# Ανάκτηση Πληροφοριών

# (Information Retrieval)

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

Διδάσκοντες:
Γιάννης Θεοδωρίδης, Επίκ. Καθηγητής (ytheod@unipi.gr)
Νίκος Ζάχαρης, ΠΔ407 (nzach@unipi.gr)

Ιστοσελίδες μαθήματος: http://thalis.cs.unipi.gr/~ir

Διαφάνειες βασισμένες στις εξής πηγές:

•διαφάνειες βιβλίου R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999 http://sunsite.dcc.uchile.cl/irbook/

•διαφάνειες μαθήματος Information Retrieval @Univ. of Alberta, instructor: Prof. Mario A. Nascimento http://www.cs.ualberta.ca/~mn/694/

# Με τι ασχολείται το μάθημα αυτό;

- Την τελευταία δεκαετία υπάρχει (εκ νέου) άνθηση της Ανάκτησης Πληροφοριών (Information Retrieval - IR) λόγω της έκρηξης στη διαθεσιμότητα κειμένων (και άλλων μέσων) που προκάλεσε το WWW.
- Αυτό το μάθημα θα σας εισάγει στην περιοχή της Ανάκτησης Πληροφοριών, που παρεμπιπτόντως δεν είναι ένα παρακλάδι των Βάσεων Δεδομένων, και θα σας επιτρέψει να κατανοήσετε και να εμβαθύνετε σε σχετικά θέματα.
- Βασικές αναφορές:
  - R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.
  - Σημειώσεις και διαφάνειες από τους διδάσκοντες

# IR versus DB

- IR is an area by itself, it is related to, but it is not "simply" database stuff !
  - IR used to be geared towards textual data, although this has changed (drastically) recently.
  - Structure:
    - a database has a well defined structure (e.g., relational model);
    - text usually has no structure (how about video ?)
  - How can we query something if we don't know its structure ?

# IR versus DB (cont.)

- Typical SQL query:

```
SELECT price FROM stock
WHERE prodname = "ski" or prodname = "snowshoes"
```

- OR

```
SELECT price FROM stock
WHERE prodid = 7657 or prodid = 7688
```

- Typical IR query:
  - price ski snowshoes
  - price AND (ski OR snowshoes)
  - price AND (ski OR snowshoes OR (snow AND shoes))

# IR versus DB (cont.)

- What if I want to sum all prices of all skis or snowshoes in stock? It is trivial using SQL:

```
SELECT SUM(price) FROM stock
WHERE prodname = "ski"  OR
             prodname = "snowshoes"
```

- In a IR context, this type of query doesn't make much sense. IR is not for that !

# IR versus DB (cont.)

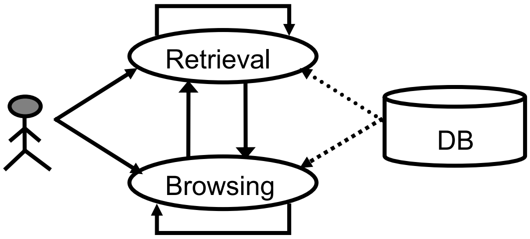
- Similarly, what if I want to search for texts which have words beginning with "bank" at most 4 words apart from the word "investment" and not containing the term "lost" ?
- That's not trivial, if possible at all, in SQL, but is trivial in IR-ish:
  - ADJ(bank*, investiment, 4) AND NOT lost
- There is no standard query language for IR.

# IR versus DB (cont.)

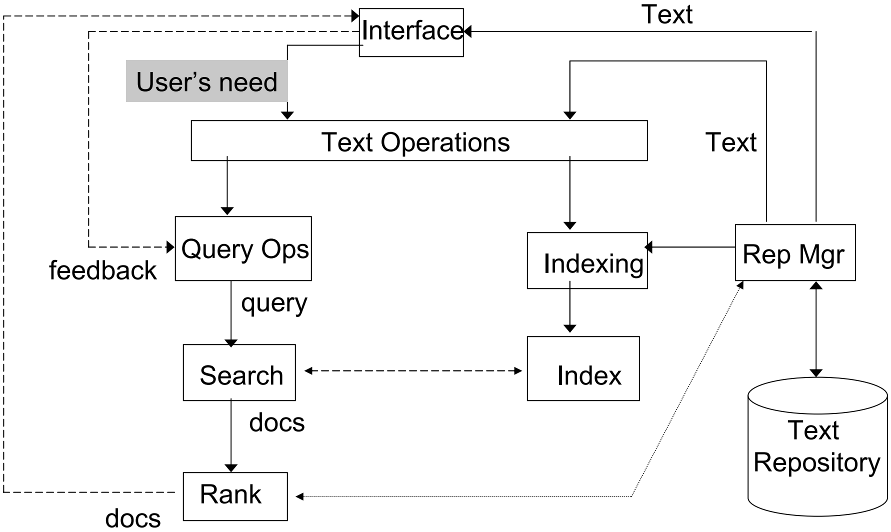
- What about the answers ?
  - DB: all answers are equally good.
  - IR: some answers are better than others.
- Is the WWW a really big DB ?
  - Does it have structure ? Not really ...
  - Is it only pure non-structured text ? Not really ...
  - Is the WWW a semi-structured DB ? (XML is hot stuff !)

# Retrieval Process

- Two alternatives to access the text (or multimedia or …) database



# Retrieval Process



# Modeling Issues

- DBs use a powerful abstraction for modeling scenarios: the ER model (conceptual level) and the Relational model (logical level).
- How can we model textual data, where there is no basic structure ?
- Several models:
  - The boolean and vector model are mostly used.
  - Probabilistic models and others also exist.

# Retrieval Evaluation

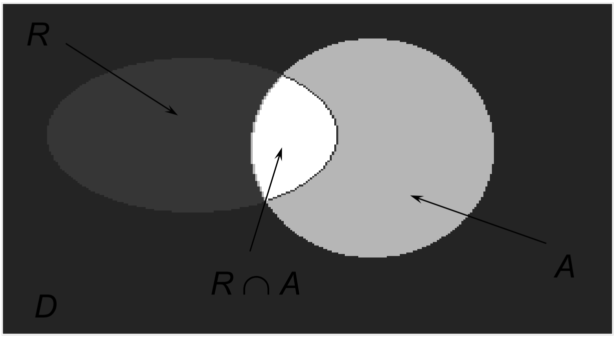
- How important is the query length ?

| Query | ~ Number of pages found |
|---|---:|
| hotel | 46,600,600 |
| hotel paris | 1,450,000 |
| hotel paris train | 227,000 |
| hotel paris train student | 31,700 |
| hotel paris train student discount | 7,140 |

# Retrieval Evaluation (cont.)

- **Precision** and **Recall** are the de facto standard metrics for retrieval evaluation.
  - Recall: the fraction of relevant documents retrieved
  - Precision: the fraction of retrieved documents being relevant

•Searching for only one term gives extremely large recall but very poor precision.
•Using all terms instead gives much better precision and likely more meaningful recall.



# Query Languages and Operations

- There is no SQL-like for IR.
- Processing natural language is still not really a choice (though a lot of effort has been put into it).
- Boolean queries are prevalent, though completeness likely yields complexity.
  - Example: "amazon OR (rain AND forest)"
- Regular expressions can be used but are not as common and not as easy to write
  - Example: “research cent{er|re}?s in Bra{z|s}il”
- The WWW suffers from the "single word query syndrome".

# Query Languages and Operations (cont.)

- Queries can also be extended "automatically" using thesauri and may give you better results
  - Example: "mall" can be automatically extended to "mall OR (shopping AND centre)"
- However, this may not be a good idea all the time.
  - Example: "amazon" could be expanded to "amazon OR (rain AND forest)". Not good if you were looking for a certain bookstore ...

# Indexing and Searching

- B-trees and Hash tables are usually the indexing methods of choices for DBs.
- They're not as good, actually not very useful, for IR.
- In IR the terms to be indexed appear many (perhaps millions of) times. This would render any B-tree/Hash table quite useless.
- Nevertheless, we cannot afford a linear scan on all documents either.

# Indexing and Searching (cont.)

- Inverted lists are widely known and usually advocated as the best bet for IR.
- Signature files have some advocates as well, but it relies mostly on the fact that inverted lists are too large - not necessarily true anymore.
- Other issues:
  - Indices on compressed data are quite popular nowadays
  - Indexing multimedia (images, audio, video) is still a challenge

# User interfaces

- Assuming we know how to query a IR system (say, www.altavista.com or www.google.com), I want to be able to understand well the response returned.
- If I am given only the first 20 of 1,000 pages found, is it any good ? Can I use that to refine my query ?
- How do the pages relate to each other ? Is a complex graph necessarily "better" than a list of URLs ranked in order of relevance ?

# Searching the WWW

- **The WWW is yet another big textual database ... NOT QUITE …**
- The WWW "data" has:
    - inherently distributed and heterogeneous nature;
    - extremely large volume;
    - high volatility and high ratio of updates;
    - poor structure (though XML tries to help there);
    - Utilizes ad hoc user interfaces;
    - etc, etc.

# Top 10 θέματα IR (το 1995)

- Το 1995 ο Croft έγραψε ένα άρθρο όπου απαριθμούσε (και επιχειρηματολογούσε για) τα 10 κορυφαία θέματα που θα ενδιέφεραν τις εταιρίες που χρησιμοποιούν ή εμπορεύονται συστήματα IR

    **`http://www.dlib.org/dlib/november95/11croft.html`**

- Κατά αύξουσα σειρά σημασίας (κατά την κρίση του), τα 10 θέματα ήταν:

    #10 Relevance Feedback
    #9 Information Extraction
    #8 Multimedia retrieval
    #7 Effective retrieval
    #6 Routing and filtering
    #5 Interfaces and Browsing
    #4 "Magic"
    #3 Efficient indexing and retrieval
    #2 Distributed IR
    #1 Integrated solutions

# Εργασία (@#%$#%$)

- Διαβάστε την εργασία του Croft και γράψτε τη γνώμη σας μέσα σε 1 σελίδα (μονά διαστήματα, μέγεθος γραμματοσειράς 12) για τα εξής:
  - Βλέποντας τα σημερινά δεδομένα, πιστεύετε ότι του «ξέφυγε» κάτι σημαντικό;
  - Υπέρ- (ή υπό-) εκτίμησε κάτι;
  - Κοιτώντας 5 χρόνια μπροστά, τι θα είναι πιο σημαντικό από εμπορική σκοπιά για το IR; (με άλλα λόγια, πού θα επενδύατε αν είχατε τα κεφάλαια)
- Στείλτε μας e-mail με την απάντησή σας σε ένα συνημμένο .pdf αρχείο πριν το επόμενο μάθημα (και προετοιμαστείτε να επιχειρηματολογήσετε σχετικά)