CS 463: Information Retrieval Systems

Εισαγωγή

Yannis Tzitzikas

University of Crete

Spring 05

Lecture : 1
Date : 22-2-2005

Διάρθρωση Διάλεξης

- *Γιατί χρειαζόμαστε Ανάκτηση Πληροφοριών (ΑΠ);*
- *Τι είναι η ΑΠ ?*
- *Πλοήγηση και Ανάκτηση*
- *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 (Google indexes 4.2 billions)
- 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

Τι να είναι η ΑΠ ?

YAHOO! Google

grep

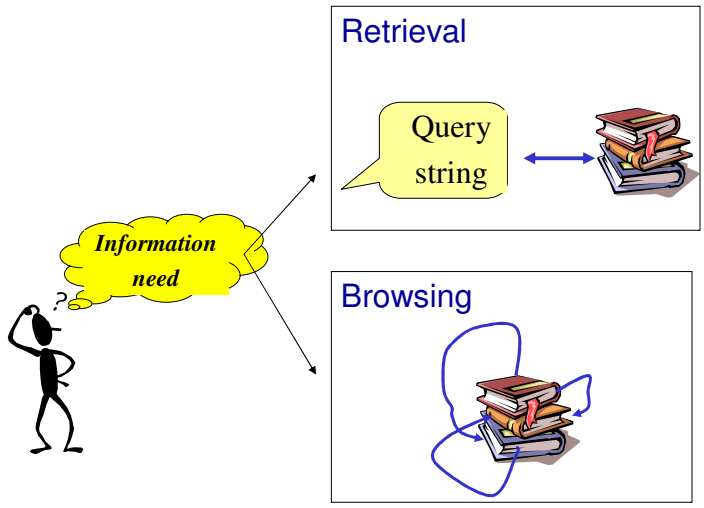
altavista Ask Jeeves

CiteSeer.IST
Scientific Literature Digital Library

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;

Retrieval vs Browsing

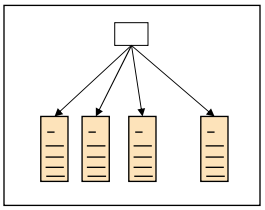


- Δεν έχουμε επερώτηση
- με σκοπό ή χωρίς σκοπό (εξερεύνηση)

Types of Browsing

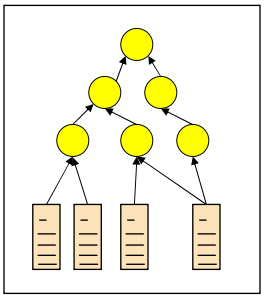
(1) Επίπεδο (flat)

- πχ. μια λίστα εγγράφων

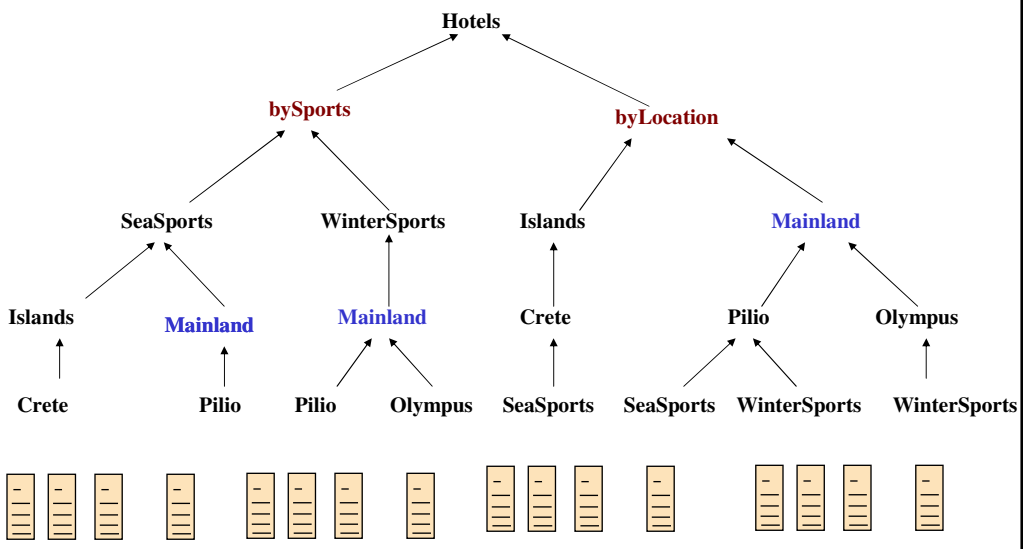


(2) Οδηγούμενο από δομή (structure guided)

- 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)

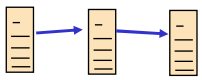


Παράδειγμα

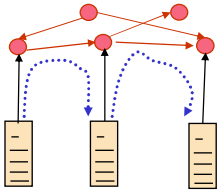


Types of Browsing (II)

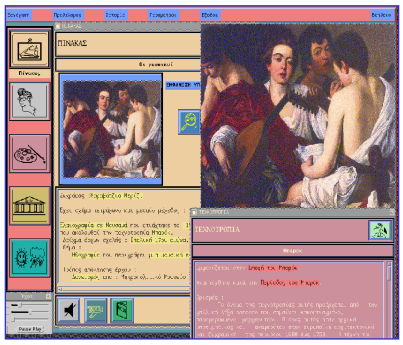
- **(3) Μη γραμμικό κείμενο** (Hypertext)
  - διευθυνόμενοι σύνδεσμοι (π.χ. HTML)
  - διπλής κατεύθυνσης
  - typed links, etc.



- (4) **Διεπίπεδο** μη γραμμικό κείμενο
  - Τα έγγραφα ταξινομούνται σε ένα εννοιολογικό σχήμα και από αυτήν την ταξινόμηση επάγονται οι συνδέσεις τους
  - πχ DOMENICUS
    [Tzitzikas & Theodorakis, Hypertext'96]

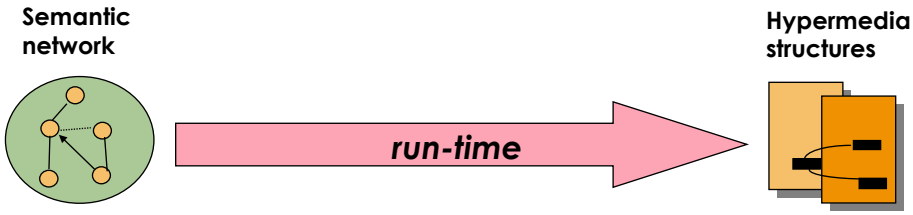


Το σύστημα Δομήνικος

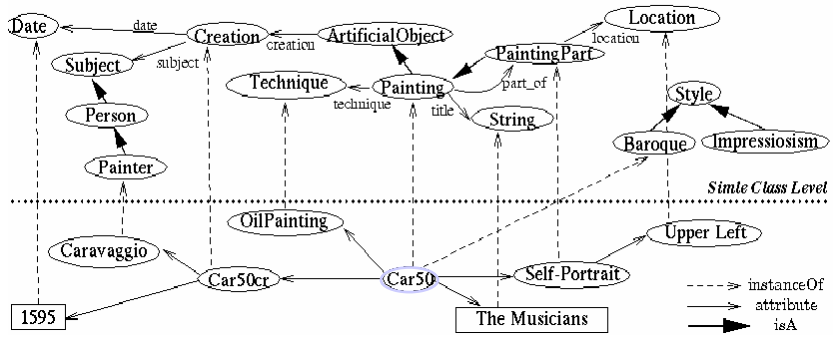


Λειτουργίες

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



Δομήνικος: Αυτόματη Παραγωγή Υπερκειμένου από Βάση Γνώσης



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

Δομήνικος

Πλεονεκτήματα διεπίπεδου hypertext

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

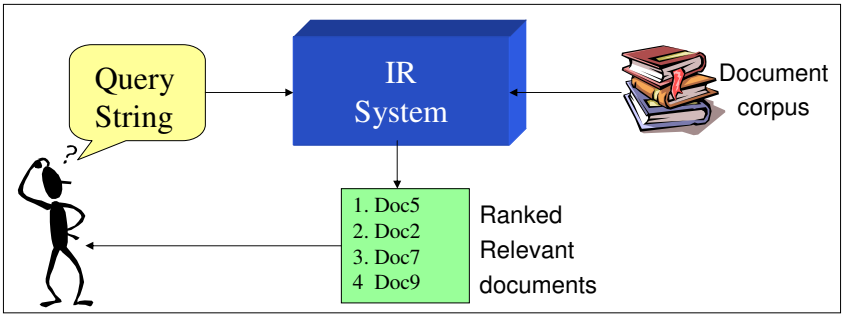
Το τυπικό πρόβλημα της Ανάκτησης Πλ.

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>**



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)

Data vs 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
  - παραγάγει διάταξη των εγγράφων βάσει της **συνάφειας**
  - *η ένοια της συνάφειας* είναι κυρίαρχο ζήτημα

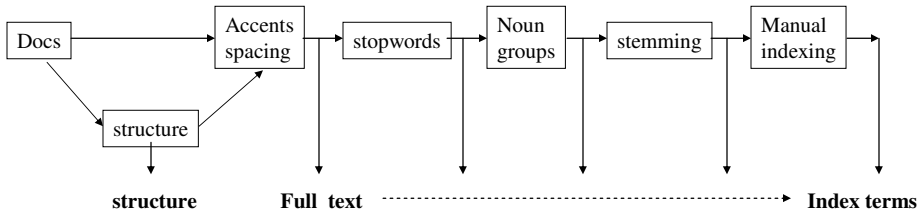
Συνάφεια (Relevance)

- **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)

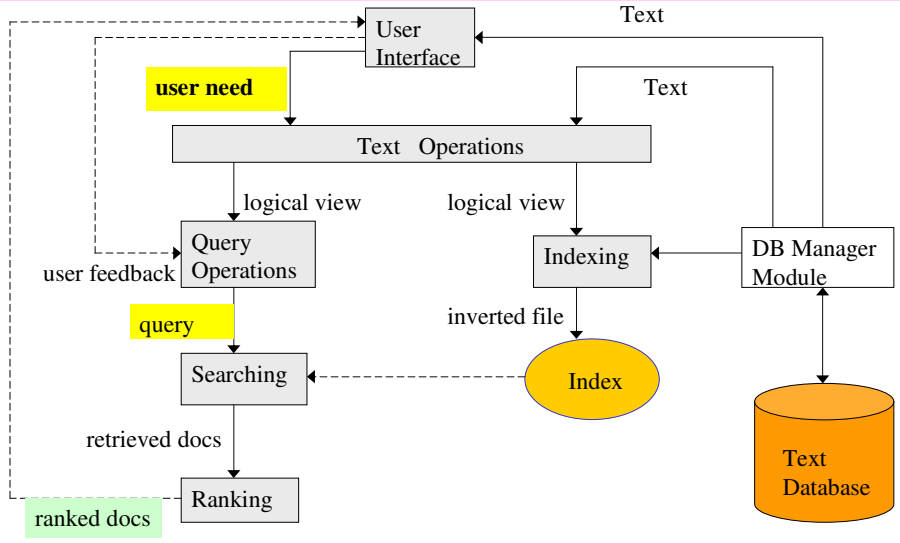
Πλήρες κείμενο => Όρους ευρετηρίου



Τα βασικά τμήματα ενός ΣΑΠ

- **Λειτουργίες Κειμένου (Text Operations)** σχηματίζουν τις λέξεις ευρετηρίου (tokens, index terms).
  - Αφαίρεση λέξεων αποκλεισμού (Stopword removal), Stemming
- **Ευρετηρίαση (Indexing)** κατασκευάζει ένα ευρετήριο (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 με βάσει μια μετρική 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
  - Μετασχηματισμός επερώτησης με ανάδραση συνάφειας

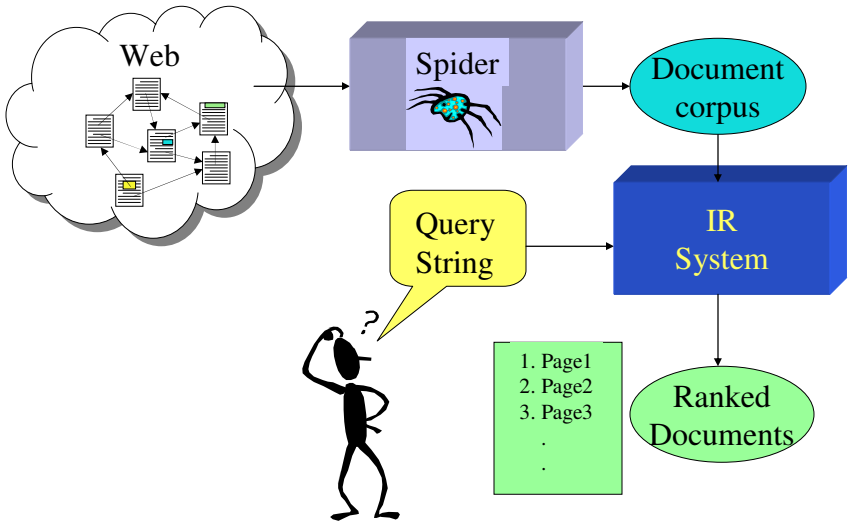
Η Αρχιτεκτονική ενός ΣΑΠ



Αναζήτηση στον Ιστό (Web Search)

- Εφαρμογή της ΑΠ σε έγγραφα HTML του Ιστού
- Διαφορές:
  - Πρέπει να **συλλέξουμε** τη συλλογή κειμένων **διασχίζοντας** (crawling/spidering) τον Ιστό
  - Μπορούμε να αξιοποιήσουμε της **δομή** της πληροφορίας στην HTML (XML).
  - Τα έγγραφα τροποποιούνται (χωρίς προειδοποίηση)
  - Μπορούμε να αξιοποιήσουμε τη **δομή των συνδέσμων** του Ιστού.

Σύστημα Αναζήτησης Ιστού



Άλλα Task σχετικά με την ΑΠ

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

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
  - ...

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

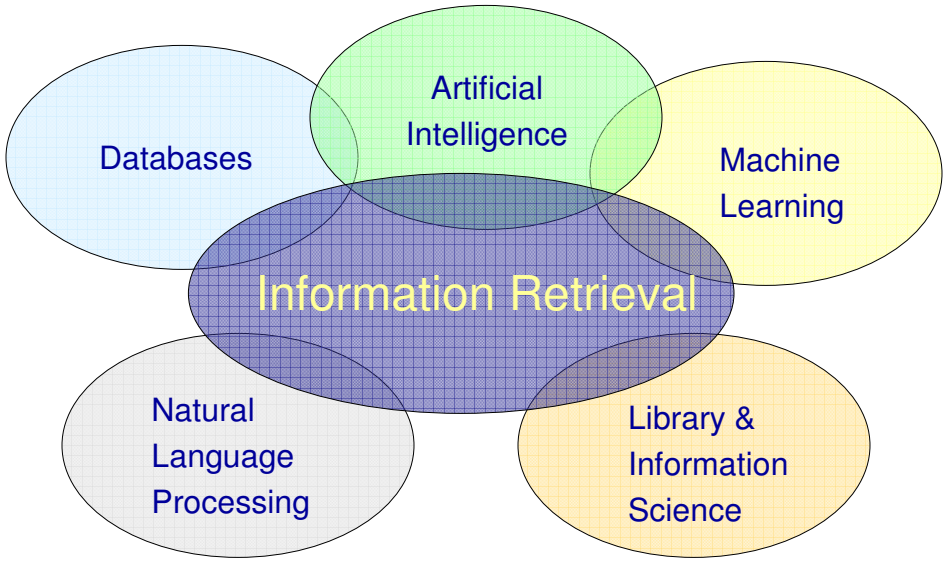
Πριν τον Ιστό η ΑΠ εθεωρείτο ότι είχε στενό πεδίο εφαρμογής

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
  - ΑΠ και context (η σωστή απάντηση για σένα εδώ και τώρα)
  - 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 παραδειγμάτων βάσει μάθησης από labeled training examples
- Μη-Επιτηρούμενη Μάθηση (Unsupervised learning)
  - Αυτόματη ομαδοποίηση unlabeled examples σε σημαντικές ομάδες (into meaningful groups).

Μηχανική Μάθηση:IR Directions

- Ταξινόμηση Κειμένου (Text Categorization)
  - Αυτόματη ιεραρχική ταξινόμηση (hierarchical classification, e.g.Yahoo).
  - Προσαρμόσιμο φιλτράρισμα/routing/σύσταση(recommending).
  - Αυτόματο φιλτράρισμα spam.
- Ομαδοποίηση Κειμένων (Text Clustering)
  - Ομαδοποίηση των αποτελεσμάτων της αναζήτησης
  - Αυτόματος σχηματισμός ιεραρχιών (Yahoo).
- Learning for Information Extraction
- Text Mining

Επεξεργασία Φυσικής Γλώσσας
Natural Language Processing

- Παραδοσιακά εστιάζει την
  - **συντακτική** (syntactic),
  - **σημασιολογική** (semantic) και
  - **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.

Επεξεργασία Φυσικής Γλώσσας (II)

- IR Directions:
  - Μέθοδοι για αποσαφήνιση 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.