ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ
(SEARCH ENGINES)

ΔΙΑΛΕΞΗ 1

ΔΙΔΑΣΚΩΝ

ΚΩΣΤΑΣ ΚΟΛΟΜΒΑΤΣΟΣ

ΔΙΑΔΙΚΑΣΤΙΚΑ

Διαδικαστικά του Μαθήματος (1/2)

- Διδασκαλία
  - Πέμπτη 1500-1800

Ιστοσελίδα Μαθήματος

- https://eclass.uth.gr/courses/CS_U_209/
- Στην ιστοσελίδα θα βρείτε
  - Σημειώσεις
  - Τις διαλέξεις του μαθήματος
  - Χρήσιμο Υλικό

Διαδικαστικά του Μαθήματος (2/2)

- Τρόπος Εξέτασης
  - Τελική Εξέταση
    - Βάρος: 70% του τελικού βαθμού της θεωρίας
  - Εργασία
    - Θα έχει βάρος 30% στον τελικό βαθμό
    - Αυστηρή ημερομηνία παράδοσης

Βιβλιογραφία

- Manning, C., Raghavan, P., Schutze, H., 'An Introduction to Information Retrieval', Cambridge University Press, 2009.
- Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni. S., 'Web Information Retrieval', Spinger, 2013
- Zhai, C., Massung, S., 'Text Data Management and Analysis', ACM, 2016.

Syllabus

- Introduction to Information Retrieval
- Information Retrieval Models
- Classification and Clustering
- Natural Language Processing
- Text Data Analysis
- Search Engines
- Link Analysis
- Recommendation Systems for Search Engines
- Meta-Search and Multi-Domain Search
- Semantic Search
- Multimedia Search

Εισαγωγή (1/7)

- Η **Ανάκτηση Πληροφορίας - ΑΠ (Information Retrieval – IR)** σχετίζεται με την αναπαράσταση, την οργάνωση και την πρόσβαση στην πληροφορία

- Στόχος της είναι να ανακτήσει πληροφορία που σχετίζεται με κάποιο χρήστη

- Η IR σχετίζεται άμεσα με δύο ορισμούς:
  - τη **συσχέτιση (relevance)** και
  - **μεγάλου όγκου μη δομημένα δεδομένα (large scale unstructured data)**

Εισαγωγή (2/7)

- Συσχέτιση / σχετικότητα / σχέση
  - Η **σχετικότητα/συσχέτιση** των αποτελεσμάτων εστιάζει στην πληροφορία που έχει ανάγκη ο χρήστης και όχι στο ίδιο το ερώτημα
  - Παράδειγμα:
    - Υποθέτουμε πως θέλουμε να βρούμε μια σοκολάτα που να μην επιβαρύνει την αρτηριακή μας πίεση
    - Έστω ότι εκφράζουμε το ερώτημα: *σοκολάτα επίπτωση αρτηριακή πίεση*
    - Θα αποτιμήσουμε το αποτέλεσμα όχι σε σχέση με το ερώτημα αλλά με την πληροφορία που θα πάρουμε
  - Γενικά όμως αν η πληροφορία που θα πάρουμε περιλαμβάνει λέξεις κλειδιά του ερωτήματος αποτελεί μια καλή αποτίμηση

Εισαγωγή (3/7)

- Η σχετικότητα/συσχέτιση έχει κάποιες ιδιότητες:
  - Είναι **υποκειμενική**
    - δύο χρήστες μπορούν να πάρουν την ίδια πληροφορία αλλά να αποτιμήσουν διαφορετικά το τελικό αποτέλεσμα (τη σχέση με τα ανακτώμενα έγγραφα)
  - Είναι **δυναμική**
    - η δυναμική έγκειται τόσο σε χώρο όσο και σε χρόνο – ο χρόνος που απεικονίζονται τα ανακτώμενα έγγραφα μπορεί να επηρεάζει την αποτίμηση
  - Είναι **πολυεπίπεδη**
    - η αποτίμηση εξαρτάται τόσο από το περιεχόμενο των αποτελεσμάτων όσο και από την αξιοπιστία, την εξαντλητικότητα, την πιο πρόσφατη πληροφορία, κ.λπ.

Εισαγωγή (4/7)

- Γενικά στην IR ισχύει:
  - Ανακτούμε ένα σύνολο εγγράφων D για ένα ερώτημα $\mathbf{q_k}$ υπολογίζοντας μια συνάρτηση σχετικότητας $\mathbf{R(q_k, d_j)}$ όπου το d_j ανήκει στο D
  - Το μοντέλο που υιοθετούμε στην IR παίζει σημαντικό ρόλο για την επιλογή της συνάρτησης R

Εισαγωγή (5/7)

- Μεγάλου όγκου μη δομημένα δεδομένα
  - Η ποσότητα πληροφορίας που είναι διαθέσιμη είναι τεράστια (exabytes, zettabyte)
  - Η ποσότητα πληροφορίας είναι μεγαλύτερη από αυτή που έχει καταγραφεί τα τελευταία 5,000 χρόνια
  - Το κλειδί είναι το γεγονός πως δεν έχουμε να κάνουμε με κάποιου είδους δόμηση στην πληροφορία

  - Η ανάκτηση των δεδομένων στα συστήματα διαχείρισης βάσεων δεδομένων ή σε μορφή XML εμπλέκει μια δομημένη μορφή
  - Σε αυτές τις περιπτώσεις τα δεδομένα μπορούν να προσπελαστούν με χρήση γλωσσών ερωταποκρίσεων
  - Επίσης, τα αποτελέσματα αποτελούν ακριβή ταιριάσματα ενώ μερικά ταιριάσματα αποκλείονται από τα αποτελέσματα

Εισαγωγή (6/7)

- Μοντέλα IR
  - Ορίζονται ως εξής:
  - IRM={D, Q, F, R(q_k, d_j)}
  - D: είναι το σύνολο των λογικών αναπαραστάσεων των εγγράφων
  - Q: είναι το σύνολο των αναπαραστάσεων των ερωτημάτων που τίθενται προς τα δεδομένα
  - F: είναι ένα πλαίσιο για την αναπαράσταση των εγγράφων, των ερωτημάτων καθώς και των σχέσεων μεταξύ τους
  - R(q_k, d_j): είναι η συνάρτηση που εξάγει ένα πραγματικό αριθμό ως το βαθμό συσχέτισης ανάμεσα σε ένα έγγραφο και ένα ερώτημα

Εισαγωγή (7/7)

- Τυπικές IR επεξεργασίες
  - **Information filtering**: αφαιρεί πλεονάζουσα πληροφορία από μια ροή (stream) πληροφορίας χρησιμοποιώντας αυτοματοποιημένες μεθόδους
  - **Document summarization**: δημιουργεί μια πιο 'σύντομη' εκδοχή ενός εγγράφου για να μειώσει την υπερφόρτωση πληροφορίας
  - **Document clustering and categorization**: ομαδοποιεί έγγραφα βάσει της 'απόστασης' μεταξύ τους ή τα κατηγοριοποιεί σε προκαθορισμένες κλάσεις
  - **Question answering**: επιλέγει τα σχετικά έγγραφα που απαντούν στα ερωτήματα των χρηστών που έχουν τεθεί σε φυσική γλώσσα
  - **Recommendation systems**: εξάγουν / προτείνουν πληροφορίες στους χρήστες με βάση τα χαρακτηριστικά τους
  - **Cross-language retrieval**: ανακτά έγγραφα που είναι σε γλώσσα διαφορετική από το ερώτημα

Αποτίμηση των IR Συστημάτων (1/9)

- Ένας αριθμός ιδιοτήτων / παραμέτρων πρέπει να λαμβάνονται υπόψιν
  - Ταχύτητα
  - Επίδοση
  - Παραδείγματα
    - Πλήθος ανακτώμενων εγγράφων ανά ώρα
    - Καθυστέρηση σε σχέση με την πολυπλοκότητα του ερωτήματος

- Χρειάζεται να βρούμε το trade off
  - Μπορεί να ανακτήσουμε ‘άχρηστα’ έγγραφα πάρα πολύ γρήγορα
- Η ικανοποίηση των χρηστών αποτελεί μια σημαντική παράμετρο
  - Ικανοποιημένοι χρήστες επιστρέφουν στην ίδια μηχανή αναζήτησης

Αποτίμηση των IR Συστημάτων (2/9)

- Όταν τα συστήματα IR επιστρέφουν μη δομημένα αποτελέσματα η επίδοσή τους μπορεί να αποτιμηθεί με τα precision & recall

- **Precision P** είναι το τμήμα των ανακτώμενων εγγράφων που σχετίζονται με το ερώτημα
- **Recall R** είναι το τμήμα των εγγράφων που σωστά έχουν ανακτηθεί

Αποτίμηση των IR Συστημάτων (3/9)

- Ορίζουμε τα ακόλουθα για ένα ένα ερώτημα q και ένα σύνολο εγγράφων D*:

$$P = \frac{|TP|}{|TP| + |FP|}$$

$$R = \frac{|TP|}{|TP| + |FN|}$$

		Predicted class: *yes*	Predicted class: *no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P′*	*N′*	*P* + *N*

- **Προφανώς κάποιος πρέπει να ορίσει το πότε ένα έγγραφο είναι σχετικό!!**
- Επιπλέον μπορούμε να ορίσουμε την **F-measure**

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

- Συνήθως β=1

*TP είναι τα σχετικά έγγραφα, FP είναι τα έγγραφα που είναι άσχετα, FN είναι τα έγγραφα που λανθασμένα δεν έχουν ανακτηθεί

Αποτίμηση των IR Συστημάτων (4/9)

- Όταν το σύστημα ελαχιστοποιεί το βαθμό επιλογής τότε εμφανίζονται άσχετα έγγραφα (false positives)
- Όσο μεγαλύτερος είναι ο αριθμός των εγγράφων τόσο μεγαλύτερη η πιθανότητα να εμπλακούν άσχετα αποτελέσματα

- Τα Precision και Recall δεν απεικονίζουν το ranking των αποτελεσμάτων
- Αν θέλουμε να κατατάξουμε τα αποτελέσματα θα πρέπει να τα ταξινομήσουμε ως προς το βαθμό συσχέτισης

- Μόνο ένα τμήμα θα παρουσιαστεί στους χρήστες

Αποτίμηση των IR Συστημάτων (5/9)

- Ένας αποδοτικός τρόπος για να απεικονίσουμε μια ταξινομημένη λίστα είναι να δούμε το τι κερδίζουμε σε precision όταν αυξάνουμε το recall

- Το μέσο precision απεικονίζει το μέσο όρο των τιμών του precision για τα top-k έγγραφα
- Συνήθως το μέσο precision υπολογίζεται ως εξής*:

$$AveP = \sum_{k=1}^{n} P(k)\Delta r(k)$$

* n είναι το πλήθος των ανακτώμενων εγγράφων, P(k) είναι το precision που υπολογίζεται όταν το σύνολο των αποτελεσμάτων αποκόπτεται στα top-k έγγραφα, Δr(k) είναι η μεταβολή του recall όταν μεταβαίνουμε από το k-1 στο k έγγραφο

Αποτίμηση των IR Συστημάτων (6/9)

- Το precision για μια IR μηχανή υπολογίζεται για ένα σύνολο ερωτημάτων
- Γενικά εφαρμόζουμε την ακόλουθη εξίσωση:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

Αποτίμηση των IR Συστημάτων (7/9)

- Πολλές μηχανές αναζήτησης εστιάζουν στο πλήθος των ‘καλών’ αποτελεσμάτων της 1η σελίδας ή των πρώτων σελίδων
- Μια κατάλληλη προσέγγιση θα ήταν να μετρήσουμε το precision για τα πρώτα 10 – 30 έγγραφα

- Αυτή η μετρική αναφέρεται ως **precision at k** ή **P@k**

- Έχει το πλεονέκτημα ότι δεν απαιτεί κάποια εκτίμηση του πλήθους των σχετικών εγγράφων
- Από την άλλη είναι η λιγότερο σταθερή μετρική αφού επηρεάζεται από από το συνολικό αριθμό των εγγράφων

Αποτίμηση των IR Συστημάτων (8/9)

- Άλλη μετρική είναι η **discounted cumulative gain (DCG)** που εφαρμόζεται στα top-k έγγραφα
- Αντίθετα με την P@k, υιοθετεί μια βαθμιαία κλιμάκωση της σχετικότητας
- Η DCG μοντελοποιεί το κέρδος ενός εγγράφου βάσει της θέσης του στη λίστα των αποτελεσμάτων
- Το κέρδος συσσωρεύεται από την κορυφή της λίστας προς το τέλος της
- Όσο πιο κάτω είναι ένα έγγραφο τόσο το μεγαλύτερο το πέναλτυ που δέχεται
- Η βαθμιαία σχετικότητα μειώνεται λογαριθμικά σε σχέση με τη θέση
- Ισχύει* $DCG = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)}$

* rel_1 είναι η σχετικότητα του 1ου εγγράφου και rel_i είναι η σχετικότητα του i-στου εγγράφου

Αποτίμηση των IR Συστημάτων (9/9)

- Διαθέσιμα σύνολα δεδομένων για την αποτίμηση των IR συστημάτων αποτελούν τα ακόλουθα:
  - Granfield collection
  - Nist dataset
  - Reuters corpus
  - Conference and labs of the Evaluation Forum