HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Γιάννης Τζίτζικας

Διάλεξη : 1
Ημερομηνία : 21-9-2009
Θέμα : Διαδικαστικά, Εισαγωγή και Επισκόπηση

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

- Διδακτικές μονάδες: 4
- Προαπαιτούμενα
  - HY240 - Δομές Δεδομένων
- Εβδομαδιαίο Πρόγραμμα :
  - **Διαλέξεις**: Δευτέρα 11-1 και Τετάρτη 11-1 στην αίθουσα PA203
  - **Φροντιστήρια**: Παρασκευή 1-3 στην αίθουσα PA201
    - (θα στέλνεται email πριν από κάθε φροντιστήριο)
- Παρακολούθηση
  - Αναμενόμενη αλλά όχι υποχρεωτική
  - Η ενεργή συμμετοχή στο μάθημα θα ληφθεί θετικά υπόψη
- Γραφτείτε (σήμερα) στη λίστα **`hy463-list`**

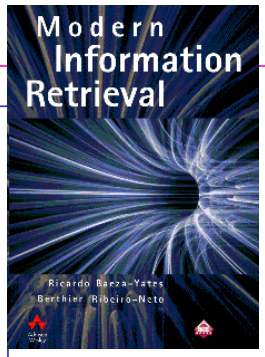
Προσωπικό

- Διδάσκων:
  - Γιάννης Τζίτζικας
  - tzitzik (at) csd.uoc.gr
  - Γραφείο: Γ107 (τηλ. 393 521)
  - Ώρες γραφείου: πριν και μετά τις διαλέξεις
- Βοηθοί:
  - Παπαδάκος Παναγιώτης
  - Νίκος Μανώλης
  - Χαράλαμπος Τζαγκαράκης
  - Υπεύθυνοι για:
    - Λύση και βαθμολόγηση ασκήσεων
    - Επίβλεψη εργασιών
    - Φροντιστήρια
    - Απάντηση ερωτήσεων

Ιστοσελίδα μαθήματος

- 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 διαδυκτιακές 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
- Όσοι επιθυμούν εκτυπώσεις των διαφανειών πρέπει να ενημερώσουν 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 μάλλον 3 σειρές ασκήσεων
  - 1. Αξιολόγηση της αποτελεσματικότητας της ανάκτησης, μοντέλα ανάκτησης και ευρετήρια
  - 2. Χρήση bazar
  - 3. Άλλα θέματα
- Βάρος: 20% του τελικού βαθμού

Πρόοδος

- Το εάν θα γίνει θα εξαρτηθεί από την συμμετοχή σας στο μάθημα.
- [Αξία: 20% τελικού βαθμού]

Εργασία μαθήματος (project)

2005:Υλοποίηση ενός Συστήματος Ανάκτησης Πληροφοριών με ψευδοανάδραση συνάφειας (pseudo relevance feedback)

- Χρονοδιάγραμμα (1 Απρίλη-Μέσα Μαΐου), ομάδες 2 ατόμων, υλοποίηση σε Java
- Βάρος: 30% Τελικού βαθμού

2006: Ανάπτυξη μια μηχανής αναζήτησης για τον παγκόσμιο ιστό.

- Κάθε ομάδα θα αναλάβει μόνο κάποια υποσυστήματα αυτής της μηχανής.
- => Groogle'2006
  - (basic functionality but too many problems)

2007: => Groogle'2007
(decent but several functionalities were missing)

2008: => Stemmer Utilities, Inverted Index (without DBMS), Crawlers

2009 άνοιξη: => Mitos'2009
(improved ranking & link analysis techniques) and whatever extra you like

2009 φθινόπωρο:
text-based image retrieval

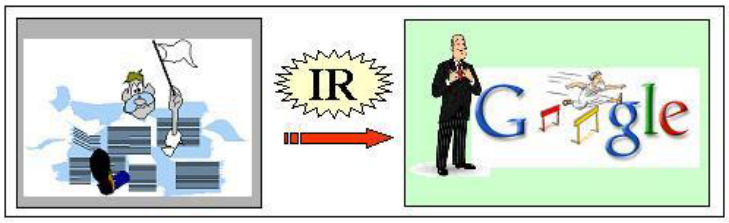
Βαθμολόγηση

- **Τελικός βαθμός**
  - **Τελικός** = 20% Ασκήσεις + 40% Εργασία + 40% ΤελικήΕξέταση
- Για να περάσετε το μάθημα χρειάζεστε
  - **Τελικός** ≥ 5 **AND** ΤελικήΕξ ≥ 4
- Σημειώσεις στην Πρόοδο/Τελική Εξέταση:
  - [Εξέταση προόδου: Κλειστές (μάλλον)]
  - Τελική εξέταση: Ανοιχτές

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

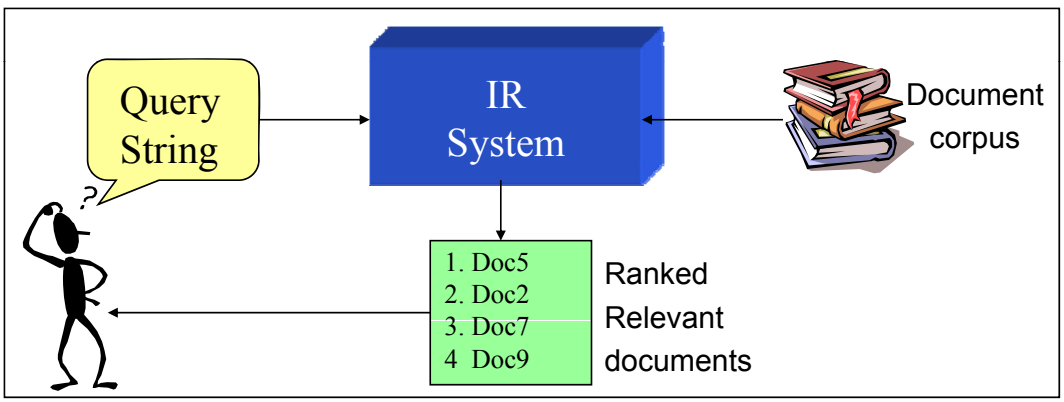
Δεδομένα Προβλήματος

- Μια συλλογή από έγγραφα με κείμενο φυσικής γλώσσας D={d1,…,dn}
- Μια επερώτηση q ενός χρήστη σε μορφή συμβολοσειράς (string)

Ζητούμενο

- Ένα διατεταγμένο σύνολο από έγγραφα που είναι συναφή με την επερώτηση

<d5,d2,d7,d9>



Περιγραφή Μαθήματος

Σκεπτικό:

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.

- *Γιατί χρειαζόμαστε Ανάκτηση Πληροφοριών (ΑΠ);*
- *Τι είναι η Ανάκτηση Πληροφοριών;*
- *Ανάκτηση, Διήθηση, Πλοήγηση*
- *Μοντέλα Πλοήγησης*
- *Το βασικό πρόβλημα στην Ανάκτηση Πληροφοριών*
- *Ανάκτηση Δεδομένων έναντι Ανάκτηση Πληροφοριών*
- *Συνάφεια*
- *Η βασική προσέγγιση & αρχιτεκτονική ενός Συστήματος Ανάκτησης Πληροφοριών (ΣΑΠ)*
- *Ανάκτηση Πληροφοριών στον Παγκόσμιο Ιστό*
- *Άλλες λειτουργίες ενός ΣΑΠ*
- *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
  - The Indexed Web contains **at least 45.84 billion pages** (Monday, 18 February, 2008).
- 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”,

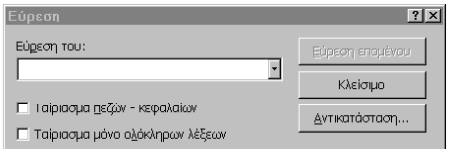
ας δούμε ποια είναι και που το πάει η Google”

Το πρόβλημα είναι σημαντικό και επίκαιρο

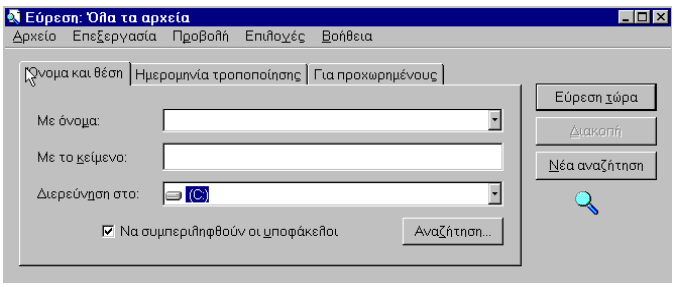
(έως και οι πολιτικοί άρχισαν να ασχολούνται με αυτό)

- 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 ευρωπαικής 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, ...
- .. Europeana

Τι να είναι η ΑΠ;



grep



YAHOO!

altavista

CiteSeer.IST
Scientific Literature Digital Library

Ask Jeeves

Google

www.vivisimo.com

Τι να είναι η ΑΠ;

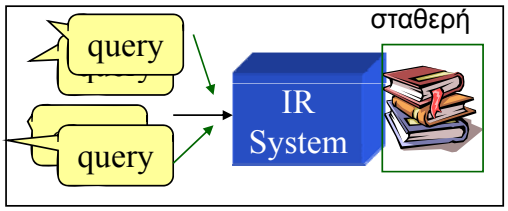
- Μήπως οι μηχανές αναζήτησης όπως το Google, Lycos ?
  - Αρκετά αποτελεσματικές (σε μερικά πράγματα)
  - Αναγνωρίσιμες και γνωστές
  - Εμπορικά επιτυχημένες (τουλάχιστον μερικές)

- 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

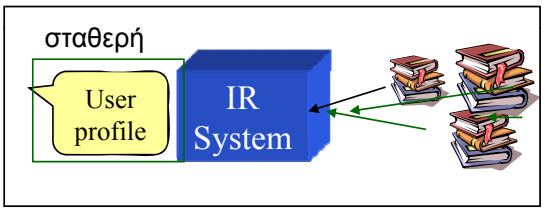
Ανάκτηση (ad hoc retrieval):

- Σταθερή συλλογή εγγράφων, μεταβαλλόμενες επερωτήσεις

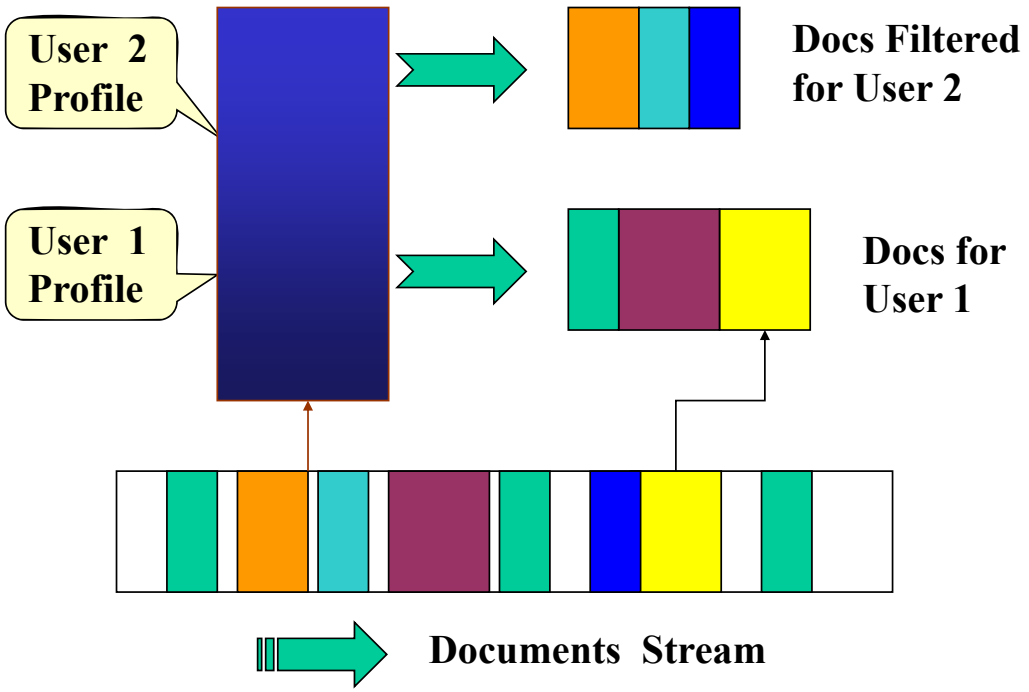


Φιλτράρισμα ή Διήθηση (Filtering):

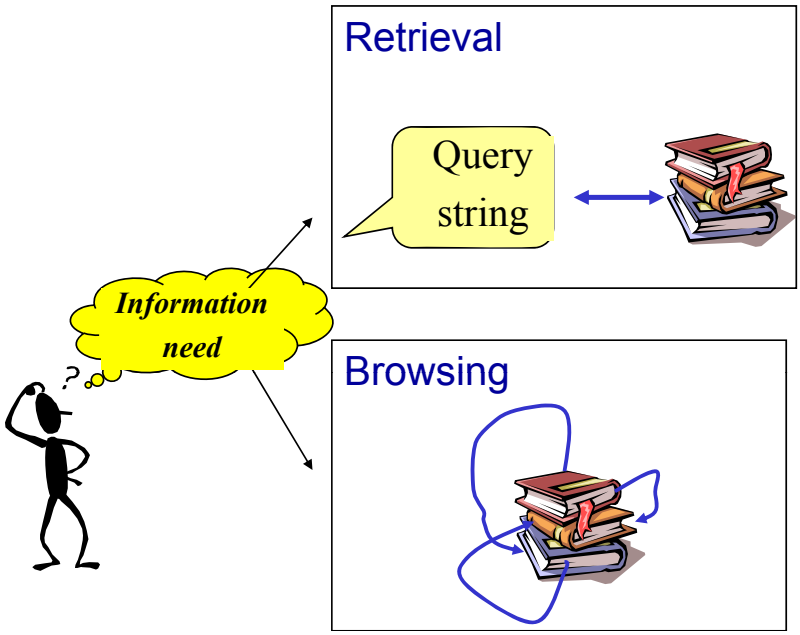
- Σταθερή επερώτηση, **ροή** νέων κειμένων
- **Προφίλ Χρήστη** = Επερώτηση που εκφράζει πιο μόνιμες προτιμήσεις
- Έμφαση στη δημιουργία/ενημέρωση του προφίλ



Φιλτράρισμα



Ανάκτηση και Πλοήγηση
(Retrieval vs Browsing)

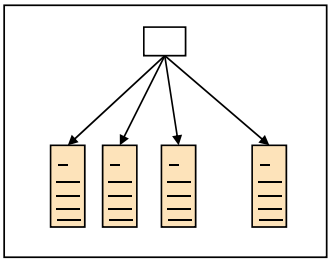


- Ο χρήστης δεν διατυπώνει επερώτηση
- Πλοήγηση με σκοπό ή χωρίς σκοπό (εξερεύνηση).

Τύποι Πλοήγησης
(Types of Browsing)

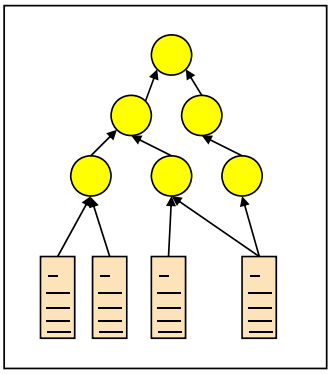
(1) Επίπεδο (flat)

- πχ. μια λίστα εγγράφων

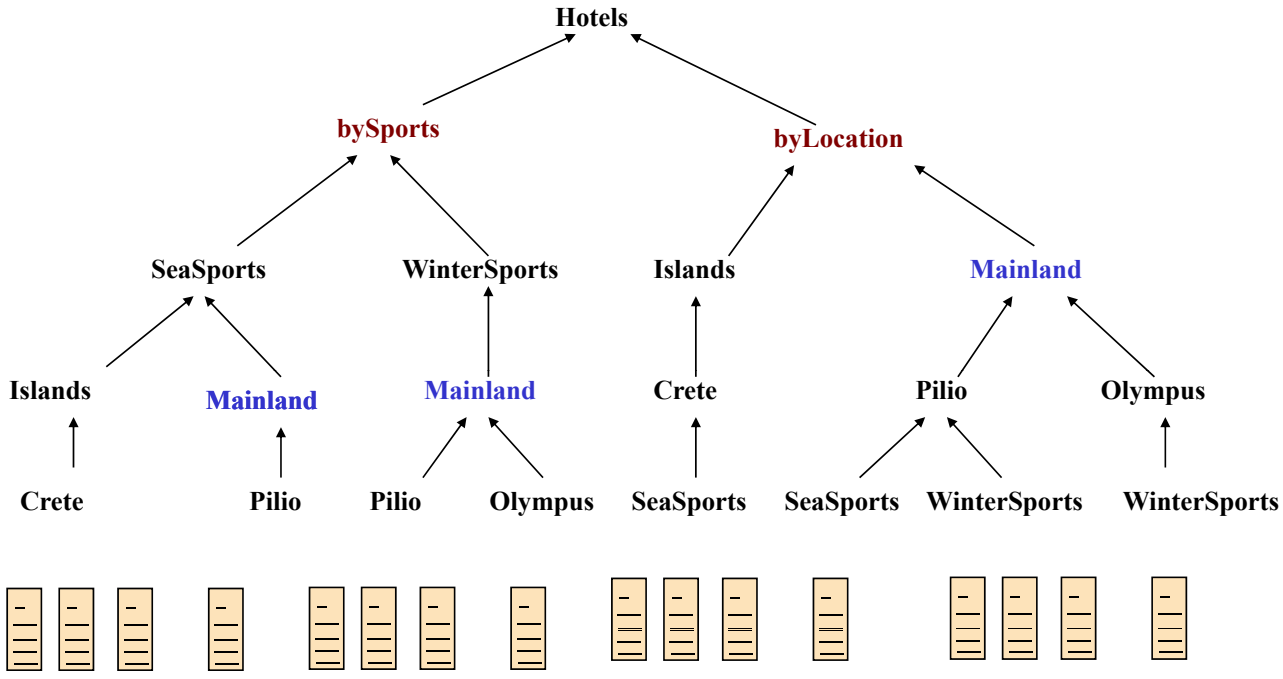


(2) Οδηγούμενο από δομή (structure guided)

- 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)

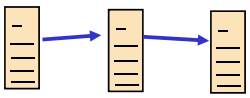


Πλοήγηση οδηγούμενη από δομή
Παράδειγμα

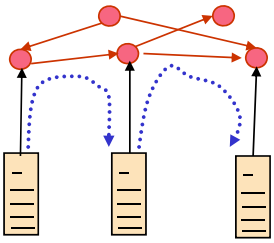


Τύποι Πλοήγησης (II)

- **(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]



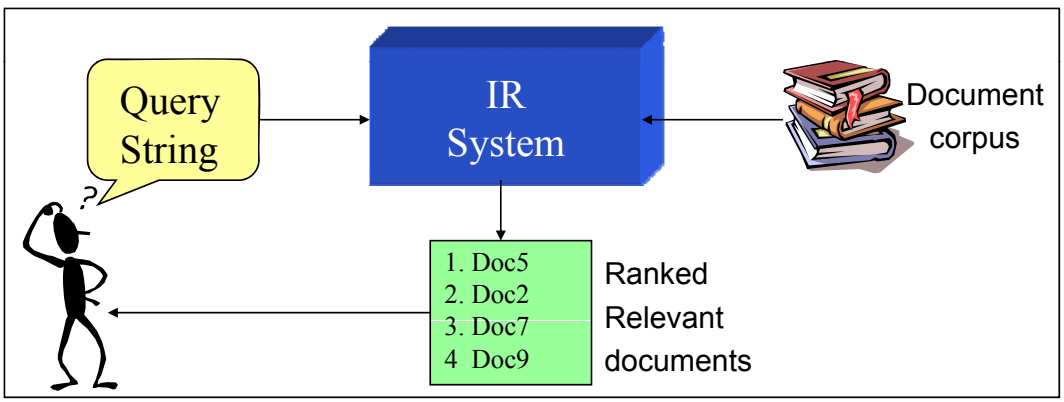
Ανάκτηση Πληροφοριών (Information Retrieval): Το τυπικό πρόβλημα

Δεδομένα

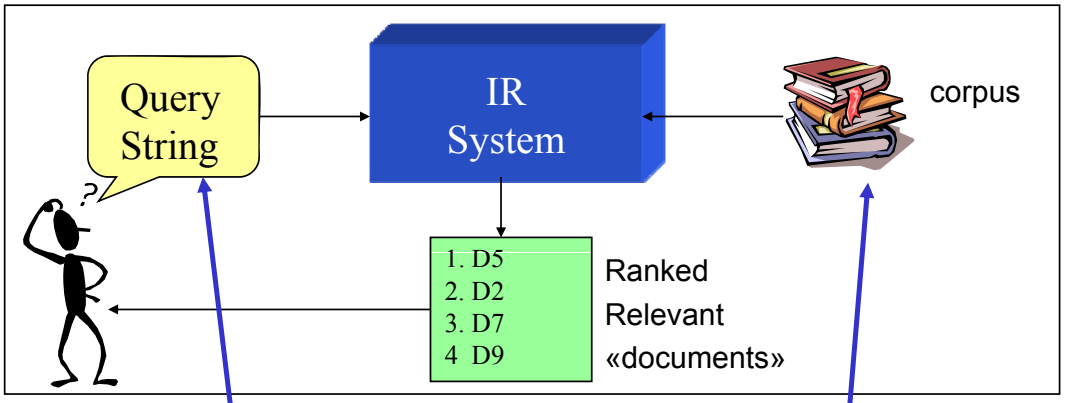
- Μια συλλογή από έγγραφα με κείμενο φυσικής γλώσσας D={d1,…,dn}
- Μια επερώτηση q ενός χρήστη σε μορφή συμβολοσειράς (string)

Ζητούμενο

- Ένα διατεταγμένο σύνολο από έγγραφα που είναι συναφή με την επερώτηση **<d5,d2,d7,d9>**



Ανάκτηση Πληροφοριών (Information Retrieval): Μερικές παραλλαγές του προβλήματος



Κείμενο φυσικής γλώσσας
Λογική έκφραση (με τελεστές and, or, not)
Με τελεστές εγγύτητας
Πολυμέσα
Δομικές συνθήκες (π.χ. εκφράσεις Xquery)

Έγγραφα κειμένου
Υπερκείμενα (π.χ. HTML)
Δομημένα έγγραφα (π.χ. XML)
Πολυμέσα (εικόνες, ήχοι, video)

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 έναντι Ανάκτηση Πληροφορίας (Data versus Information Retrieval)

- 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 *έννοια 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)

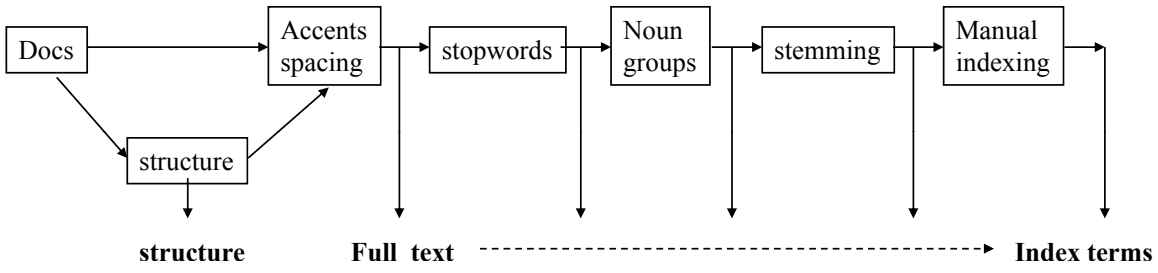
Πως βλέπουμε ένα έγγραφο;

- 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 (και τον τύπο των επερωτήσεων που μπορούμε να απαντήσουμε).

- Σνωμφύα με μια ένυερα του Κέμπριτζ η σιερά των γμμάαρωτν σε μια λξέη δεν έεχι σησίμαα. Ακρεί το πώτρο και το ταίυελετο γμαράμ να είανι στη σστωή σεριά.

- Σύμφωνα με μια έρευνα του Κέμπριτζ η σειρά 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

Documents	k_1	k_2	...	k_j	...	k_t
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 di και στον όρο kj, το οποίο μπορεί να περιέχει:

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

Δημιουργία του Ευρετηρίου

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

Indexing Items

Documents	k_1	k_2	. . .	k_j	. . .	k_t
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

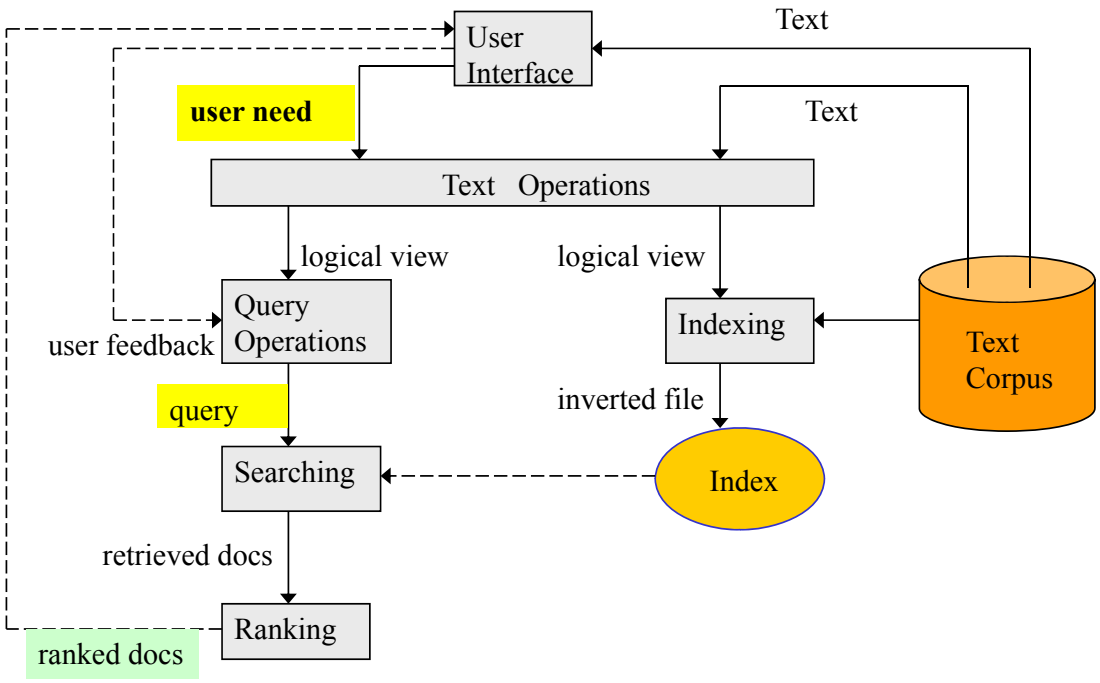
query

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

Indexing Items

Documents	k_1	k_2	. . .	k_j	. . .	k_t
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	$c_{1,N}$	$c_{2,N}$	. . .	$c_{i,N}$	. . .	$c_{t,N}$

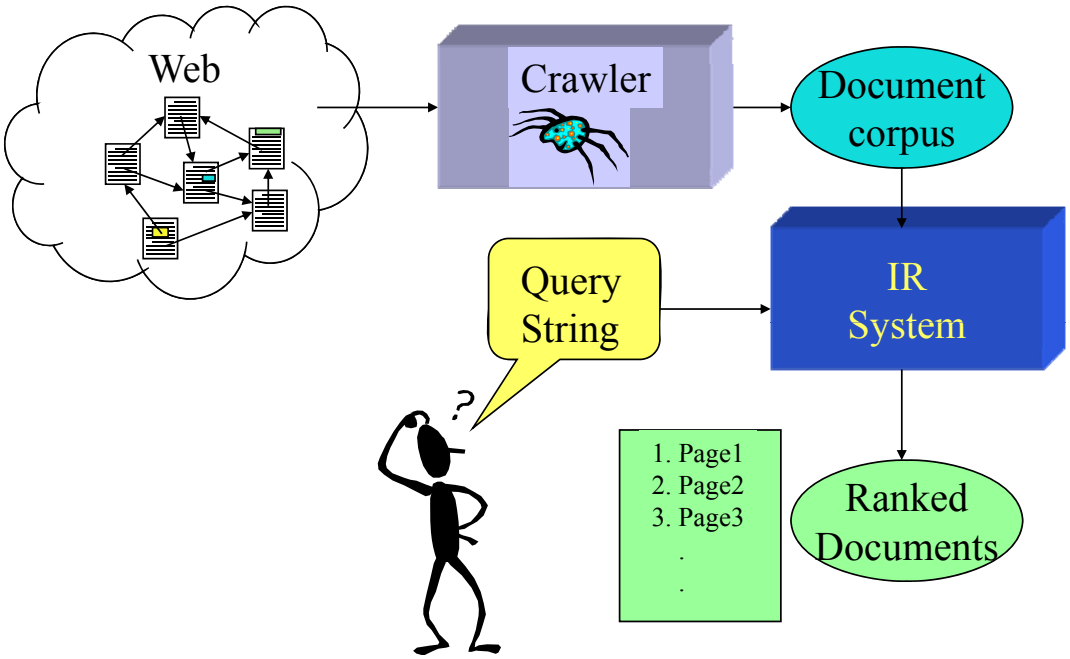
Η Αρχιτεκτονική ενός ΣΑΠ



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 **ενήμερη** 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>

Σύστημα Αναζήτησης Ιστού



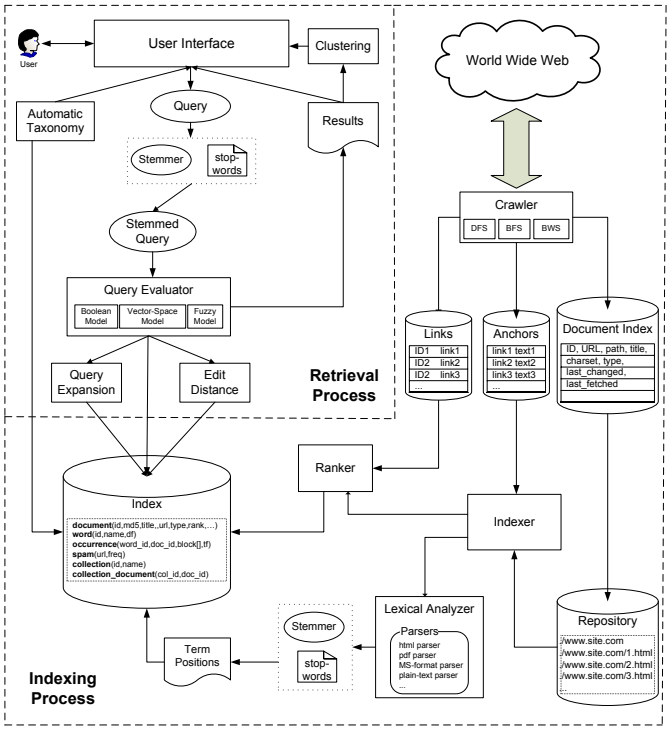
Γενική μορφή ενός Ευρετηρίου για ανάκτηση πληροφοριών από τον Ιστό

- **Crawling («έρπειν»)**

Web pages	Indexing Items k_1	k_2	...	k_j	...	k_t
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	$c_{1.N}$	$c_{2.N}$	...	$c_{i.N}$	...	$c_{t.N}$

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
  - Inquery, 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, Inquira
  - vivisimo (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 Συμπίεση 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

15. Ανάκτηση Δομημένων Εγγράφων

Ευρετηρίαση και ανάκτηση εγγράφων XML

16. Διεπαφές Χρήσης και Οπτικοποίηση (≈ 1 διάλεξη)

HY463: Θεματικές Ενότητες (IV)

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 Recent Advances*
-

Ιστορική Αναδρομή

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

Ιστορική Αναδρομή (II)

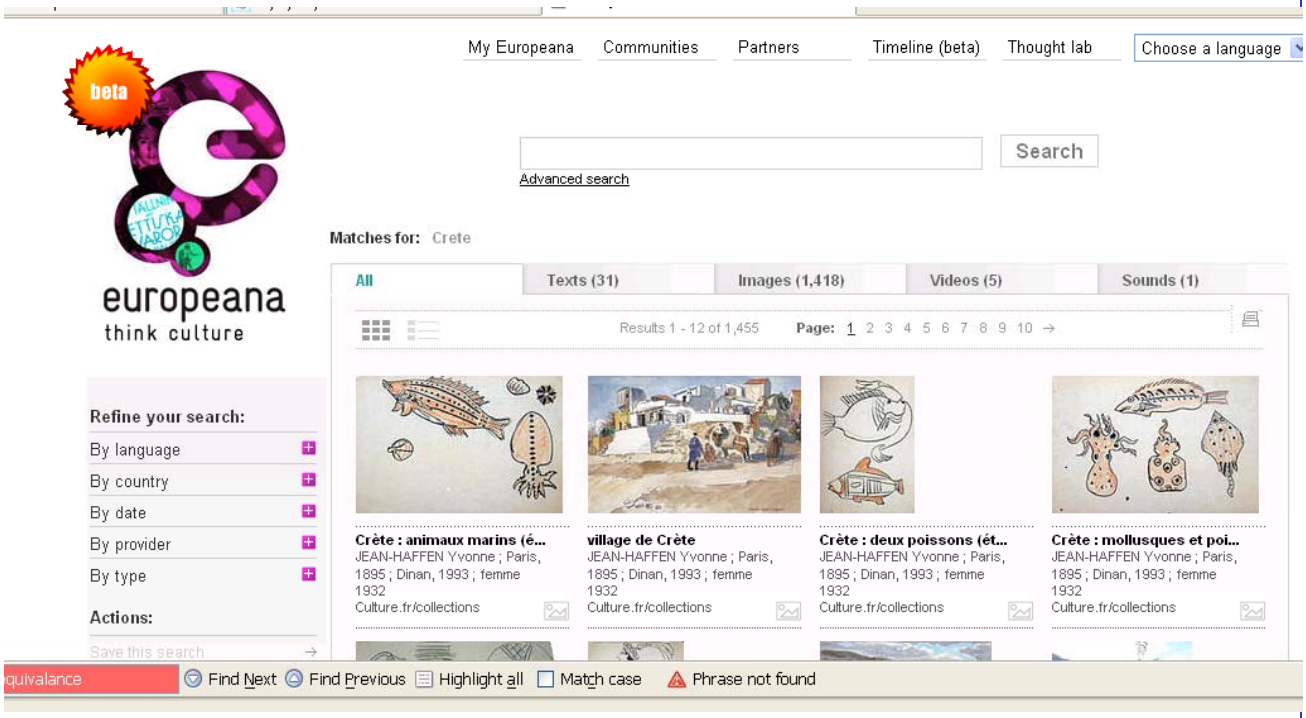
- 1990's:
  - Searching FTPable documents on the Internet
    - Archie
    - WAIS
  - Searching the World Wide Web
    - Lycos
    - Yahoo
    - Altavista
  - Organized Competitions
    - NIST TREC
  - Recommender Systems
    - Ringo
    - Amazon
    - NetPerceptions
  - Automated Text Categorization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

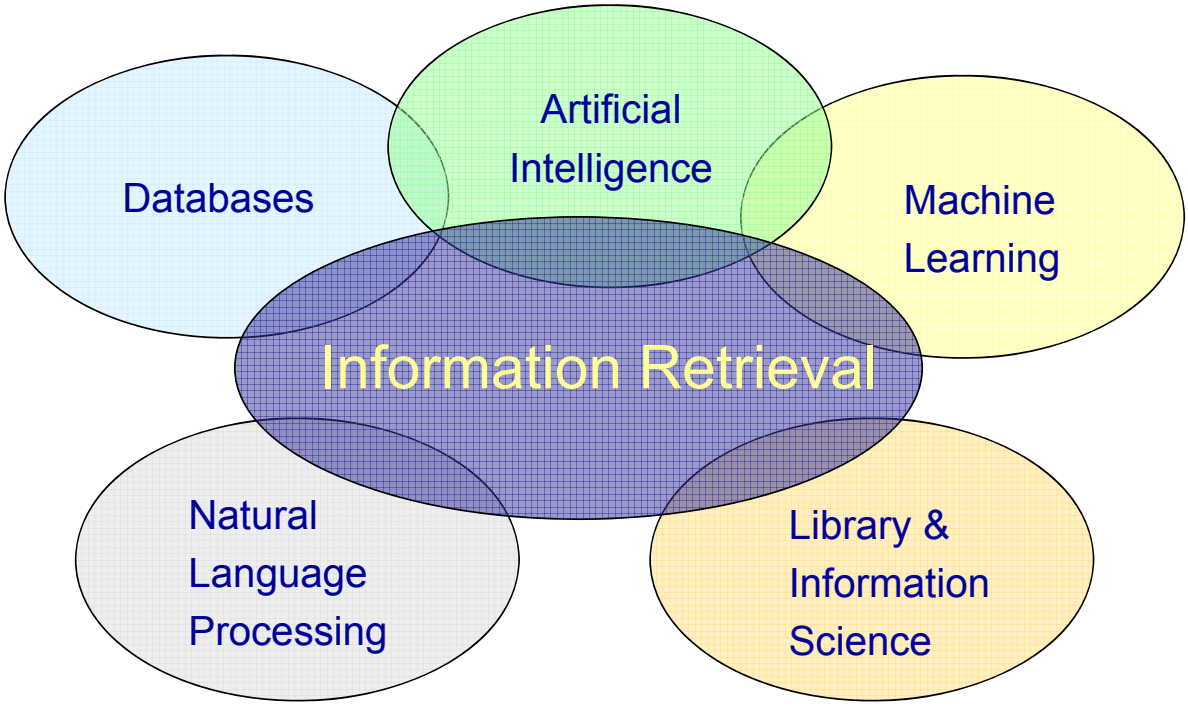
Μετά την επινόηση του Web αυτό άλλαξε για τα καλά:
- οικουμενική δεξαμενή γνώσης
- ελεύθερη (και φθηνή) καθολική πρόσβαση
- έλλειψη κεντρικού ελέγχου σύνταξης



«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

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

Τεχνητή Νοημοσύνη (Artificial Intelligence)

- Παραδοσιακά εστιάζει στην
  - παράσταση γνώσης (knolwedge 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).

Επεξεργασία Φυσικής Γλώσσας
Natural Language Processing

- Παραδοσιακά εστιάζει την
  - **συντακτική** (syntactic) ανάλυση,
  - **σημασιολογική** (semantic) ανάλυση και
  - **πραγματολογική (pragmatic)** ανάλυση

  της 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

Library and Information Science

- 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.