `Ποιοί είναι οι ασθενείς του γιατρού κ.ΑΤάδε΄,
- `Ποιές εργαστηρικές εξετάσεις έκανε ο κ.ΒΤάδε΄,
- `Ποιές εργαστηρικές εξετάσεις έκαναν οι ασθενείς του κ.ΑΤάδε΄, και
- `Ποιά τα αποτελέσματα των εργαστηρικών εξετάσεων χοληστερίνης΄.

<**3**> Για έναν αριθμό a χαρακτηριστικών να βρεθούν οι αλγόριθμοι δημιουργίας:

- των συνδυασμένων καταλόγων,
- των μειωμένων συνδυασμένων καταλόγων, και
- των τροποποιημένων συνδυασμένων καταλόγων.

<**4**> Να δημιουργηθεί ένα δισδιάστατο δένδρο εισάγοντας τα δεδομένα του Πίνακα 7.2 κατά την αντίστροφη σειρά. Να σχεδιασθούν οι υποδιαίρεσεις του χώρου και να γίνει η κατανομή των εγγραφών στους κατάλληλους κάδους.

<**5**> Πόσες αντιπροσωπευτικές εγγραφές από ένα αρχείο 1000 εγγραφών χρειάζονται για να δημιουργηθεί ένα τρισδιάστατο δένδρο και ποιό είναι το βάθος του; Ας υποτεθεί ότι η χωρητικότητα των κάδων είναι 10 εγγραφές και ότι το δένδρο είναι περίπου ζυγισμένο.

<**6**> Να δημιουργηθεί ένα δικτυωτό αρχείο με τα δεδομένα του Πίνακα 7.2 θεωρώντας ότι η χωρητικότητα κάδων είναι τρεις εγγραφές. Να σχεδιασθούν οι υποδιαίρεσεις του δικτυωτού καταλόγου και να γίνει η κατανομή των εγγραφών στους κατάλληλους κάδους.

<**7**> Να δημιουργηθεί ένα δικτυωτό αρχείο εισάγοντας τα δεδομένα του Πίνακα 7.2 κατά την αντίστροφη σειρά. Να σχεδιασθούν οι υποδιαίρεσεις του δικτυωτού καταλόγου και να γίνει η κατανομή των εγγραφών στους κατάλληλους κάδους.