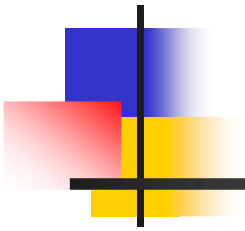
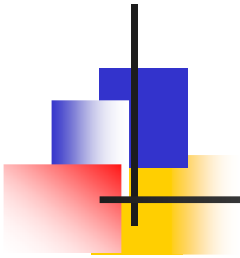


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ - ΤΜΗΥΠ
ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΙΙ

Β. Μεγαλοοικονόμου

Βάσεις Δεδομένων Κειμένου

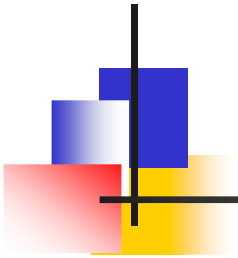
(παρουσίαση βασισμένη εν μέρη σε σημειώσεις των Silberchatz, Korth και Sudarshan και του C. Faloutsos)



Κείμενο – Δομή διάλεξης

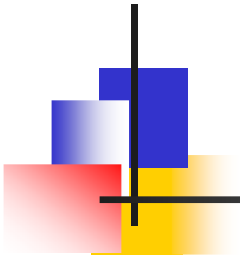
Κείμενο

- Πρόβλημα
- Σάρωση πλήρους κειμένου
- Αναστροφή
- Αρχεία υπογραφής
- Ομαδοποίηση
- Φιλτράρισμα πληροφορίας και LSI



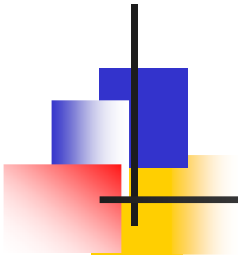
Πρόβλημα- Κίνητρο

- Π.χ., να βρεθούν έγγραφα τα οποία περιέχουν τις λέξεις “data” και “*retrieval*”
- Εφαρμογές:
  - Ιστός
  - Δικηγορικά γραφεία και γραφεία ευρεσιτεχνειών
  - Ψηφιακές βιβλιοθήκες
  - Φιλτράρισμα πληροφορίας



Πρόβλημα- Κίνητρο

- Τύποι ερωτημάτων:
  - Λογικοί (‘data’ AND ‘retrieval’ AND NOT ...)



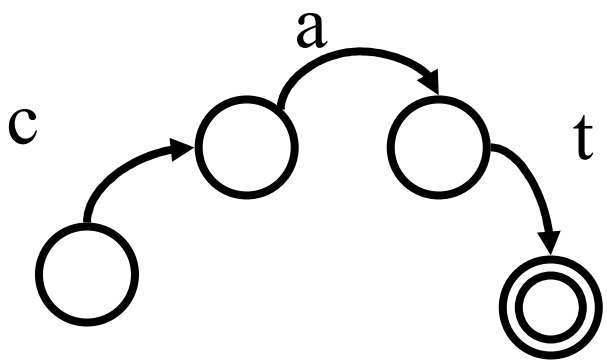
Πρόβλημα- Κίνητρο

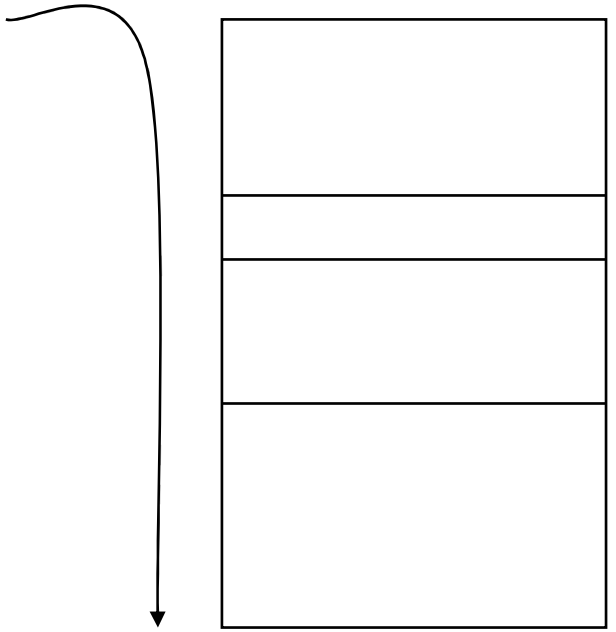
- Τύποι ερωτημάτων:
  - Λογικοί
    (‘data’ AND ‘retrieval’ AND NOT ...)
  - Επιπλέον χαρακτηριστικά
    (‘data’ ADJACENT ‘retrieval’)
  - Ερωτήματα λέξεων κλειδιών
    (‘data’, ‘retrieval’)

- Πώς γίνεται η αναζήτηση σε μία μεγάλη συλλογή κειμένων;

Σάρωση πλήρους κειμένου

- Κατασκευή ενός FSA, Σάρρωση



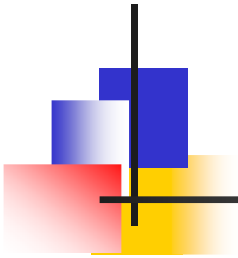


Σάρωση πλήρους κειμένου

- Για έναν όρο:
  - (απλά: O(N*M))

ABRACADABRA κείμενο

CAB πρότυπο



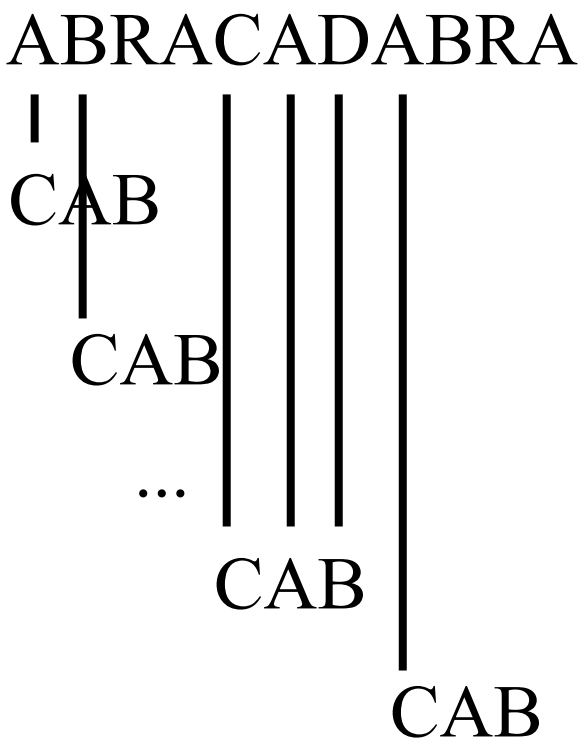
Σάρωση πλήρους κειμένου

- Για έναν όρο:
  - (απλά: O(N*M))
  - Knuth Morris & Pratt ('77)
    - Κατασκευή μικρού FSA, επίσκεψη κάθε γράμματος του κειμένου μόνο μία φορά, με προσεκτική ολίσθηση περισσοτέρων από ένα βημάτων

ABRACADABRA κείμενο

CAB πρότυπο

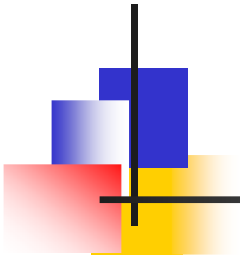
Σάρωση πλήρους κειμένου



κείμενο

πρότυπο

Σάρωση πλήρους κειμένου

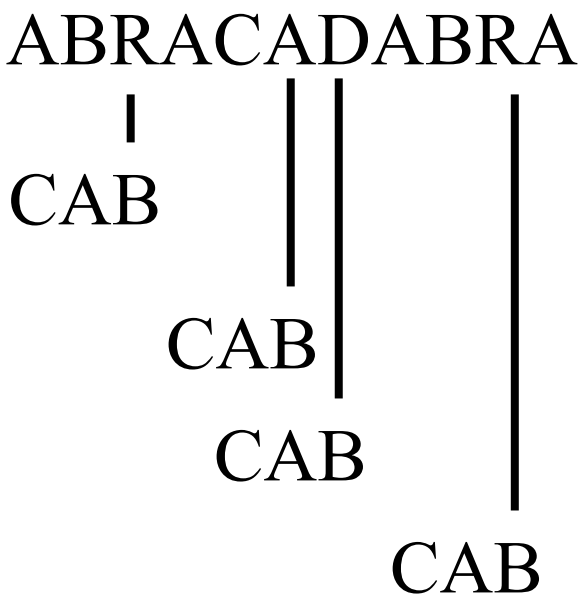


- Για έναν όρο :
  - (απλός τρόπος: O(N*M))
  - Knuth Morris & Pratt (‘77)
  - Boyer & Moore (‘77)
    - Προεπεξεργασία προτύπου, ξεκινά από δεξιά προς τα αριστερά και προσπερνά!

ABRACADABRA κείμενο

CAB πρότυπο

Σάρωση πλήρους κειμένου



κείμενο

πρότυπο

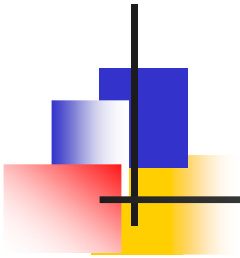
Σάρωση πλήρους κειμένου

ABRACADABRA — κείμενο

OMINOUS — πρότυπο

OMINOUS

Boyer+Moore: γρηγορότερος, στην πράξη
Sunday (‘90): κάποιες βελτιώσεις



Σάρωση πλήρους κειμένου

- Για πολλαπλούς όρους (w/o "don't care" characters): Aho+Corasic ('75)
  - πάλι, κατασκεύασε ένα απλό FSA σε O(M) χρόνο
- Πιθανοτικοί αλγόριθμοι: 'fingerprints' (Karp + Rabin '87)
- Προσεγγιστικό ταίριασμα: 'agrep' [Wu+Manber, Baeza-Yates+, '92]

Σάρωση πλήρους κειμένου

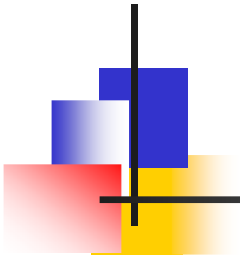
- Προσεγγιστικό ταίριασμα- Απόσταση μετασχηματισμού συμβολοσειράς (**string editing** distance):

  d(‘survey’, ‘surgery’) = 2

  = ελάχιστος # ενθέσεων, διαγραφών, αντικαταστάσεων για τον μετασχηματισμό της πρώτης συμβολοσειράς στην δεύτερη

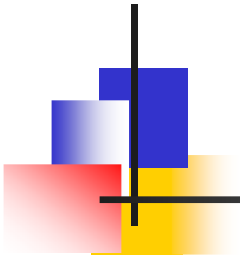
SURVEY

SURGERY



Σάρωση πλήρους κειμένου

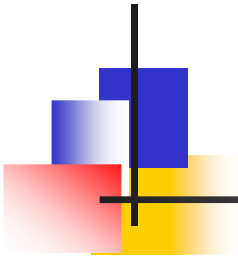
- **string editing** distance – Πως υπολογίζεται;
- Α:



Σάρωση πλήρους κειμένου

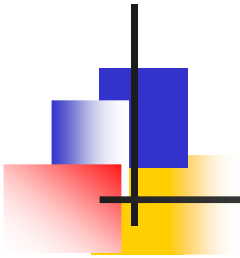
- **string editing** distance – Πως υπολογίζεται;
- Α: Δυναμικός προγραμματισμός

  cost(i, j) = το κόστος ταιριάσματος του προθέματος μήκους i της πρώτης συμβολοσειράς s με το πρόθεμα μήκους j της δεύτερης συμβολοσειράς t



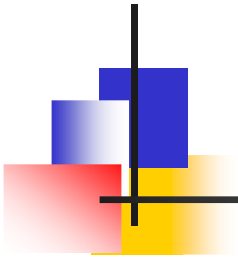
Σάρωση πλήρους κειμένου

```
if s[i] = t[j] then
        cost( i, j ) = cost(i-1, j-1)
else
        cost(i, j ) = min (
                        1 + cost(i, j-1) // deletion
                        1 + cost(i-1, j-1) // substitution
                        1 + cost(i-1, j) // insertion
                            )
```



Σάρωση πλήρους κειμένου

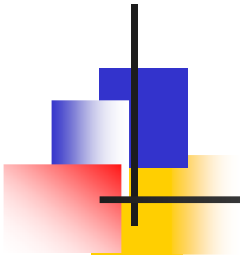
Πολυπλοκότητα: O(M*N) (όταν χρησιμοποιείται πίνακας για την απομνημόνευση των επιμέρους αποτελεσμάτων)



Σάρωση πλήρους κειμένου

Συμπεράσματα:

- Η σάρωση πλήρους κειμένου δεν χρειάζεται επιπλέον χώρο, αλλά είναι αργή για μεγάλα σύνολα δεδομένων

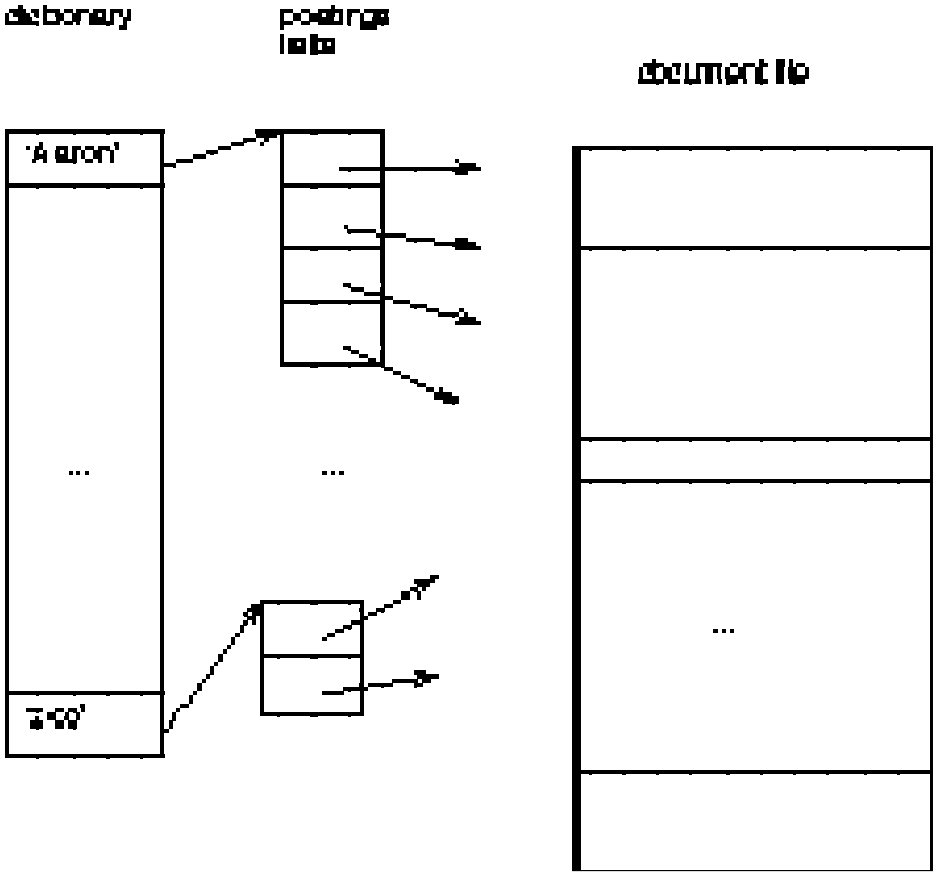


Κείμενο – Δομή διάλεξης

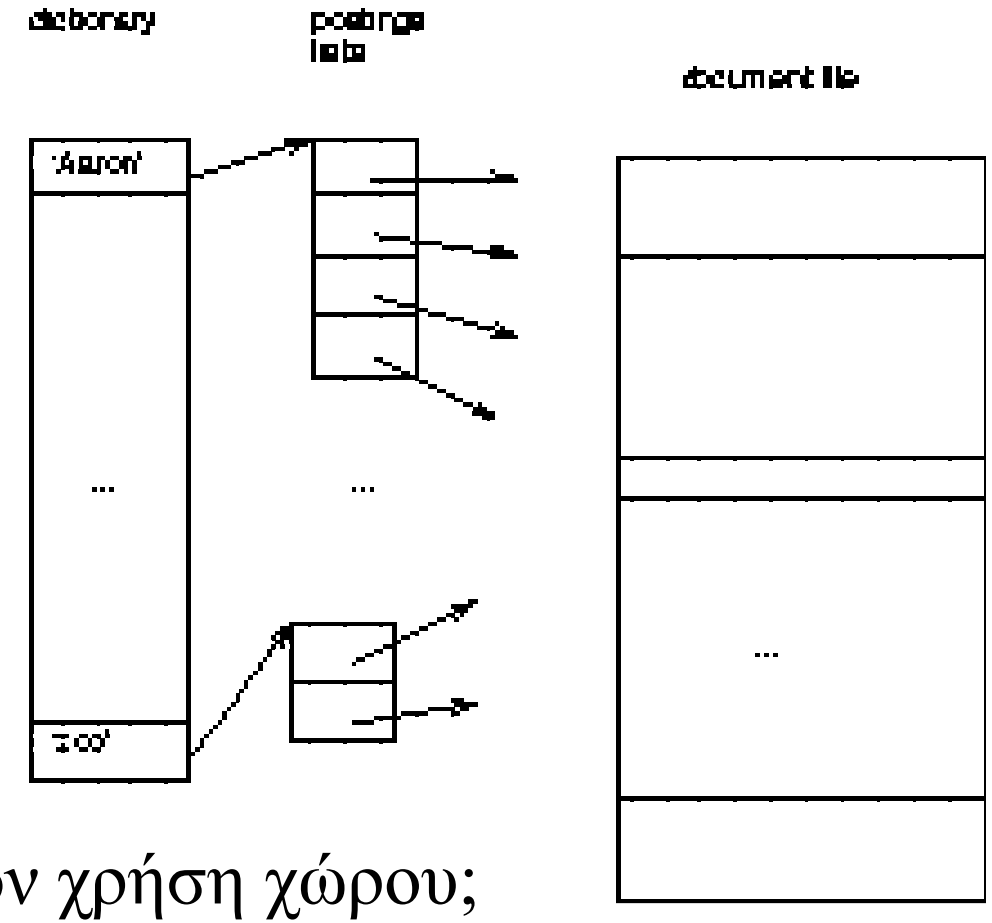
Κείμενο

- Πρόβλημα
- Σάρωση πλήρους κειμένου
- Αναστροφή
- Αρχεία υπογραφών
- Ομαδοποίηση
- Φιλτράρισμα πληροφορίας και LSI

Κείμενο- Αναστροφή

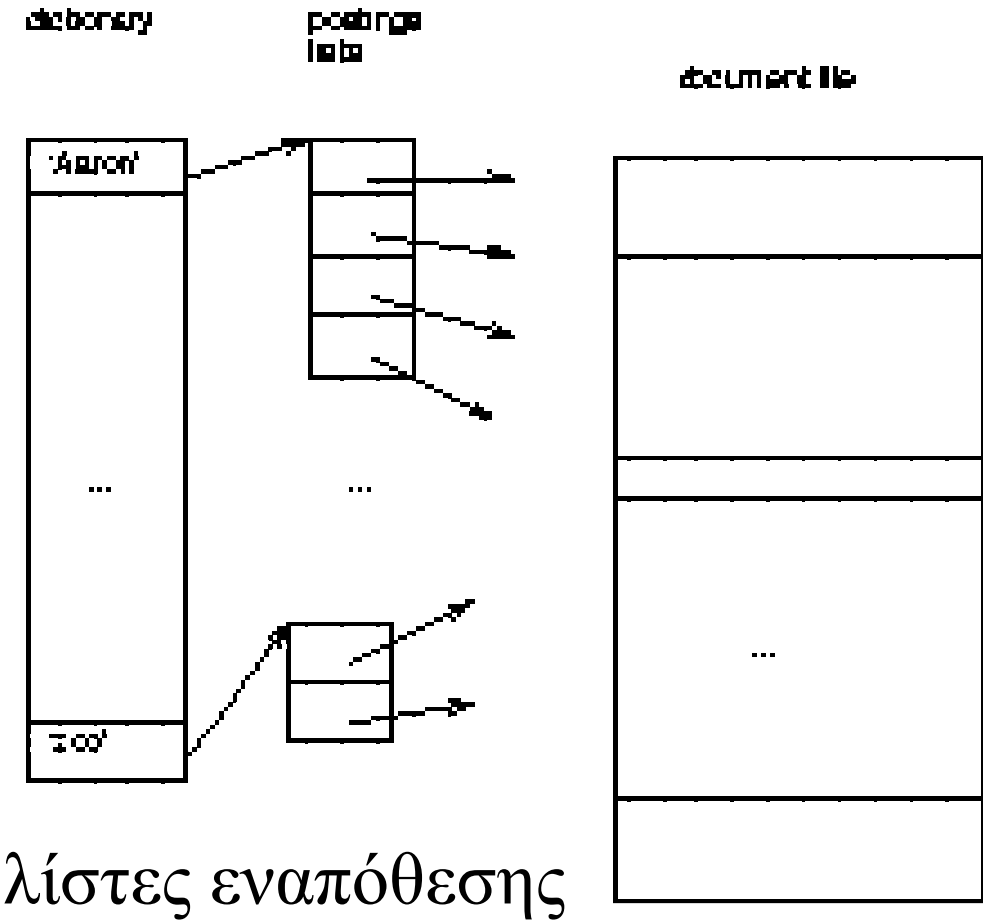


Κείμενο- Αναστροφή

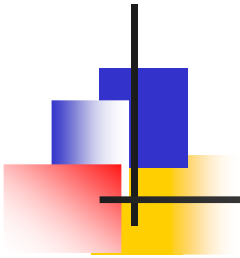


Q: επιπλέον χρήση χώρου;

Κείμενο- Αναστροφή



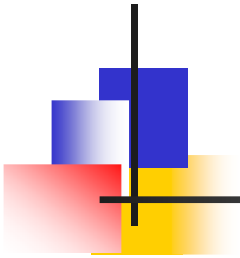
Α: κυρίως, λίστες εναπόθεσης



Κείμενο- Αναστροφή

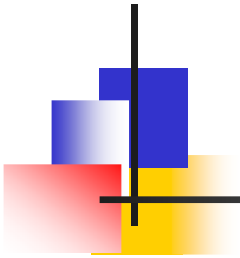
- Πώς γίνεται η οργάνωση λεξικού;

- stemming – N/O;
- Είσοδοι;



Κείμενο- Αναστροφή

- Πώς γίνεται η οργάνωση λεξικού;
  - B-tree, hashing, TRIEs, PATRICIA trees, ...
- stemming – N/O;
- Είσοδοι

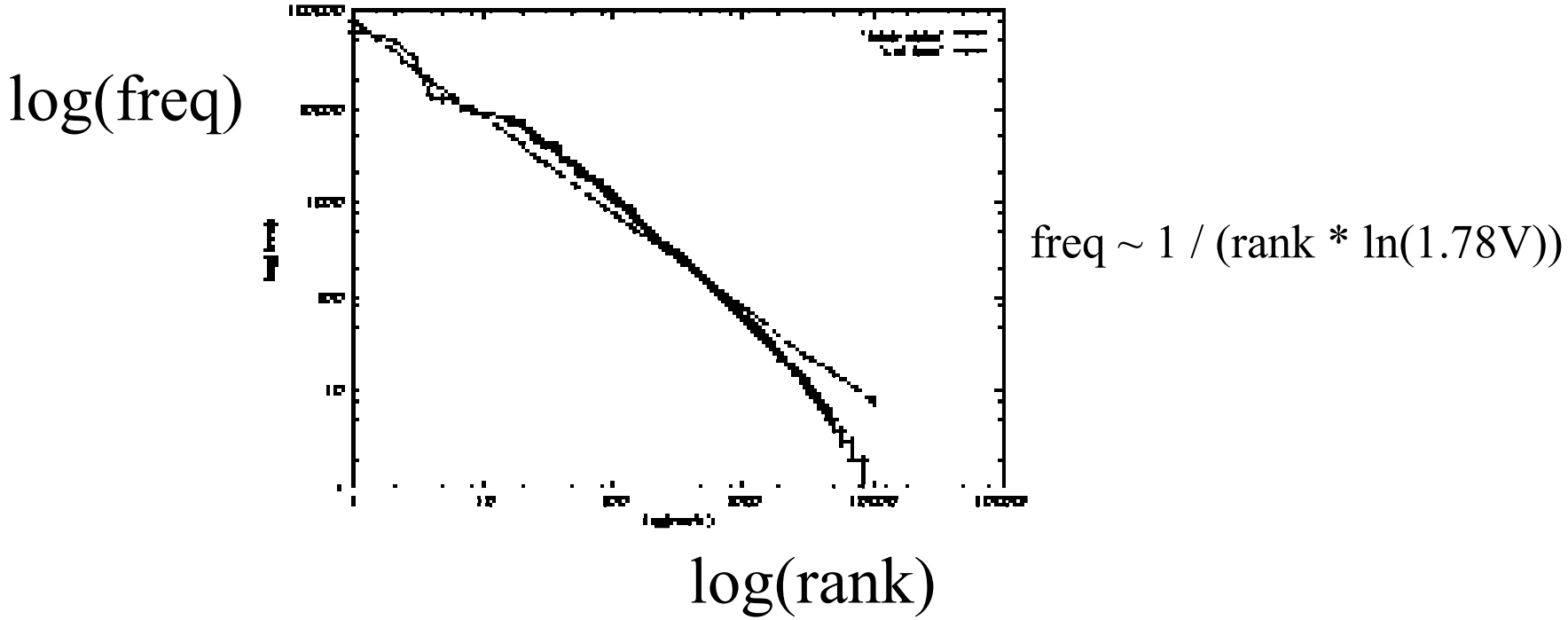


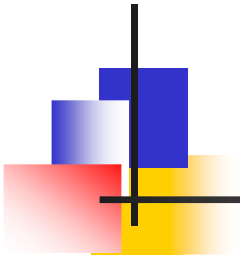
Κείμενο- Αναστροφή

- Νέα θέματα:
- Παραλληλισμός[Tomasic+,93]
- Είσοδοι [Tomasic+94], [Brown+]
  - ‘zipf’ διανομές
- Προσεγγιστική αναζήτηση (‘glimpse’ [Wu+])

Κείμενο- Αναστροφή

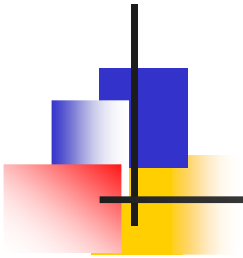
- postings list – more Zipf distr.: eg., rank-frequency plot of 'Bible'





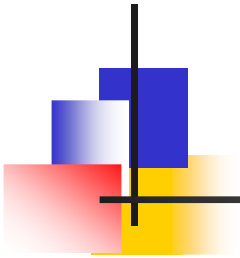
Κείμενο- Αναστροφή

- postings lists
  - Cutting+Pedersen
    - (κράτησε τα πρώτα 4 in B-tree leaves)
  - Πως γίνεται η δέσμευση χώρου: [Faloutsos+92]
    - Γεωμετρική πρόοδος
  - Συμπίεση (Elias codes) [Zobel+] – μόλις 2% επιπλέον!



Συμπεράσματα

- Συμπεράσματα: χρειάζεται επιπλέον χώρος (2%-300%), αλλά έχουμε καλύτερη ταχύτητα



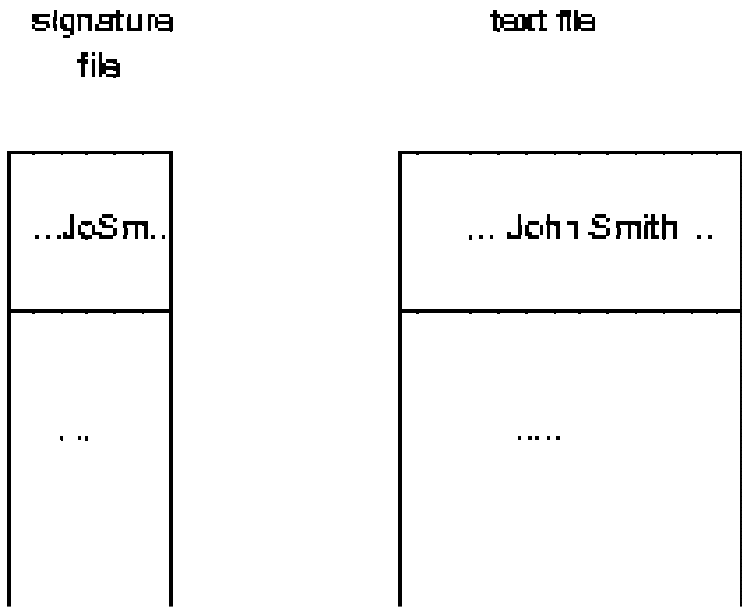
Κείμενο – Δομή διάλεξης

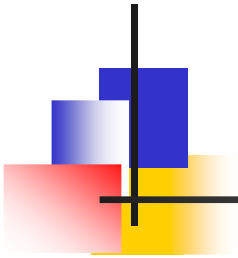
Κείμενο

- Πρόβλημα
- Σάρωση πλήρους κειμένου
- Αναστροφή
- Αρχεία υπογραφής
- Ομαδοποίηση
- Φιλτράρισμα πληροφορίας και LSI

Αρχεία υπογραφής

- Ιδέα: ‘quick & dirty’ filter





Αρχεία υπογραφής

- Ιδέα: «γρήγορο και βρώμικο» φιλτράρισμα
- Έπειτα, σάρωσε ακολουθιακά το αρχείο υπογραφής και εντόπισε τις «ενδείξεις σφάλματος»
- Πλεονέκτημα: εύκολες ενθέσεις
- Μειονέκτημα: αναζήτηση σε O(N)

  (με μικρή σταθερά)
- Q: Πώς γίνεται η εξαγωγή υπογραφών;

Αρχεία υπογραφής

- Α: κωδικοποίηση υπέρθεσης!! [Mooers49], ...

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

m (=4 bits/word) ~ (=4 bits παίρνουν την τιμή “1”
Και τα υπόλοιπα παραμένουν “0”)
F (=12 bits sign. size)
Τα πρότυπα των bits σχηματίζουν την υπογραφή του κειμένου

Αρχεία υπογραφής

- Α: κωδικοποίηση υπέρθεσης!! [Mooers49], ...

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

data

Ταίριασμα

Αρχεία υπογραφής

- Α: κωδικοποίηση υπέρθεσης!! [Mooers49], ...

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

retrieval 

Αποτυχία

Αρχεία υπογραφής

- Α: κωδικοποίηση υπέρθεσης!! [Mooers49], ...

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

nucleotic

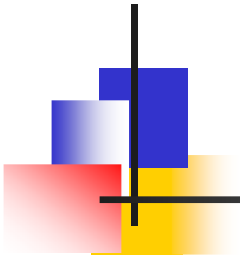
Ένδειξη σφάλματος (‘false drop’)

Αρχεία υπογραφής

- Α: Κωδικοποίηση υπέρθεσης!! [Mooers49], ...

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

‘ΝΑΙ’ είναι ‘ΙΣΩΣ’
‘ΟΧΙ’ είναι‘ΟΧΙ’



Αρχεία υπογραφής

- Q1: Πώς επιλέγονται τα *F* και *m;*
- Q2: Για ποιο λόγο καλείται ‘false drop’;
- Q3: Άλλες εφαρμογές των αρχείων υπογραφής;

Αρχεία υπογραφής

- Q1: Πώς επιλέγονται τα *F* και *m;*

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

m (=4 bits/word)
F (=12 bits sign. size)

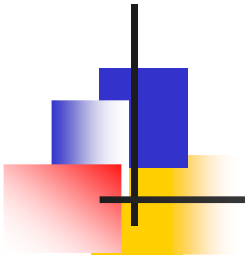
Αρχεία υπογραφής

- Q1: Πώς επιλέγονται τα *F* και *m;*
- A: Έτσι ώστε η υπογραφή κειμένου να είναι 50% πλήρης

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011

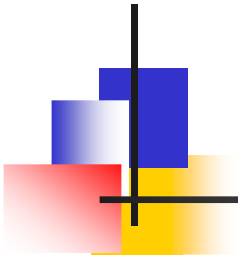
m (=4 bits/word)

F (=12 bits sign. size)



Αρχεία υπογραφής

- Q1: Πώς επιλέγονται τα *F* και *m;*
- Q2: Για ποιο λόγο καλείται ‘false drop’;
- Q3: Άλλες εφαρμογές των αρχείων υπογραφής;

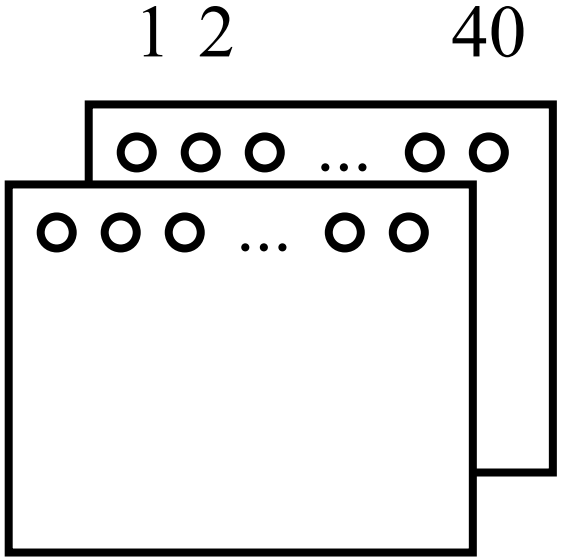


Αρχεία υπογραφής

- Q2: Για ποιο λόγο καλείται ‘false drop’;
- Παλιά, αλλά ενδιαφέρουσα ιστορία [1949]
  - Πώς γίνεται η αναζήτηση βιβλίων (με την λέξη τίτλου, και/ή τον συγγραφέα, και/ή την λέξη κλειδί)
  - Σε χρόνο O(1);
  - *Χωρίς υπολογιστές*

Αρχεία υπογραφής

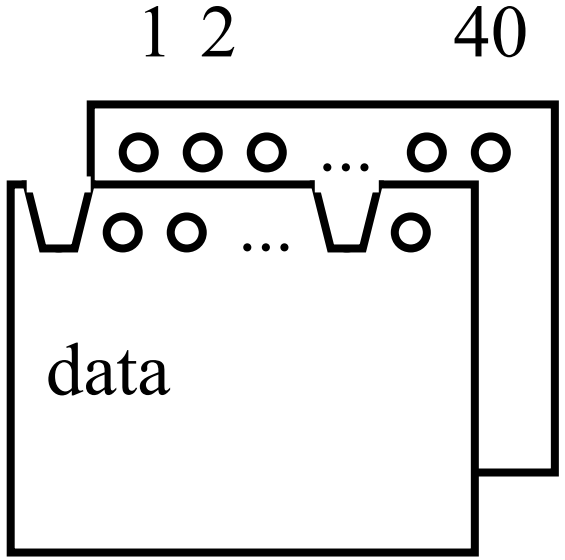
- Λύση: "Edge-notched cards"



•Κάθε λέξη τίτλου αντιστοιχίζεται σε *m* αριθμούς (πώς;)
•Και οι αντίστοιχες τρύπες αποκόπτονται:

Αρχεία υπογραφής

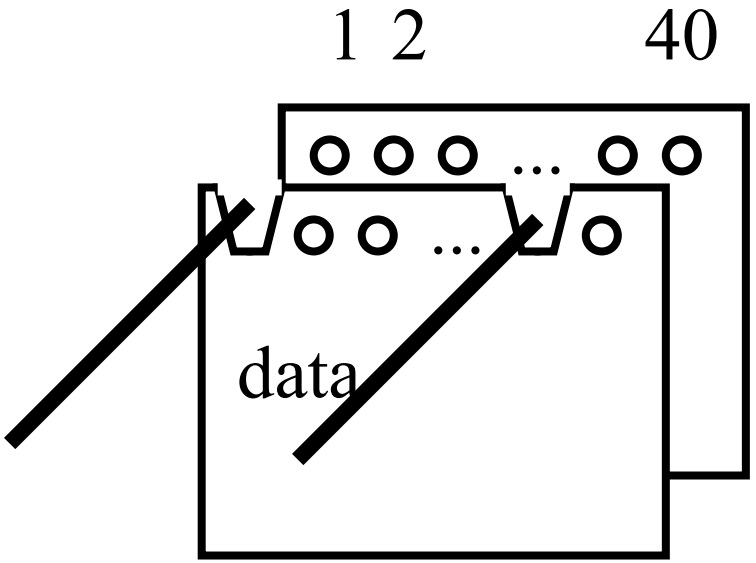
- Λύση: "Edge-notched cards"



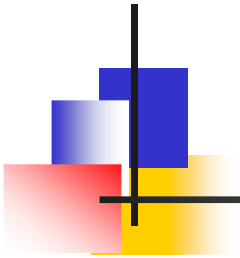
'data' -> #1, #39

Αρχεία υπογραφής

- Ψάξε, π.χ., για ‘data’: ενεργοποίησε την βελόνα #1, #39 και ανακάτεψε τον σωρό των καρτών!

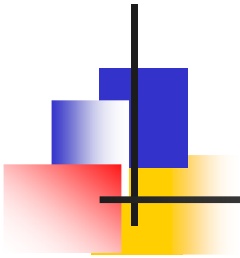


‘data’ -> #1, #39



Αρχεία υπογραφής

- Γνωστή επίσης ως ‘zatocoding’, ‘Zator’



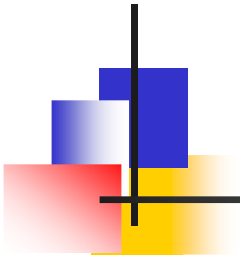
Αρχεία υπογραφής

- Q1: Πώς επιλέγονται τα *F* και *m;*
- Q2: Για ποιο λόγο καλείται ‘false drop’;
- Q3: Άλλες εφαρμογές των αρχείων υπογραφής;

Αρχεία υπογραφής

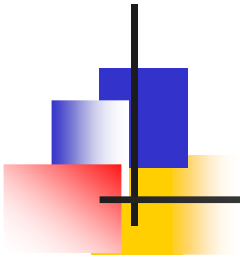
- Q3: Άλλες εφαρμογές των αρχείων υπογραφής;
- A: Ο,τιδήποτε έχει να κάνει με "membership testing": Ανήκει η λέξη 'data' στο σύνολο λέξεων του κειμένου;

Word	Signature
data	001 000 110 010
base	000 010 101 001
doc. signature	001 010 111 011



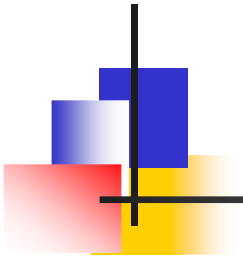
Αρχεία υπογραφής

- UNIX's early 'spell' system [McIlroy]
- Bloom-joins in System R* [Mackert+] & 'active disks' [Riedel99]
- Διαφορετικά αρχεία [Severance+Lohman]



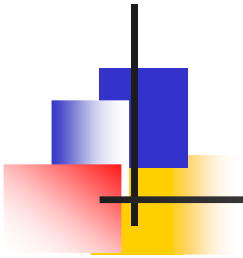
Αρχεία υπογραφής- Συμπερασματα

- Εύκολες ενθέσεις, πιο αργά από αναστροφή
- Εξαιρετική η ιδέα του ‘quick and dirty’ φίλτρου: απομακρύνεται γρήγορα η πλειοψηφία των άσχετων στοιχείων και γίνεται εστίαση στα υπόλοιπα



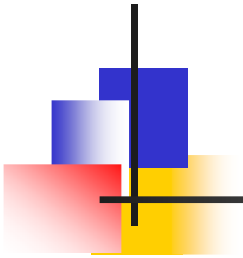
Αναφορές

- Aho, A. V. and M. J. Corasick (June 1975). "Fast Pattern Matching: An Aid to Bibliographic Search." CACM 18(6): 333-340.
- Boyer, R. S. and J. S. Moore (Oct. 1977). "A Fast String Searching Algorithm." CACM 20(10): 762-772.
- Brown, E. W., J. P. Callan, et al. (March 1994). Supporting Full-Text Information Retrieval with a Persistent Object Store. Proc. of EDBT conference, Cambridge, U.K., Springer Verlag.



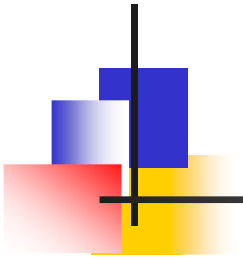
Αναφορές

- Faloutsos, C. and H. V. Jagadish (Aug. 23-27, 1992). On B-tree Indices for Skewed Distributions. 18th VLDB Conference, Vancouver, British Columbia.
- Karp, R. M. and M. O. Rabin (March 1987). "Efficient Randomized Pattern-Matching Algorithms." IBM Journal of Research and Development 31(2): 249-260.
- Knuth, D. E., J. H. Morris, et al. (June 1977). "Fast Pattern Matching in Strings." SIAM J. Comput 6(2): 323-350.



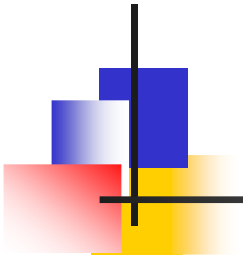
Αναφορές

- Mackert, L. M. and G. M. Lohman (August 1986). R* Optimizer Validation and Performance Evaluation for Distributed Queries. Proc. of 12th Int. Conf. on Very Large Data Bases (VLDB), Kyoto, Japan.
- Manber, U. and S. Wu (1994). GLIMPSE: A Tool to Search Through Entire File Systems. Proc. of USENIX Techn. Conf.
- McIlroy, M. D. (Jan. 1982). "Development of a Spelling List." IEEE Trans. on Communications COM-30(1): 91-99.



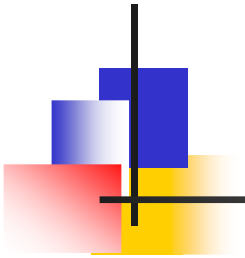
Αναφορές

- Mooers, C. (1949). Application of Random Codes to the Gathering of Statistical Information
- Bulletin 31. Cambridge, Mass, Zator Co.
- Pedersen, D. C. a. J. (1990). Optimizations for dynamic inverted index maintenance. ACM SIGIR.
- Riedel, E. (1999). Active Disks: Remote Execution for Network Attached Storage. ECE, CMU. Pittsburgh, PA.



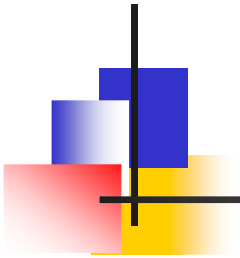
Αναφορές

- Severance, D. G. and G. M. Lohman (Sept. 1976). "Differential Files: Their Application to the Maintenance of Large Databases." ACM TODS 1(3): 256-267.
- Tomasic, A. and H. Garcia-Molina (1993). Performance of Inverted Indices in Distributed Text Document Retrieval Systems. PDIS.
- Tomasic, A., H. Garcia-Molina, et al. (May 24-27, 1994). Incremental Updates of Inverted Lists for Text Document Retrieval. ACM SIGMOD, Minneapolis, MN.



Αναφορές

- Wu, S. and U. Manber (1992). "AGREP- A Fast Approximate Pattern-Matching Tool." .
- Zobel, J., A. Moffat, et al. (Aug. 23-27, 1992). An Efficient Indexing Technique for Full-Text Database Systems. VLDB, Vancouver, B.C., Canada.

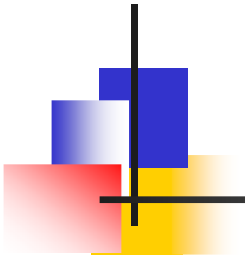


Κείμενο – Δομή διάλεξης

Κείμενο

- Πρόβλημα
- Σάρωση πλήρους κειμένου
- Αναστροφή
- Αρχεία υπογραφής
- Ομαδοποίηση
- Φιλτράρισμα πληροφορίας και LSI

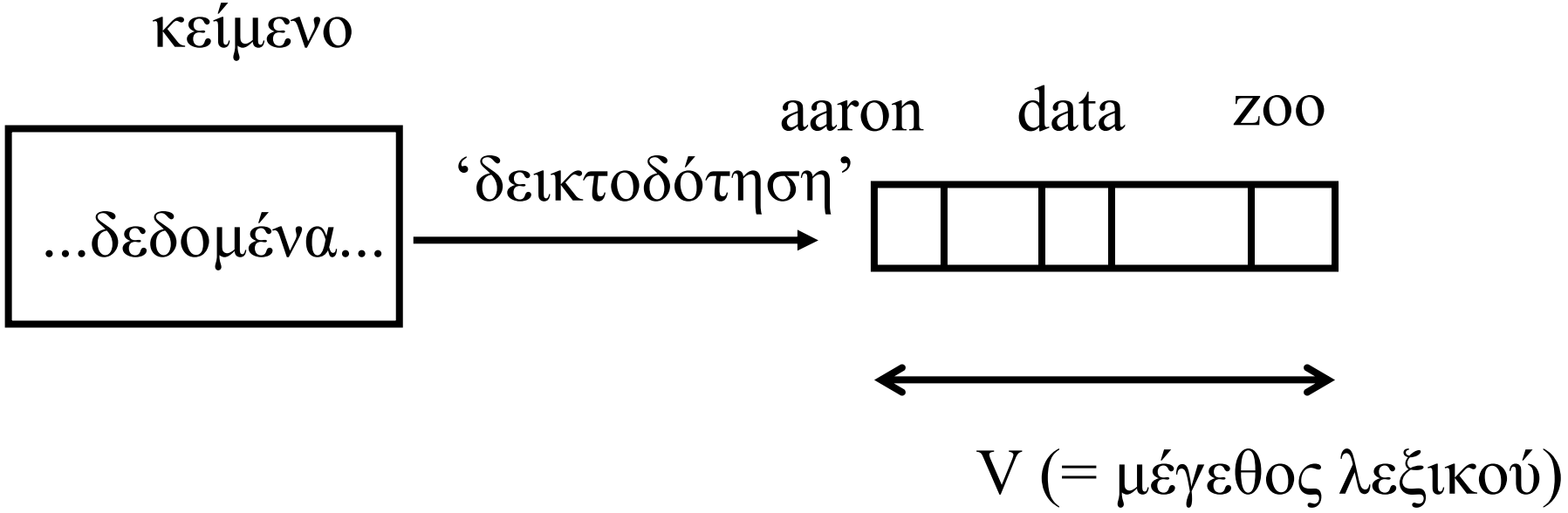
Μοντέλο διανυσματικού χώρου και ομαδοποίηση

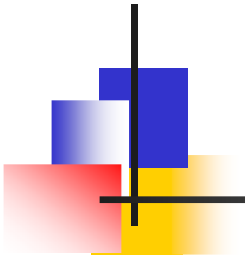


- Ερωτήματα λέξεων κλειδιών (vs Boolean)
- Κάθε κείμενο: -> διάνυσμα (ΠΩΣ;)
- Κάθε ερώτημα: -> διάνυσμα
- Αναζήτηση ‘όμοιων’ διανυσμάτων

Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Κεντρική Ιδέα:





Μοντέλο διανυσματικού χώρου και ομαδοποίηση

Έπειτα, ομαδοποίησε τα πλησιεστερα διανύσματα

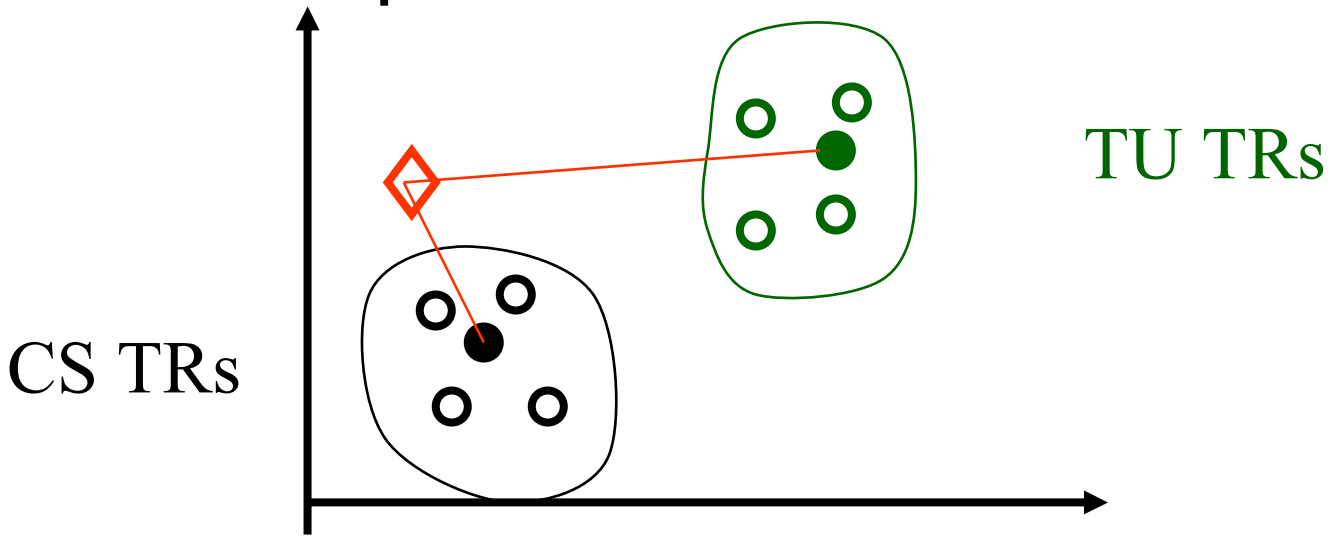
- Q1: Αναζήτηση ομάδας;
- Q2: Παραγωγή ομάδας;

Σημαντική συνεισφορά:

- Ταξινομημένη έξοδος
- Ανατροφοδότηση συνάφειας

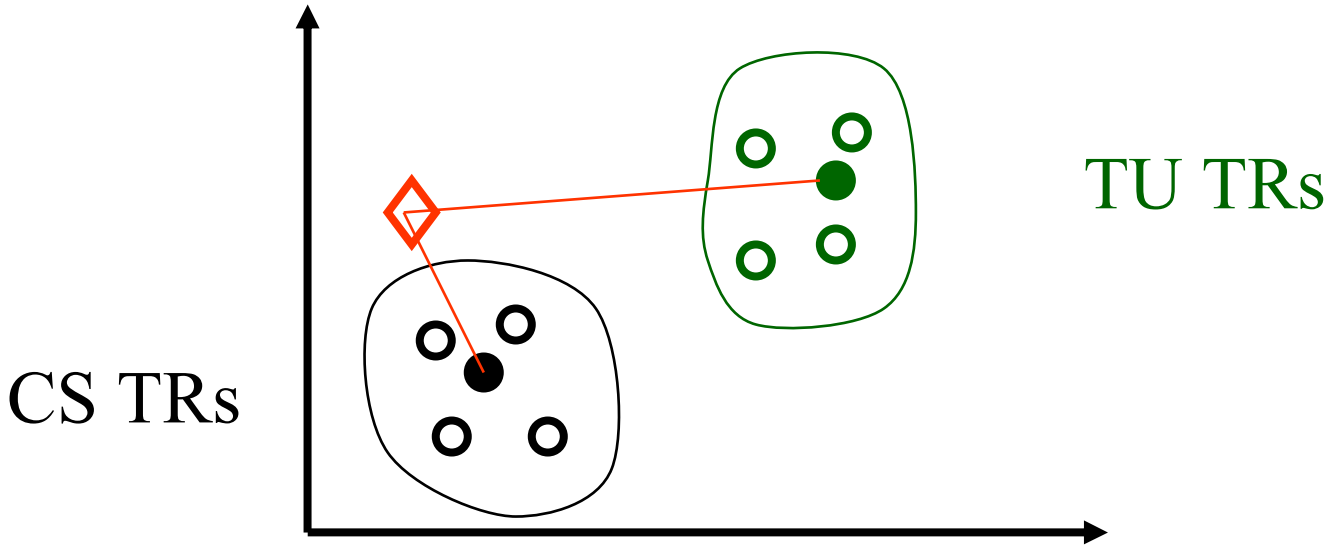
Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Αναζήτηση ομάδας: επισκέψου τις (k) πλησιέστερες υπερομάδες, συνέχισε επαναληπτικά



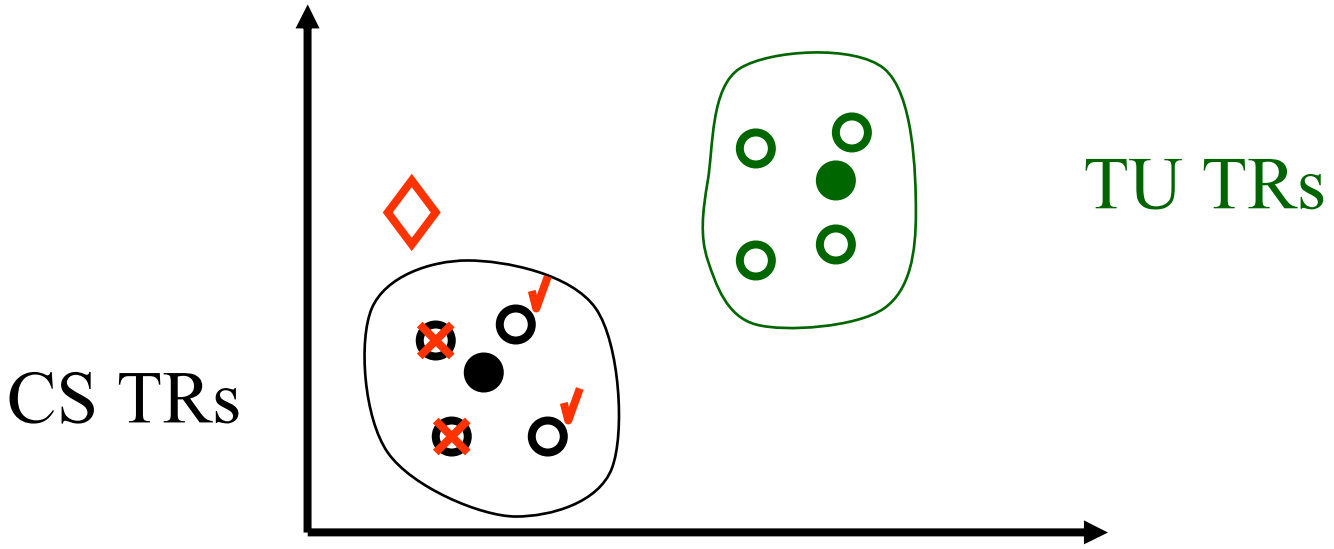
Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Ταξινομημένη έξοδος: εύκολο!



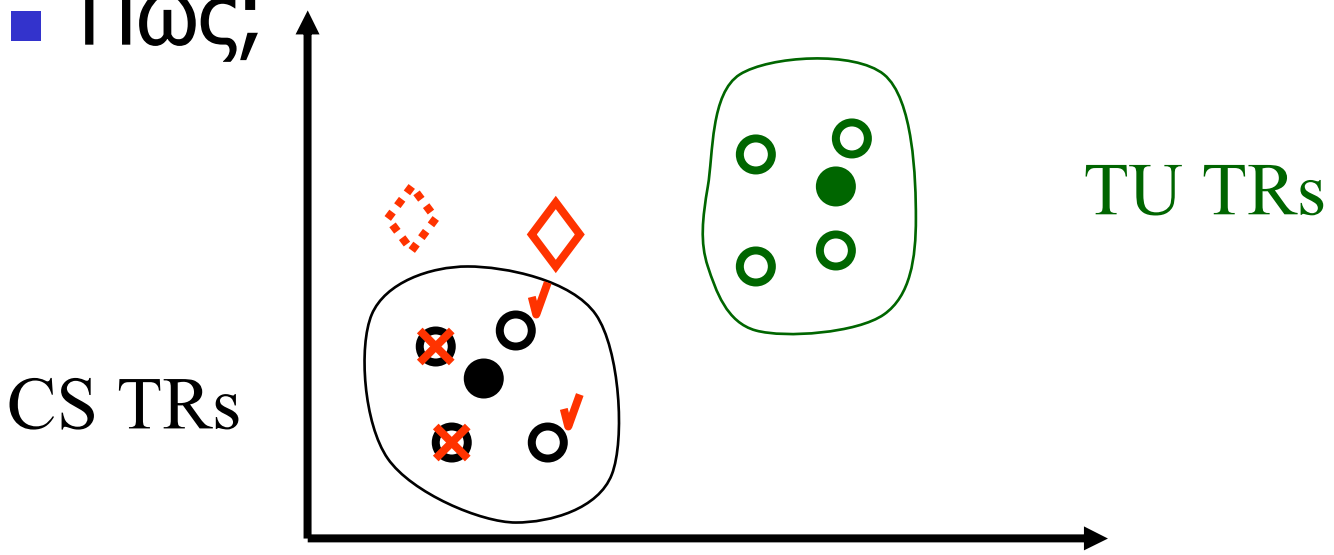
Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Ανατροφοδότηση συνάφειας (εξαιρετική ιδέα) [Roccio'73]



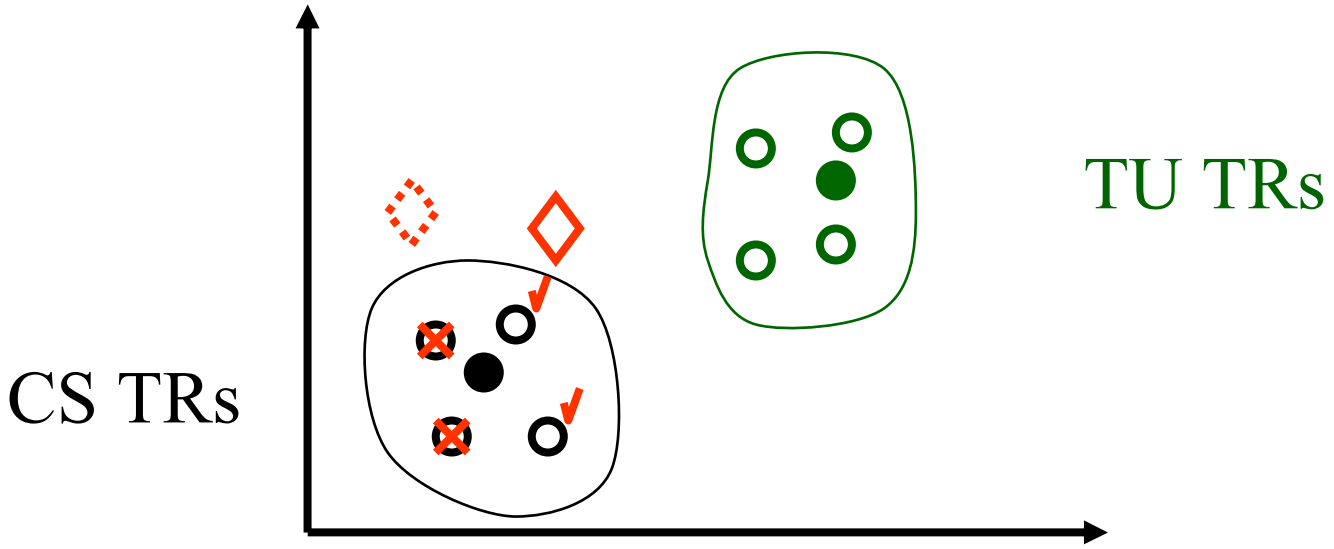
Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Ανατροφοδότηση συνάφειας (εξαιρετική ιδέα) [Roccio'73]
- Πώς;

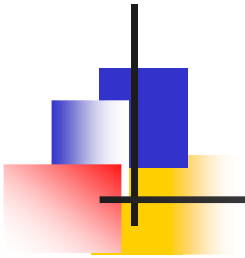


Μοντέλο διανυσματικού χώρου και ομαδοποίηση

- Πώς; Α: Προσθέτοντας 'καλά' διανύσματα και αφαιρώντας τα 'κακά'



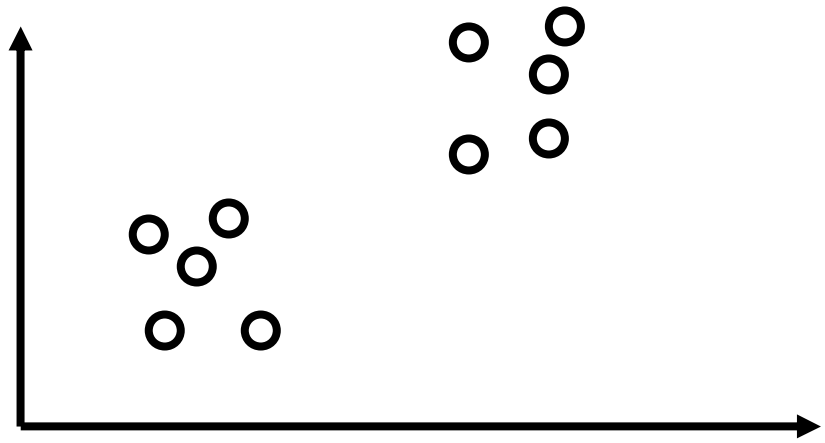
Σχεδιάγραμμα-Λεπτομερές



- Βασική ιδέα
- Αναζήτηση ομάδας
- Παραγωγή ομάδας
- Αξιολόγηση

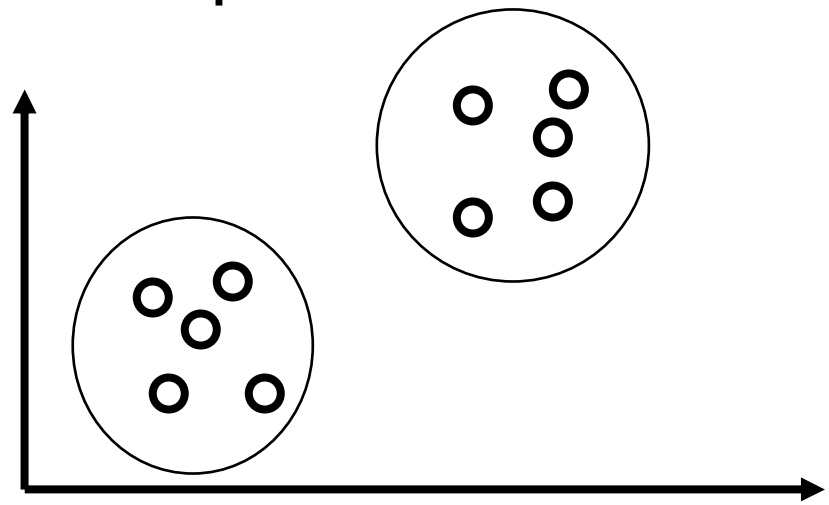
Παραγωγή ομάδας

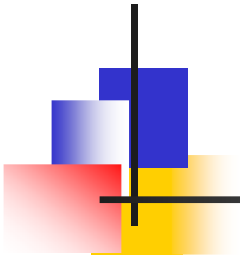
- Πρόβλημα:
  - Δεδομένων Ν σημείων σε V διαστάσεις,
  - ομαδοποίησέ τα



Παραγωγή ομάδας

- Πρόβλημα:
  - Δεδομένων Ν σημείων σε V διαστάσεις,
  - ομαδοποίησέ τα

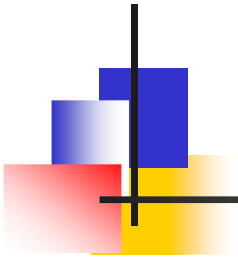




Παραγωγή ομάδας

Χρειαζόμαστε

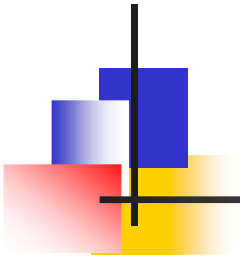
- Q1: ομοιότητα κειμένου με κείμενο
- Q2: ομοιότητα κειμένου με ομάδα



Παραγωγή ομάδας

Q1: Ομοιότητα κειμένου με κείμενο
(βλέπε: αναπαράσταση ‘σάκου με λέξεις’)

- D1: {‘data’, ‘retrieval’, ‘system’}
- D2: {‘lung’, ‘pulmonary’, ‘system’}
- Συναρτήσεις απόστασης/ομοιότητας;



Παραγωγή ομάδας

A1: # κοινών λέξεων

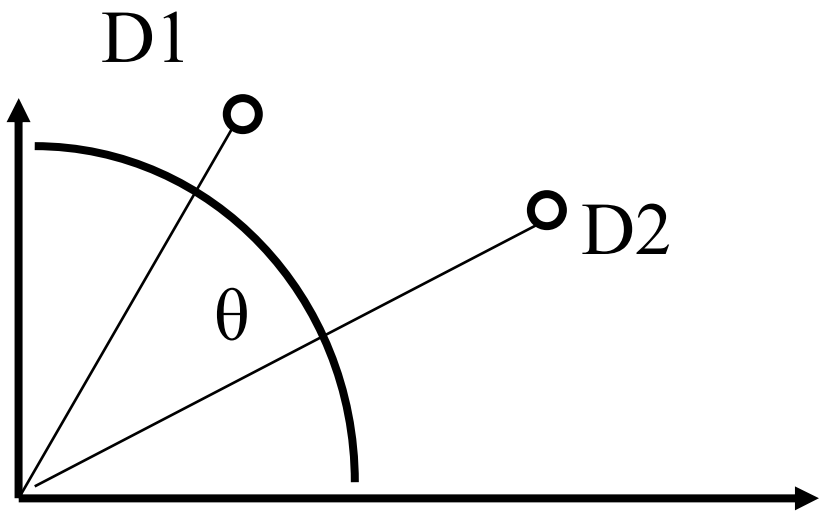
A2: ..... κανονικοποίηση με βάση τα μεγέθη του λεξικού

A3: .... κτλ

Παραγωγή ομάδας

Ομοιότητα συνημιτόνου:

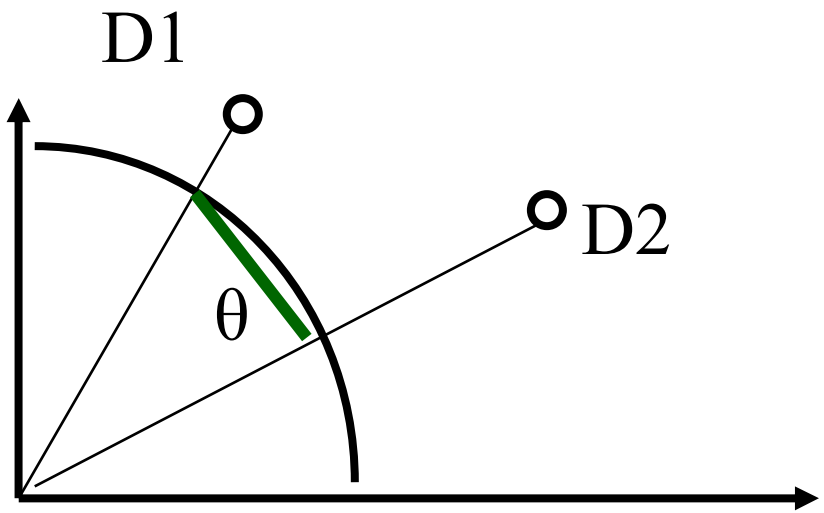
$$\text{similarity}(D1, D2) = \cos(\theta) = \text{sum}(v_{1,i} * v_{2,i}) / \text{len}(v_1) / \text{len}(v_2)$$

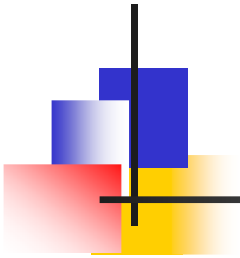


Παραγωγή ομάδας

Ομοιότητα συνημιτόνου- Παρατηρήσεις:

- Σχετίζεται με την Ευκλείδια απόσταση
- Ζυγίζει το $v_{i,j}$: σύμφωνα με την σχέση tf/idf





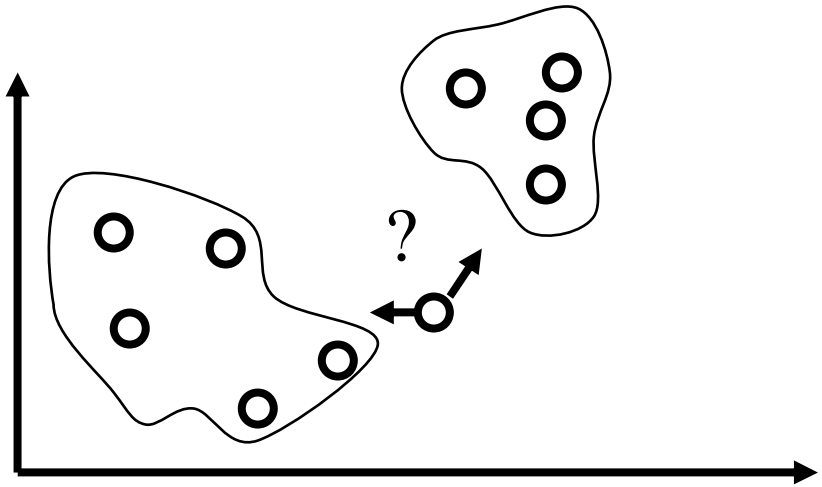
Παραγωγή ομάδας

- tf (‘συχνότητα όρου’)
  - Υψηλή, στην περίπτωση που ο όρος παρουσιάζεται πολύ συχνά μέσα στο κείμενο
- idf (‘αντίστροφη συχνότητα κειμένου’)
  - Δίνει λιγότερη σημασία στις ‘κοινές’ λέξεις, που εμφανίζονται σχεδόν σε κάθε κείμενο

Παραγωγή ομάδας

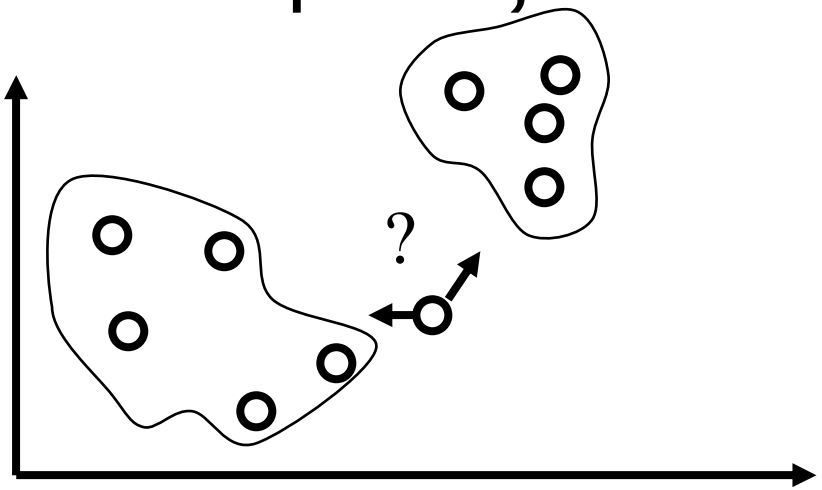
Χρειαζόμαστε

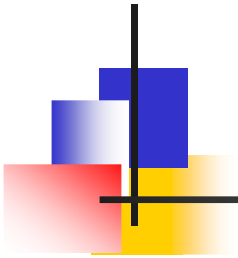
- Q1: ομοιότητα κειμένου με κείμενο
- Q2: ομοιότητα κειμένου με ομάδα



Παραγωγή ομάδας

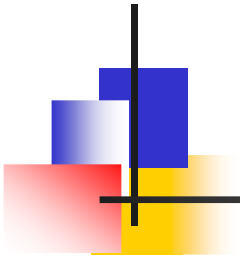
- A1: ελάχιστη απόσταση (‘single-link’)
- A2: μέγιστη απόσταση (‘all-link’)
- A3: μέση απόσταση
- A4: απόσταση από το κεντροϊδές





Παραγωγή ομάδας

- A1: ελάχιστη απόσταση (‘single-link’)
  - Οδηγεί σε μεγαλύτερες ομάδες
- A2: μέγιστη απόσταση (‘all-link’)
  - Πολλές, μικρές, αυστηρές ομάδες
- A3: μέση απόσταση
  - Μεταξύ των δύο ανωτέρω περιπτώσεων
- A4: απόσταση από το κεντροϊδές
  - Γρήγορο στον υπολογισμό

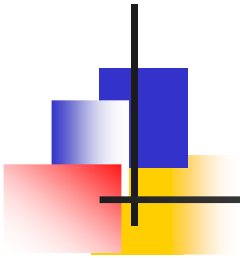


Παραγωγή ομάδας

Έχουμε

- Ομοιότητα κειμένου με κείμενο
- Ομοιότητα κειμένου με ομάδα

Q: Πώς γίνεται η ομαδοποίηση κειμένων σε ‘φυσικές’ ομάδες



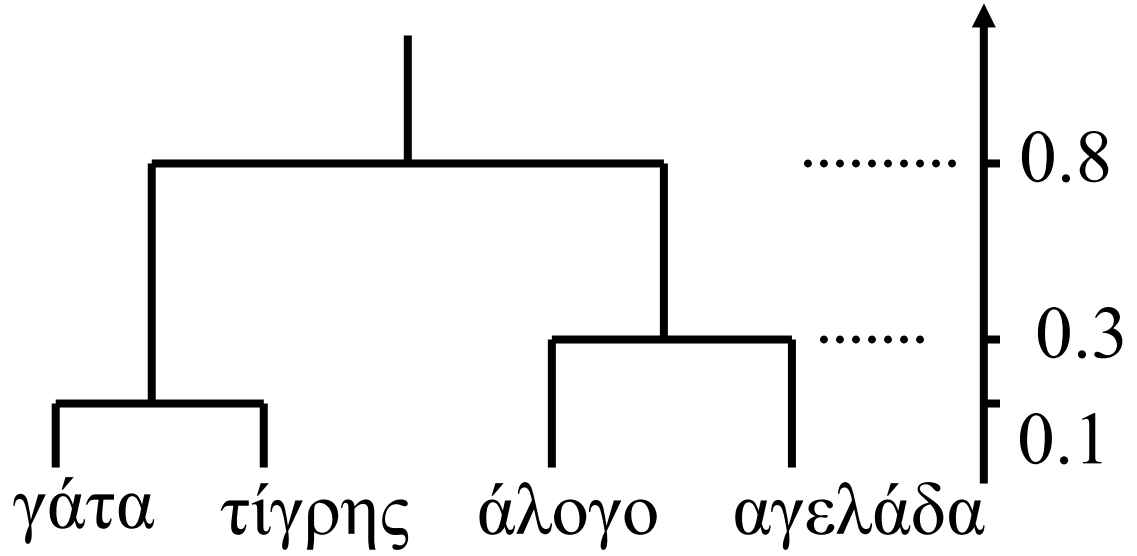
Παραγωγή ομάδας

Α: *many-many* αλγόριθμοι– σε δύο σύνολα [VanRijsbergen]:

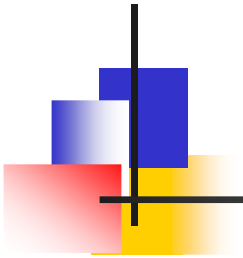
- Theoretically sound (O(*N^2*))
  - Ανεξάρτητα από την σειρά εισαγωγής
- Επαναληπτικός (O(*M*), O(*N* log(*M*))

Παραγωγή ομάδων- 'sound' methods

- Προσέγγιση #1: δενδρογράμματα- δημιούργησε μία ιεραρχία (από κάτω προς τα πάνω ή το αντίστροφο) – επέλεξε ένα όριο(πώς;) και περιόρισε



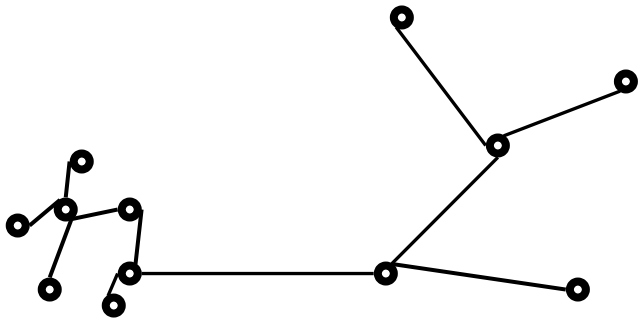
Παραγωγή ομάδας- ‘sound’ methods



- Προσέγγιση#2: ελαχιστοποίησε κάποιο στατιστικό κριτήριο (πχ., το άθροισμα τετραγώνων από τα κέντρα των ομάδων)
  - Όπως ‘k-means’
  - Όμως πώς επιλέγεται το ‘k’;

Παραγωγή ομάδας- ‘sound’ methods

- Προσέγγιση#3: Θεωρητικός γράφος[Zahn]:
  - Κατασκεύασε MST,
  - Διέγραψε τις ακμές με μήκος μεγαλύτερο του 2.5* std του τοπικού μέσου όρου

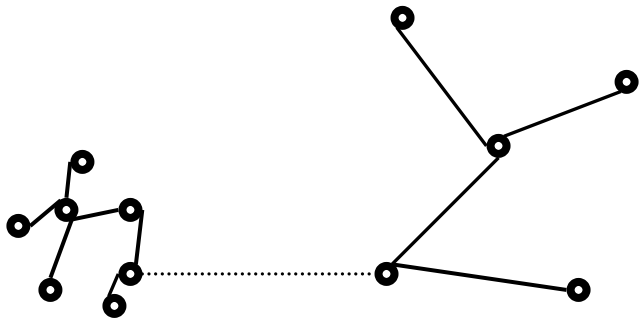


Παραγωγή ομάδας- ‘sound’ methods

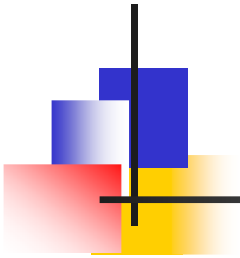
- Αποτέλεσμα:

• Παραλλαγές

•Πολυπλοκότητα;



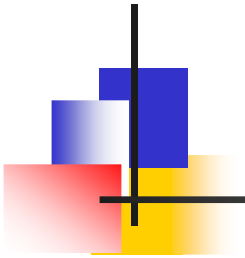
Παραγωγή ομάδας- ‘επαναληπτικές’ μέθοδοι



Γενικό σχεδιάγραμμα:

- Επέλεξε ‘φύτρα’ (Πώς;)
- Αντιστοίχισε κάθε διάνυσμα στην πλησιέστερη φύτρα (πιθανόν προσαρμόζοντας το κεντροϊδές της ομάδας)
- Πιθανόν, να ανατεθούν ξανά κάποια διανύσματα για να βελτιωθούν οι ομάδες

Γρήγορο και πρακτικό, αλλά ‘απρόβλεπτο’



Παραγωγή ομάδας

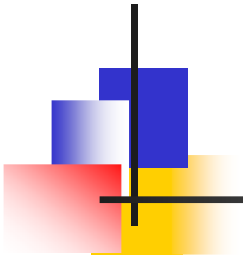
Ένας τρόπος να εκτιμηθεί το πλήθος των ομάδων *k*: 'cover coefficient' [Can+] ~ SVD

Σχεδιάγραμμα-Λεπτομερές

- Βασική ιδέα
- Αναζήτηση ομάδας
- Παραγωγή ομάδας
- Αξιολόγηση

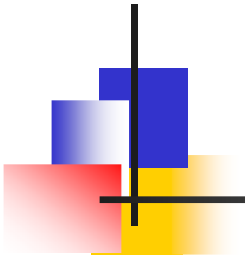
Αξιολόγηση

- Q: πώς μετριέται η ‘υπεροχή’ μίας συνάρτησης απόστασης σε σχέση με κάποια άλλη;
- A: σε πρώτη φάση από τους ανθρώπους και
  - ‘ακρίβεια’ & ‘επανάκληση’



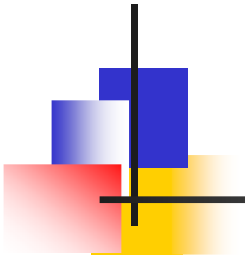
Αξιολόγηση

- Ακρίβεια= (ανακτηθείσα & σχετική) / ανακτηθείσα
  - 100% ακρίβεια-> χωρίς ενδείξεις σφάλματος
- Επανάκληση= (ανακτηθείσα & σχετική)/ανακτηθείσα
  - 100% επανάκληση-> χωρίς απώλεια σφάλματος



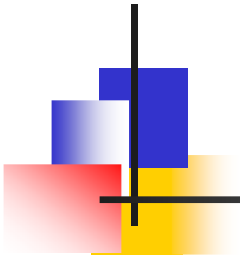
Αναφορές

- Can, F. and E. A. Ozkarahan (Dec. 1990). "Concepts and Effectiveness of the Cover-Coefficient-Based Clustering Methodology for Text Databases." ACM TODS 15(4): 483-517.
- Noreault, T., M. McGill, et al. (1983). A Performance Evaluation of Similarity Measures, Document Term Weighting Schemes and Representation in a Boolean Environment. Information Retrieval Research, Butterworths.
- Rocchio, J. J. (1971). Relevance Feedback in Information Retrieval. The SMART Retrieval System - Experiments in Automatic Document Processing. G. Salton. Englewood Cliffs, New Jersey, Prentice-Hall Inc.



Αναφορές

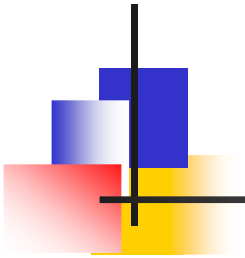
- Salton, G. (1971). The SMART Retrieval System - Experiments in Automatic Document Processing. Englewood Cliffs, New Jersey, Prentice-Hall Inc.
- Salton, G. and M. J. McGill (1983). Introduction to Modern Information Retrieval, McGraw-Hill.
- Van-Rijsbergen, C. J. (1979). Information Retrieval. London, England, Butterworths.
- Zahn, C. T. (Jan. 1971). "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters." IEEE Trans. on Computers C-20(1): 68-86.



Κείμενο – Δομή διάλεξης

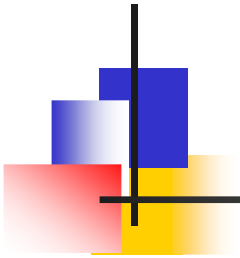
Κείμενο

- Πρόβλημα
- Σάρωση πλήρους κειμένου
- Αναστροφή
- Αρχεία υπογραφής
- Ομαδοποίηση
- Φιλτράρισμα πληροφορίας και LSI



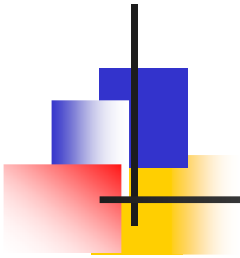
LSI – Λεπτομερές σχεδιάγραμμα

- LSI
  - Ορισμός προβλήματος
  - Βασική ιδέα
  - Πειράματα



Φιλτράρισμα πληροφορίας+ LSI

- [Foltz+,'92] Στόχος:
  - οι χρήστες προσδιορίζουν τα ενδιαφέροντά τους (= λέξεις κλειδιά)
  - Το σύστημα τα εντοπίζει, σε κατάλληλα νέα κείμενα
- Μέγιστη συνεισφορά: LSI = Latent Semantic Indexing
  - Κρυμμένες ('hidden') ιδέες



Φιλτράρισμα πληροφορίας+ LSI

Βασική ιδέα

- Αντιστοίχισε κάθε κείμενο σε κάποιες ‘ιδέες’
- Αντιστοίχισε κάθε όρο σε κάποιες ‘ιδέες’

‘Ιδέα’ (‘concept’):~ ένα σύνολο όρων, με βάρη, π.χ.

- “data” (0.8), “system” (0.5), “retrieval” (0.6) -> DBMS_concept

Φιλτράρισμα πληροφορίας+ LSI

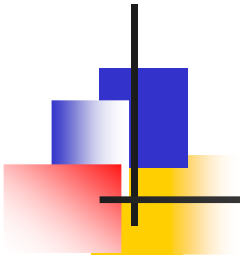
Απεικόνιση: πίνακας όρων-κειμένου (BEFORE)

	'data'	'system'	'retrieval'	'lung'	'ear'
TR1	1	1	1		
TR2	1	1	1		
TR3				1	1
TR4				1	1

Information Filtering + LSI

Απεικόνιση: πίνακας ιδέας-κειμένου και...

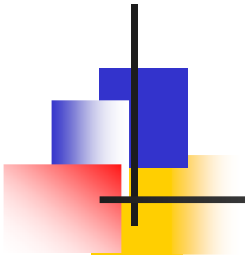
	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1



Φιλτράρισμα πληροφορίας+ LSI

... και πίνακας όρου-κειμένου

	'DBMS-concept'	'medical-concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1



Φιλτράρισμα πληροφορίας+ LSI

Q: Πώς γίνεται η αναζήτηση, πχ., για το ‘system’;

Φιλτράρισμα πληροφορίας+ LSI

Α: βρες την αντίστοιχη ιδέα(ες) και τα αντίστοιχα κείμενα

	'DBMS-concept'	'medical-concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1

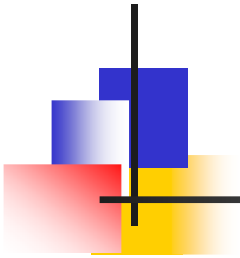
	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1

Φιλτράρισμα πληροφορίας+ LSI

Α: βρες την αντίστοιχη ιδέα(ες) και τα αντίστοιχα κείμενα

	'DBMS-concept'	'medical-concept'
data	1	
system	1	
retrieval	1	
lung		1
ear		1

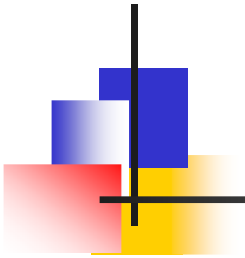
	'DBMS-concept'	'medical-concept'
TR1	1	
TR2	1	
TR3		1
TR4		1



Φιλτράρισμα πληροφορίας+ LSI

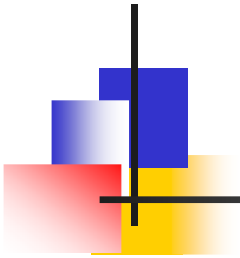
Με τον τρόπο αυτό λειτουργεί σαν (αυτόματη κατασκευή) θησαυρός:

Μπορεί να ανακτήσουμε κείμενα τα οποία ΔΕΝ έχουν τον όρο ‘system’, αλλά περιέχουν σχεδόν όλους τους άλλους (‘data’, ‘retrieval’)



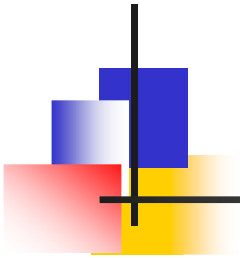
LSI – Λεπτομερές σχεδιάγραμμα

- LSI
  - Ορισμός προβλήματος
  - Βασική ιδέα
  - Πειράματα



LSI - Πειράματα

- 150 Tech Memos (TM) / μήνα
- 34 ‘προφίλ’ χρηστών (6-66 λέξεις ανά προφίλ)
- 100-300 ιδέες

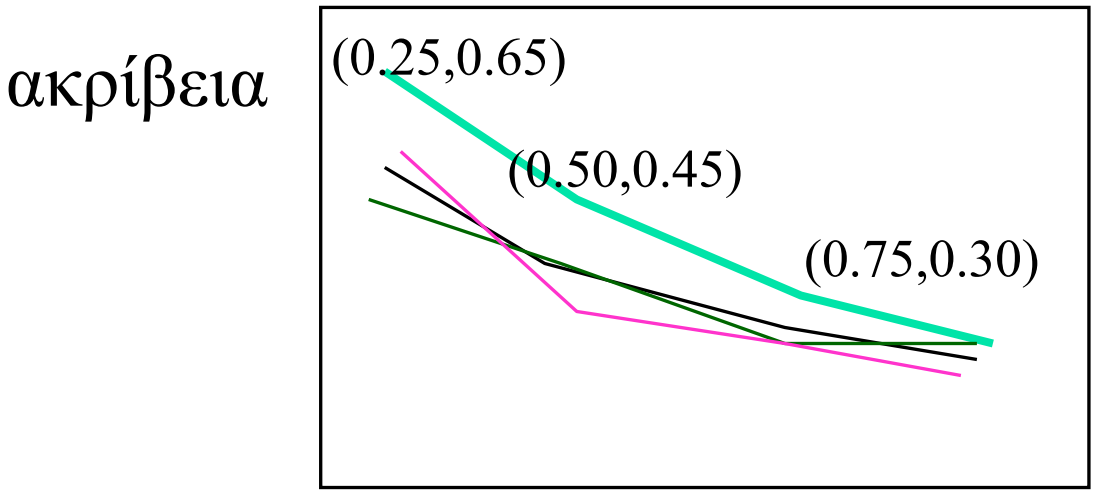


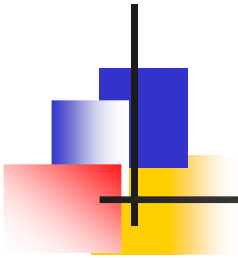
LSI - Πειράματα

- Τέσσερεις μέθοδοι, προϊόν των ακολούθων:
  - Vector-space / LSI, για την μέτρηση ομοιότητας
  - Λέξεις κλειδιά ή δείγμα κειμένου, για τον προσδιορισμό του προφίλ
- Μετρημένη : ακρίβεια/ανάκληση

LSI - Πειράματα

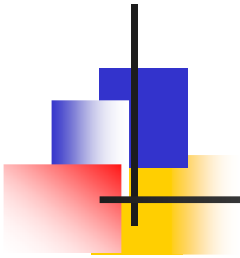
- LSI, με προφίλ βάσει κειμένων, καλύτερη





LSI - Συζήτηση- Συμπεράσματα

- Σημαντική ιδέα,
  - Η παραγωγή ‘ιδεών’ από τα κείμενα
  - Η αυτόματη παραγωγή ‘στατιστικού θησαυρού’
  - Η μείωση των διαστάσεων
- Συχνά οδηγεί σε καλύτερη σχέση ακρίβειας/ανάκλησης
- Αλλά:
  - Χρειάζεται ένα σύνολο κειμένων προετοιμασίας (‘training’ set)
  - Τα διανύσματα ‘ιδεών’ είναι πλέον πλήρη



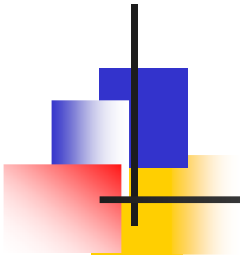
LSI – Συζήτηση- Συμπεράσματα

Παρατηρήσεις

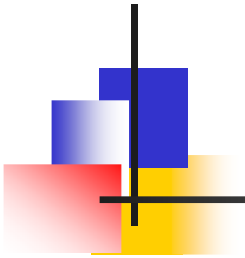
- Bellcore (-> Telcordia): έχει μία πατέντα
- Χρησιμοποιείται για πολύγλωσση ανάκτηση

SVD: Πώς λειτουργεί ακριβώς;

Δεικτοδότηση- Λεπτομερές σχεδιάγραμμα



- Δεικτοδότηση πρωτεύοντος κλειδιού
- Δεικτοδότηση δευτερεύοντος κλειδιού/multi-key
- Μέθοδοι χωρικής προσπέλασης
- fractals
- Κείμενο
- SVD: ένα εργαλείο με μεγάλες δυνατότητες
- πολυμέσα
- ...



Αναφορές

- Foltz, P. W. and S. T. Dumais (Dec. 1992). "Personalized Information Delivery: An Analysis of Information Filtering Methods." Comm. of ACM (CACM) 35(12): 51-60.