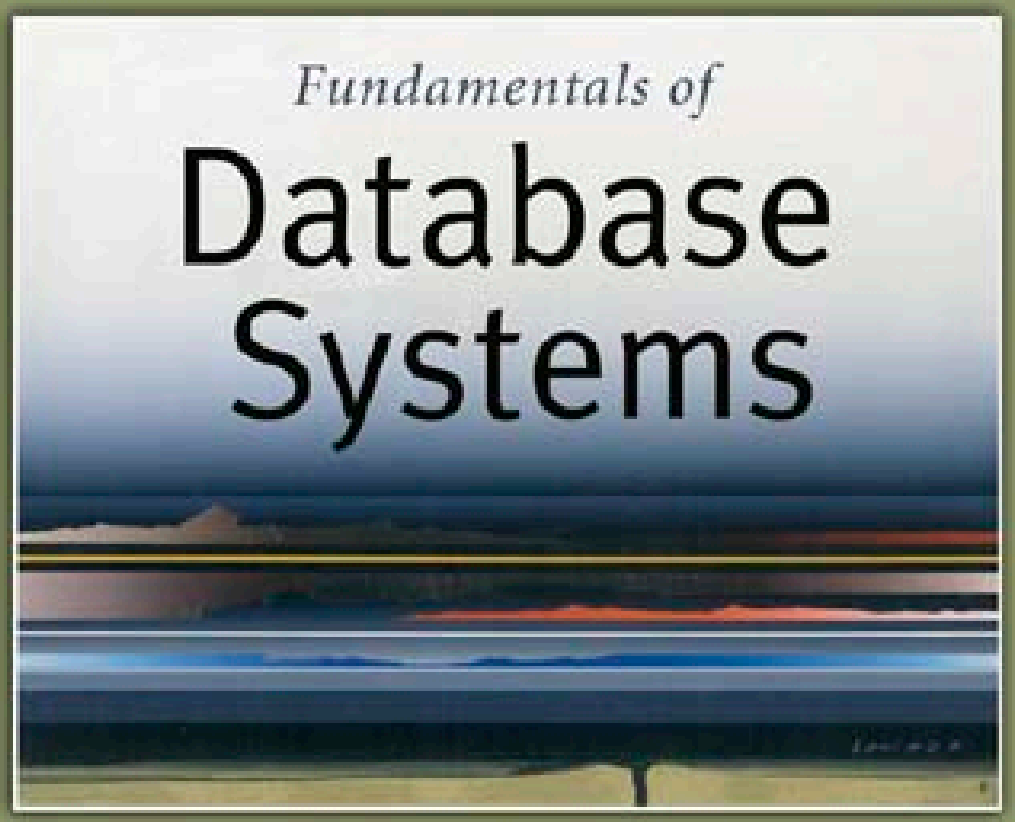
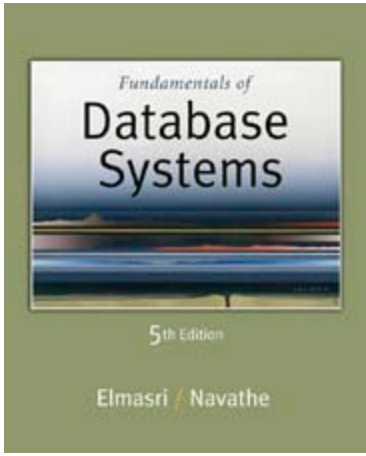


5th Edition

Elmasri / Navathe

Κεφάλαιο 14

Δομές Ευρετηρίων για Αρχεία



Θα μιλήσουμε για

- Τύποι Ταξινομημένων Ευρετηρίων ενός επιπέδου
  - Πρωτεύοντα Ευρετήρια
  - Ευρετήρια Συστάδες
  - Δευτερεύοντα Ευρετήρια
- Ευρετήρια Πολλών Επιπέδων
- Δυναμικά Ευρετήρια Πολλών Επιπέδων με χρήση B-Δένδρων και B+-Δένδρων
- Ευρετήρια σε Πολλαπλά Κλειδιά

Ευρετήρια σαν Μέθοδοι Προσπέλασης

- Ένα ευρετήριο ενός επιπέδου είναι ένα βοηθητικό αρχείο που κάνει πιο αποτελεσματική την αναζήτηση μιας εγγραφής σε ένα αρχείο δεδομένων.
- Το ευρετήριο συνήθως ορίζεται σε ένα πεδίο του αρχείου (αν και μπορεί να ορισθεί σε πολλά πεδία)
- Μια μορφή ευρετηρίου είναι ένα αρχείο με καταχωρήσεις <**τιμή πεδίου, δείκτης στην εγγραφή**>, που είναι ταξινομημένο με τιμή πεδίου
- Το ευρετήριο λέγεται δρόμος προσπέλασης στο πεδίο.

Ευρετήρια σαν Μέθοδοι Προσπέλασης(συν.)

- Το αρχείο του ευρετηρίου συνήθως καταλαμβάνει σημαντικά μικρότερο πλήθος μπλοκ από ότι το αρχείο δεδομένων επειδή οι καταχωρήσεις του είναι κατά πολύ μικρότερες
- Μια δυαδική αναζήτηση στο ευρετήριο παράγει ένα δείκτη στην εγγραφή του αρχείου
- Τα ευρετήρια μπορούν επίσης να χαρακτηρισθούν σαν πυκνά ή αραιά
  - Ένα **πυκνό ευρετήριο** έχει μια καταχώρηση ευρετηρίου για κάθε αναζήτηση τιμής κλειδιού (και επομένως κάθε εγγραφή) στο αρχείο δεδομένων.
  - Ένα **αραιό (ή μη πυκνό) ευρετήριο**, από την άλλη, έχει καταχωρήσεις ευρετηρίου μόνο για κάποιες από τις τιμές αναζήτησης.

Ευρετήρια σαν Μέθοδοι Προσπέλασης(συν.)

- Παράδειγμα: Έστω το αρχείο δεδομένων ΕΡΓΑΖΟΜΕΝΟΣ(ΟΝΟΜΑ, ΑΡ_ΤΑΥΤ, ΔΙΕΥΘΥΝΣΗ, ΕΡΓΑΣΙΑ, ΜΙΣΘΟΣ, ...)
- Υποθέστε ότι:
  - μέγεθος εγγραφής R=150 bytes μέγεθος μπλοκ B=512 bytes r=30000 εγγραφές
- Τότε, έχουμε:
  - παράγοντας ομαδοποίησης Bfr= B div R= 512 div 150= 3 εγγραφές/μπλοκ
  - πλήθος μπλοκ αρχείου b= (r/Bfr)= (30000/3)= 10000 μπλοκ
- Για ένα ευρετήριο στο πεδίο ΑΡ_ΤΑΥΤ, υποθέστε μέγεθος πεδίου $V_{ΑρΤΑΥΤ}$=9 bytes, υποθέστε μέγεθος δείκτη εγγραφής P_R=7 bytes. Τότε:
  - μέγεθος καταχώρησης ευρετηρίου $R_I=(V_{SSN}+ P_R)=(9+7)=16$ bytes
  - παράγοντας ομαδοποίησης ευρετηρίου Bfr_I= B div R_I= 512 div 16= 32 κατχ/μπλοκ
  - πλήθος μπλοκ ευρετηρίου b= (r/ Bfr_I)= (30000/32)= 938 μπλοκ
  - η δυαδική αναζήτηση απαιτεί $\log_2 bI= \log_2 938$= 10 μπλοκ προσπελάσεις
  - Σε σχέση με το κόστος της μέσης γραμμικής αναζήτησης που είναι:
    - (b/2)= 10000/2= 5000 μπλοκ προσπελάσεις
  - Αν οι εγγραφές του αρχείου είναι ταξινομημένες, η δυαδική αναζήτηση θα ήταν:
    - $\log_2 b= \log_2 10000$= 14 μπλοκ προσπελάσεις

Τα ευρετήρια ορίζονται σε ένα ή περισσότερα πεδία που ονομάζονται πεδίο(α) ή γνώρισμα(τα) ευρετηριοποίησης.

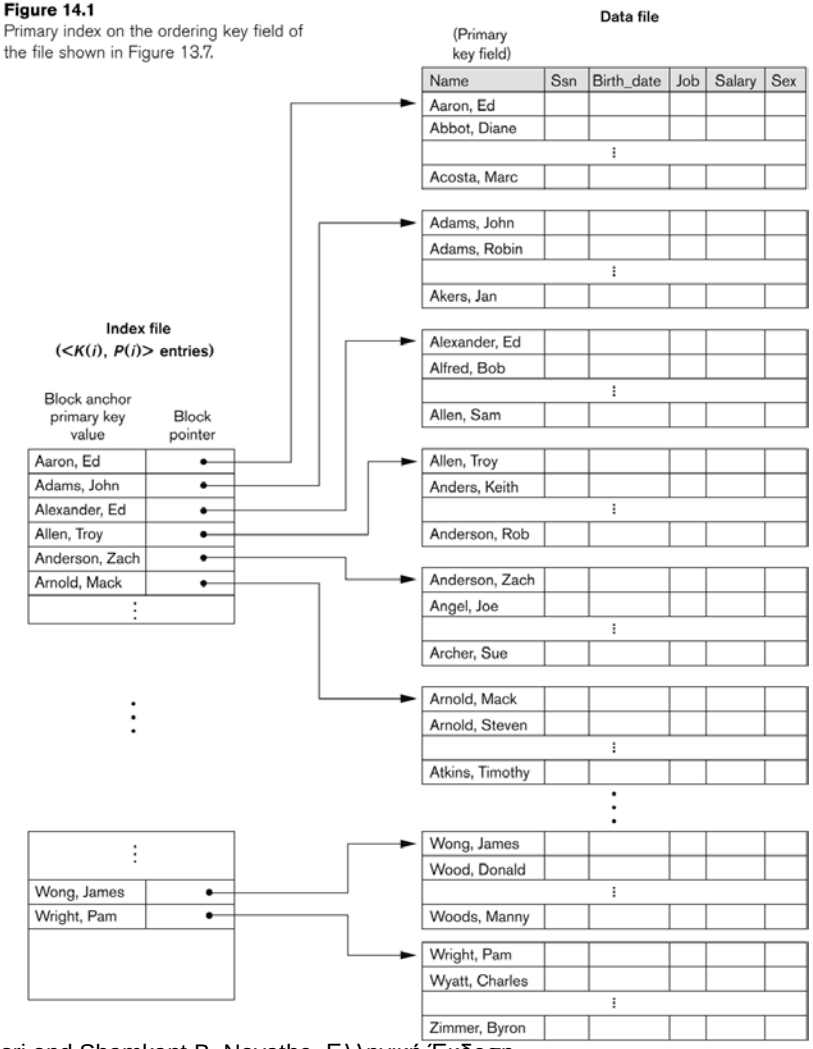
- **Πρωτεύον Ευρετήριο**
- **Ευρετήριο Συστάδων**
- **Δευτερεύον Ευρετήριο**

Τύποι Ευρετηρίων ενός Επιπέδου

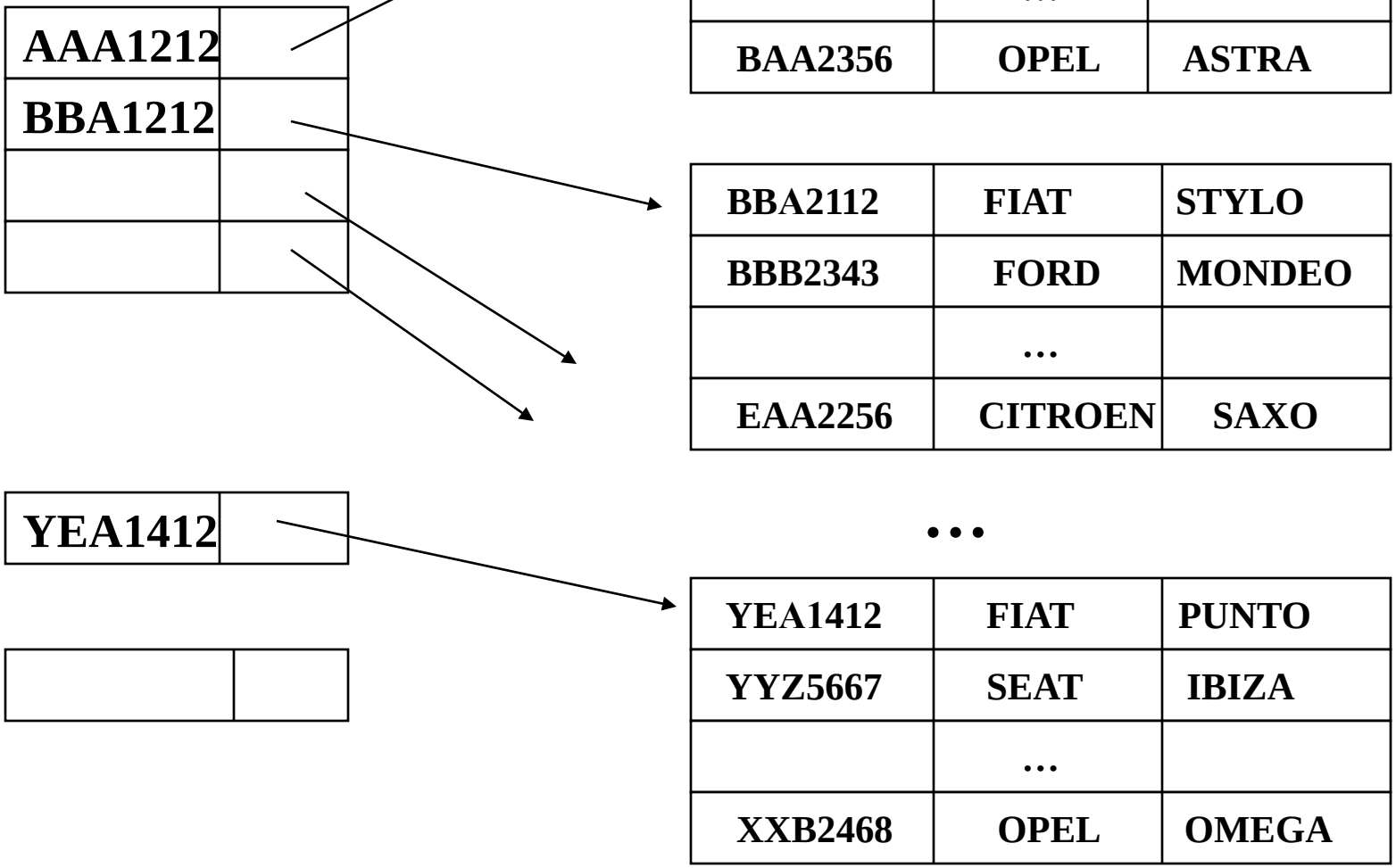
- Πρωτεύον Ευρετήριο
  - Ορίζεται σε ένα ταξινομημένο αρχείο δεδομένων
  - Το αρχείο δεδομένων είναι ταξινομημένο σε ένα **πεδίο κλειδί**
  - Περιλαμβάνει μια καταχώρηση ευρετηρίου *για κάθε μπλοκ* στο αρχείο δεδομένων· η καταχώρηση ευρετηρίου έχει την τιμή του πεδίου κλειδιού για την *πρώτη εγγραφή* στο μπλοκ, που ονομάζεται *άγκυρα του μπλοκ*
  - Ένα παρόμοιο σχήμα μπορεί να χρησιμοποιεί την *τελευταία εγγραφή* σε ένα μπλοκ.
  - Ένα πρωτεύον ευρετήριο είναι ένα μη πυκνό (αραιό) ευρετήριο, αφού έχει μια καταχώρηση για κάθε μπλοκ του αρχείου δεδομένων στο δίσκο και στα κλειδιά των εγγραφών άγκυρα παρά για κάθε τιμή αναζήτησης.

Πρωτεύον ευρετήριο στο πεδίο κλειδί ταξινόμησης

Figure 14.1
Primary index on the ordering key field of the file shown in Figure 13.7.



Ευρετήριο

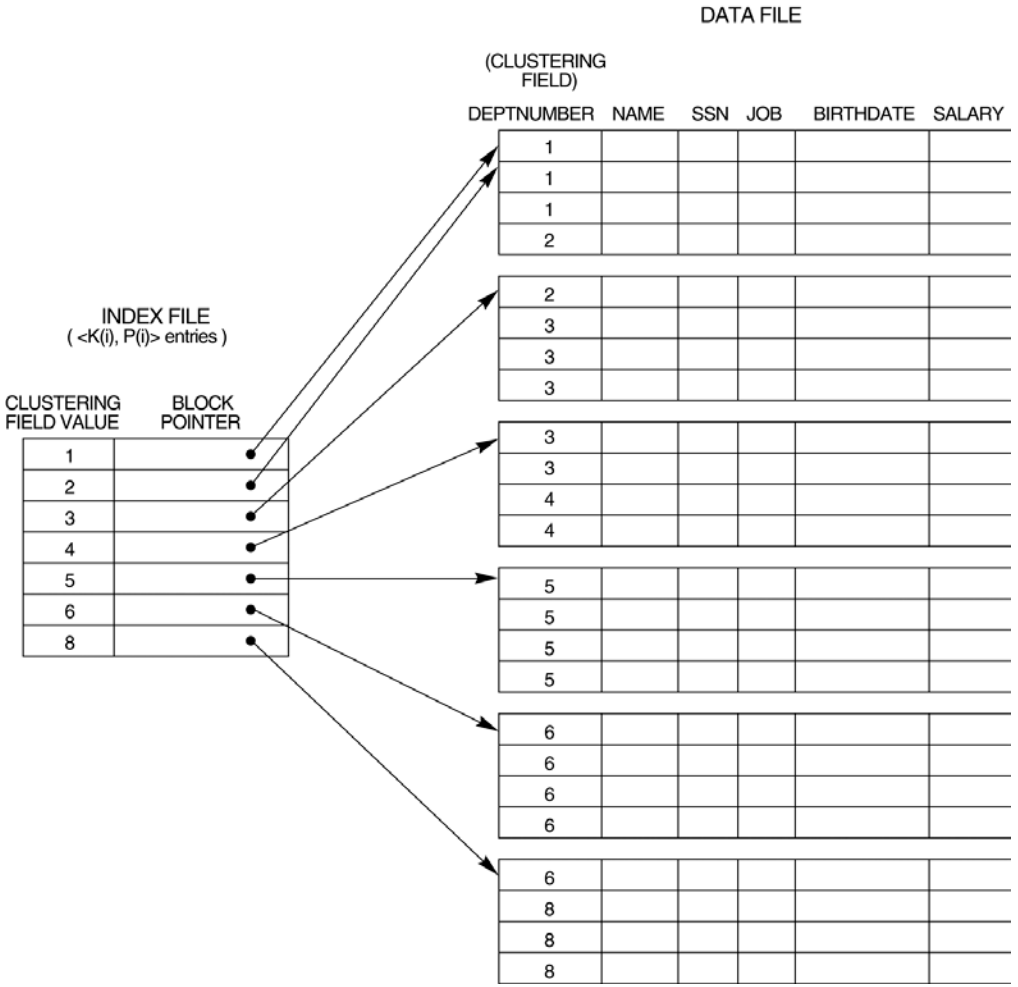


Τύποι Ευρετηρίων ενός Επιπέδου

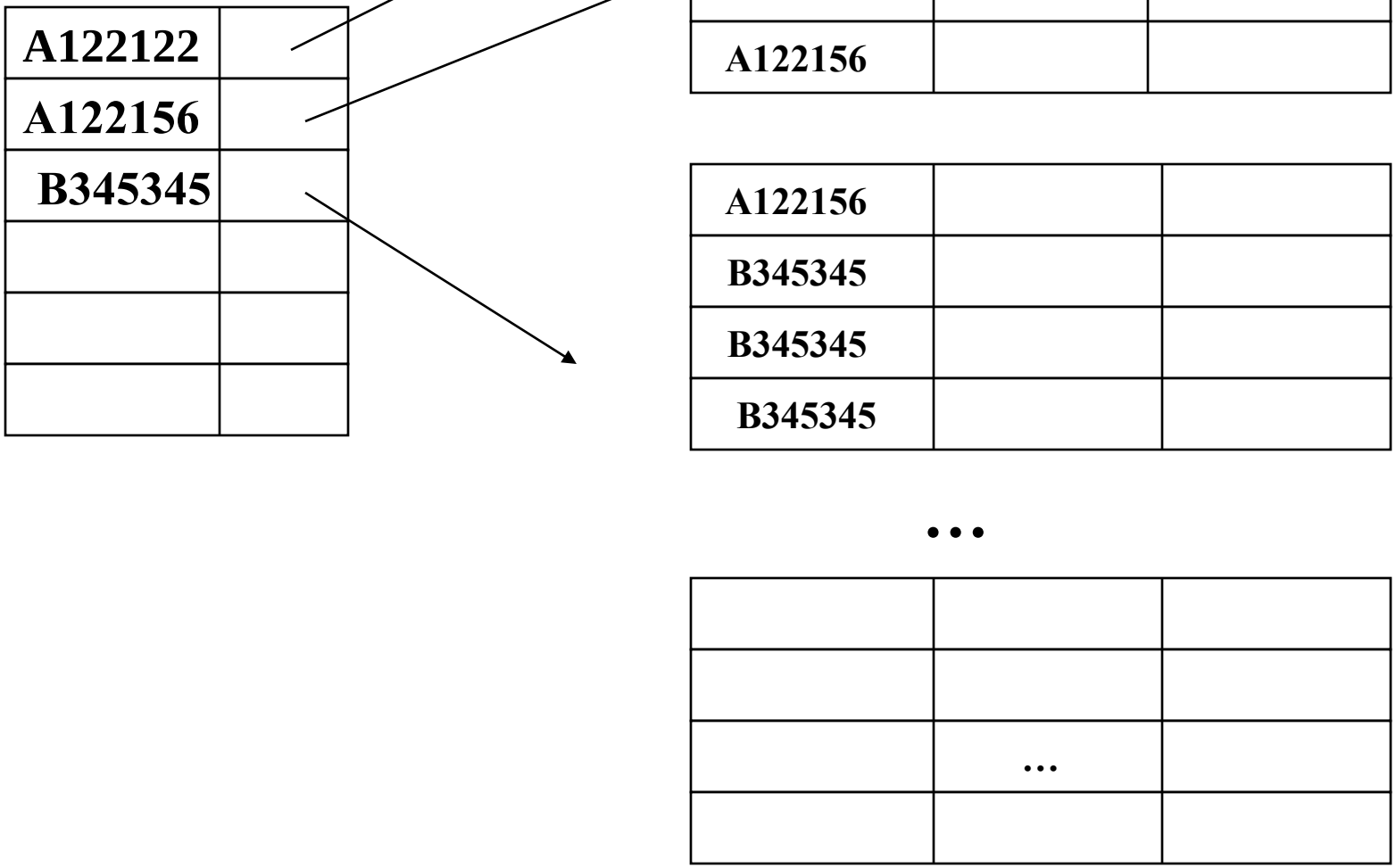
- Ευρετήριο Συστάδα
  - Ορίζεται σε ένα ταξινομημένο αρχείο
  - Το αρχείο δεδομένων είναι ταξινομημένο σε ένα *πεδίο που δεν είναι κλειδί*, σε αντίθεση από το πρωτεύον ευρετήριο, που απαιτεί ότι το πεδίο ταξινόμησης στο αρχείο δεδομένων έχει μια διακριτή τιμή για κάθε εγγραφή.
  - Περιλαμβάνει μια καταχώρηση ευρετηρίου *για κάθε διακριτή τιμή* του πεδίου· η καταχώρηση ευρετηρίου δείχνει στο πρώτο μπλοκ δεδομένων που περιέχει εγγραφές με αυτή την τιμή πεδίου.
  - Είναι ένα ακόμη παράδειγμα *μη πυκνού* ευρετηρίου όπου η εισαγωγή και η διαγραφή είναι σχετικά εύκολη με ένα ευρετήριο συστάδα.

Ένα παράδειγμα ευρετηρίου συστάδας

Ένα ευρετήριο συστάδα στο πεδίο ταξινόμησης ΚΩΔ_ΤΜΗΜ που δεν είναι κλειδί στο αρχείο ΕΡΓΑΖΟΜΕΝΟΣ.



Ευρετήριο



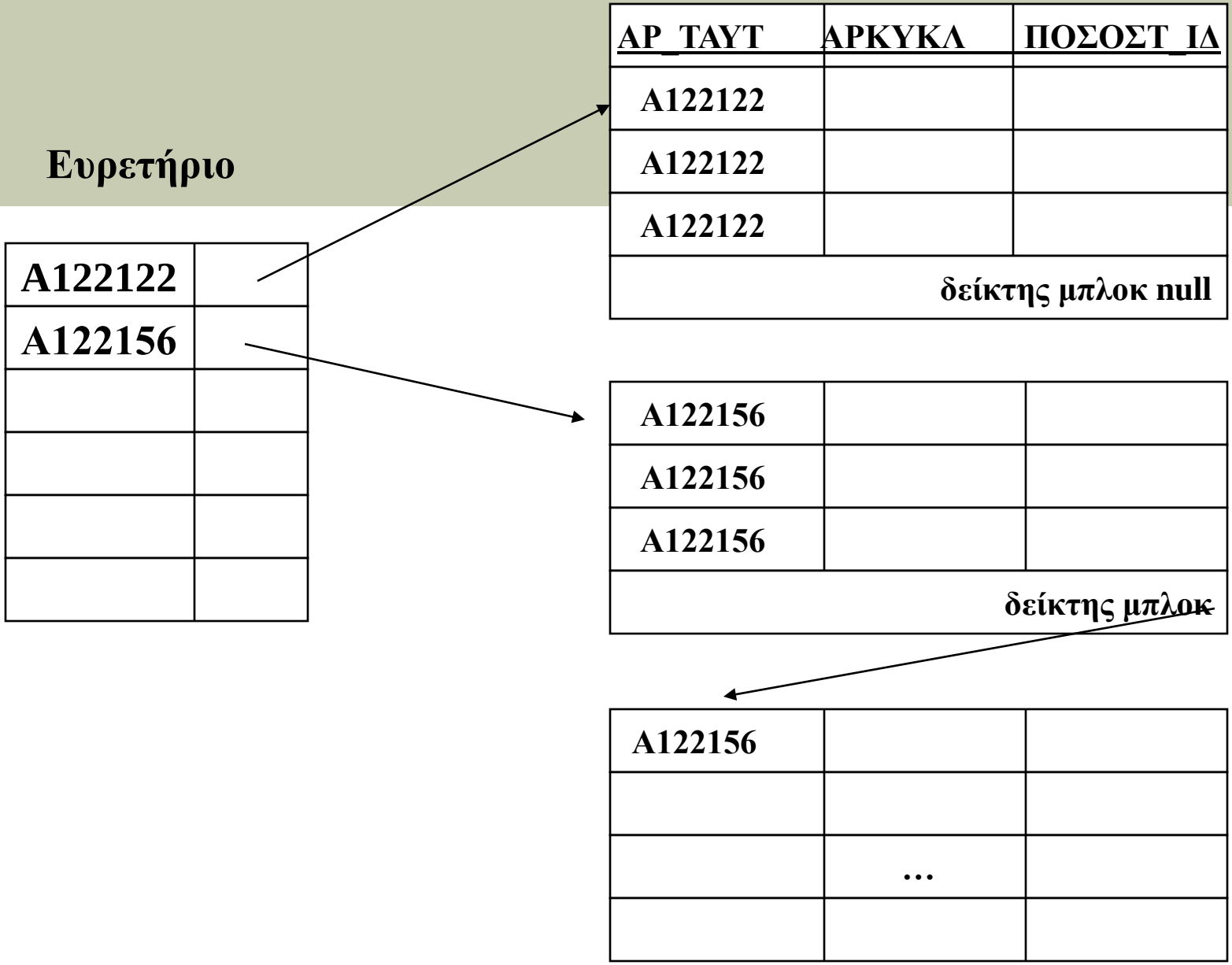
A122122	
A122156	
B345345	

AP_ΤΑΥΤ	ΑΡΚΥΚΛ	ΠΟΣΟΣΤ_ΙΔ
A122122		
A122122		
A122122		
A122156		

A122156		
B345345		
B345345		
B345345		

• • •

	…	



Ένα ακόμη παράδειγμα ευρετηρίου συστάδας

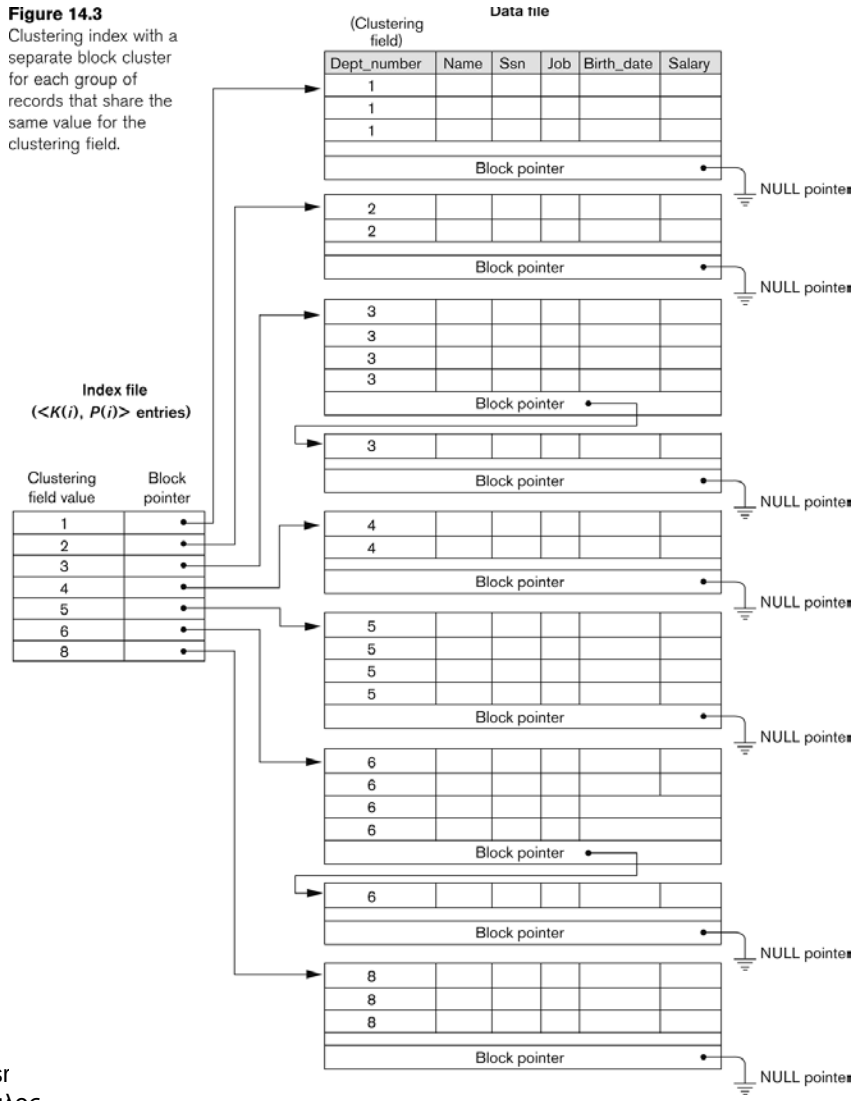


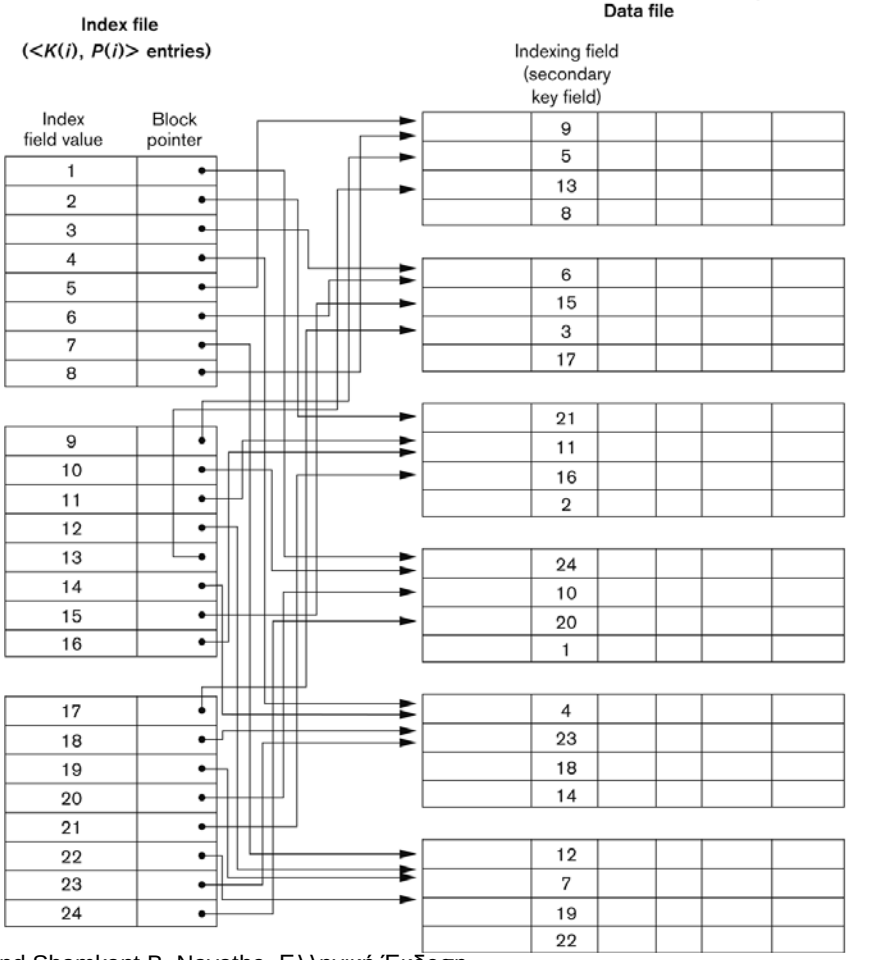
Figure 14.3
Clustering index with a separate block cluster for each group of records that share the same value for the clustering field.

Τύποι Ευρετηρίων ενός Επιπέδου

- Δευτερεύον Ευρετήριο
  - Ένα δευτερεύον ευρετήριο υποστηρίζει ένα δευτερεύοντα τρόπο προσπέλασης ενός αρχείου για το οποίο υπάρχει ήδη πρωτεύουσα οργάνωση.
  - Το δευτερεύον ευρετήριο μπορεί να είναι σε ένα πεδίο που είναι υποψήφιο κλειδί και έχει μοναδική τιμή σε κάθε εγγραφή, ή ένα πεδίο που δεν είναι κλειδί με διπλές τιμές.
  - Το ευρετήριο είναι ένα ταξινομημένο αρχείο με δύο πεδία.
    - Το πρώτο πεδίο είναι ίδιου τύπου δεδομένων με κάποιο **πεδίο που δεν είναι κλειδί** του αρχείου δεδομένων και είναι το πεδίο ευρετηρίου.
    - Το δεύτερο πεδίο είναι ή δείκτης **μπλοκ** ή δείκτης εγγραφής .
    - Μπορεί να υπάρχουν *πολλά* δευτερεύοντα ευρετήρια (και επομένως, πεδία ευρετηρίασης) για το ίδιο αρχείο.
  - Περιλαμβάνει μια καταχώρηση *για κάθε εγγραφή* στο αρχείο δεδομένων· επομένως, είναι ένα *πυκνό ευρετήριο.*

Παράδειγμα ενός Πυκνού Πρωτεύοντος Ευρετηρίου

Figure 14.4
A dense secondary index (with block pointers) on a nonordering key field of a file.



Ένα Παράδειγμα Δευτερεύοντος Ευρετηρίου

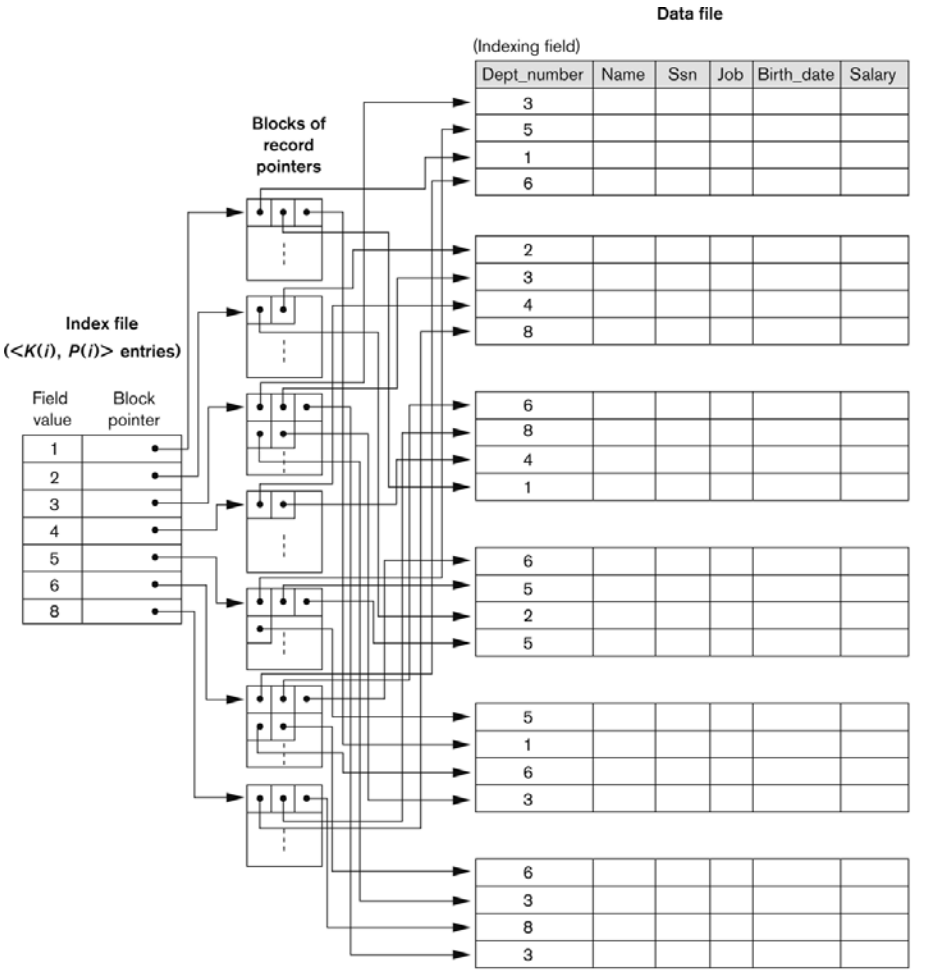


Figure 14.5
A secondary index (with record pointers) on a nonkey field implemented using one leve of indirection so that index entries are of fixed length and have unique field values

Ιδιότητες Τύπων Ευρετηρίων

Τύπος Ευρετηρίου	Πλήθος Καταχωρήσεων του Ευρετηρίου (Πρώτου Επιπέδου)	Πυκνό η Μη Πυκνό	Άγκυρες μπλοκ στο Αρχείο Δεδομένων
Πρωτεύον Ευρετήριο	Πλήθος μπλοκ του αρχείου δεδομένων	Μη πυκνό	Ναι
Ευρετήριο Συστάδων	πλήθος διαφορετικών τιμών του πεδίου ευρετηρίασης	Μη Πυκνό	Ναι/όχι[α]
Δευτερεύον Ευρετήριο πάνω σε πεδίο-κλειδί	Πλήθος εγγραφών αρχείου δεδομένων	Πυκνό	όχι
Δευτερεύον Ευρετήριο πάνω σε πεδίο-μη κλειδί	Πλήθος εγγραφών[β] η πλήθος διαφορετικών τιμών του πεδίου ευρετηριοποίησης[γ]	Πυκνό ή μη πυκνό	όχι

Ευρετήρια Πολλών Επιπέδων

- Επειδή ένα ευρετήριο ενός επιπέδου είναι ένα ταξινομημένο αρχείο, μπορούμε να δημιουργήσουμε ένα πρωτεύον ευρετήριο *στο ίδιο το ευρετήριο*·
  - Στην περίπτωση αυτή, το αρχικό αρχείο ευρετηρίου ονομάζεται *πρώτο επίπεδο ευρετηρίου* και το ευρετήριο του ευρετηρίου ονομάζεται *δεύτερο επίπεδο ευρετηρίου*.
- Μπορούμε να επαναλάβουμε τη διαδικασία αυτή, δημιουργώντας ένα τρίτο, τέταρτο ... επίπεδο, μέχρι που όλες οι καταχωρήσεις στο *ψηλότερο επίπεδο* να χωράνε σε ένα μπλοκ μνήμης.
- Ένα ευρετήριο πολλών επιπέδων μπορεί να δημιουργηθεί οποιοδήποτε τύπο ευρετηρίου πρώτου επιπέδου (πρωτεύον, δευτερεύον, συστάδα) όσο το πρώτο επίπεδο ευρετηρίου αποτελείται από *περισσότερα από ένα* μπλοκ δίσκου

Πρωτεύον Ευρετήριο δύο επιπέδων

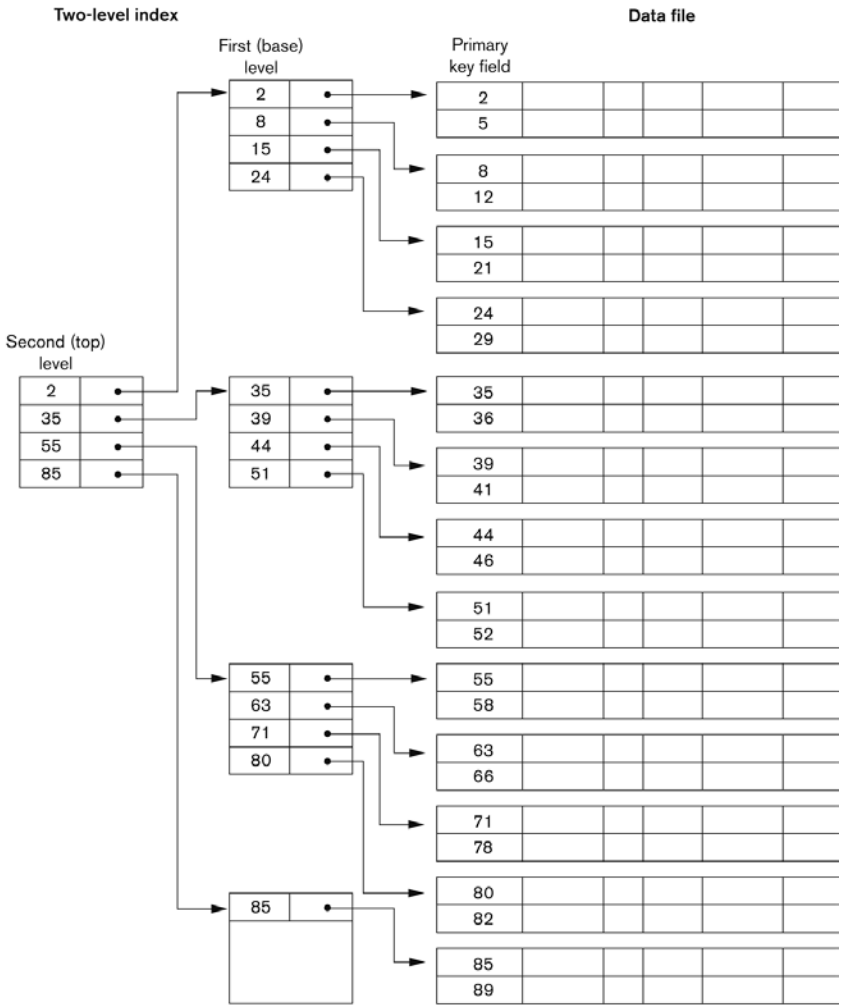
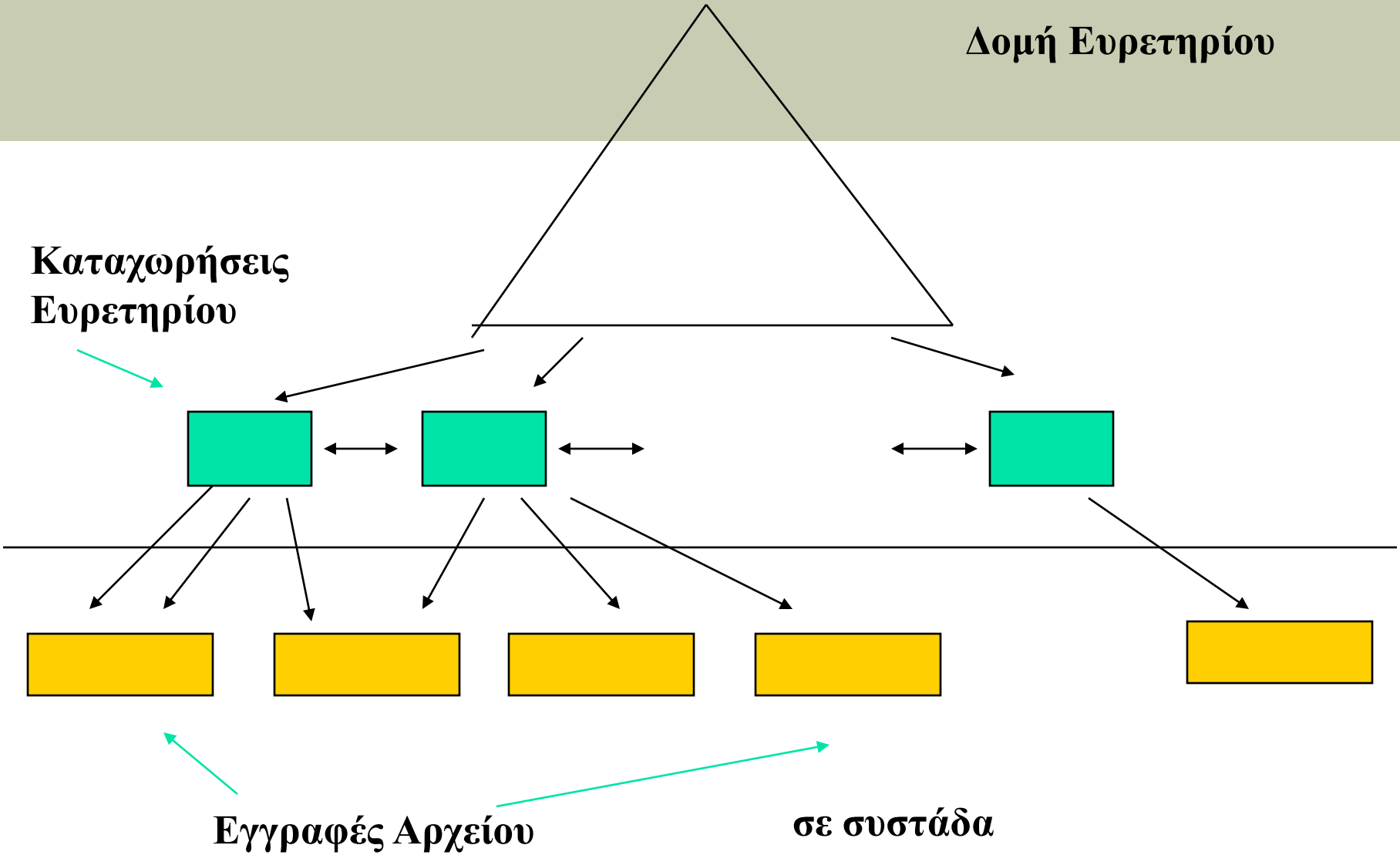


Figure 14.6
A two-level primary index resembling ISAM (Index Sequential Access Method) organization.

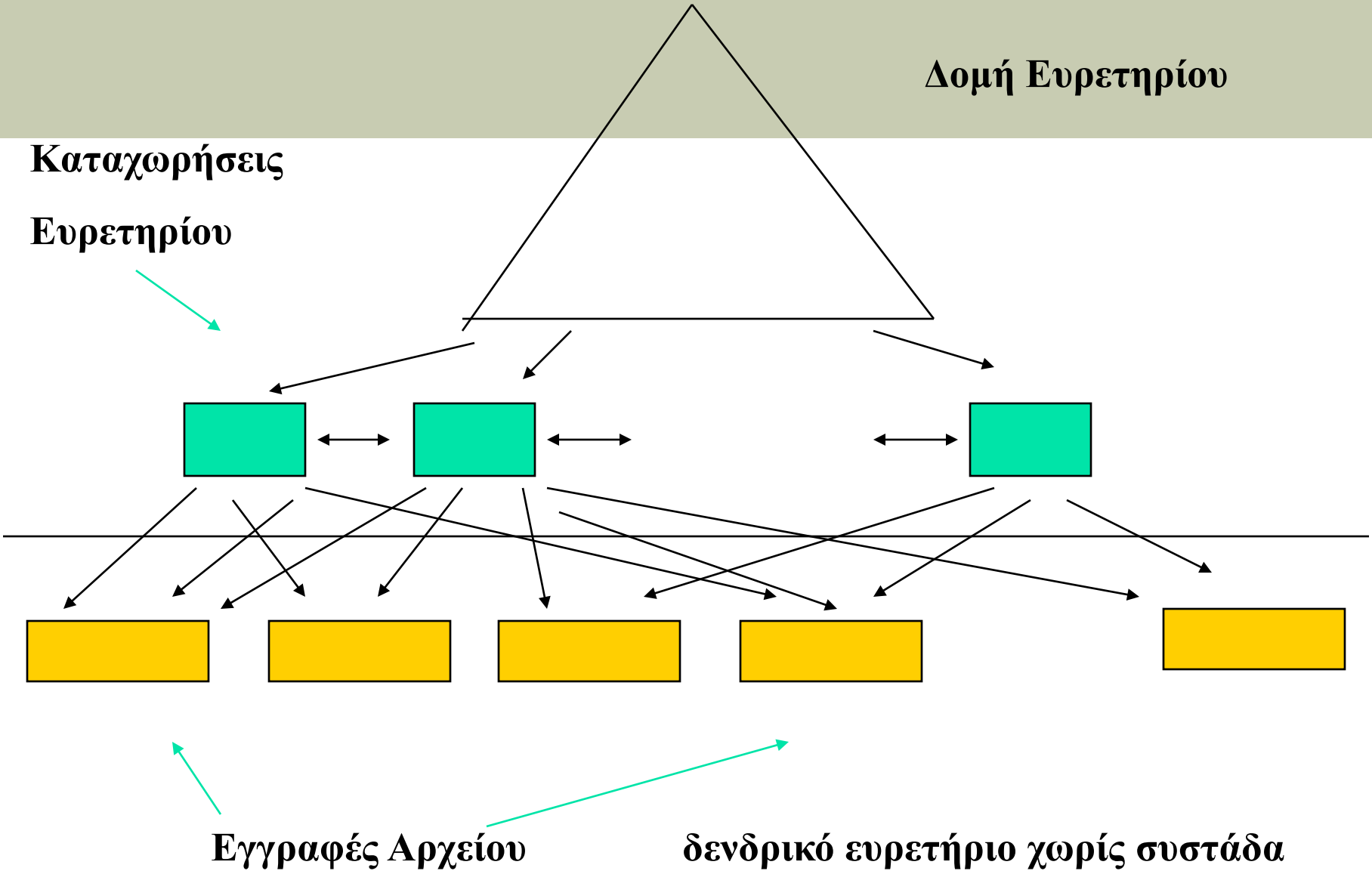
Ευρετήρια Πολλών Επιπέδων

- Ένα τέτοιο ευρετήριο πολλών επιπέδων είναι μια μορφή *δένδρου αναζήτησης*
  - Ωστόσο, η εισαγωγή και η διαγραφή καταχωρήσεων στο ευρετήριο αποτελεί σοβαρό πρόβλημα επειδή κάθε επίπεδο του ευρετηρίου είναι ένα *ταξινομημένο αρχείο*.

Δομή Ευρετηρίου



Δομή Ευρετηρίου



Μπορούμε να κατασκευάσουμε ευρετήρια σε δευτερεύοντα πεδία. Στην περίπτωση αυτή έχουμε διαφορετικούς τρόπους υλοποίησης

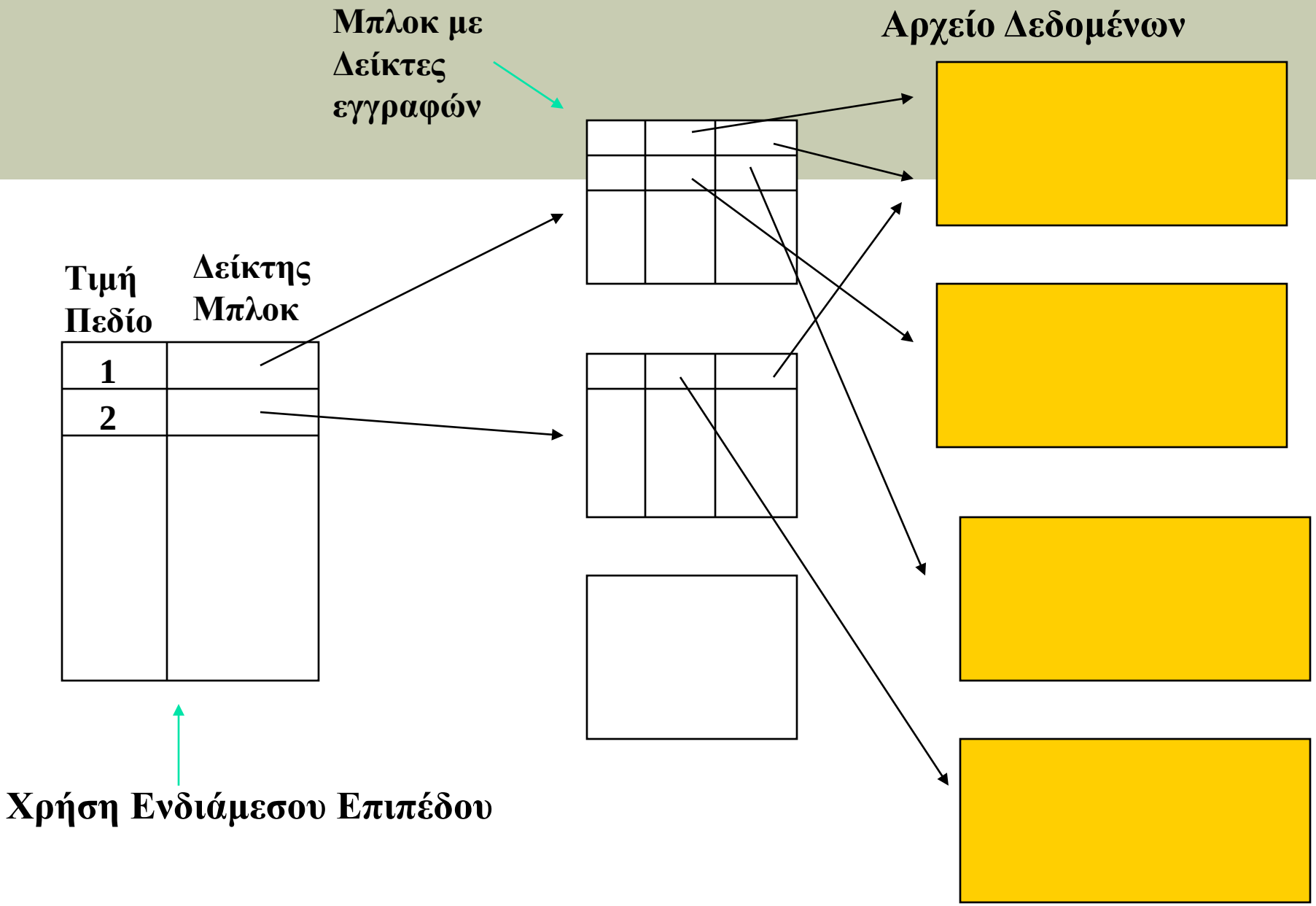
- Πυκνό ευρετήριο. Μια καταχώρηση για κάθε εγγραφή.
- Μεταβλητού μήκους εγγραφές που για κάθε τιμή του πεδίου περιέχουν μια λίστα από διευθύνσεις.
- Χρησιμοποίηση ενδιάμεσου επιπέδου ευρετηρίου (που χρησιμοποιείται και πιο συχνά). Υπάρχει μια καταχώρηση για κάθε διαφορετική τιμή του πεδίου με δείκτη σε δεύτερο επίπεδο που υπάρχει μια λίστα από διευθύνσεις

Τιμή_Πεδίου	Διευθ_Μπλοκ

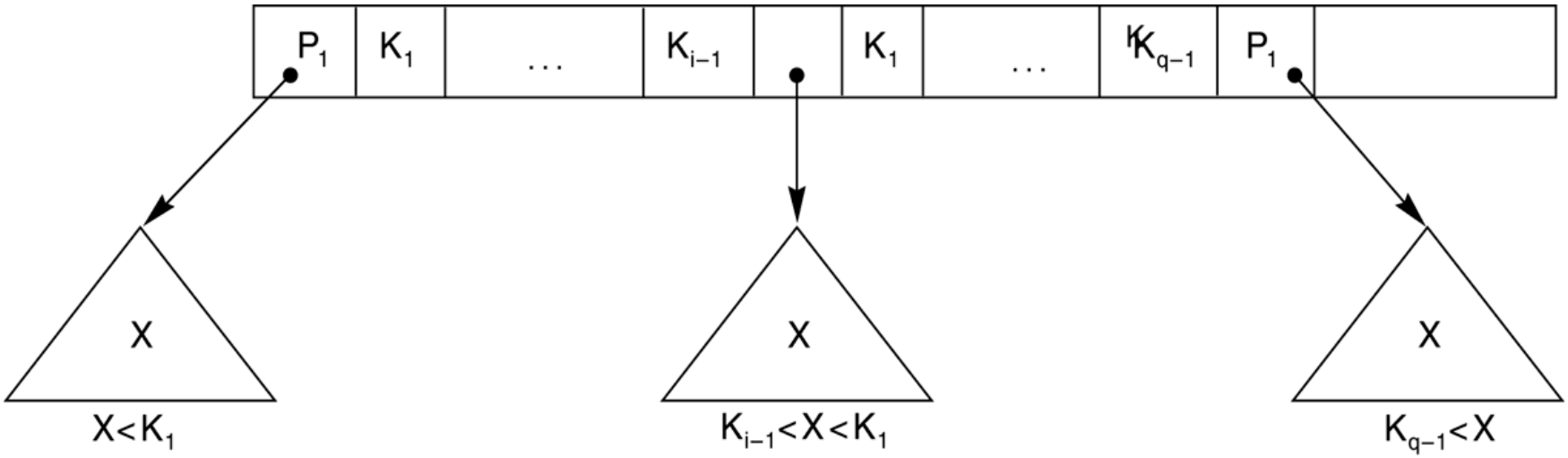
Ξεχωριστή καταχώρηση για κάθε εγγραφή

Επαναλαμβανόμενο πεδίο με τις διευθύνσεις

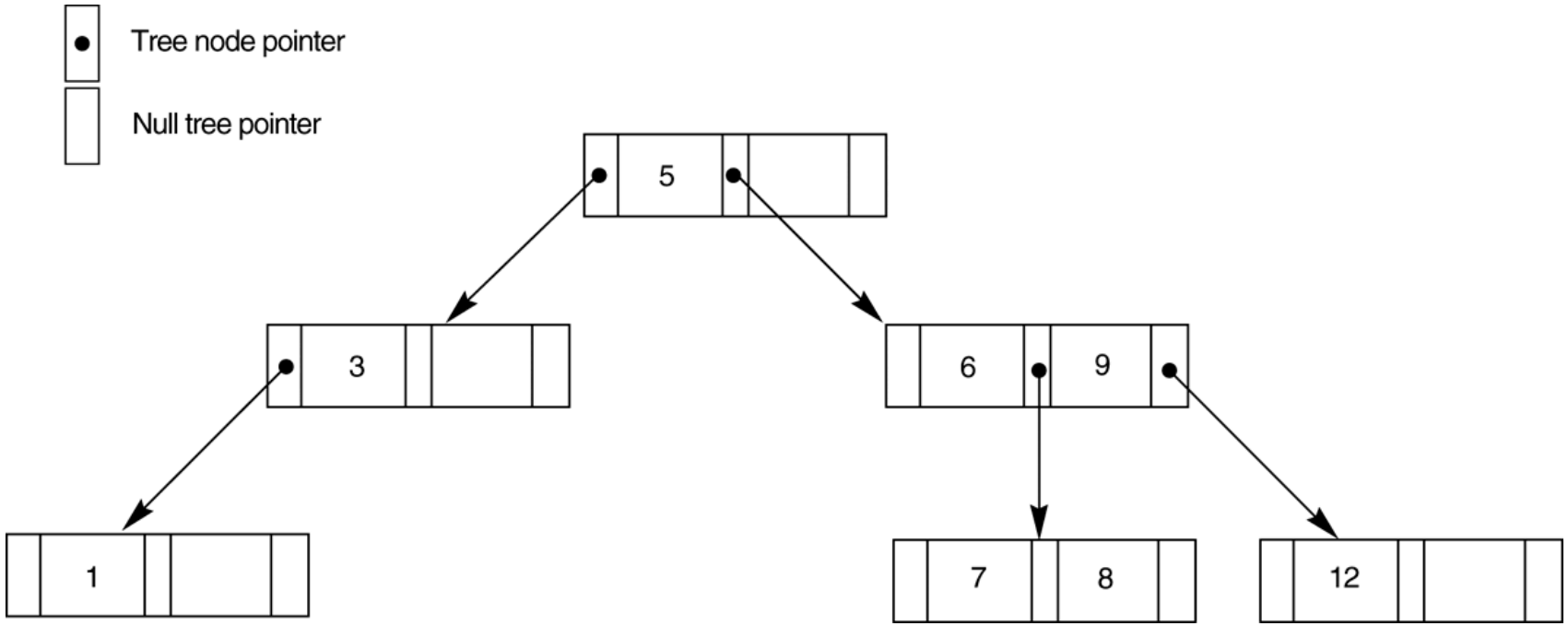
Τιμή_Πεδίου	Διευθύνσεις_Μπλοκ



Ένας κόμβος σε ένα δένδρο αναζήτησης με δείκτες στα υποδένρα του



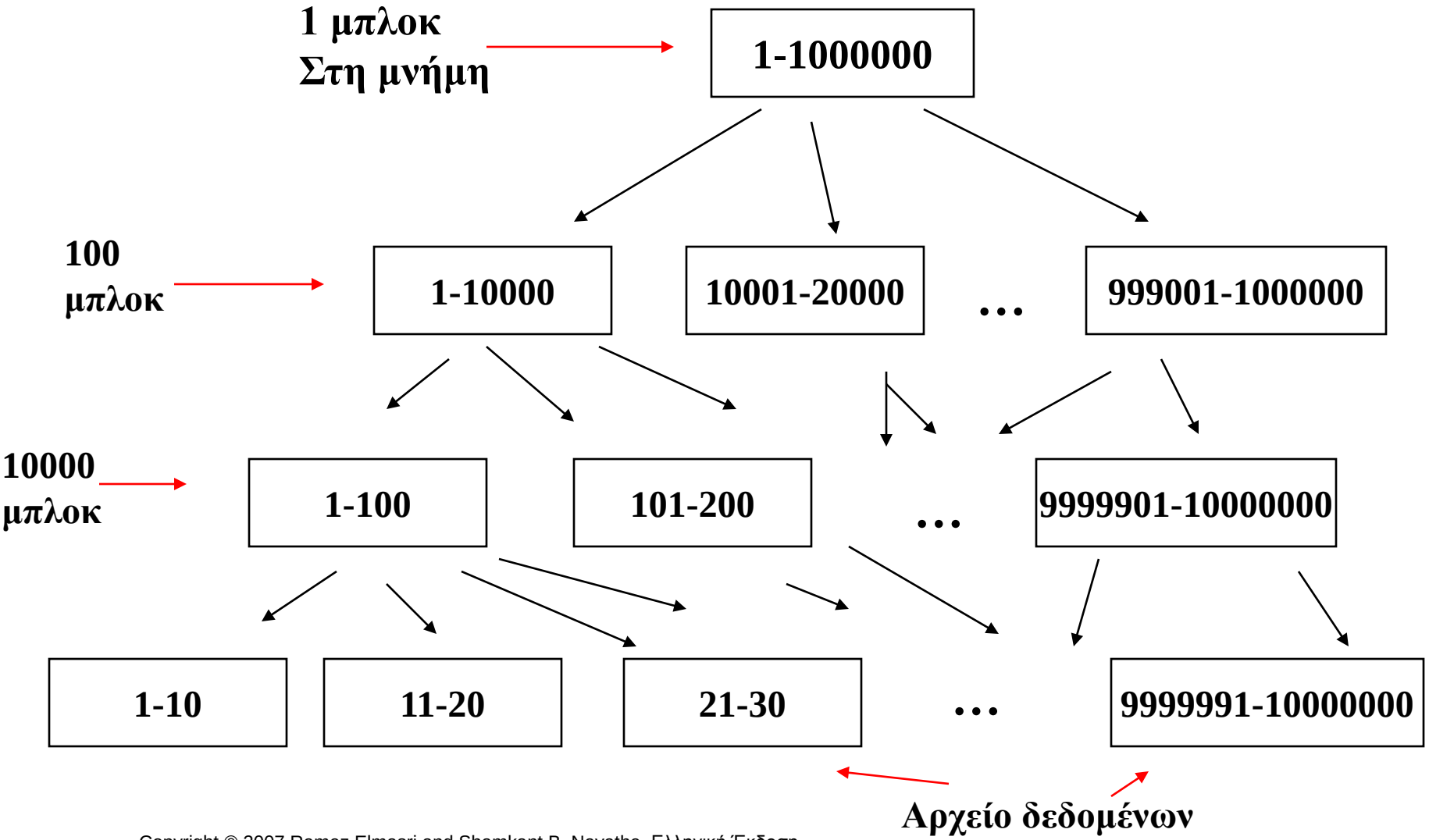
Ένα δένδρο αναζήτησης τάξεως p = 3.



ISAM αρχεία (Indexed Sequential Access Method)

- Δημιουργούμε ένα δεύτερο αρχείο με μια εγγραφή για κάθε σελίδα του αρχικού αρχείου, της μορφής <πρώτο κλειδί στη σελίδα, δείκτης στη σελίδα> ταξινομημένο προς το κλειδί
- Το ζευγάρι <key, pointer> αναφέρονται σαν καταχώρηση. Κάθε κλειδί στο ευρετήριο αποτελεί διαχωριστή για τα περιεχόμενα των σελίδων που δείχνουν οι δείκτες.

Πολυεπίπεδα Ευρετήρια



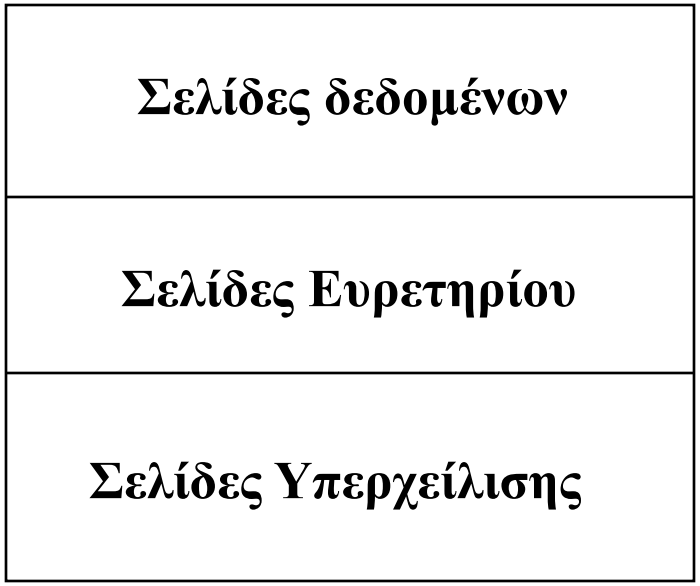
Αρχεία ISAM (Indexed Sequential Access Method)

Σελίδα του ευρετηρίου

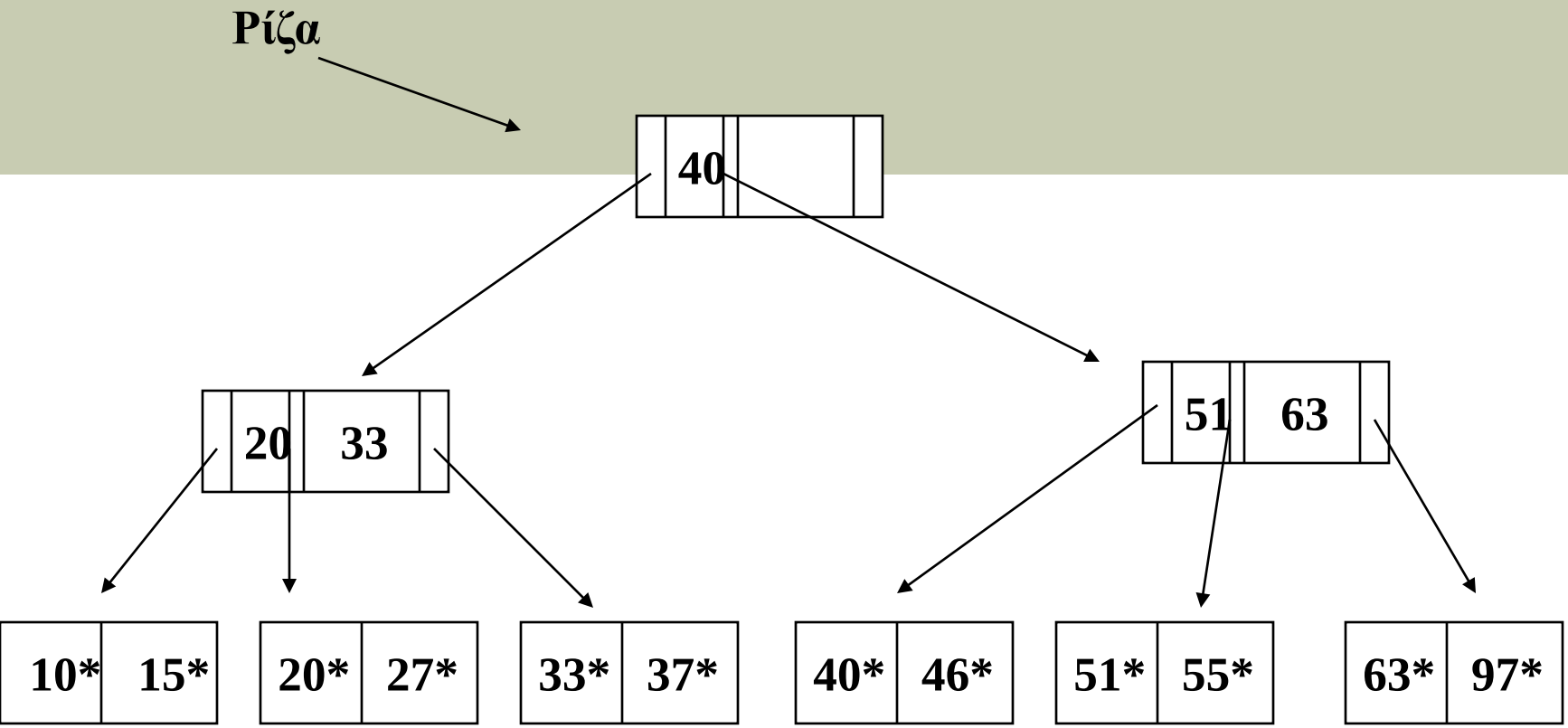
Το ζεύγος κλειδί, δείκτης αποτελεί μια καταχώρηση του ευρετηρίου

(k_i, p_i)

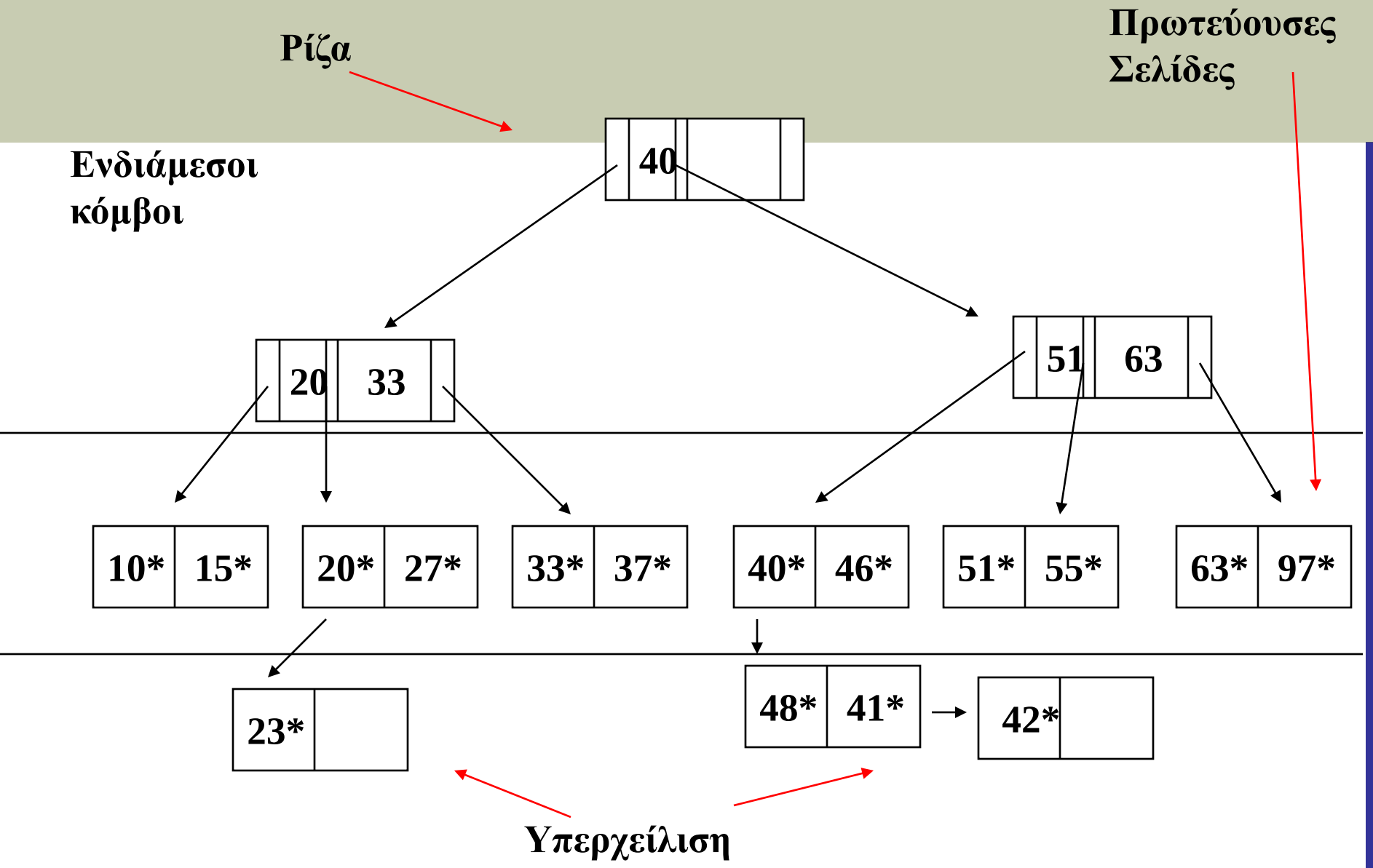
Αντί να γίνεται αναζήτηση στο αρχείο δεδομένων η αναζήτηση γίνεται στο ευρετήριο. Εκεί μπορεί να γίνει δυαδική αναζήτηση σε μικρότερο αρχείο. Η ιδέα είναι αν μπορώ να μεταφέρω την οργάνωση του ευρετηρίου αναδρομικά μέχρι το αρχικό βοηθητικό αρχείο να χωράει στη μνήμη.



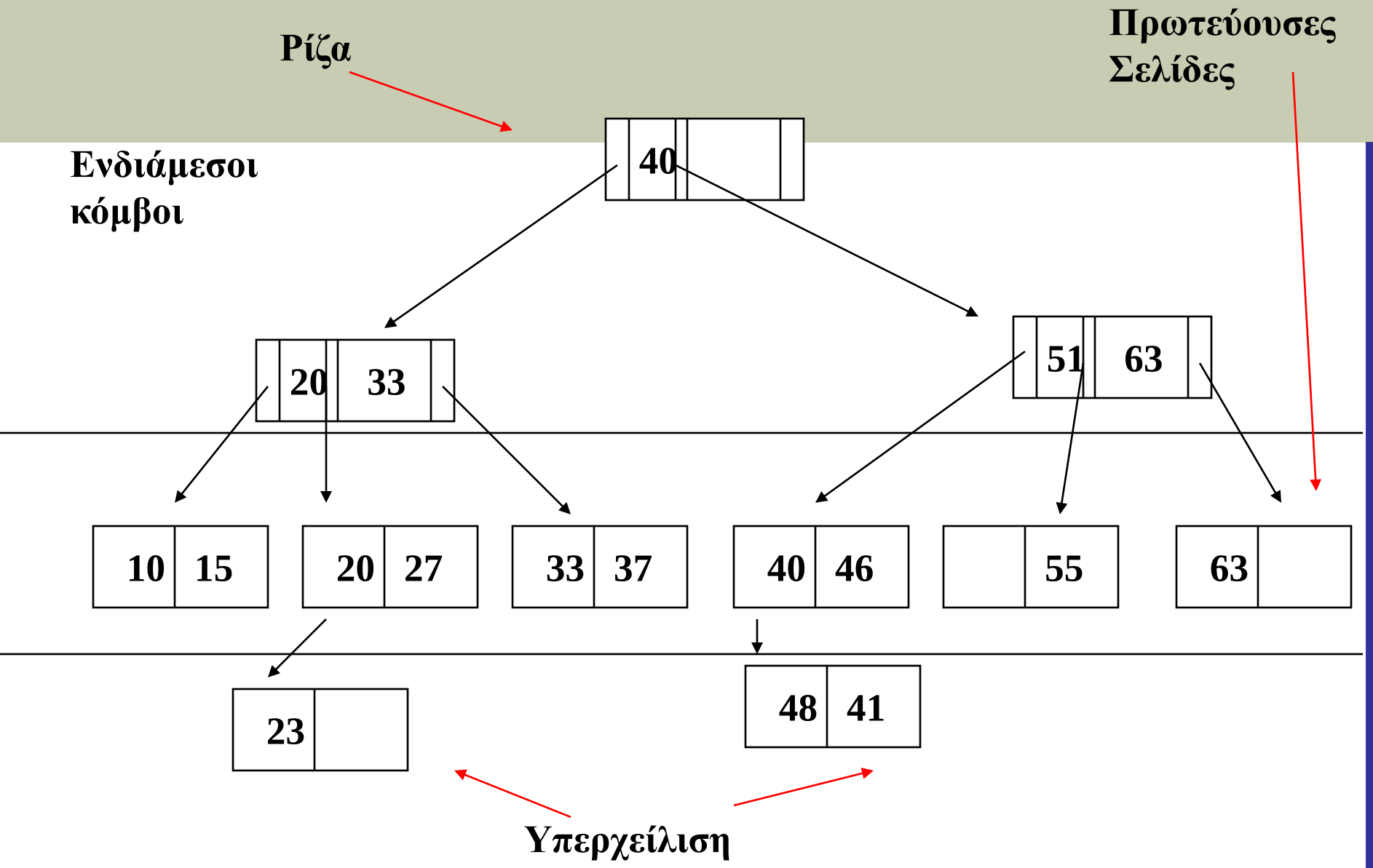
Εναλλακτικά μπορούμε να υλοποιήσουμε την μέθοδο θέτοντας στα φύλα μόνο τιμή κλειδιού και διεύθυνση στο αρχείο.



Εισαγωγή 23, 48, 41, 42



Διαγραφή 42, 51, 97



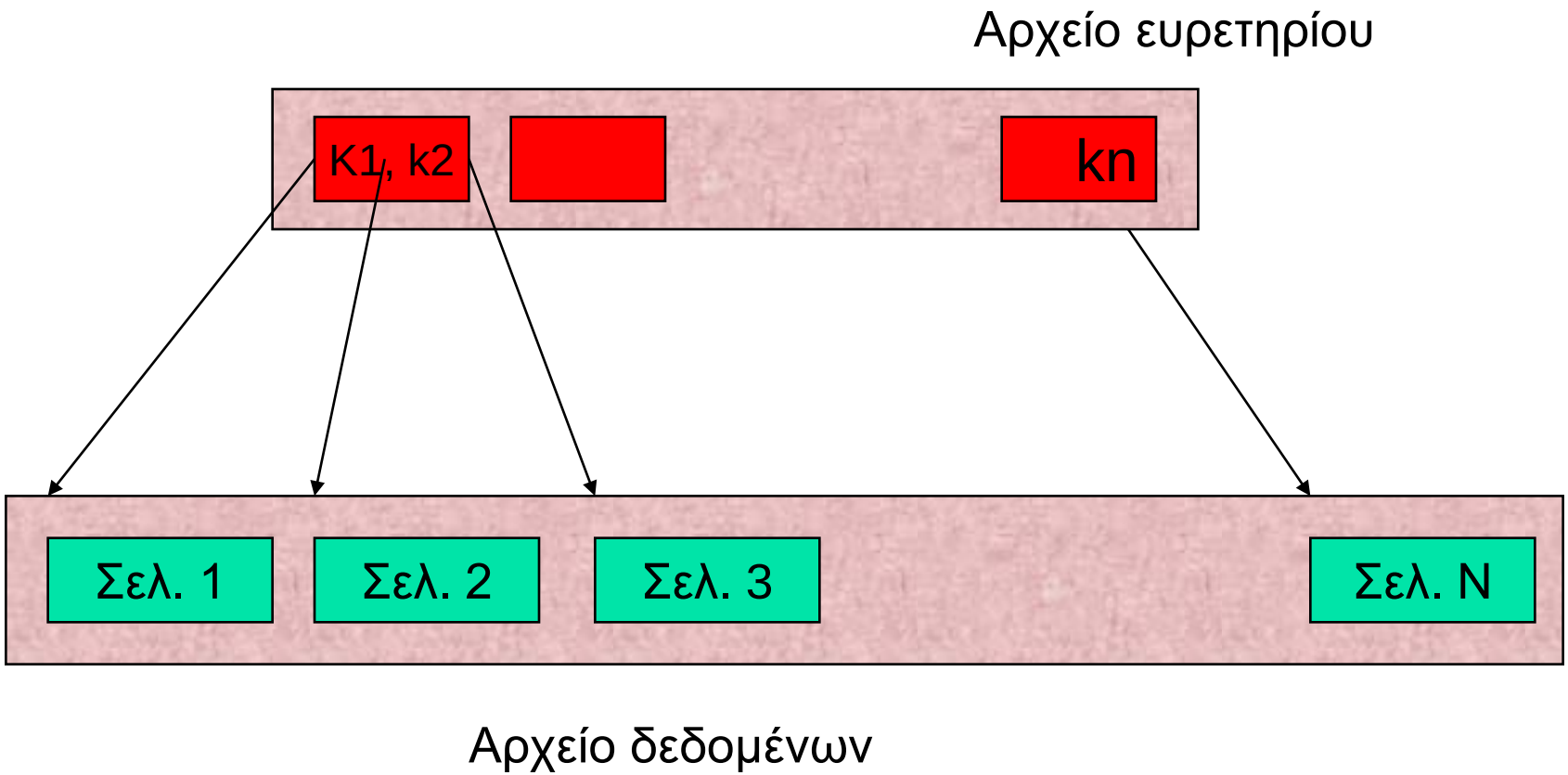
Το κόστος της αναζήτησης σε προσπελάσεις στο δίσκο είναι όσο και το ύψος του ευρετηρίου συν την προσπέλαση στη σελίδα των δεδομένων συν τις προσπελάσεις στην υπερχείλιση.
Όταν δημιουργηθεί το ISAM αρχείο οι διαγραφές και οι εισαγωγές επηρεάζουν τα φύλα. Σαν αποτέλεσμα δημιουργούνται μακρές λίστες από σελίδες υπερχείλισης.

Η αναζήτηση στο ευρετήριο εξαρτάται από το ύψος του. Αν επομένως το πλήθος των πεδίων ανά σελίδα ευρετηρίου είναι F (λέγεται fun out) και αν N το πλήθος των πρωτευόντων σελίδων φύλων τότε η αναζήτηση απαιτεί $\log_F N$ προσπελάσεις.

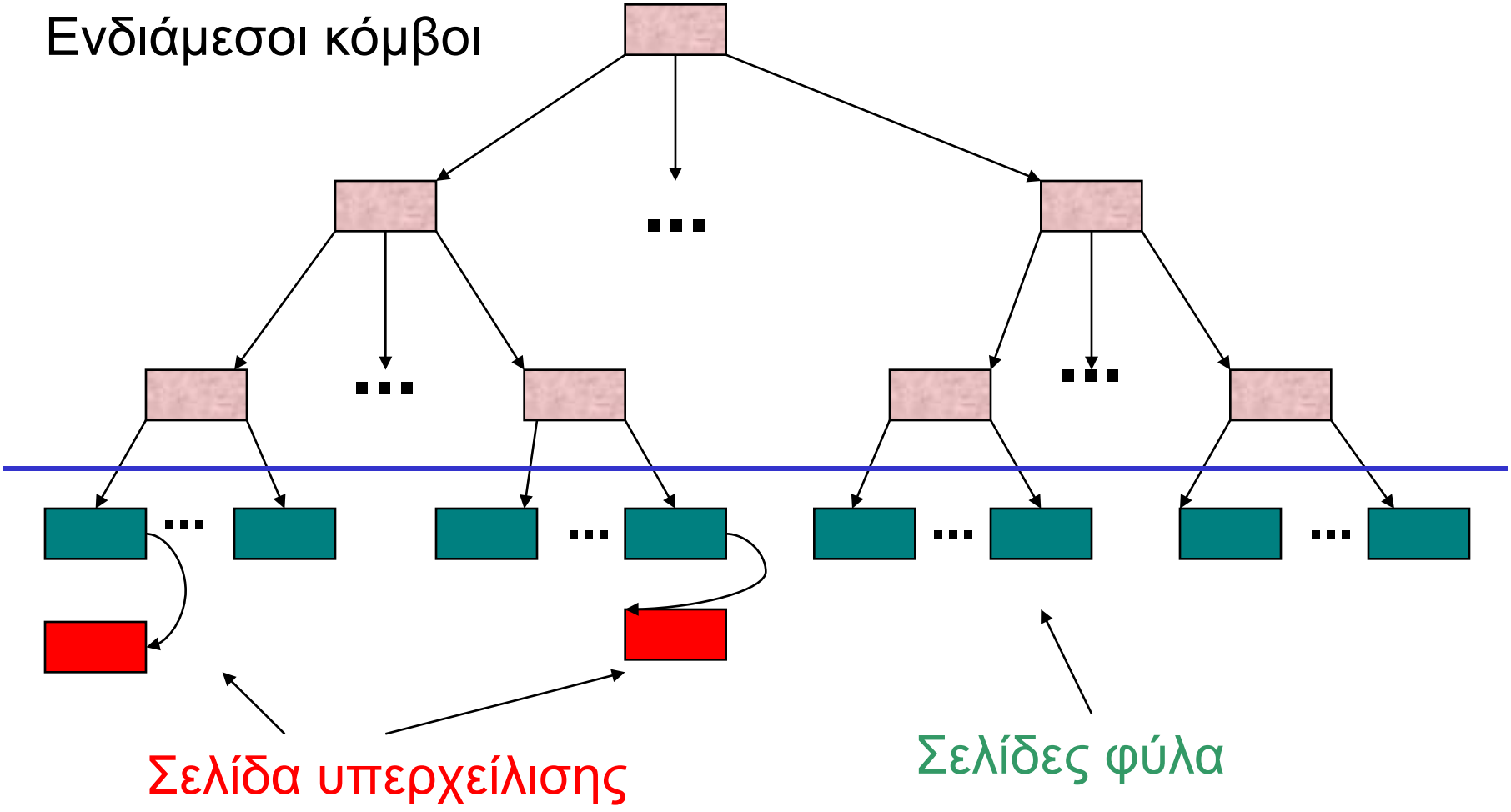
Σαν μια λύση στο πρόβλημα αυτό είναι η φόρτωση των φύλων μόνο σε ένα ποσοστό (80% για παράδειγμα).

Μια τέτοια οργάνωση είναι τα VSAM (Virtual Storage Access Method) αρχεία που αναπτύχθηκαν από την IBM και θεωρούνται προκάτοχοι των B-δένδρων.

ISAM αρχεία (συν)



ISAM αρχεία (συν). Το προηγούμενο βήμα μπορεί να επεκταθεί στο κτίσιμο ενός ευρετηρίου πολλών επιπέδων



Δυναμικά Πολυεπίπεδα Ευρετήρια με Χρήση Β-Δένδρων και Β+-Δένδρων

- Τα περισσότερα ευρετήρια πολλών επιπέδων χρησιμοποιούν δομές δεδομένων Β-δένδρων ή Β+-δένδρων λόγω του προβλήματος εισαγωγής και διαγραφής
  - Αυτό αφήνει σε κάθε κόμβο του δένδρου (μπλοκ στο δίσκο) ελεύθερο χώρο για νέες καταχωρήσεις
- Αυτές οι δομές δεδομένων αποτελούν παραλλαγές των δένδρων αναζήτησης που υποστηρίζουν αποτελεσματική εισαγωγή και διαγραφή νέων τιμών αναζήτησης.
- Στις δομές δεδομένων Β-δένδρων και Β+-δένδρων, κάθε κόμβος αντιστοιχεί σε ένα μπλοκ του δίσκου
- Κάθε κόμβος είναι από γεμάτος ή μέχρι τη μέση

Δυναμικά Πολυεπίπεδα Ευρετήρια με Χρήση Β-Δένδρων και Β+-Δένδρων(συν.)

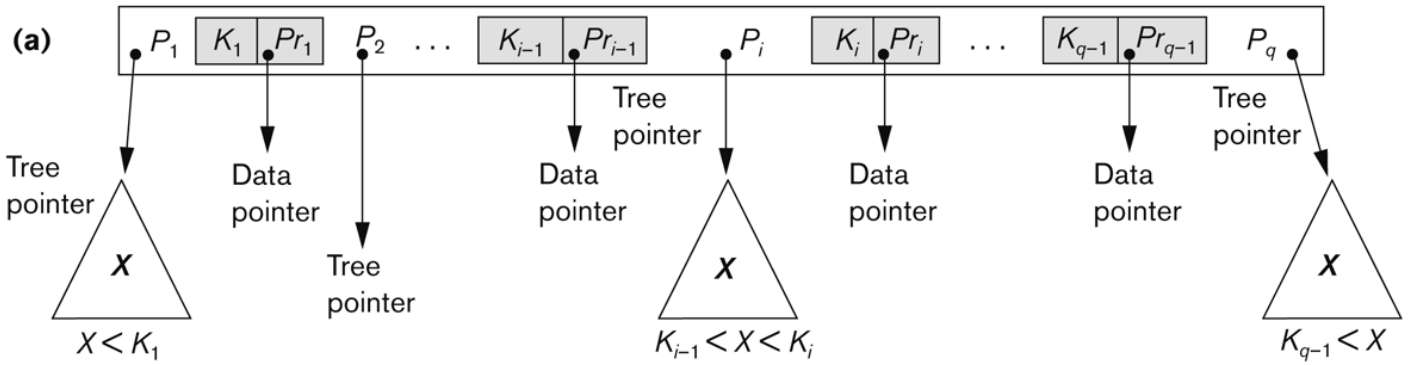
- Η εισαγωγή σε ένα κόμβο που δεν είναι γεμάτος είναι πολύ αποδοτική
  - Αν ένας κόμβος είναι γεμάτος η εισαγωγή προκαλεί διάσπαση σε δύο κόμβους
- Η διάσπαση μπορεί να διαδοθεί σε άλλα επίπεδα του δένδρου
- Μια διαγραφή είναι πολύ αποδοτική αν ένας κόμβος δεν πέφτει κάτω από το μισό
- Αν μια διαγραφή οδηγήσει σε κόμβο με λιγότερες από μισές καταχωρήσεις, πρέπει να συνενωθεί με γειτονικούς κόμβους

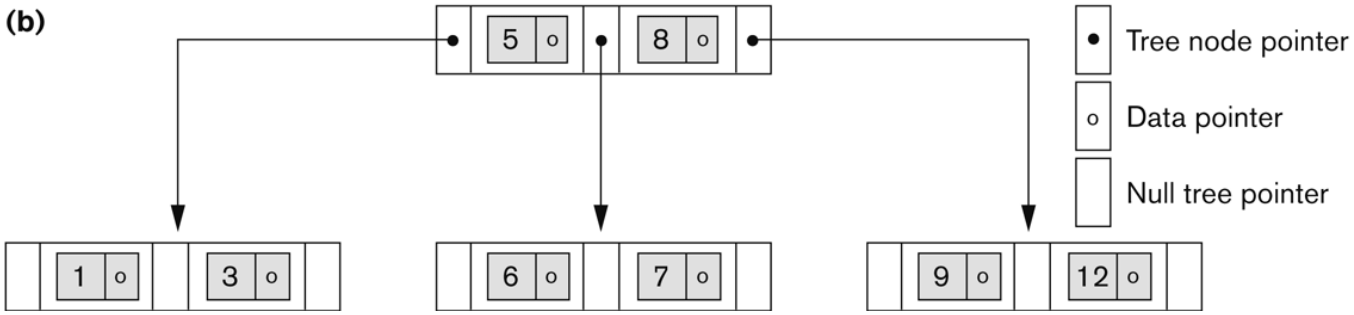
Διαφορά μεταξύ B-δένδρου και B+-δένδρου

- Σε ένα B-δένδρου, υπάρχουν δείκτες προς εγγραφές δεδομένων σε όλα τα επίπεδα του δένδρου
- Σε ένα B+-δένδρο, όλοι οι δείκτες προς τις εγγραφές δεδομένων υπάρχουν στους κόμβους φύλα
- Ένα B+-δένδρο μπορεί να έχει λιγότερα επίπεδα (ή μεγαλύτερη χωρητικότητα τιμών αναζήτησης) από το αντίστοιχο B-δένδρο

Δομές B-δένδρου

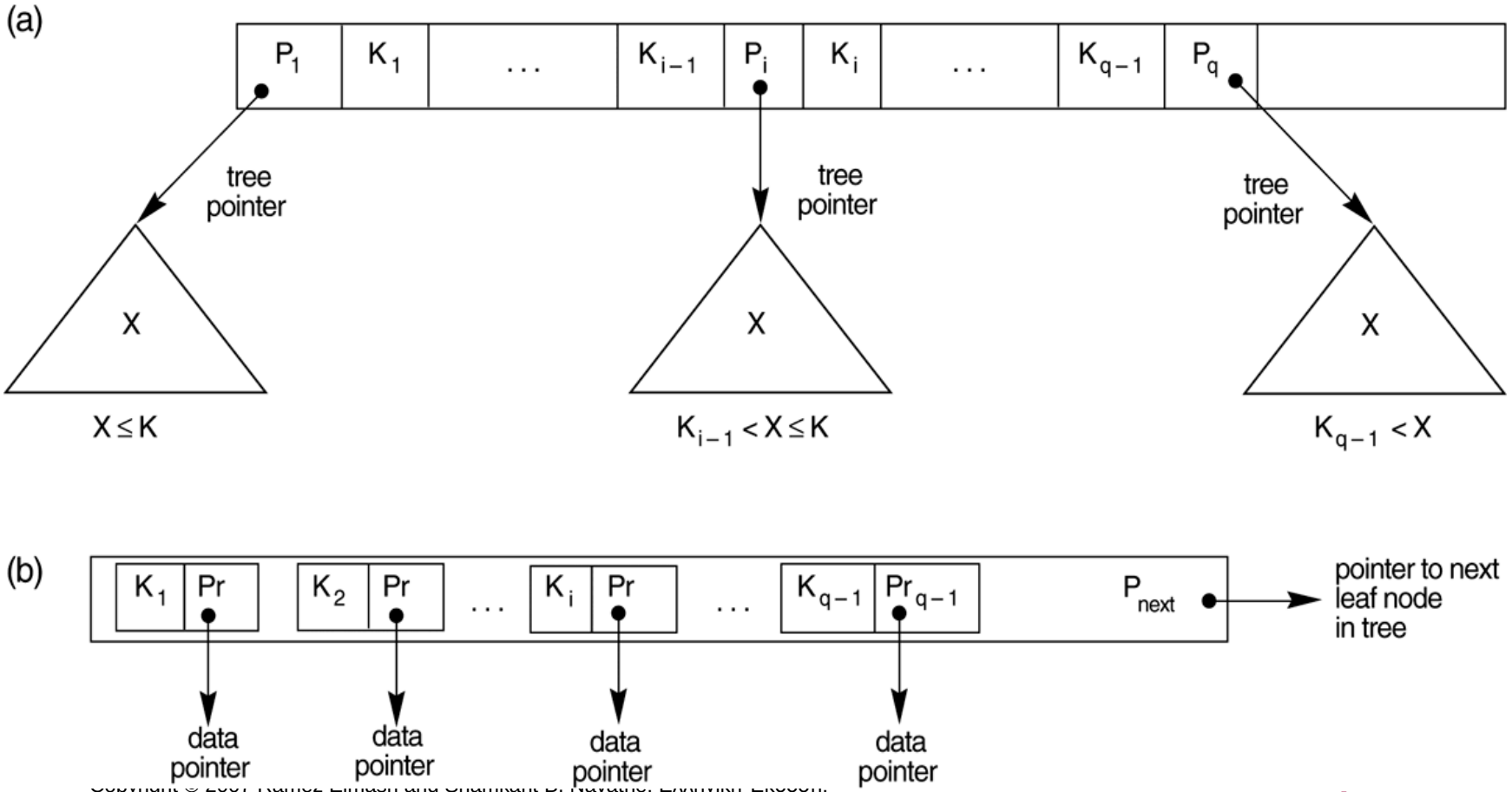
Figure 14.10
B-Tree structures. (a) A node in a B-tree with $q - 1$ search values. (b) A B-tree of order $p = 3$.The values were inserted in the order 8, 5, 1, 7, 3, 12, 9, 6.



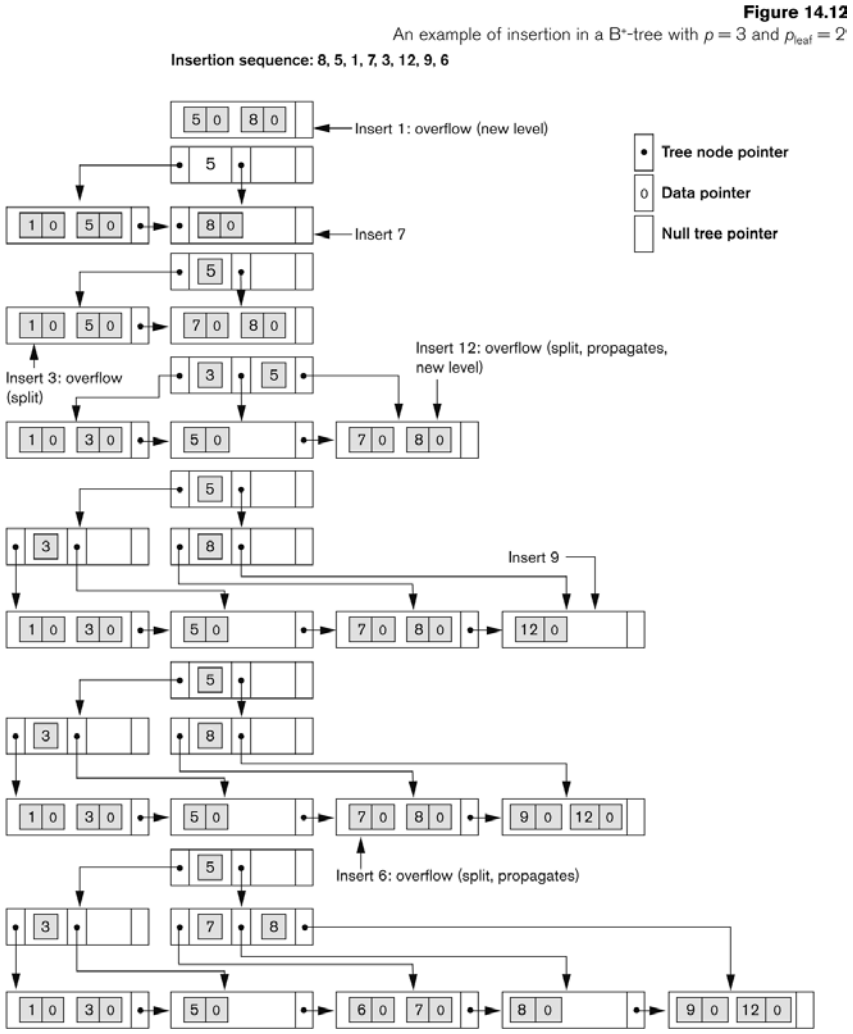


Οι κόμβοι ενός B+-δένδρου

- FIGURE 14.11 The nodes of a B+-tree
  - (a) Internal node of a B+-tree with q –1 search values.
  - (b) Leaf node of a B+-tree with q – 1 search values and q – 1 data pointers.



Ένα παράδειγμα εισαγωγής σε B+-δένδρο



Ένα παράδειγμα διαγραφής σε ένα B+-δένδρο

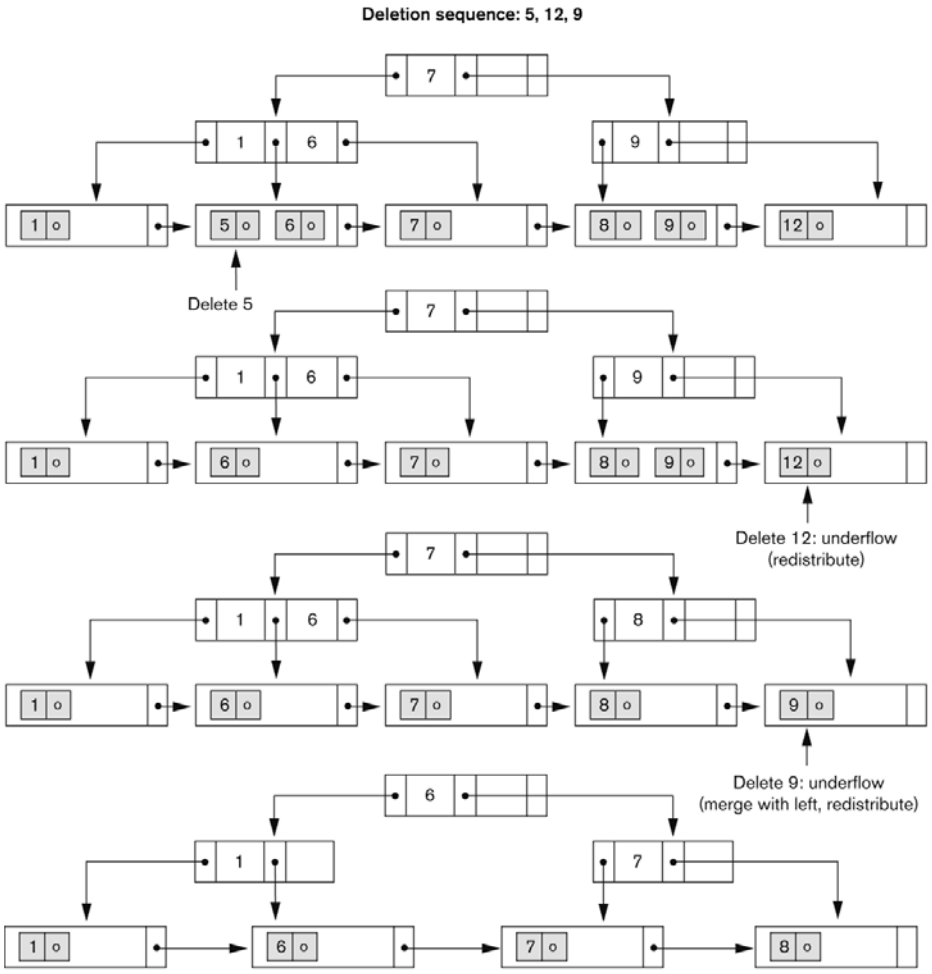


Figure 14.13
An example of deletion from a B⁺-tree.

Σύνοψη

- Τύποι διατεταγμένων ευρετηρίων ενός επιπέδου
  - Πρωτεύοντα ευρετήρια
  - Ευρετήρια Συστάδες
  - Δευτερεύοντα Ευρετήρια
- Ευρετήρια πολλών επειπέδων
- Δυναμικά ευρετήρια πολλών επιπέδων με χρήση B-δένδρων και B+-δένδρων
- Ευρετήρια σε πολλά κλειδιά