



## HY463 - Συστήματα Ανάκτησης Πληροφοριών Information Retrieval (IR) Systems

Γιάννης Τζίτζικας

Διάλεξη : 1

Ημερομηνία : 22-2-2006

Θέμα : Διαδικαστικά και Εισαγωγή



## HY463 – Συστήματα Ανάκτησης Πληροφοριών (CS463 - Information Retrieval Systems)

- Διδακτικές μονάδες: 4
- Προαπαιτούμενα
  - HY240 - Δομές Δεδομένων
- Εβδομαδιαίο Πρόγραμμα (**ίσως αλλάξει**):
  - **Διαλέξεις**: Τετάρτη 3-5 και Πέμπτη 5-7 στην αίθουσα PA201
  - **Φροντιστήρια**: Παρασκευή 5-7 στην αίθουσα PA201
    - (θα στέλνεται email πριν από κάθε φροντιστήριο)
- Παρακολούθηση
  - Αναμενόμενη αλλά όχι υποχρεωτική
  - Η ενεργή συμμετοχή στο μάθημα θα ληφθεί θετικά υπόψη
- Γραφείτε (σήμερα) στη λίστα **hy463-list**



## Προσωπικό

- **Διδάσκων:**
  - Γιάννης Τζιτζίκας
  - tzizik (at) csd.uoc.gr
  - Γραφείο: Γ111
  - Ώρες γραφείου: πριν τις διαλέξεις (Τετάρτη 2-3, Πέμπτη 3-5)
- **Βοηθοί:**
  - Κώστας Βανδίκας
  - Γιάννης Θεοχάρης
  - Υπεύθυνοι για:
    - Λύση και βαθμολόγηση ασκήσεων
    - Επίβλεψη εργασιών
    - Φροντιστήρια,
    - Απάντηση ερωτήσεων

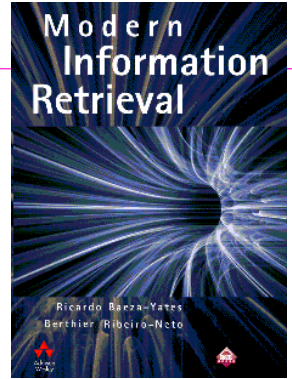


## Ιστοσελίδα μαθήματος

- [www.csd.uoc.gr/~hy463](http://www.csd.uoc.gr/~hy463)
  - Τελευταίες Ανακοινώσεις
  - Περιγραφή Μαθήματος - Διδακτέα Ύλη
  - Πρόγραμμα Διαλέξεων
  - Διαφάνειες Διαλέξεων, Πρόγραμμα Μελέτης
  - Ασκήσεις, Λύσεις, Βαθμολογίες
  - Ύλη Μαθήματος
  - Συνδέσμους σε συμπληρωματικό διδακτικό υλικό (βιβλία, άρθρα, σχετικές διαδικτυακές πύλες, ανάλογα μαθήματα σε άλλα Παν/μια, κλπ).



## Διδακτικό Ύλικό



- **Κύριο Βιβλίο**
  - *Modern Information Retrieval*, by Baeza-Yates and Ribeiro-Neto
- **Πρόσθετα Βιβλία και Ερευνητικά Άρθρα**
  - θα αναρτώνται στην ιστοσελίδα (ήδη υπάρχουν κάποια)
- **Φωτοτυπίες κεφαλαίων από το κύριο βιβλίο**
  - συνεννοηθείτε με τους βοηθούς



## Σειρές Ασκήσεων

- **Σκοπός:**
  - η κατανόηση και εμπέδωση της ύλης, και η συνεχής επαφή με το μάθημα κατά τη διάρκεια του εξαμήνου
- **Θα δοθούν 4 ή 5 σειρές ασκήσεων**
  - 1. Αξιολόγηση της αποτελεσματικότητας της ανάκτησης
  - 2. Μοντέλα ανάκτησης
  - 3. Οργάνωση ευρετηρίων για έγγραφα
  - 4. Κατανεμημένη ανάκτηση πληροφοριών
  - 5 ...
- **Βάρος: 20% του τελικού βαθμού**



## Πρόοδος

- Υποχρεωτική (20% τελικού βαθμού) και θα γίνει μάλλον στις αρχές του Απριλίου



## Εργασία (project) του 2005

- Θέμα:
  - Υλοποίηση ενός Συστήματος Ανάκτησης Πληροφοριών με ψευδοανάδραση συνάφειας (pseudo relevance feedback)
- Χρονοδιάγραμμα
  - Έναρξη: 1 Απρίλη
  - Πέρας: Μέσα Μάη
- Ομάδες 2 ατόμων
- Υλοποίηση σε Java
- Βάρος: 30% Τελικού βαθμού



## Εργασία (project) του 2006: ... GRoogle2006

- Ανάπτυξη μια μηχανής αναζήτησης για τον παγκόσμιο ιστό.
- Κάθε ομάδα θα αναλάβει μόνο κάποια υποσυστήματα αυτής της μηχανής.



## Βαθμολόγηση

- **Τελικός βαθμός**
  - **Total** = 20% Ασκήσεις + 30% Εργασία + 20% Πρόοδος + 30% Τελική Εξ
- Για να περάσετε το μάθημα χρειάζεστε
  - **Total**  $\geq 5$  **AND** Τελική Εξ  $\geq 4$

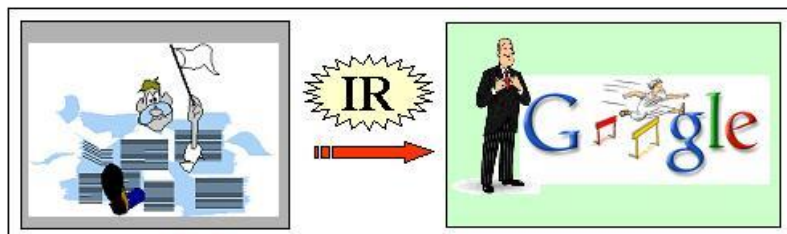


# Εντιμότητα

- Αντιγραφή ή άλλες μορφές κλοπής θα σημάνουν αυτόματα αποτυχία στο μάθημα
- Συμβουλές
  - μην αντιγράφετε ή δίνετε τις εργασίες σας σε άλλους
  - προστατέψτε τα αρχεία και τα έγγραφά σας
  - πάντα να αναφέρετε τις πηγές σας (άτομα, βιβλία, Web)



# Ανάκτηση Πληροφοριών (Information Retrieval): Το τυπικό πρόβλημα





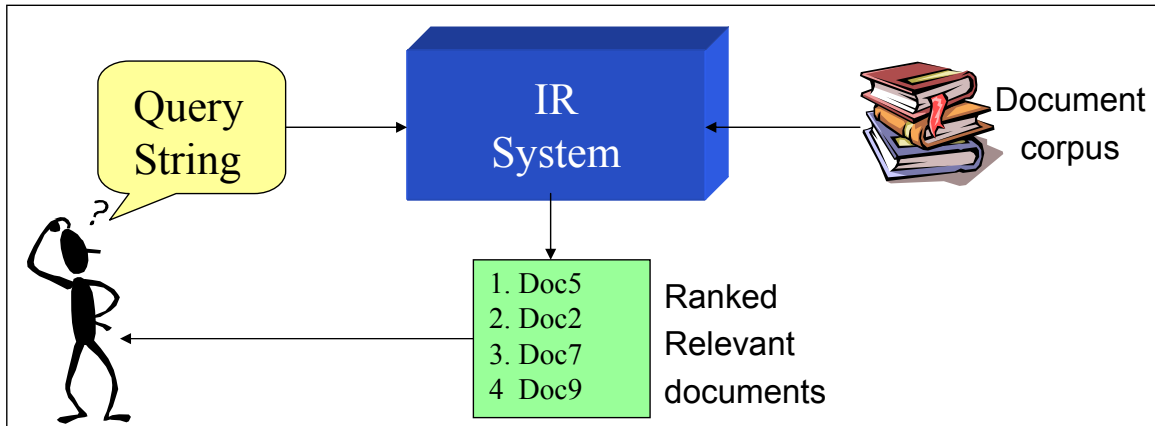
# Ανάκτηση Πληροφοριών (Information Retrieval): Το τυπικό πρόβλημα

## Δεδομένα Προβλήματος

- Μια συλλογή από έγγραφα με κείμενο φυσικής γλώσσας  $D=\{d_1, \dots, d_n\}$
- Μια επερώτηση  $q$  ενός χρήστη σε μορφή συμβολοσειράς (string)

## Ζητούμενο

- Ένα διατεταγμένο σύνολο από έγγραφα που είναι συναφή με την επερώτηση  $\langle d_5, d_2, d_7, d_9 \rangle$



# Περιγραφή Μαθήματος

## ΣΚΕΠΤΙΚΟ:

Τα Συστήματα Ανάκτησης Πληροφοριών (Information Retrieval systems) επιτρέπουν την πρόσβαση σε μεγάλους όγκους πληροφοριών αποθηκευμένων με τη μορφή κειμένου, φωνής, video, ή σε σύνθετη μορφή όπως Ιστοσελίδες.

**Σκοπός** των συστημάτων αυτών είναι η ανάκτηση μόνο εκείνων των εγγράφων που είναι συναφή με αυτό που αναζητεί ο χρήστης. Για να το επιτύχουν πρέπει να αντιμετωπίσουν την αβεβαιότητα ως προς το τι πραγματικά αναζητεί ο χρήστης και ποιο το θέμα ενός εγγράφου.

## Σκοπός του μαθήματος

Εισαγωγή στην περιοχή των συστημάτων ανάκτησης πληροφοριών και εξέταση των θεωρητικών και πρακτικών ζητημάτων που σχετίζονται με την σχεδίαση, υλοποίηση και αξιολόγηση τέτοιων συστημάτων.



## Στόχοι του μαθήματος

- Μετά το πέρας αυτού του μαθήματος πρέπει να:
  - έχετε κατανοήσει τη θεωρητική βάση των καθιερωμένων μοντέλων ανάκτησης (Boolean, Vector Space, Probabilistic, Logical Models),
  - έχετε κατανοήσει τεχνικές παράστασης και ανάκτησης εγγράφων, εικόνων, ομιλίας, κλπ,
  - έχετε μάθει να υλοποιείτε και να αξιολογείτε ένα σύστημα ανάκτησης πληροφοριών,
  - να έχετε κατανοήσει τους καθιερωμένους τρόπους ευρετηρίασης και ανάκτησης του Παγκόσμιου Ιστού,
  - να έχετε γνωρίσει ποικίλους αλγόριθμους και συστήματα.



## Εισαγωγή στην Ανάκτηση Πληροφοριών Διάρθρωση

- *Γιατί χρειαζόμαστε Ανάκτηση Πληροφοριών (ΑΠ);*
- *Τι είναι η ΑΠ ?*
- *Ανάκτηση, Διήθηση, Πλοήγηση*
- *Μοντέλα Πλοήγησης*
- *Το βασικό πρόβλημα ΑΠ*
- *Ανάκτηση Δεδομένων και Ανάκτηση Πληροφοριών*
- *Συνάφεια*
- *Η βασική προσέγγιση & αρχιτεκτονική ενός ΣΑΠ*
- *ΑΠ στον Παγκόσμιο Ιστό*
- *Άλλες λειτουργίες ΑΠ*
- *Ιστορική Αναδρομή*
- *Σχετικές Περιοχές*





## Γιατί χρειαζόμαστε ΑΠ ?

- Για να μπορούμε να ... **βρίσκουμε ψύλλους στ' άχυρα**
- **Πόσο εύχρηστος θα ήταν ο Ιστός χωρίς μηχανές αναζήτησης;**
  - Ο Ιστός περιέχει δισεκατομμύρια σελίδες (η μηχανή Google έχει ευρετηριάσει περίπου **8 δισεκατομμύρια** σελίδες)
- Ο "κόσμος" παράγει περίπου **2 exabytes** ( $2^{60}$ ) νέας πληροφορίας το χρόνο, 90% της οποίας είναι σε ψηφιακή μορφή και με 50% ετήσια αύξηση



## Το πρόβλημα δεν είναι νέο

"There is a growing mountain of research... The investigator is staggered by the findings and conclusions of thousands of other workers - conclusions which he cannot find time to grasp, much less remember. The summation of human experience is being expanded at a prodigious rate and the means we use for threading through the consequent maze to the momentarily important item is the same that was used in the days of the square rigged ships."

V. Bush 1945



## Το πρόβλημα είναι σημαντικό και επίκαιρο

(Εφημερίδα: Το ΒΗΜΑ 22/1/2006)

“Μέσα σε μόλις επτά χρόνια μια παγκόσμια αυτοκρατορία εξαπλώθηκε.

Όχι δεν έχει στρατό και πλοία.

Είναι μια εξουσία της γνώσης: η μεγαλύτερη μηχανή διύλισης - για την ακρίβεια - των πληροφοριών που κυκλοφορούν στο Διαδίκτυο.

Είναι δωρεάν και προσφέρει απλόχερα τις αγαθοεργούς υπηρεσίες της εν είδει “καθολικής και αποστολικής εκκλησίας της γνώσης”.

Και όπως κάθε παγκόσμια εκκλησία, έχει θησαυρίσει.

Με δεδομένη την καχυποψία μας για κάθε αυτοκρατορική εξουσία και με τη φθονερή βεβαιότητα ότι ... “ουδέν καλόν αμιγές κακού”,  
ας δούμε ποια είναι και που το πάει η Google”



## Το πρόβλημα είναι σημαντικό και επίκαιρο

(έως και οι πολιτικοί άρχισαν να ασχολούνται με αυτό)

- Ο πρόεδρος της Γαλλίας σήμανε προσκλητήριο για μια ευρωπαϊκή μηχανή αναζήτησης που θα απέκρουε τον αγγλοσαξονικό πολιτισμικό ιμπεριαλισμό.
- Εξήγγειλε ως βασική προτεραιότητα του για το 2006 το Project Quaero (“Ερευνώ” στα λατινικά), την υλοποίηση δηλαδή μιας ευρωπαϊκής μηχανής αναζήτησης
  - 30/8/2005: “Βρισκόμαστε στο μέσον ενός παγκόσμιου ανταγωνισμού για τεχνολογική υπεροχή. Στη Γαλλία, στην Ευρώπη, διακυβεύεται η αυτοκυριαρχία μας.”
  - 1/1/2006: “Σήμερα χαράσσεται η νέα γεωγραφία της γνώσης και των πολιτισμών. Αύριο εκείνο που δεν είναι ευρέσιμο στο Διαδίκτυο κινδυνεύει να είναι αθέατο από τον κόσμο.”
- Project Quaero
  - Συνεταίροι: Thomson, France Telecom, Deutsche Telekom, CNRS, RWTH (Aachen), INRIA, Bertelsmann, ...
  - Θα επεκταθεί η υπάρχουσα μηχανή Exalead
    - αυτόματη μετάφραση, καταλογογράφηση, ...



## Τι να είναι η ΑΠ;

Εύρεση του:

Ταίριασμα πρεζών - κεφαλαίων  
 Ταίριασμα μόνο ολόκληρων λέξεων

Εύρεση: Όλα τα αρχεία  
Αρχείο Επεξεργασία Προβολή Επιλογές Βοήθεια

Όνομα και θέση Ημερομηνία τροποποίησης Για προχωρημένους

Με όνομα:   
Με τα κείμενα:   
Διερεύνηση στο:

Να συμπεριληφθούν οι μπαράκελοι

grep

YAHOO!

altavista

CiteSeer.IST  
Scientific Literature Digital Library

Ask Jeeves

Google

www.vivisimo.com



## Τι να είναι η ΑΠ;

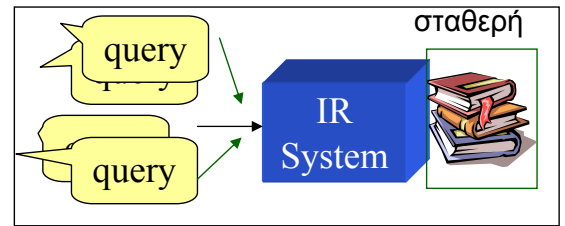
- Μήπως οι μηχανές αναζήτησης όπως το Google, Lycos ?
  - Αρκετά αποτελεσματικές (σε μερικά πράγματα)
  - Αναγνωρίσιμες και γνωστές
  - Εμπορικά επιτυχημένες (τουλάχιστον μερικές)
- Τι συμβαίνει όμως **πίσω** από τη σκηνή ;
- **Πως** δουλεύουν?
- Πως μπορούμε να κρίνουμε αν **δουλεύουν καλά**;
- Πως μπορούμε να τις κάνουμε **πιο αποτελεσματικές**;
- Πως μπορούμε να τις κάνουμε να λειτουργούν **πιο γρήγορα**;
- Υπάρχει τίποτα παραπάνω από αυτό που βλέπουμε στον Ιστό;



# Ανάκτηση και Φιλτράρισμα

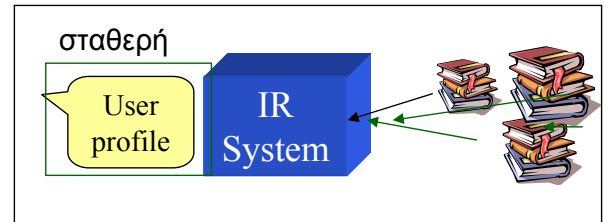
## Ανάκτηση (ad hoc retrieval):

- Σταθερή συλλογή εγγράφων, μεταβαλλόμενες επερωτήσεις

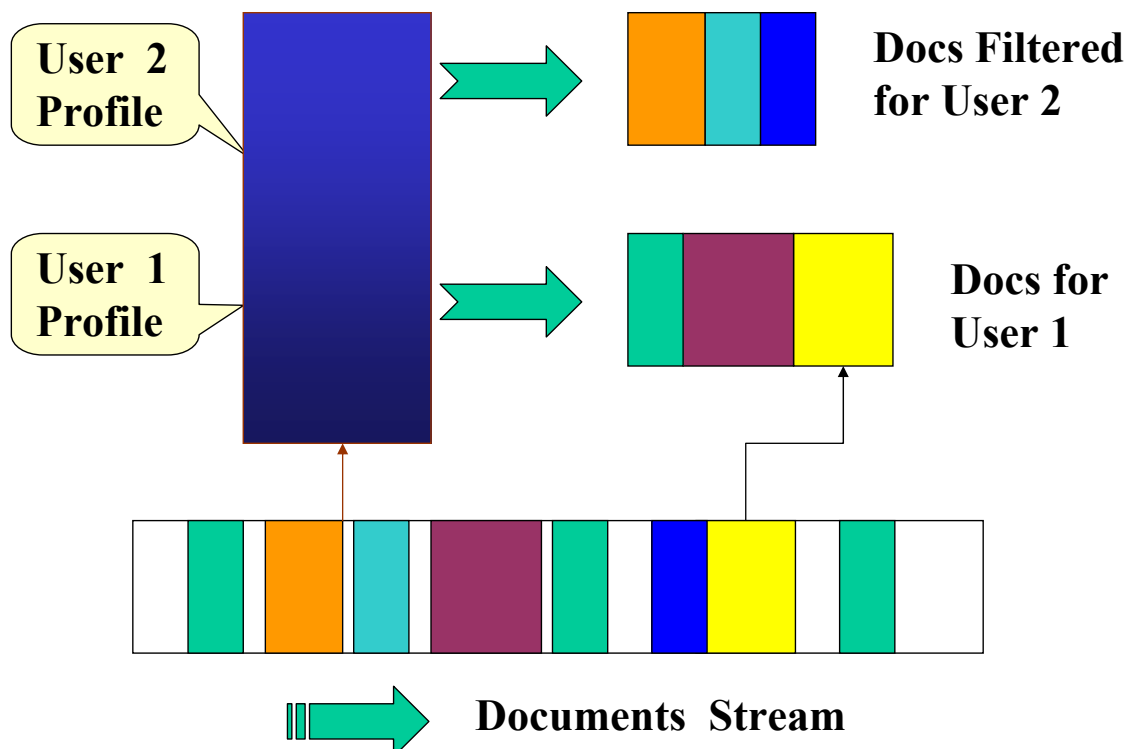


## Φιλτράρισμα ή Διήθηση (Filtering):

- Σταθερή επερώτηση, ροή νέων κειμένων
- **Προφίλ Χρήστη** = Επερώτηση που εκφράζει πιο μόνιμες προτιμήσεις
- Έμφαση στη δημιουργία/ενημέρωση του προφίλ

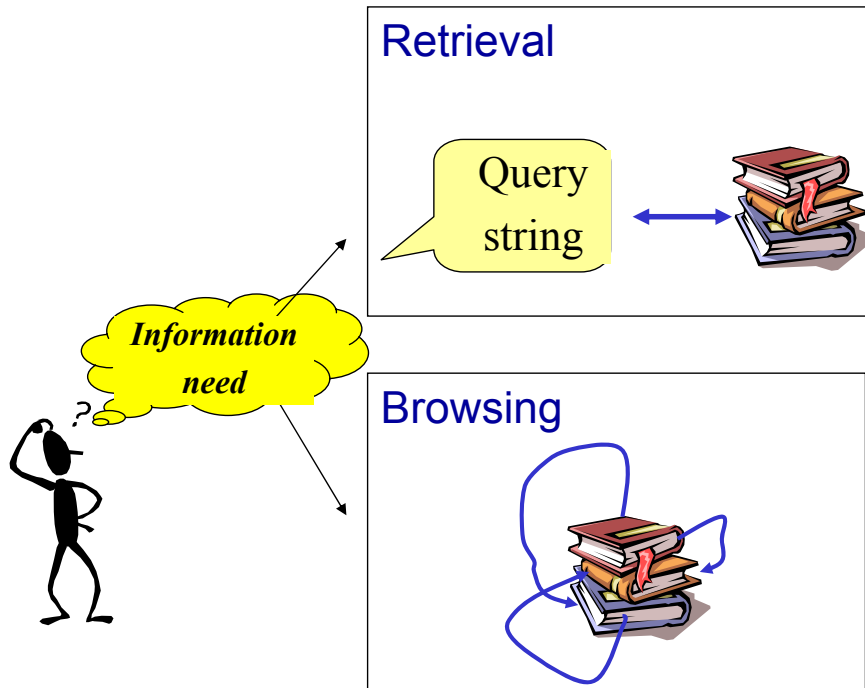


# Φιλτράρισμα





# Ανάκτηση και Πλοήγηση (Retrieval vs Browsing)



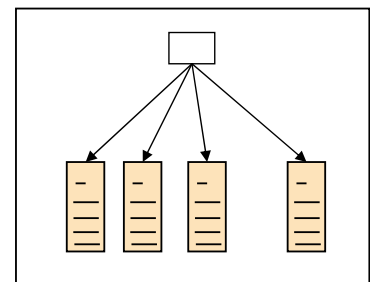
- Ο χρήστης δεν διατυπώνει επερώτηση
- Πλοήγηση με σκοπό ή χωρίς σκοπό (εξερεύνηση).



# Τύποι Πλοήγησης (Types of Browsing)

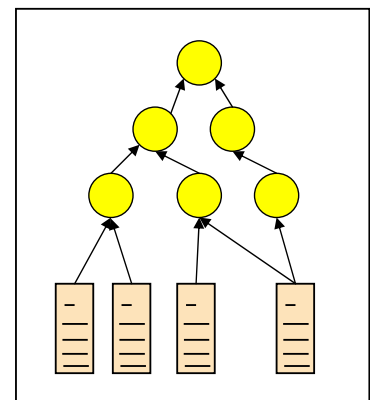
## (1) Επίπεδο (flat)

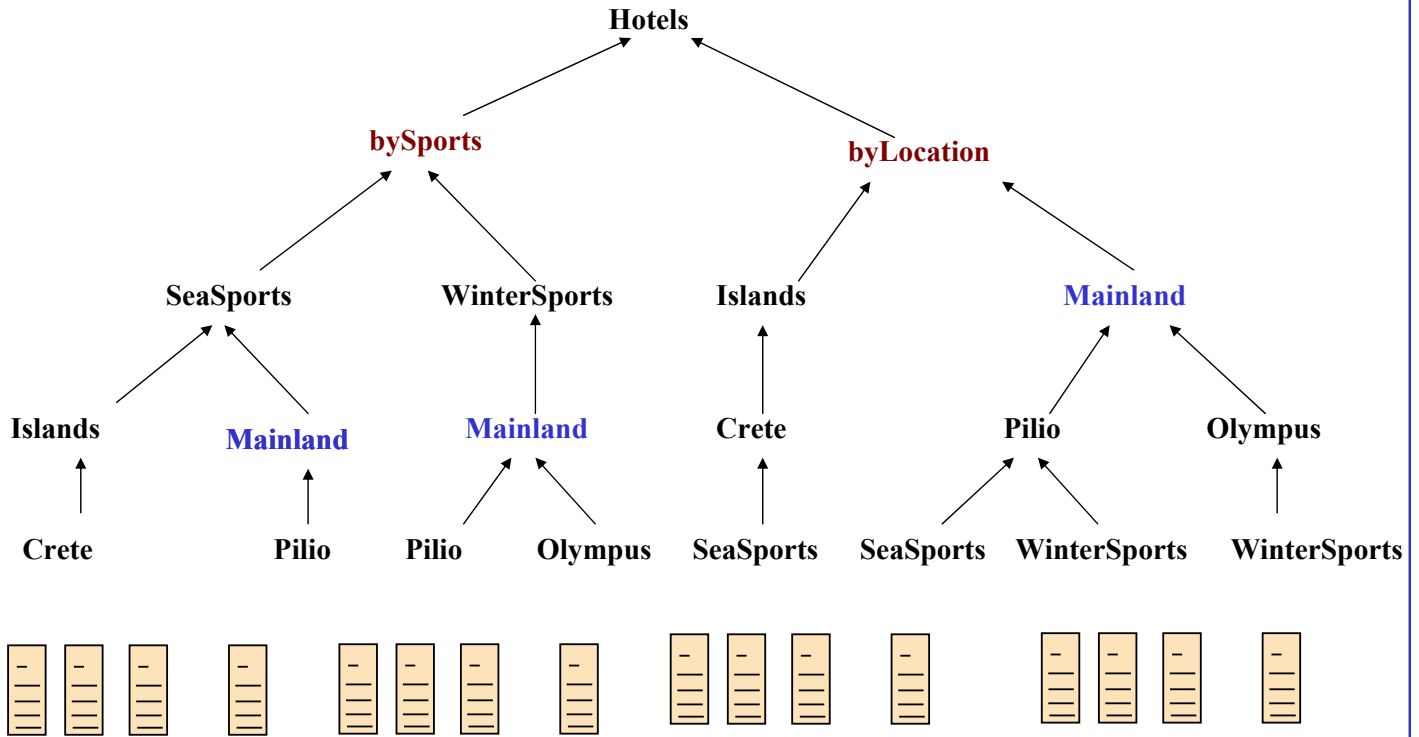
- πχ. μια λίστα εγγράφων



## (2) Οδηγούμενο από δομή (structure guided)

- Υπάρχει δομή (συνήθως ιεραρχική)
- Παραδείγματα
  - η οργάνωση αρχείων σε φακέλους
  - το ευρετήριο του Yahoo! ή του ODP
- Δομή μπορεί να υπάρχει και στο επίπεδο των εγγράφων
  - πχ abstract, section 1, ..., αναφορές)

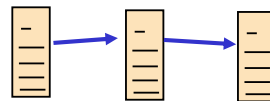




## Τύποι Πλοήγησης (II)

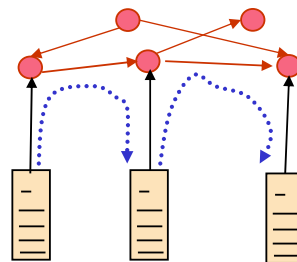
### • (3) Μη γραμμικό κείμενο (Hypertext)

- διευθυνόμενοι σύνδεσμοι (π.χ. HTML)
- σύνδεσμοι διπλής κατεύθυνσης
- τύποι συνδέσμων (typed links)



### • (4) Διεπίπεδο μη γραμμικό κείμενο

- Τα έγγραφα ταξινομούνται σε ένα εννοιολογικό σχήμα και από αυτήν την ταξινόμηση επάγονται οι συνδέσεις τους
- Παράδειγμα: σύστημα DOMENICUS [Tzitzikas & Theodorakis, Hypertext'96]

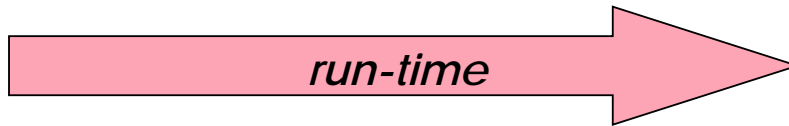
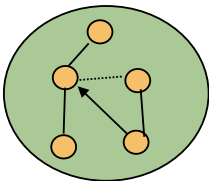




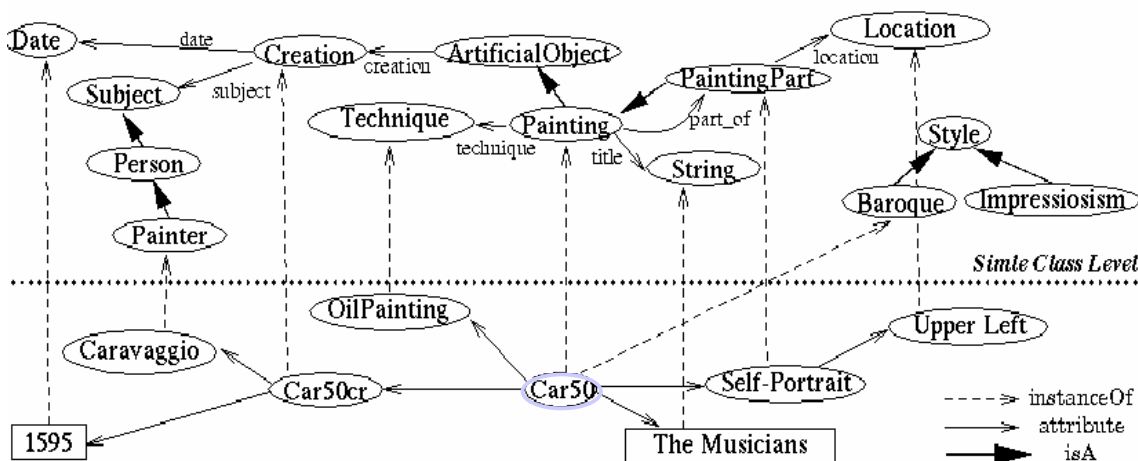
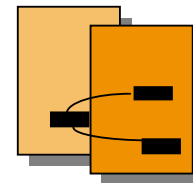
## Λειτουργίες

- Subject catalog
- Alphabetic lists
- Guided tours
- Query cards
- Schema-based generation of hyperlinks

Semantic network



Hypermedia structures



### Curators

Painting title: "The Musicians".  
OilPainting created by **Caravaggio**  
on 1595, belonging to the style of **Baroque**.  
The upper left part of the painting  
represents a **Self-Portrait**.

### Children

This **OilPainting** has the title  
"The Musicians" and has been  
created by **Caravaggio**



## Πλεονεκτήματα διεπίπεδου hypertext

- Αυτόματη διασύνδεση βάσει περιεχομένου
- Δυνατότητα πολλαπλών παρουσιάσεων
- Συνέπεια, ακεραιότητα δεδομένων και συνδέσμων
- Ισχυρή γλώσσα επερώτησης
- Εύκολη και γρήγορη εισαγωγή και ενημέρωση στοιχείων



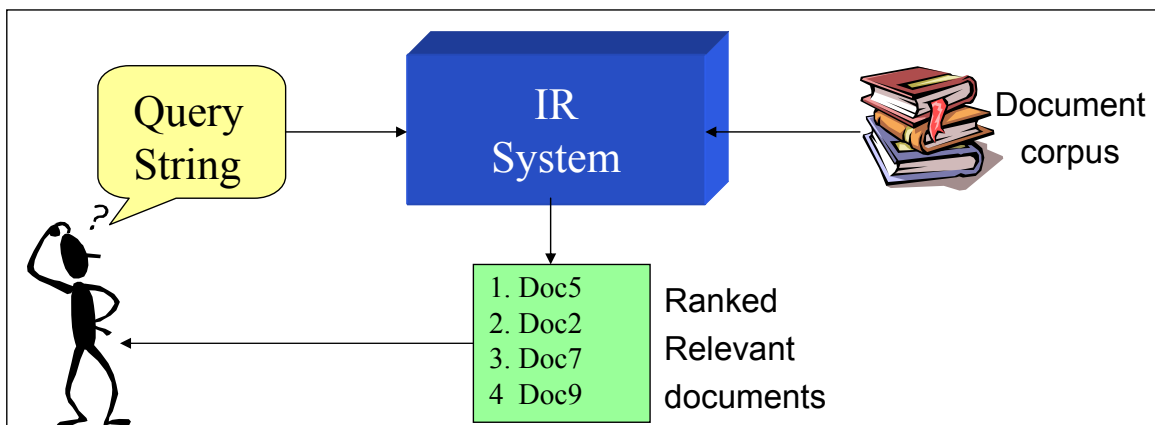
## Ανάκτηση Πληροφοριών (Information Retrieval): Το τυπικό πρόβλημα

### Δεδομένα

- Μια συλλογή από έγγραφα με κείμενο φυσικής γλώσσας  $D=\{d_1, \dots, d_n\}$
- Μια επερώτηση  $q$  ενός χρήστη σε μορφή συμβολοσειράς (string)

### Ζητούμενο

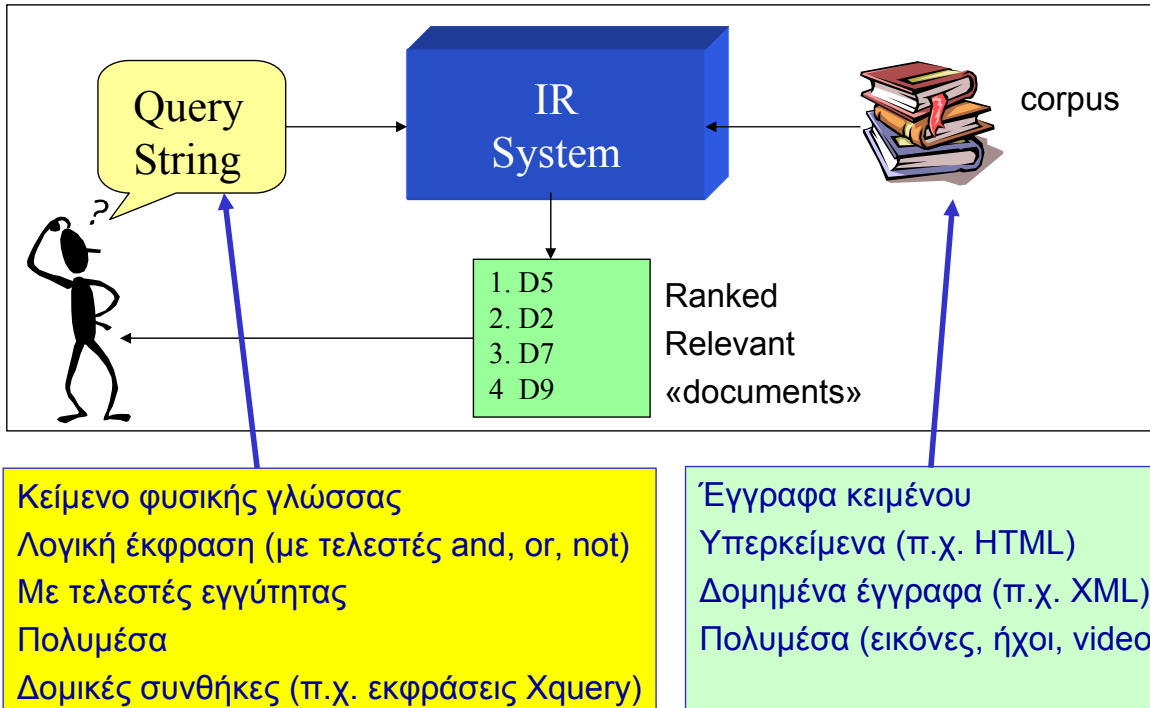
- Ένα διατεταγμένο σύνολο από έγγραφα που είναι συναφή με την επερώτηση  $\langle d_5, d_2, d_7, d_9 \rangle$







## Ανάκτηση Πληροφοριών (Information Retrieval): Μερικές παραλλαγές του προβλήματος



## Πληροφοριακές Ανάγκες Χρήστη (User Information Need)



- Παράδειγμα
  - **Find all docs containing information on college tennis teams which: (1) are maintained by a USA university and (2) participate in the NCAA tournament.**
- Έμφαση στην ανάκτηση πληροφορίας (όχι δεδομένων)



- Ανάκτηση Δεδομένων
  - ποια έγγραφα περιέχουν αυτές τις λέξεις ;
  - Καλά ορισμένη σημασιολογία (δεδομένων και επερωτήσεων)
  - ένα λάθος αντικείμενο ισοδυναμεί με αποτυχία
    - ορθότητα (soundness), πληρότητα (completeness)
- Ανάκτηση Πληροφορίας
  - βρες πληροφορίες σχετικές με αυτό το θέμα
  - η σημασιολογία είναι αρκετά χαλαρή
  - ανοχή σε μικρά σφάλματα

### Σύστημα Ανάκτησης Πληροφορίας (ΣΑΠ) :

- προσπαθεί να ερμηνεύσει το περιεχόμενο των εγγράφων και επερωτήσεων και να παράξει μια διάταξη των εγγράφων βάσει του βαθμού **συνάφειας** τους με την επερώτηση. Η *ένοια της συνάφειας* είναι κυρίαρχο ζήτημα.



## Συνάφεια (Relevance)

- **Δεν υπάρχει τυπικός ορισμός της συνάφειας !**
- Η συνάφεια είναι σε μεγάλο βαθμό **υποκειμενική**.
- **Συναφές έγγραφο** μπορεί να σημαίνει:
  - στο σωστό **θέμα**
  - **επίκαιρο** (timely)
  - **έγκυρο** (από αξιόπιστη πηγή).
  - Ικανό να ικανοποιήσει τους **σκοπούς** του χρήστη (τη επιθυμητή χρήση της αναζητούμενης πληροφορίας) (**information need**)
  - ...



## Η βασική προσέγγιση ΑΠ

- Οι πιο επιτυχημένες προσεγγίσεις είναι οι **στατιστικές**
- Γιατί όχι επεξεργασία φυσικής γλώσσας;
- Χειρονακτικά προσδιορισμένες επικεφαλίδες (headings)
  - e.g. Library of Congress headings, Dewey Decimal headings
  - η χειρονακτική ευρετηρίαση είναι ακριβή
  - η χειρονακτική ευρετηρίαση απαιτεί συμφωνία (human agreement)



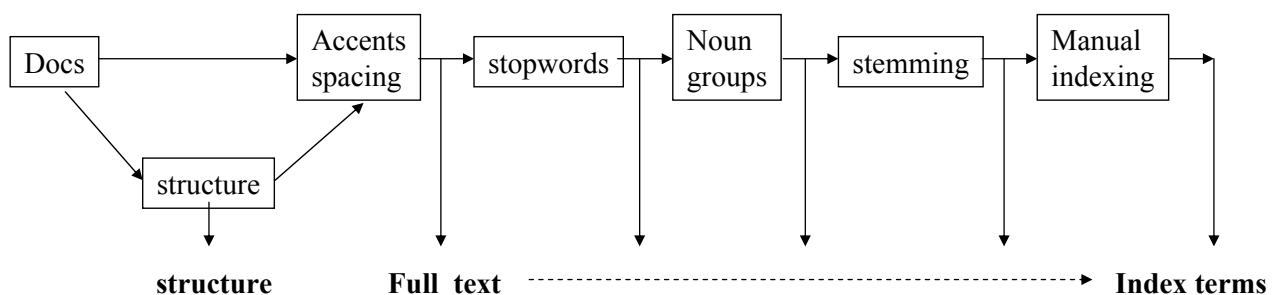
## Πως βλέπουμε ένα έγγραφο;

- Πως βλέπουμε ένα έγγραφο;
  - Ως έχει (full text);
  - Αγνοώντας λέξεις που δεν φέρουν νόημα (π.χ. τα άρθρα) ;
  - Ως σάκο (bag) όρων ευρετηρίου (bag of index terms),  
δηλαδή αγνοώντας τη σειρά με την οποία εμφανίζονται οι λέξεις στο κείμενο;
  - Ως σύνολο όρων ευρετηρίου (set of Index terms)
  - Ως δομημένο έγγραφο (π.χ. hypertext, XML)
- Η απάντηση σε αυτό το ερώτημα θα καθορίσει τη μορφή του ευρετηρίου που πρέπει να κατασκευάσουμε.

- Σνωμφύα με μια ένυερα του Κέμπριτζ η σιερά των γμμάρωτων σε μια λξή δεν έεχι σησίμα. Ακρεί το πώτρο και το ταίυελετο γμαράμ να είανι στη σστωή σεριά.
- Σύμφωνα με μια έρευνα του Κέμπριτζ η σειά των γραμμάτων σε μια λέξη δεν έχει σημασία. Αρκεί το πρώτο και το τελευταίο γράμμα να είναι στη σωστή σειρά.



## Πως βλέπουμε ένα έγγραφο;





# Οι βασικές λειτουργικές μονάδες ενός ΣΑΠ

- **Λειτουργίες Κειμένου (Text Operations)** σχηματίζουν τις λέξεις ευρετηρίου (tokens, index terms).
  - Αφαίρεση λέξεων αποκλεισμού (Stopword removal), Stemming
- **Ευρετηριασμός (Indexing)** κατασκευάζει ένα ευρετήριο (συνήθως inverted index) με δείκτες από τις λέξεις προς τα έγγραφα
- **Αναζήτηση (Searching)** ανακτά τα έγγραφα που περιέχουν μια λέξη (της επερώτησης) από το inverted index.
- **Κατάταξη (Ranking)** διαβαθμίζει όλα τα ανακτημένα αρχεία βάσει μιας μετρικής συνάφειας.
- **Διεπαφή (User Interface)** διευθύνει την αλληλεπίδραση με το χρήστη
- **Λειτουργίες επερώτησης (Query Operations)** μετασχηματίζουν την επερώτηση για βελτίωση της ανάκτησης:
  - Επέκταση επερώτησης χρησιμοποιώντας έναν θησαυρό
  - Επέκταση επερώτησης βάσει τοπικής ή καθολικής ανάλυσης
  - Μετασχηματισμός επερώτησης με ανάδραση συνάφειας
  - ...



# Γενική μορφή ενός ευρετηρίου

		Indexing Items					
		$k_1$	$k_2$	...	$k_j$	...	$k_t$
D o c u m e n t s	$d_1$	$c_{1,1}$	$c_{2,1}$	...	$c_{i,1}$	...	$c_{t,1}$
	$d_2$	$c_{1,2}$	$c_{2,2}$	...	$c_{i,2}$	...	$c_{t,2}$
	...	...	...	...	...	...	...
	$d_i$	$c_{1,j}$	$c_{2,j}$	...	$c_{i,j}$	...	$c_{t,j}$
	...	...	...	...	...	...	...
	$d_N$	$c_{1,N}$	$c_{2,N}$	...	$c_{i,N}$	...	$c_{t,N}$

$c_{ij}$ : το κελί που αντιστοιχεί στο έγγραφο  $d_i$  και στον όρο  $k_j$ , το οποίο μπορεί να περιέχει:

- ένα  $w_{ij}$  που να δηλώνει την παρουσία ή απουσία του  $k_j$  στο  $d_i$  (ή τη σπουδαιότητα του  $k_j$  στο  $d_i$ )
- τις θέσεις στις οποίες ο όρος  $k_j$  εμφανίζεται στο  $d_i$  (αν πράγματι εμφανίζεται)



# Δημιουργία του Ευρετηρίου

- **Λειτουργίες Κειμένου (Text Operations)** σχηματίζουν τις λέξεις ευρετηρίου (tokens, index terms).

		Indexing Items					
		$k_1$	$k_2$	...	$k_j$	...	$k_t$
D o c u m e n t s	$d_1$	$c_{1,1}$	$c_{2,1}$	...	$c_{i,1}$	...	$c_{t,1}$
	$d_2$	$c_{1,2}$	$c_{2,2}$	...	$c_{i,2}$	...	$c_{t,2}$
	...	...	...	...	...	...	...
	$d_i$	$c_{1,j}$	$c_{2,j}$	...	$c_{i,j}$	...	$c_{t,j}$
	...	...	...	...	...	...	...
	$d_N$	$c_{1,N}$	$c_{2,N}$	...	$c_{i,N}$	...	$c_{t,N}$

- **Ευρετηρίαση (Indexing)** κατασκευάζει ένα ευρετήριο (inverted index) με δείκτες από τις λέξεις προς τα έγγραφα



# Χρήση του Ευρετηρίου

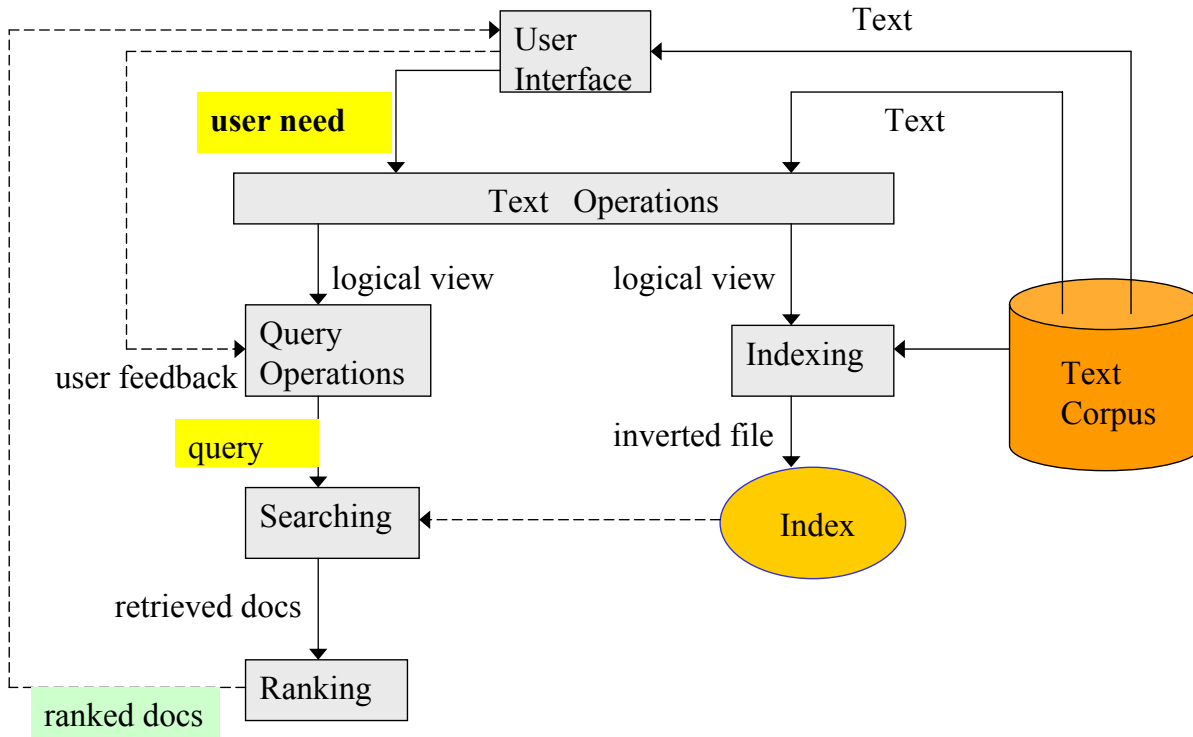
query

- **Αναζήτηση (Searching)** ανακτά τα έγγραφα που περιέχουν μια λέξη (της επερώτησης) από το inverted index.
- **Κατάταξη (Ranking)** διαβαθμίζει όλα τα ανακτημένα αρχεία με βάση μια μετρική συνάφειας.

		Indexing Items					
		$k_1$	$k_2$	...	$k_j$	...	$k_t$
D o c u m e n t s	$d_1$	$c_{1,1}$	$c_{2,1}$	...	$c_{i,1}$	...	$c_{t,1}$
	$d_2$	$c_{1,2}$	$c_{2,2}$	...	$c_{i,2}$	...	$c_{t,2}$
	...	...	...	...	...	...	...
	$d_i$	$c_{1,j}$	$c_{2,j}$	...	$c_{i,j}$	...	$c_{t,j}$
	...	...	...	...	...	...	...
	$d_N$	$c_{1,N}$	$c_{2,N}$	...	$c_{i,N}$	...	$c_{t,N}$

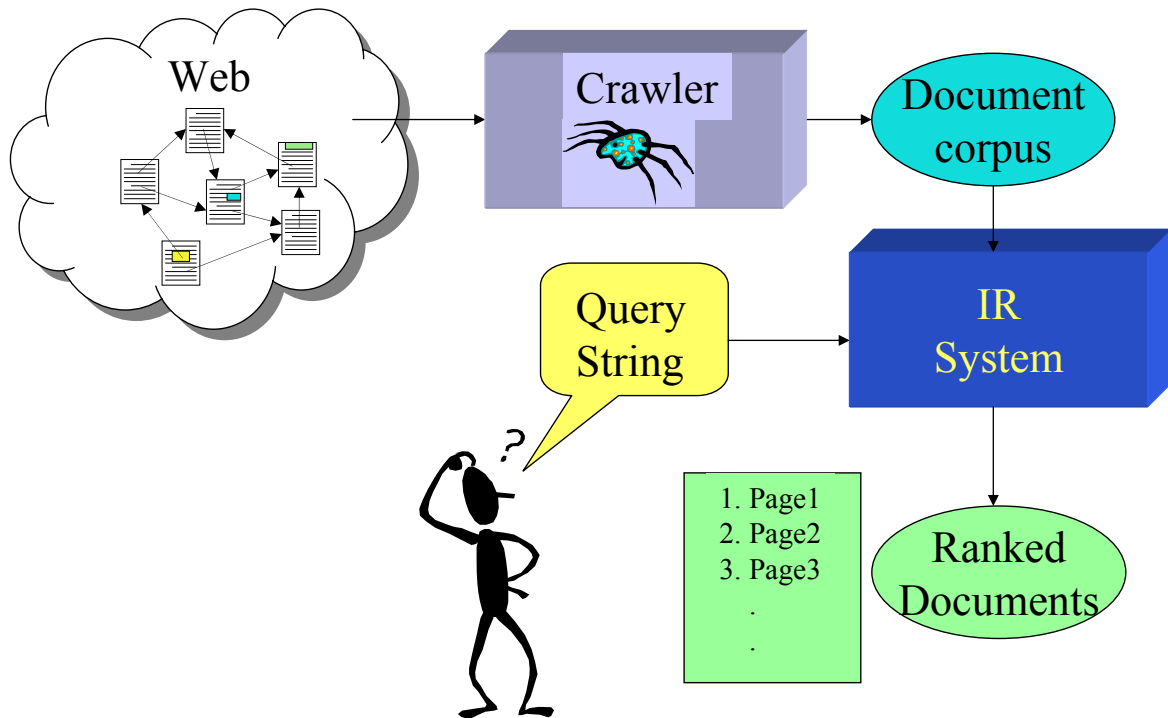


## Η Αρχιτεκτονική ενός ΣΑΠ



## Αναζήτηση στον Ιστό (Web Search)

- Εφαρμογή της ΑΠ σε έγγραφα HTML του Ιστού
- Διαφορές:
  - Εδώ πρέπει να **συλλέξουμε** τη συλλογή των εγγράφων **διασχίζοντας** (crawling/spidering) τον Ιστό και να την κρατάμε **ενημέρη** διότι οι σελίδες τροποποιούνται/διαγράφονται χωρίς προειδοποίηση.
  - Μπορούμε να καταγράψουμε και να αξιοποιήσουμε τη **δομή των συνδέσμων** του Ιστού.
  - Μπορούμε να αξιοποιήσουμε τη **δομή** της πληροφορίας των HTML (ή XML) εγγράφων, π.χ. οι λέξεις που εμφανίζονται μεταξύ `<h1>.. </h1>` μπορεί να θεωρηθούν «σπουδαιότερες» από αυτές που εμφανίζονται μεταξύ `<h3>.. </h3>`



• **Crawling («έρπειν»)**

W e b  p a g e s	$d_1$	$k_1$	$k_2$	$\dots$	Indexing Items $k_j$	$\dots$	$k_t$
	$d_2$	$c_{1,1}$	$c_{2,1}$	$\dots$	$c_{i,1}$	$\dots$	$c_{t,1}$
	$\dots$	$c_{1,2}$	$c_{2,2}$	$\dots$	$c_{i,2}$	$\dots$	$c_{t,2}$
	$d_i$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$\dots$	$c_{1,j}$	$c_{2,j}$	$\dots$	$c_{i,j}$	$\dots$	$c_{t,j}$
	$d_N$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
		$c_{1,N}$	$c_{2,N}$	$\dots$	$c_{i,N}$	$\dots$	$c_{t,N}$

From	To
d2	d3
d2	d4
d4	d1
d10	d20

• **Ευρετηρίαση (Indexing)**





## Άλλες λειτουργίες που σχετίζονται με την ΑΠ

- Question answering (απάντηση ερωτήσεων)
- Recommender systems (συστήματα συστάσεων)
- Automatic clustering (αυτόματη ομαδοποίηση)
- Cross-language retrieval (διαγλωσσική ανάκτηση)
- Data and information mining (εξόρυξη δεδομένων και πληροφοριών)
- Information integration (εννοποίηση πληροφοριών)
- Knowledge management (διαχείριση γνώσης)
- Meta-search (multi-database searching) (μέτα-αναζήτηση)
- Summarization (αυτόματη περίληψη)
- Agents (filtering, routing)
- ...



## Ενδεικτικά Συστήματα

- **IR Systems**
  - Verity, Fulcrum, Excalibur, Eurospider
  - Hummingbird, Documentum
  - Inquiry, Smart, Okapi, Lemur, Indri
- **Web search and in-house systems**
  - West, LEXIS/NEXIS, Dialog
  - Lycos, AltaVista, Excite, Yahoo, Google, Nothern Light, Teoma, HotBot, Direct Hit, ...
  - Ask Jeeves
  - eLibrary, Inquiria
  - vivisimo ([www.vivisimo.com](http://www.vivisimo.com))
  - ...



## HY463: Θεματικές Ενότητες



## HY463: Θεματικές Ενότητες

### *1. Εισαγωγή*

Τι είναι η Ανάκτηση Πληροφοριών, Βασικές έννοιες, Ιστορική αναδρομή

### *2. Αξιολόγηση Αποτελεσματικότητας (≈ 1-2 διαλέξεις)*

Ακρίβεια, Ανάκληση, Εναλλακτικά μέτρα, Συλλογές αναφοράς

### *3. Μοντέλα Ανάκτησης Πληροφοριών (≈ 3 διαλέξεις)*

Boolean, Διανυσματικό, Πιθανοκρατικό, Εναλλακτικά μοντέλα

### *4. Προχωρημένες Λειτουργίες Επερώτησης (≈ 1 διάλεξη)*

Επέκταση επερώτησης, Ανάδραση συνάφειας, Αυτόματη τοπική/καθολική ανάλυση

### *5. Γλώσσες Επερώτησης για Ανάκτηση Πληροφοριών (≈ 1 διάλεξη)*

Λέξεις κλειδιά, Λογικές επερωτήσεις, Επερωτήσεις συμφραζομένων, Επερωτήσεις φυσικής γλώσσας, Δομημένες επερωτήσεις, Ευρετηρίαση και Ανάκτηση XML εγγράφων

### *6. Ομαδοποίηση Εγγράφων (Clustering) (≈ 1 διάλεξη)*



## HY463: Θεματικές Ενότητες (II)

7. *Ευρετηρίαση, Προεπεξεργασία και Οργάνωση Αρχείων Κειμένου (≈ 2 δ)*  
Λέξεις αποκλεισμού (stopwords), stemming (στελέχωση κειμένου), θησαυροί όρων  
Ανεστραμμένα Αρχεία (inverted files), Δένδρα Καταλήξεων (suffix trees), Αρχεία  
Υπογραφών (signature files)
8. *Στατιστικά και Συμπύεση Κειμένου (≈ 1 διάλεξη)*
9. *Αναζήτηση σε Κείμενα*  
Αλγόριθμοι Knuth-Morris-Pratt, Boyer-Moore, Αυτόματο καταλήξεων (suffix  
automaton), Φράσεις και εγγύτητα
10. *Ανάκτηση Πολυμέσων (≈ 2 διαλ.)*  
Μοντέλα και γλώσσες, Ευρετηρίαση και Αναζήτηση
11. *Παράλληλη και Κατανεμημένη Ανάκτηση Πληροφοριών (≈ 3 διαλέξεις)*  
Αρχιτεκτονικές MIMD, SIMD, Peer-2-Peer (P2P), Διαμερισμός συλλογών, Επιλογή  
πηγής, Επεξεργασία επερωτήσεων, Ανάκτηση Πληροφοριών σε P2P



## HY463: Θεματικές Ενότητες (III)

12. *Τεχνικές μετα-Κατάταξης (meta-ranking) (≈ 1 διάλεξη)*  
Ενοποιημένες και απομονωμένες μέθοδοι, Παρεμβολή, Ψηφοφορία
13. *Αναζήτηση στον Παγκόσμιο Ιστό (≈ 3 διαλέξεις)*  
Ευρετηρίαση ιστοσελίδων, Διάσχιση του ιστού (crawling), Τεχνικές ανάλυσης  
συνδέσμων (link analysis), PageRank, HITS
14. *Εξατομικευμένη Ανάκτηση και Διήθηση*  
Προφίλ χρηστών, Συνεργατική Ανάκτηση και Διήθηση T
15. *Ανάκτηση Δομημένων Εγγράφων*  
Ευρετηρίαση και ανάκτηση εγγράφων XML
16. *Διεπαφές Χρήσης και Οπτικοποίηση (≈ 1 διάλεξη)*



*Άλλα σχετικά ζητήματα που ίσως προλάβουμε να θίξουμε:*

- *Cross language retrieval*
- *Information Extraction*
- *Text Categorization*
- *Digital Libraries Video Retrieval*
  
- *Generalized Interaction Models*
- *Faceted Classification Theory and Advances*
- 



## Ιστορική Αναδρομή



- 1960-70's:
  - Initial exploration of text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
  - Development of the basic Boolean and vector-space models of retrieval.
  - Prof. Salton and his students at Cornell University are the leading researchers in the area.
- 1980's:
  - Large document database systems, many run by companies:
    - Lexis-Nexis
    - Dialog
    - MEDLINE



- 1990's:
  - Searching FTPable documents on the Internet
    - Archie
    - WAIS
  - Searching the World Wide Web
    - Lycos
    - Yahoo
    - Altavista
  - Organized Competitions
    - NIST TREC
  - Recommender Systems
    - Ringo
    - Amazon
    - NetPerceptions
  - Automated Text Categorization & Clustering



## Ιστορική Αναδρομή (III)



- 2000's
  - Link analysis for Web Search
    - Google
  - Automated Information Extraction
    - Whizbang
    - Fetch
    - Burning Glass
  - Question Answering
    - TREC Q/A track
  - Multimedia IR
    - Image, Video, Audio and music
  - Cross-Language IR
    - DARPA Tides
  - Document Summarization

Πριν τον Ιστό η ΑΠ εθεωρείτο ότι είχε στενό πεδίο εφαρμογής

Μετά την επινόηση του Web αυτό άλλαξε για τα καλά:

- οικουμενική δεξαμενή γνώσης
- ελεύθερη (και φθηνή) καθολική πρόσβαση
- έλλειψη κεντρικού ελέγχου σύνταξης

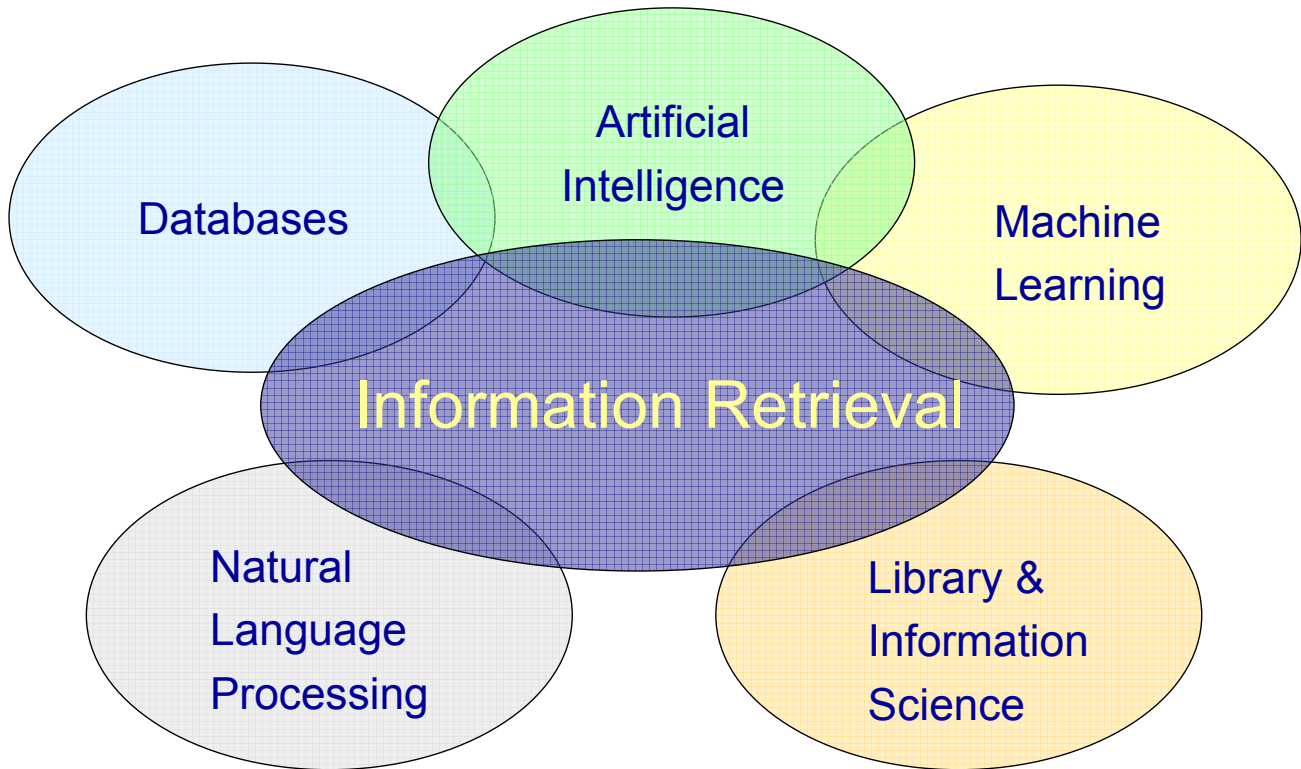


## «Ιστορική» Αναδρομή (IV)

- Στο μέλλον
  - Στόχος: **εύρεση της «σωστής» απάντησης για σένα εδώ και τώρα**
    - Εξατομίκευση (personalization), περίσταση (context)
  - Επεξεργασία φυσικής γλώσσας
  - Ενοποίηση με άλλες τεχνολογίες
  - Κατανεμημένη, ετερογενή ΑΠ



# Σχετικές Περιοχές



# Comparing IR to Databases

	<b>Databases</b>	<b>IR</b>
Data	<b>Structured</b>	<b>Unstructured</b>
Fields	<b>Defined</b> (e.g. age, price)	<b>No fields</b> (other than text)
Queries	<b>Defined</b> (e.g. SQL)	<b>Free text</b> (natural language), Boolean
Matching	<b>Exact</b> (results are always «correct»)	<b>Imprecise</b> (need to measure effectiveness)



## Τεχνητή Νοημοσύνη (Artificial Intelligence)

- Παραδοσιακά εστιάζει στην
  - παράσταση γνώσης (knowledge representation) και τον συλλογισμό (reasoning).
- Φορμαλισμοί για παράσταση γνώσης και επερωτήσεων:
  - First-order Predicate Logic
  - Bayesian Networks
- Η πρόσφατη δουλειά σε **web ontologies** και **intelligent information agents** την φέρνει πιο κοντά στην ΑΠ



## Μηχανική Μάθηση (Machine Learning)

- Εστιάζει στην ανάπτυξη υπολογιστικών συστημάτων που βελτιώνουν τις επιδόσεις τους με το χρόνο (αξιοποιώντας πρωθύστερη εμπειρία)
- Επιτηρούμενη Μάθηση (Supervised learning)
  - Αυτόματη ταξινόμηση μέσω μάθησης από παραδείγματα (labeled training examples)
- Μη-Επιτηρούμενη Μάθηση (Unsupervised learning)
  - Αυτόματη ομαδοποίηση
- Μηχανική μάθηση και Ανάκτηση Πληροφοριών
  - Κατηγοριοποίηση Κειμένων (Text Categorization)
    - Αυτόματη ιεραρχική ταξινόμηση (hierarchical classification, e.g. Yahoo).
    - Προσαρμόσιμη διήθηση (filtering) / δρομολόγηση (routing) / συστάσεις (recommending).
    - Αυτόματος εντοπισμός spam.
  - Ομαδοποίηση Κειμένων (Text Clustering)
    - Ομαδοποίηση των αποτελεσμάτων της αναζήτησης
    - Αυτόματος σχηματισμός ιεραρχιών (Yahoo).





- Παραδοσιακά εστιάζει την
  - **συντακτική** (syntactic) ανάλυση,
  - **σημασιολογική** (semantic) ανάλυση και
  - **πραγματολογική (pragmatic)** ανάλυσητης φυσικής γλώσσας και ομιλίας
- Η ανάλυση του συντακτικού (δομή φράσεων) και της σημασιολογίας θα μπορούσε να επιτρέψει την ανάκτηση μέσω νοήματος, αντί λέξεων.
- Σχετικά θέματα:
  - Μέθοδοι αποσαφήνισης του νοήματος των διφορούμενων λέξεων βάσει των συμφραζομένων (*word sense disambiguation*).
  - Μέθοδοι αναγνώρισης συγκεκριμένων τμημάτων πληροφορίας σε ένα έγγραφο (*information extraction*).
  - Μέθοδοι απάντησης επερωτήσεων φυσικής γλώσσας από συλλογές κειμένου



- Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization).
- Concerned with effective categorization of human knowledge.
- Concerned with citation analysis and *bibliometrics* (structure of information).
- Recent work on *digital libraries* brings it closer to CS & IR.



---

End of Lecture 1

---