ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Ανάκτηση πληροφορίας

Ενότητα 7: Κατάλογοι Υπογραφών

Απόστολος Παπαδόπουλος
Τμήμα Πληροφορικής

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Ο Κατάλογος Υπογραφών (Signature File)

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

Βασικές μέθοδοι εξαγωγής υπογραφών

- WS (word signatures)
- SC (superimposed coding)

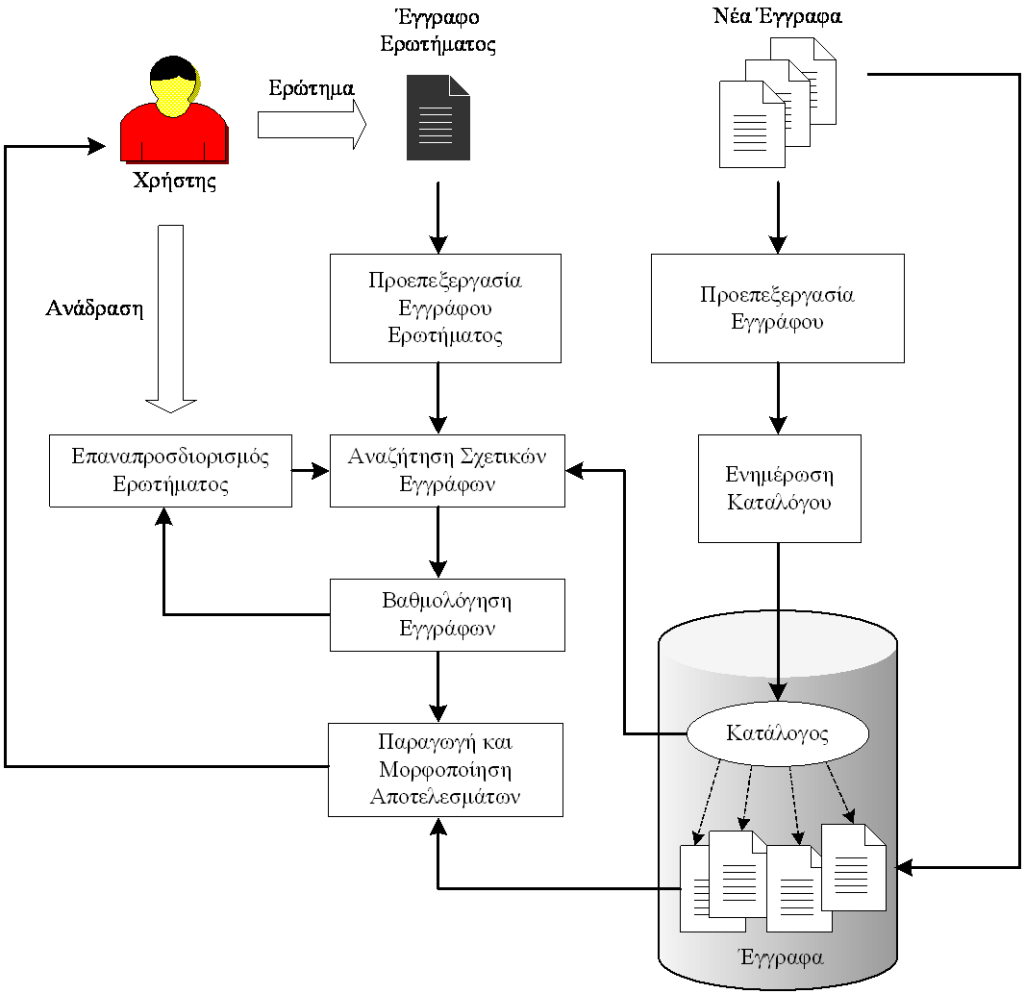
Εξαγωγή υπογραφών με συμπίεση

- BC (bit-block compression)
- RL (run-length encoding)
- VBC (variable bit-block compression)

Οργάνωση αρχείου υπογραφών

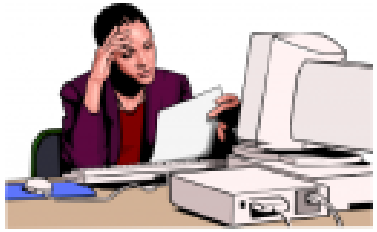
- SSF (sequential signature file)
- BSSF (bit-sliced signature file)
- CBS (compressed bit slices)
- DCBS (doubly compressed bit slices)
- NFD (no false drops)
- μέθοδοι οριζόντιου διαμερισμού

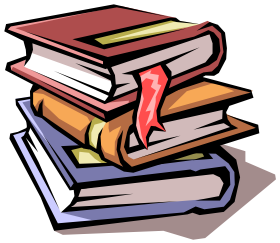
Δομή ενός ΣΑΠ



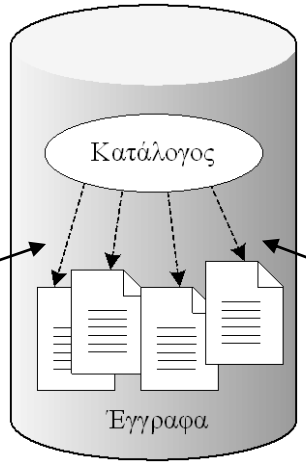
Χρήση Καταλόγων

- Τα συστήματα ανάκτησης σπάνια αναζητούν την πληροφορία απευθείας στη συλλογή εγγράφων. Συνήθως, χρησιμοποιούνται **κατάλογοι** οι οποίοι **επιταχύνουν** τη διαδικασία αναζήτησης.





συλλογή εγγράφων



αναζήτηση καταλόγου

δημιουργία καταλόγου

Υπογραφές

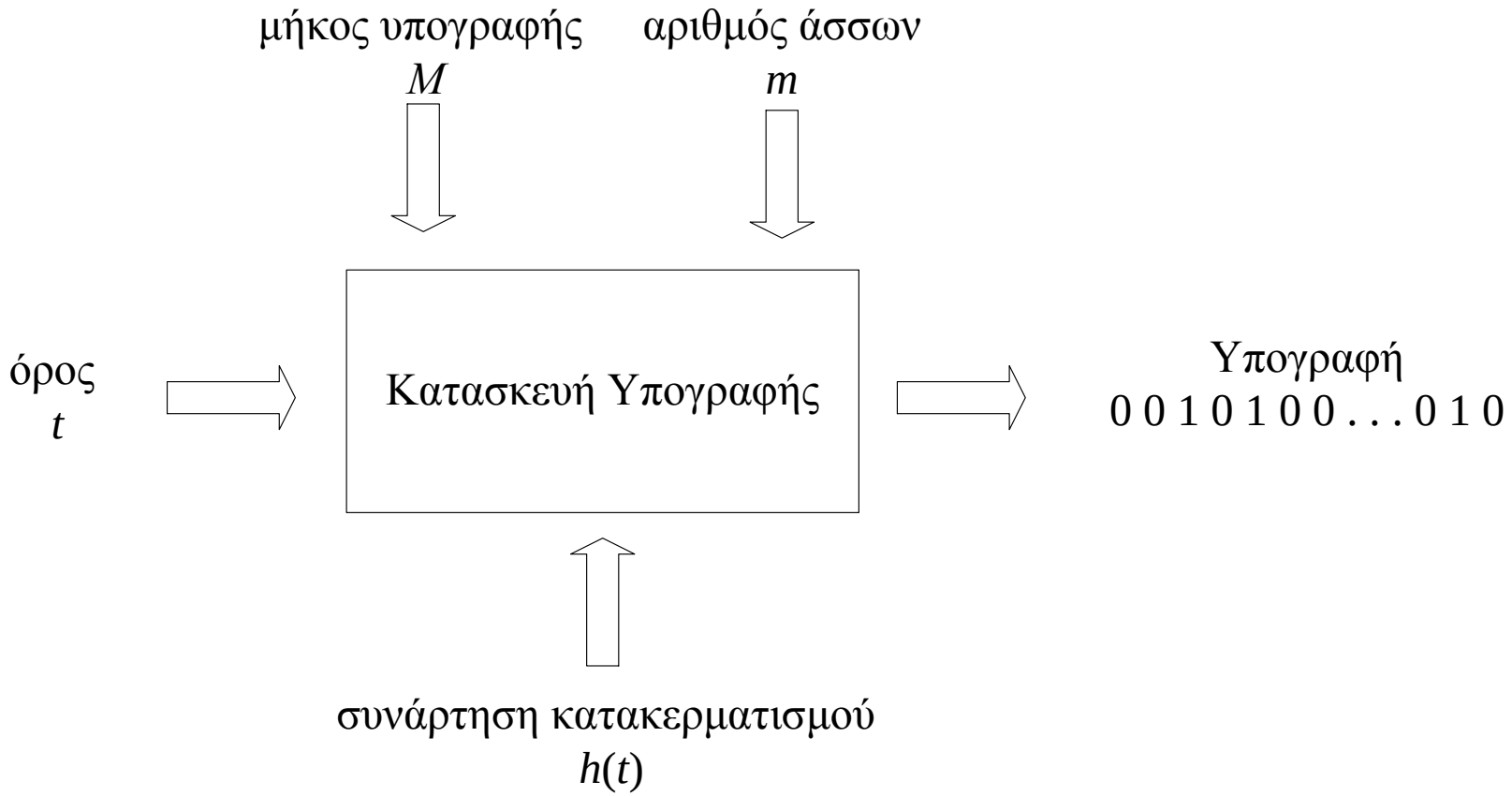
Μία υπογραφή χαρακτηρίζεται από δύο βασικά στοιχεία:

- το μέγεθος (μήκος) της υπογραφής M και
- το πλήθος των δυαδικών ψηφίων που είναι μονάδα m.

Οι τιμές των παραμέτρων αυτών μπορούν να διαφέρουν και εξαρτώνται από την υλοποίηση ή από τις σχεδιαστικές επιλογές.

Αρχικά όλα τα bits της υπογραφής είναι 0 και στη συνέχεια κάποια από αυτά γίνονται 1 χρησιμοποιώντας **συναρτήσεις κατακερματισμού**.

Εξαγωγή Υπογραφής



Βασικές Μέθοδοι Εξαγωγής Υπογραφών-1

- Μία από τις πρώτες μεθόδους παραγωγής υπογραφών για την επεξεργασία εγγράφων κειμένου προτάθηκε από τους Tsichritzis και Christodoulakis (1983).
- Από τον κάθε όρο t του εγγράφου εξάγεται μία υπογραφή *TS(t)* μήκους *f*. Η υπογραφή του συνολικού εγγράφου d, που συμβολίζεται με *DS(d)*, προκύπτει με τη **συνένωση** (concatenation) όλων των υπογραφών των όρων που συναντούμε στο έγγραφο.
- Αυτή η μέθοδος εξαγωγής υπογραφών είναι γνωστή ως **WS** (word signatures).

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-2

- Έστω tq ένας όρος που βρίσκεται στο ερώτημα. Αρχικά, εξάγεται η υπογραφή $TS(t_q)$ και στη συνέχεια ελέγχονται οι υπογραφές των εγγράφων.
- Σε περίπτωση που βρεθεί μία υπογραφή *DS(d)* που αντιστοιχεί στο έγγραφο *d* και περιέχει την υπογραφή $TS(t_q)$ τότε αυτό σημαίνει ότι το έγγραφο *d* **ενδεχομένως** να περιέχει τον όρο tq.
- Στην περίπτωση αυτή, το *d* θεωρείται ως **υποψήφιο έγγραφο**.

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-3

- Ένας άλλος τρόπος εξαγωγής υπογραφών προτάθηκε από τους Faloutsos και Christodoulakis (1984).
- Η μέθοδος εξαγωγής καλείται **SC** (superimposed coding)
- Το έγγραφο χωρίζεται σε **λογικά τμήματα** και το κάθε τμήμα περιέχει ένα μέρος του εγγράφου που αποτελείται από *T* όρους.
- Μας ενδιαφέρουν οι μοναδικοί όροι του λογικού τμήματος και επομένως δε λαμβάνονται υπόψη οι πολλαπλές εμφανίσεις των όρων.
- Από κάθε όρο t υπολογίζεται η υπογραφή του όρου *TS(t)* μήκους F_{SC}.

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-4

- Στη συνέχεια, χρησιμοποιείται **υπέρθεση** (superposition) σύμφωνα με την οποία εφαρμόζεται ο λογικός τελεστής OR σε ένα προς ένα τα δυαδικά ψηφία των υπογραφών και προκύπτει η υπογραφή του τμήματος.
- Η υπογραφή του συνολικού εγγράφου προκύπτει με τη συνένωση των υπογραφών των τμημάτων.

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-5

1. Η αναζήτηση σύμφωνα με τη μέθοδο SC είναι παρόμοια με αυτή της WS.
2. Έστω t_q ένας όρος στο ερώτημα. Αρχικά, εξάγεται η υπογραφή του όρου $TS(t_q)$ και στη συνέχεια προσδιορίζονται οι υπογραφές των τμημάτων των οποίων οι θέσεις που έχουν άσσους ταυτίζονται με τις αντίστοιχες θέσεις των άσσων της υπογραφής $TS(t_q)$.
3. Στην περίπτωση αυτή, το συγκεκριμένο τμήμα **μπορεί** να περιέχει τον όρο t_q και επομένως το αντίστοιχο έγγραφο θεωρείται **υποψήφιο**.

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-1

Παράδειγμα

d7 : Ο Άρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

Για τη μέθοδο WS θεωρούμε ότι το μήκος της κάθε υπογραφής είναι $f = 5$ και το πλήθος των 1 στην υπογραφή είναι $m = 2$.

ο	0 1 0 0 1	του	1 1 0 0 0
Άρης	0 0 0 1 1	ηλιακού	0 0 1 0 1
είναι	0 1 0 0 1	μας	1 0 0 0 1
ένας	0 0 1 1 0	συστήματος	0 1 1 0 0
πλανήτης	1 0 1 0 0		

Άρα, σύμφωνα με τη μέθοδο WS

DS(d7) = 01001 00011 01001 00110 10100 11000 00101 10001 01100

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-2

Παράδειγμα (συνέχεια)

Για τη μέθοδο SC, θα θεωρήσουμε ότι οι υπογραφές έχουν μήκος *FSC*=12 ενώ ο αριθμός των άσσων της κάθε υπογραφής πρέπει να είναι *m*=4. Οι υπογραφές δίνονται στο παρακάτω πίνακα:

Όροι	Υπογραφές
Ο Άρης είναι	000 010 110 001 010 000 101 010 101 001 100 000 **111 011 111 011** (υπογραφή 1ου τμήματος)
ένας πλανήτης του	011 000 010 001 100 110 100 000 000 100 101 010 **111 110 111 011** (υπογραφή 2ου τμήματος)
ηλιακού μας συστήματος	110 100 010 000 010 010 010 010 100 100 100 100 **110 110 110 110** (υπογραφή 3ου τμήματος)

Βασικές Μέθοδοι Εξαγωγής Υπογραφών-3

Παραλλαγή της μεθόδου SC

Για να είναι δυνατή η αναζήτηση μέρους (και όχι ολόκληρου) του όρου, οι Faloutsos και Christodoulakis πρότειναν την ακόλουθη παραλλαγή:

i. στον όρο *t* εισάγονται δύο κενοί χαρακτήρες στην αρχή και στο τέλος του όρου,

ii. δημιουργούνται συνεχόμενες και επικαλυπτόμενες τριάδες χαρακτήρων,

iii. η κάθε τριάδα μέσω του κατακερματισμού ενεργοποιεί ένα συγκεκριμένο δυαδικό ψηφίο της υπογραφής και

iv. εάν ο αριθμός ψ των δυαδικών ψηφίων που ενεργοποιούνται είναι μεγαλύτερος από m, τότε μόνο m δυαδικά ψηφία θα ενεργοποιηθούν, διαφορετικά (αν $\psi < m$) τότε τα υπόλοιπα $m-\psi$ δυαδικά ψηφία ενεργοποιούνται χρησιμοποιώντας μία γεννήτρια τυχαίων αριθμών με φίτρο (seed) που ισούται με μία αριθμητική αναπαράσταση του συγκεκριμένου όρου.

Όπως και προηγουμένως, το έγγραφο χωρίζεται σε τμήματα και η διαδικασία εκτελείται για όλους τους όρους του κάθε τμήματος. Στη συνέχεια, δημιουργούνται οι υπογραφές των τμημάτων με χρήση υπέρθεσης και τέλος κατασκευάζεται η υπογραφή του εγγράφου με συνένωση (concatanation) των υπογραφών των τμημάτων.

Εξαγωγή Υπογραφών με Συμπίεση-1

Παράδειγμα παραλλαγής SC

Έστω ο όρος t = πλανήτης. Υποθέτουμε ότι το σύμβολο _ δηλώνει τον κενό χαρακτήρα και το τοποθετούμε στην αρχή και το τέλος του όρου: **_πλανήτης_**

Οι διαφορετικές συνεχόμενες και επικαλυπτόμενες τριάδες χαρακτήρων που προκύπτουν είναι οι εξής:

πλ, πλα, λαν, ανή, νήτ, ήτη, της, ης

Κάθε τριάδα χαρακτήρων κατακερματίζεται σε μία συγκεκριμένη θέση μέσα στην υπογραφή του όρου πλανήτης και θέτει το αντίστοιχο δυαδικό ψηφίο σε 1.

Εξαγωγή Υπογραφών με Συμπίεση-2

- Η επόμενη μέθοδος εξαγωγής υπογραφών, που καλείται **BC** (bit-block compression), βασίζεται στη συμπίεση και προτάθηκε το 1985.
- Το έγγραφο χωρίζεται σε τμήματα. Για κάθε τμήμα σχηματίζεται μία υπογραφή μεγάλου μεγέθους που αποτελείται από B δυαδικά ψηφία.
- Ο κατακερματισμός του κάθε όρου του τμήματος θα ενεργοποιήσει ένα ή περισσότερα δυαδικά ψηφία της υπογραφής.
- Το διάνυσμα δυαδικών ψηφίων που προκύπτει χαρακτηρίζεται ως αραιό (περιέχει λίγους άσσους σε σχέση με τα μηδενικά) και επομένως μπορεί να συμπιεστεί κατάλληλα.

Εξαγωγή Υπογραφών με Συμπίεση-3

- Η προτεινόμενη μέθοδος συμπίεσης χρησιμοποιεί τμήματα δυαδικών ψηφίων (bit-blocks).
- Το αραιό διάνυσμα που έχει προκύψει χωρίζεται σε τμήματα δυαδικών ψηφίων.
- Το μέγεθος των τμημάτων επιλέγεται έτσι ώστε να βελτιστοποιείται η απόδοση της μεθόδου.

Εξαγωγή Υπογραφών με Συμπίεση-4

Για κάθε τμήμα bb_i δημιουργείται μία νέα υπογραφή μεταβλητού μήκους που αποτελείται από **τρία το πολύ μέρη**:

Το **πρώτο μέρος** της υπογραφής αποτελείται από ένα δυαδικό ψηφίο το οποίο είναι 1 αν υπάρχει τουλάχιστον ένας άσσος στο τμήμα bb_i ή 0 διαφορετικά. Αν ισχύει το δεύτερο, τότε η μέθοδος σταματά εδώ.

Το **δεύτερο μέρος** της υπογραφής που προκύπτει από το bb_i δηλώνει τον αριθμό των άσσων που περιέχονται στο bb_i. Ο αριθμός αυτός κωδικοποιείται χρησιμοποιώντας το μοναδιαίο κώδικα, βάσει του οποίου ένας αριθμός x κωδικοποιείται με x-1 άσσους και ένα μηδενικό στο τέλος.

Το **τρίτο τμήμα** της υπογραφής αποθηκεύει τις θέσεις των άσσων στο τμήμα bb_i χρησιμοποιώντας την απόσταση του ψηφίου από την αρχή του bb_i. Εάν το μέγεθος του τμήματος bbi είναι b δυαδικά ψηφία, για την καταχώρηση της θέσης ενός άσσου απαιτούνται log b δυαδικά ψηφία.

Εξαγωγή Υπογραφών με Συμπίεση-5

Για το σχηματισμό της τελικής υπογραφής του τμήματος bb_i έχουμε δύο εναλλακτικές λύσεις:

i. γίνεται συνένωση όλων των τμηματικών υπογραφών και

ii. παραθέτουμε πρώτα τα πρώτα μέρη, μετά τα δεύτερα και τέλος τα τρίτα από κάθε τμηματική υπογραφή.

Εξαγωγή Υπογραφών με Συμπίεση-6

Θα δουλέψουμε με το έγγραφο *d7* της συλλογής μας το οποίο θεωρούμε ότι χωρίζεται σε 3 τμήματα. Αν υποθέσουμε ότι το μήκος της υπογραφής του τμήματος είναι $B = 20$ τότε ένα παράδειγμα της μορφής που μπορούν να έχουν οι υπογραφές των τμημάτων είναι:

Όροι	Υπογραφές
Ο Άρης είναι	0000 0100 0000 0000 0000 1000 0000 0000 0000 0000 0000 0010 0000 1000 0000 **1000 0110 0000 0000 0000** (υπογραφή 1ου τμήματος)
ένας πλανήτης του	0000 0000 0000 1000 0000 0000 0000 0000 0010 0000 0000 0000 0000 0000 1000 **0000 0000 0000 1010 1000** (υπογραφή 2ου τμήματος)
ηλιακού μας συστήματος	0000 0000 0000 0000 0010 0000 0000 0000 0000 0001 0000 0000 1000 0010 0000 **0000 0000 1000 0000 0011** (υπογραφή 3ου τμήματος)

Εξαγωγή Υπογραφών με Συμπίεση-7

Θα εξηγήσουμε τον τρόπο κωδικοποίησης των υπογραφών για την υπογραφή του τρίτου τμήματος που είναι:

0000 0000 1000 0000 0011

Θα θεωρήσουμε ότι το μέγεθος του κάθε τμήματος δυαδικών ψηφίων είναι $b = 4$. Επομένως, η υπογραφή θα χωριστεί σε πέντε διαφορετικά τμήματα δυαδικών ψηφίων, που είναι τα 0000, 0000, 1000, 0000 και 0011.

Για κάθε ένα από τα τμήματα αυτά θα πρέπει να εφαρμοστεί η μέθοδος εύρεσης της τελικής υπογραφής, σύμφωνα με τα τρία βήματα που αναπτύχθηκαν προηγουμένως.

Εξαγωγή Υπογραφών με Συμπίεση-8

τμήμα	1ο μέρος	2ο μέρος	3ο μέρος	
0000	0	-	-	0
0000	0	-	-	0
1000	1	0	00	1 0 00
0000	0	-	-	0
0011	1	10	10 11	1 10 10 11
	0 0 1 0 1	0 10	00 10 11	

Εξαγωγή Υπογραφών με Συμπίεση-9

- Μία ακόμη μέθοδος εξαγωγής υπογραφών που βασίζεται στη συμπίεση,
- χρησιμοποιεί **κωδικοποίηση μήκους** (run-length encoding) και καλείται **RL**.
- Η μέθοδος είχε προταθεί αρχικά από τον McIlroy (1982) για διαφορετικό περιβάλλον αλλά προσαρμόστηκε για την εξαγωγή υπογραφών.
- Το αραιό διάνυσμα μπορεί να συμπιεστεί κωδικοποιώντας τον αριθμό των μηδενικών που διαχωρίζουν δύο συνεχόμενους άσσους. Για την κωδικοποίηση χρησιμοποιήθηκε η μέθοδος Golomb.

Εξαγωγή Υπογραφών με Συμπίεση-10

- Μία ακόμη μέθοδος που έχει προταθεί προσπαθεί να περιορίσει την επίδραση του αριθμού των όρων ανά τμήμα στην απόδοση της μεθόδου BC.
- Με τη χρήση αυτής της μεθόδου δεν απαιτείται πλέον ο διαχωρισμός του εγγράφου σε τμήματα, ενώ η επεξεργασία των πολύπλοκων ερωτημάτων γίνεται απλούστερη.
- Η μέθοδος καλείται **VBC** (variable bit-block compression) και η βασίζεται στην επιλογή διαφορετικού μήκους για τα τμήματα δυαδικών ψηφίων του κάθε εγγράφου. Το μήκος αυτό εξαρτάται από το πλήθος των μοναδικών όρων του κάθε εγγράφου.

Ψευδείς Συναγερμοί-1

- Το κοινό χαρακτηριστικό όλων των μεθόδων εξαγωγής υπογραφών είναι το γεγονός ότι μπορεί να δώσουν λανθασμένο αποτέλεσμα ως προς το αν ένας όρος περιέχεται ή όχι σε ένα έγγραφο.
- Έστω ένα όρος *t* του ερωτήματος με υπογραφή 00110.
- Έστω τώρα ότι χρησιμοποιώντας τη μέθοδο εξαγωγής υπογραφών με υπέρθεση (SC) έχουμε εντοπίσει ένα τμήμα του εγγράφου με υπογραφή 10110.
- Το γεγονός ότι η υπογραφή του τμήματος έχει άσσους στις θέσεις όπου εμφανίζονται οι άσσοι στην υπογραφή του όρου δε σημαίνει ότι ο όρος σίγουρα θα περιέχεται στο τμήμα.

Ψευδείς Συναγερμοί-2

$$logFAP_{ws} = logT - \frac{F_{ws}}{T}$$

$$\log FAP_{SC} = logT - \frac{F_{SC}}{T \cdot loge} = -0.693 \cdot \frac{F_{SC}}{T}$$

$$\log FAP_{BC} = 1.913 \cdot n - \frac{F_{SC}}{T}$$

$$\log FAP_{RL} = 1.528 \cdot n - \frac{F_{RL}}{T}$$

Ποιοτική Σύγκριση Μεθόδων

- Ως προς την ταχύτητα ελέγχου των υπογραφών, η μέθοδος SC εκτελεί τις λιγότερες συγκρίσεις μεταξύ δυαδικών ψηφίων. Υπενθυμίζεται ότι για να χαρακτηριστεί ένα λογικό τμήμα ως υποψήφιο θα πρέπει οι θέσεις των άσσων στην υπογραφή του όρου να ταυτίζονται με τις θέσεις των άσσων στην υπογραφή του τμήματος. Συνήθως, ο αριθμός m των δυαδικών ψηφίων που θέτει η μέθοδος SC είναι μικρός (π.χ., 10). Αντίθετα, οι μέθοδοι BC και RL απαιτούν πολύ περισσότερες συγκρίσεις. Τέλος η μέθοδος WS απαιτεί την εξέταση ολόκληρης της υπογραφής (του εγγράφου ή του τμήματος) για να διαπιστωθεί εάν περιέχει ή όχι τον όρο.
- Όλες οι μέθοδοι αναμένεται να έχουν καλή επίδοση κατά την επεξεργασία συζευκτικών ερωτημάτων (τύπου AND). Αυτό ισχύει διότι το λογικό AND μεταξύ των υπογραφών των όρων έχει ως αποτέλεσμα τη μείωση των άσσων στην υπογραφή που προκύπτει μετά την υπέρθεση. Αυτό έχει ως άμεσο αποτέλεσμα τη μείωση του κόστους εξέτασης των υπογραφών.
- Από τις μεθόδους που εξετάστηκαν μόνο η SC έχει τη δυνατότητα να υποστηρίξει ερωτήματα που αφορούν σε τμήμα του όρου. Αυτό επιτυγχάνεται χρησιμοποιώντας επικαλυπτόμενες τριάδες συνεχόμενων χαρακτήρων. Κάθε τριάδα ενεργοποιεί και ένα δυαδικό ψηφίο της υπογραφής του όρου.
- Η μόνη μέθοδος που διατηρεί τη σειρά των όρων μέσα στο έγγραφο είναι η WS. Αυτό είναι ιδιαίτερα βολικό διότι διευκολύνει την αναζήτηση φράσεων όπου οι όροι στο ερώτημα πρέπει να εμφανίζονται συνεχόμενοι στα έγγραφα.

Αλγόριθμος αναζήτησης

Αλγόριθμος SignatureSearch (*t*)
t: όρος αναζήτησης

1. υπολογισμός της υπογραφής *TS(t)* του όρου *t*

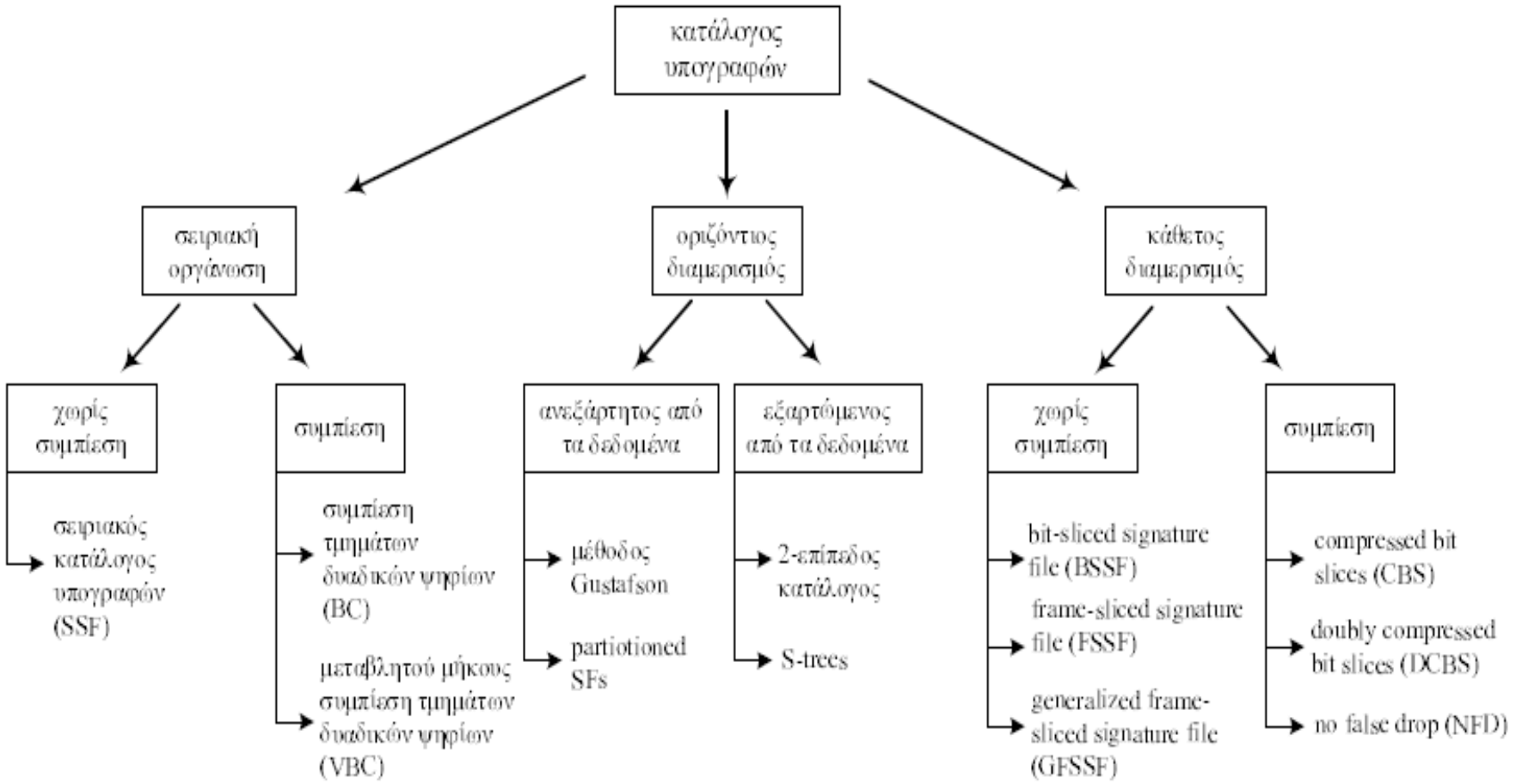
2. αναζήτηση των λογικών τμημάτων lbi για τα οποία ισχύει $LBS(lb_i)$ AND $TS(t) = TS(t)$

3. εισαγωγή της υπογραφής $sig_i = LBS\ (lb_i)$ στο σύνολο υποψηφίων *C*

4. για κάθε υπογραφή $sig_i \in C$
   4.1 έλεγχος αν το λογικό τμήμα lb_i περιέχει τον όρο *t*
   4.2 αν όχι, τότε επανάληψη από το βήμα 4.
   4.3 αν ναι, τότε το έγγραφο d που περιέχει το lb_i προστίθεται στο σύνολο *A*

5. Τα έγγραφα με κωδικούς που ανήκουν στο A επιστρέφονται στο χρήστη

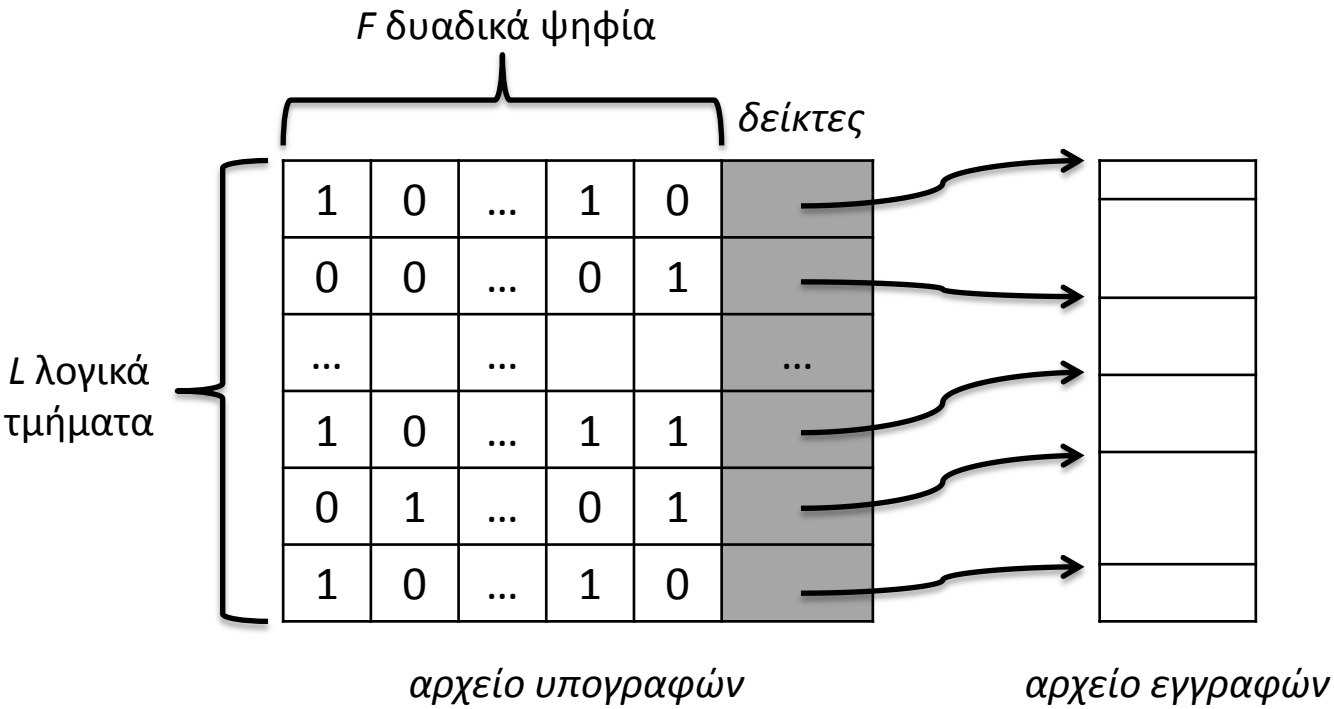
Οργάνωση Υπογραφών



Σειριακή Οργάνωση Υπογραφών-1

- Η πιο απλή μορφή καταλόγου βασίζεται στη σειριακή παράθεση των υπογραφών σε ένα αρχείο που καλείται **σειριακό αρχείο υπογραφών** (sequential signature file - SSF).
- Το αρχείο υπογραφών είναι στην ουσία ένας πίνακας *L* x *F* με *L* γραμμές (πλήθος λογικών τμημάτων) και *F* στήλες (πλήθος δυαδικών ψηφίων ανά υπογραφή).
- Σε κάθε υπογραφή αντιστοιχεί και ένα δείκτης (pointer) που δείχνει στην αρχή του λογικού τμήματος του εγγράφου. Σε περίπτωση που
- Εάν οι υπογραφές έχουν παραχθεί με την απλή μέθοδο της υπέρθεσης (SC) τότε το μήκος όλων των υπογραφών είναι κοινό.
- Εάν έχει χρησιμοποιηθεί μία από τς μεθόδους BC ή VBC τότε στη γενική περίπτωση τα μήκη δύο υπογραφών μπορεί να είναι διαφορετικά.

Σειριακή Οργάνωση Υπογραφών-2



Σειριακή Οργάνωση Υπογραφών-3

- Ο κατάλογος SSF υποστηρίζει αναζητήσεις, εισαγωγές και διαγραφές. Για την αναζήτηση ενός όρου, αρχικά εξάγεται η υπογραφή του όρου και στη συνέχεια προσπελαύνεται το αρχείο υπογραφών με στόχο να βρεθούν οι σχετικές υπογραφές των λογικών τμημάτων.
- Στη συνέχεια, ακολουθούνται οι δείκτες που οδηγούν στα λογικά τμήματα των εγγράφων. Στην τελική φάση της αναζήτησης, ο όρος αναζητείται μέσα σε κάθε υποψήφιο λογικό τμήμα χρησιμοποιώντας μεθόδους αναζήτησης συμβολοσειράς.
- Οι εισαγωγές και οι διαγραφές υποστηρίζονται εύκολα.Για την εισαγωγή ενός νέου εγγράφου αρχικά το έγγραφο διαχωρίζεται σε λογικά τμήματα, στη συνέχεια εξάγονται οι υπογραφές των τμημάτων και τέλος ενημερώνεται το αρχείο υπογραφών, το αρχείο εγγράφων και η λίστα των δεικτών. Για τη διαγραφή, εντοπίζονται όλα τα λογικά τμήματα του εγγράφου τα οποία διαγράφονται από το αρχείο εγγράφων και στη συνέχεια διαγράφονται οι αντίστοιχες υπογραφές από το αρχείο υπογραφών.
- Η λειτουργία της ενημέρωσης ενός μέρους του εγγράφου είναι πιο πολύπλοκη, καθώς θα πρέπει ενδεχομένως να επαναπροσδιοριστούν οι υπογραφές των λογικών τμημάτων που έπονται του τμήματος που έχει μεταβληθεί.

Οργάνωση Υπογραφών

- Με βάση τον τρόπο λειτουργίας του καταλόγου SSF προκύπτει ότι για την αναζήτηση ενός και μόνο όρου θα πρέπει να εξεταστούν όλες οι υπογραφές των λογικών τμημάτων.

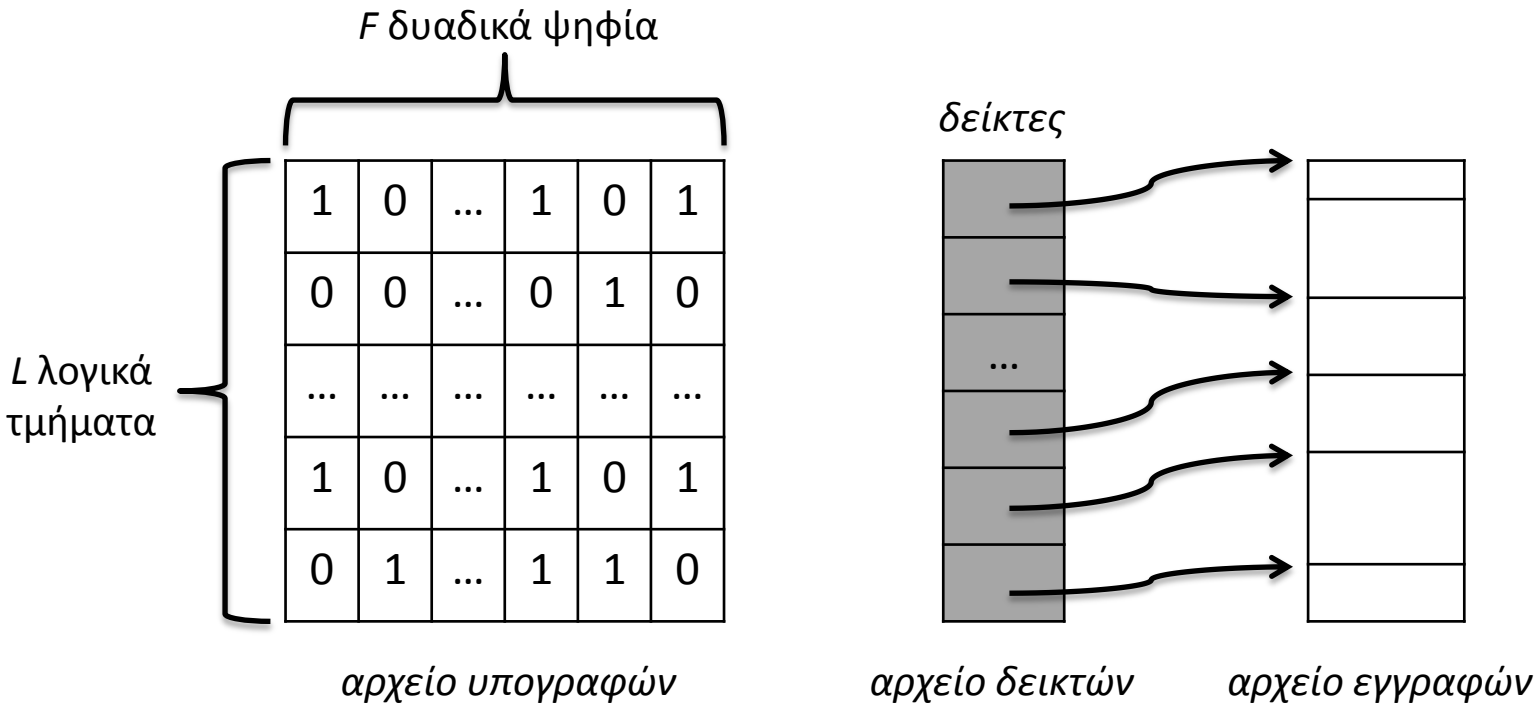
- Ένα από τα θέματα που απασχόλησαν του ερευνητές ήταν το πως θα βελτιωθεί ο χρόνος επεξεργασίας.

- Προς αυτήν την κατεύθυνση έχουν προταθεί εναλλακτικές μορφές οργάνωσης του αρχείου υπογραφών.

Κάθετος Διαμερισμός: BSSF-1

- Η πρώτη από τις μεθόδους που θα εξετάσουμε βασίζεται στον τεμαχισμό (slicing) του πίνακα υπογραφών (1988) και καλείται **BSSF** (bit-sliced signature file).
- Πρόκειται για μία μέθοδο που στηρίζεται στον **κάθετο διαμερισμό** του πίνακα υπογραφών.
- Η αποθήκευση του πίνακα γίνεται κατά στήλες (και όχι κατά γραμμές όπως στη μέθοδο SSF).
- Ο πίνακας υπογραφών του αντιστρέφεται, και αποκτά διαστάσεις *F x L* (*F* γραμμές και *L* στήλες). Η κάθε γραμμή του αντεστραμμένου πίνακα καλείται **τεμάχιο** (slice) και αποτελείται από τα δυαδικά ψηφία που βρίσκονται στην ίδια θέση σε όλες τις υπογραφές των λογικών τμημάτων.
- Για να μπορεί η δομή να υποστηρίξει εισαγωγές και διαγραφές αποδοτικά, η κάθε γραμμή του αντεστραμμένου πίνακα αποθηκεύεται σε ξεχωριστό αρχείο.

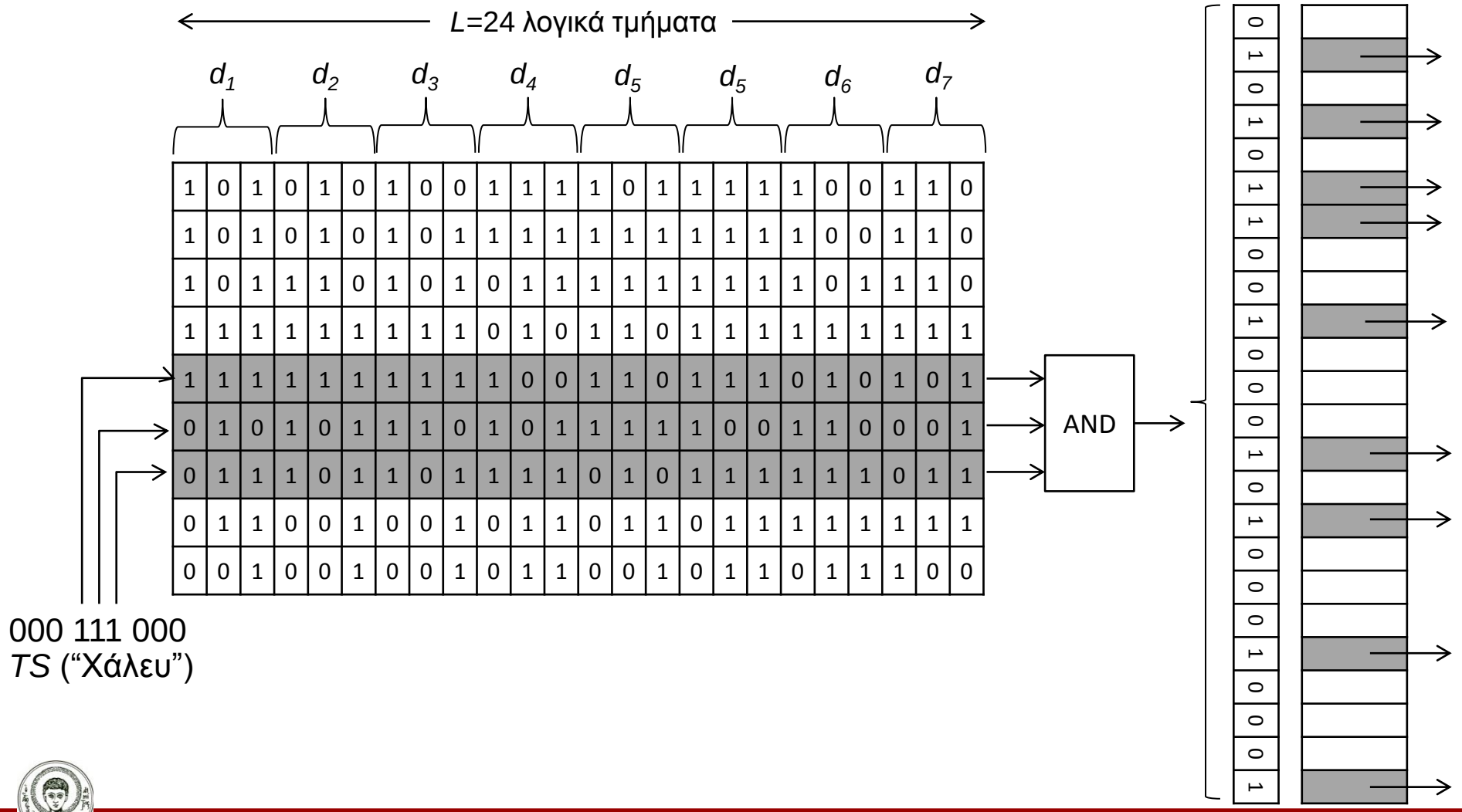
Κάθετος Διαμερισμός: BSSF-2



Κάθετος Διαμερισμός: BSSF-3

- Η αναζήτηση ενός όρου στη δομή BSSF ξεκινά με τον υπολογισμό της υπογραφής του όρου.
- Η υπογραφή του όρου θα περιέχει άσσους σε ακριβώς m δυαδικά ψηφία. Επομένως, σε αντίθεση με τη δομή SSF, απαιτείται η εξέταση *m* τεμαχίων (γραμμών του αντεστραμμένου πίνακα).
- Τα δυαδικά ψηφία των *m* γραμμών συνδυάζονται με τη χρήση υπέρθεσης (λογικό AND) και προκύπτει ένα διάνυσμα *L* θέσεων. Στη συνέχεια, λαμβάνονται υπόψη οι θέσεις των άσσων στο διάνυσμα αυτό και προσπελαύονται οι αντίστοιχοι δείκτες του αρχείου δεικτών για να οδηγηθούμε τελικά στα λογικά τμήματα των εγγράφων.
- Για την εισαγωγή ενός νέου εγγράφου, αρχικά προσδιορίζονται τα νέα λογικά τμήματα και οι αντίστοιχες υπογραφές. Στη συνέχεια, για κάθε νέο λογικό τμήμα πραγματοποιείται τεμαχισμός της υπογραφής του και κάθε ένα από τα *F* διαφορετικά αρχεία λαμβάνει και ένα δυαδικό ψηφίο της υπογραφής που αποθηκεύεται στο τέλος.

Κάθετος Διαμερισμός: BSSF-4



Κάθετος Διαμερισμός: BSSF-5

- Η μέθοδος BSSF είναι πιο αποδοτική από την SSF ως προς τη λειτουργία της αναζήτησης. Ωστόσο, υπάρχει επιπλέον χώρος για βελτίωση που οφείλεται σε δύο κυρίως λόγους:

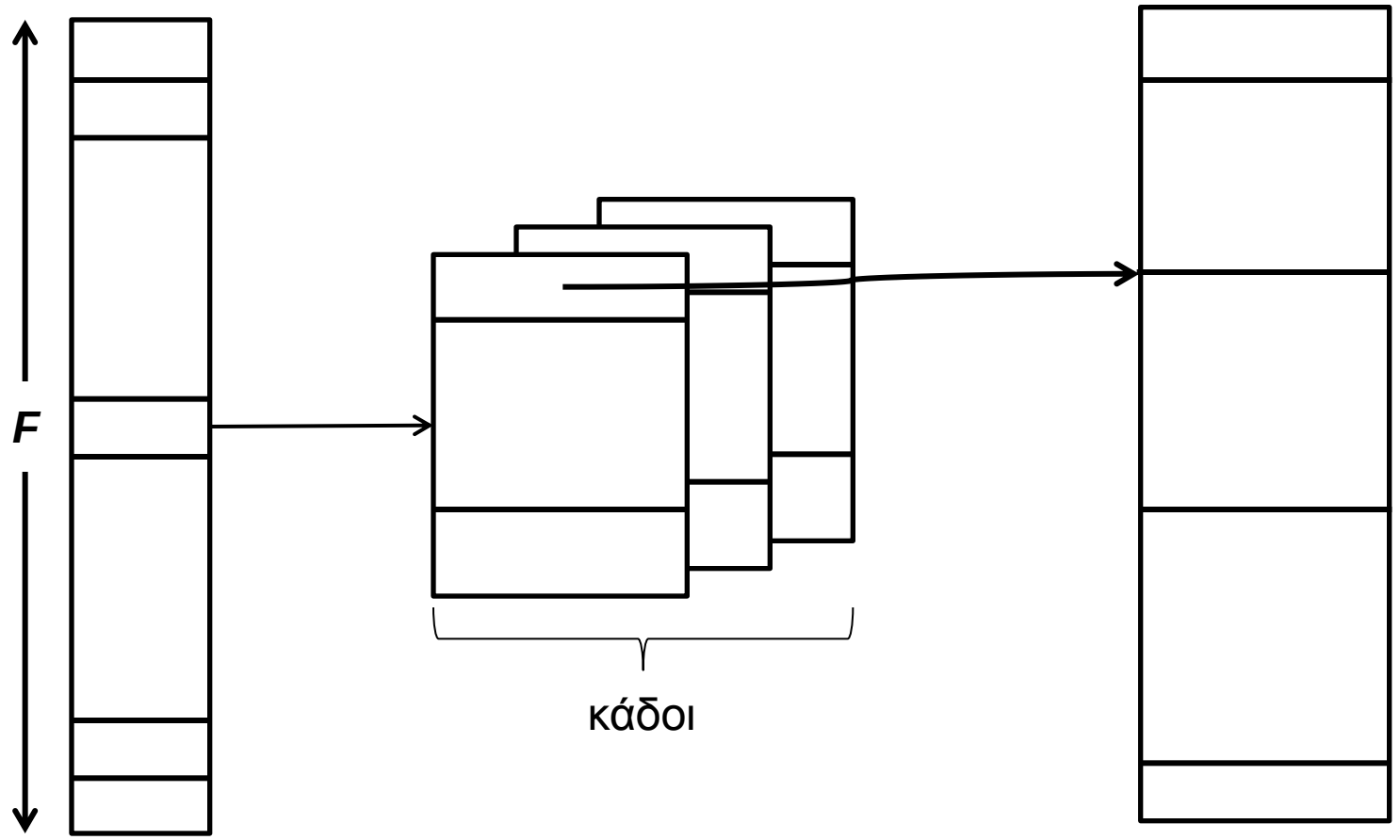
- Η αναζήτηση ενός όρου επιβάλλει την προσπέλαση m τεμαχίων, όπου *m* είναι ο αριθμός των άσσων στην υπογραφή του όρου. Αν *m*=1 τότε θα μπορούσε να αυξηθεί η απόδοση της μεθόδου.

- Η εισαγωγή ενός νέου λογικού τμήματος απαιτεί ένα μεγάλο αριθμό προσπελάσεων που ρυθμίζεται από τον αριθμό των δυαδικών ψηφίων της υπογραφής του λογικού τμήματος *F*. Αν η τιμή της παραμέτρου *F* είναι μεγάλη (π.χ. 1000) τότε αυξάνεται σημαντικά το κόστος εισαγωγής.

Κάθετος Διαμερισμός: CBS-1

- Εάν θέσουμε *m* = 1, τότε θα πρέπει να αυξηθεί σημαντικά το μήκος της υπογραφής ώστε η πιθανότητα ψευδών συναγερμών να μην αυξηθεί. Αυτό έχει ως αποτέλεσμα, ο πίνακας διαστάσεων *F x L* που θα προκύψει να χαρακτηρίζεται ως **αραιός**, διότι το ποσοστό των άσσων σε σχέση με αυτό των μηδενικών είναι μικρό. Άρα, μπορούν να εφαρμοστούν μέθοδοι συμπίεσης με στόχο τη μείωση του μεγέθους του κάθε τεμαχίου.
- Η πιο απλή μέθοδος που μπορεί να εφαρμοστεί είναι να αποθηκεύονται οι θέσεις των άσσων σε κάθε τεμάχιο. Με τον τρόπο αυτό, το μέγεθος του κάθε τεμαχίου δεν είναι σταθερό, οπότε το κάθε αρχείο αποθηκεύεται σε έναν ή περισσότερους κάδους (buckets) οι οποίοι συνδέονται με τη μορφή συνδεδεμένης λίστας.
- Το μέγεθος του κάθε κάδου (*K*) αποτελεί σχεδιαστική παράμετρο. Η μέθοδος αυτή προτάθηκε το 1988 CBS (compressed bit slices). Εκτός από το ότι κάθε όρος ενεργοποιεί μόνο ένα δυαδικό ψηφίο, η δομή CBS δε χρειάζεται το αρχείο δεικτών. Αντί να αποθηκεύεται η θέση του κάθε άσσου, αποθηκεύεται απευθείας ο δείκτης στο αρχείο εγγράφων.

Κάθετος Διαμερισμός: CBS-2



Σύνοψη-1

- Οι κατάλογοι υπογραφών αποτελούν μία διαφορετική προσέγγιση για την οργάνωση μίας συλλογής εγγράφων.
- Το βασικό χαρακτηριστικό των καταλόγων αυτών είναι ότι στηρίζονται στη δημιουργία υπογραφών από τους όρους των εγγράφων.
- Μία υπογραφή είναι μία ακολουθία δυαδικών ψηφίων (bits) τα οποία περιέχουν άσσους σε συγκεκριμένες θέσεις που καθορίζονται από τη συνάρτηση κατακερματισμού που χρησιμοποιείται.

Σύνοψη-2

- Σύμφωνα με πειραματικές μελέτες σχετικά με την επίδοση των καταλόγων υπογραφών σε σχέση με τους αντεστραμμένους καταλόγους, έχει επαληθευτεί ότι οι κατάλογοι που στηρίζονται στην αντιστροφή έχουν γενικά καλύτερες επιδόσεις από τους καταλόγους που στηρίζονται σε υπογραφές.
- Ωστόσο, οι κατάλογοι υπογραφών έχουν μερικές πολύ καλές ιδιότητες (π.χ., ευκολία στον παραλληλισμό) και επομένως η μελέτη τους θεωρείται χρήσιμη.

Βιβλιογραφία

TSICHRITZIS, D., and S. CHRISTODOULAKIS. 1983. "Message Files." *ACM Trans. on Office Information Systems,* 1 (1), 88-98.

FALOUTSOS, C., and S. CHRISTODOULAKIS. 1984. "Signature Files: An Access Method for Documents and its Analytical Performance Evaluation." *ACM Trans. on Office Information Systems,* 2 (4), 267-88.

MCILROY, M. D. 1982. "Development of a Spelling List." *IEEE Trans. on Communications,* COM-30, (1), 91-99.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά - Μη Εμπορική Χρήση - Όχι Παράγωγα Έργα 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Ως **Μη Εμπορική** ορίζεται η χρήση:

- που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
- που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
- που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

Τέλος ενότητας

Επεξεργασία: <Μαυρίδης Απόστολος>
Θεσσαλονίκη, <Εαρινό εξάμηνο 2013-2014>

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.