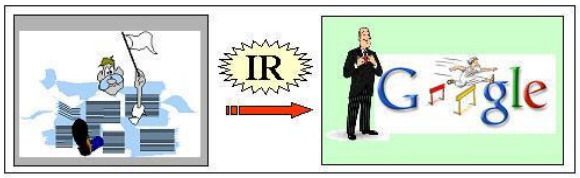
HY-463
Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval Systems

Πανεπιστήμιο Κρήτης, Άνοιξη 2005



Γιάννης Τζίτζικας

Lecture : 1
Date : 22-2-2005
Title : Administration

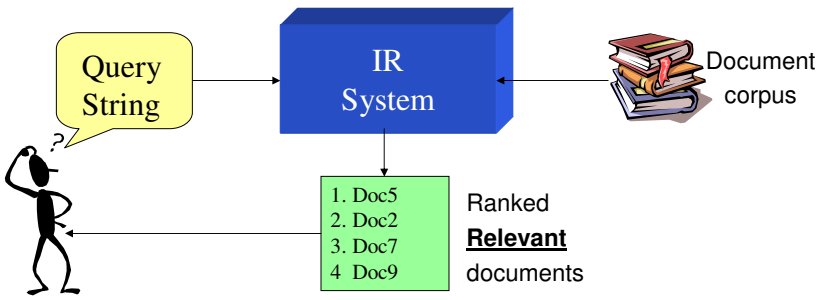
Το Αντικείμενο του Μαθήματος

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>**



CS-463: Information Retrieval Systems

- Διδακτικές μονάδες: 4
- Προαπαιτούμενα
  - CS-240 Δομές Δεδομένων
- Εβδομαδιαίο Πρόγραμμα :
  - **Διαλέξεις**: Τρίτη 9-11 και Πέμπτη 1-3 (αίθουσα PA201)
  - **Φροντιστήρια**: Δευτέρα 3-5 (αίθουσα Λ206)
- Παρακολούθηση
  - Αναμενόμενη αλλά όχι υποχρεωτική
- Γραφτείτε (σήμερα) στη λίστα **`hy463-list`**

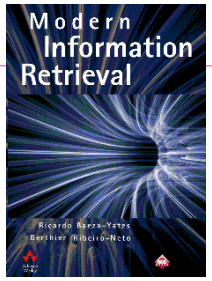
Προσωπικό

- Διδάσκων:
  - Γιάννης Τζίτζικας
  - tzitzik@csd.uoc.gr
  - Γραφείο: Γ111
  - Ώρες γραφείου: μετά τις διαλέξεις (Τρίτη 11-1, Πέμπτη 3-4)
- Βοηθοί:
  - Akkus Zebide
  - Σταύρος Σαχτούρης
  - Περικλής Τζιάβας
    - Υπεύθυνοι για
      - Λύση και Βαθμολόγηση ασκήσεων, Επίβλεψη projects
      - 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
  - δείτε ιστοσελίδα
- Φωτοτυπίες

Βαθμολόγηση

- **Τελικός βαθμός**
  - **Total** = 20% Ασκήσεις + 30% Έργο + 20% Πρόοδος + 30%ΤελικήΕξ
- Για να περάσετε το μάθημα χρειάζεστε
  - **Total** ≥ 5 **AND** ΤελικήΕξ ≥ 5
- Σημειώσεις στην Πρόοδο/Τελική Εξέταση:
  - Κλειστά.

Εντιμότητα

- Αντιγραφή ή άλλες μορφές κλοπής θα σημάνουν 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)

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 IR system,
  - να έχετε κατανοήσει τους καθιερωμένους τρόπους ευρετηρίασης και ανάκτησης του Παγκόσμιου Ιστού,
  - να έχετε γνωρίσει ποικίλους αλγόριθμους και συστήματα.

Οργάνωση Περιεχομένου

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. Γλώσσες Επερώτησης για Ανάκτηση Πληροφοριών (1 διάλεξη)

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

5. Προχωρημένες Λειτουργίες Επερώτησης (1 διάλεξη)

Επέκταση επερώτησης, Ανάδραση συνάφειας, Αυτόματη τοπική/καθολική ανάλυση

Οργάνωση Περιεχομένου (II)

6. Ευρετηρίαση, Προεπεξεργασία και Οργάνωση Αρχείων Κειμένου (1 δ)

Λέξεις αποκλεισμού (stopwords), stemming (στελέχωση κειμένου) , θησαυροί όρων

Ανεστραμμένα Αρχεία (inverted files), Δένδρα Καταλήξεων (suffix trees), Αρχεία Υπογραφών (signature files)

7. Στατιστικά και Συμπίεση Κειμένου (1 διάλεξη)

8. Αναζήτηση σε Κείμενα

Αλγόριθμοι Knuth-Morris-Pratt, Boyer-Moore, Αυτόματο καταλήξεων (suffix automaton), Φράσεις και εγγύτητα

9. Ομαδοποίηση Εγγράφων (Clustering) (1 διάλεξη)

10. Ανάκτηση Πολυμέσων (1 διάλεξη)

Μοντέλα και γλώσσες, Ευρετηρίαση και Αναζήτηση

Οργάνωση Περιεχομένου (III)

11.Παράλληλη και Κατανεμημένη Ανάκτηση Πληροφοριών (2 διαλέξεις)

Αρχιτεκτονικές MIMD, SIMD, Peer-2-Peer,

Διαμερισμός συλλογών, Επιλογή πηγής ,Επεξεργασία επερωτήσεων

11. Τεχνικές μετα-Κατάταξης (meta-ranking) (1 διάλεξη)

Ενοποιημένες και απομονωμένες μέθοδοι, Παρεμβολή, Ψηφοφορία

12. Αναζήτηση στον Παγκόσμιο Ιστό (3 διαλέξεις)

Ιστορική αναδρομή, Ευρετηριασμός ιστοσελίδων, Διάσχιση του ιστού (crawling), Τεχνικές ανάλυσης συνδέσμων (link analysis), PageRank, HITS

13. Διεπαφές Χρήσης και Οπτικοποίηση (1 διάλεξη)

14. Μελέτη Περιπτώσεων (Case Studies) (…)

Οργάνωση Περιεχομένου (IV)

15. Άλλα σχετικά ζητήματα

- *User Profiles*
- *Cross language retrieval*
- *Collaborative Filtering*
- *Generalized Interaction Models*
- *Faced Classification Theory and Advances*
- *Information Extraction (Stanford)*
- *Text Categorization (Stanford)*
- *Digital Libraries Video Retrieval*
-

Σειρές Ασκήσεων

- Στόχος: η κατανόηση και εμπέδωση της ύλης, και η επαφή με το μάθημα κατά τη διάρκεια του εξαμήνου
- Θα δοθούν 4 σειρές ασκήσεων
  - 1. Αξιολόγηση (θεωρητική)
  - 2. Μοντέλα Ανάκτησης (θεωρητική)
  - 3. Οργάνωση αρχείων κειμένου (θεωρητική)
  - 4. Ιδιότητες κειμένου (προγραμματιστική)
- Βάρος: 20% του τελικού βαθμού

Πρόοδος

- Υποχρεωτική (20% τελικού βαθμού) και θα γίνει πριν την 25η Μαρτίου

Project

- 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