HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Αξιολόγηση Ανάκτησης
Retrieval Evaluation

Γιάννης Τζίτζικας
Διάλεξη : 2/3

Διάρθρωση Διάλεξης

- *Τι εξυπηρετεί η αξιολόγηση;*
  - *αξιολόγηση αποτελεσματικότητας*
- *Δυσκολίες της αξιολόγησης*
- *Αξιολόγηση μέσω Χειρονακτικά Μαρκαρισμένων Συλλογών*
- *Μέτρα αξιολόγησης αποτελεσματικότητας*
  - *Ανάκληση & Ακρίβεια & (Recall & Precision)*
  - *Καμπύλες Ακρίβειας/Ανάκλησης*
    - *Σύγκριση Συστημάτων*
  - *Εναλλακτικά μέτρα*
    - *R-Precision (Precision Histograms)*
    - *F-Measure*
    - *E-Measure*
    - *Fallout, Expected Search Length*
  - *User-Oriented Measures*
- *Δοκιμασίες Αποτελεσματικότητας-Συλλογές Αναφοράς (TREC)*

Τύποι Αξιολόγησης

- Αποδοτικότητας (efficiency)
  - Εδώ αξιολογούμε τις επιδόσεις του συστήματος (system performance)
  - Μέτρα: χρόνος απόκρισης, αποθηκευτικός χώρος, …
- Κόστους
  - Κόστος ανάπτυξης
    - σχεδιασμού, υλοποίησης, δοκιμών (testing), αξιολόγησης
  - Λειτουργικά έξοδα
    - εξοπλισμού, προσωπικού, κτλ
- **Αποτελεσματικότητας (effectiveness)**

Τι εξυπηρετεί η αξιολόγηση Αποτελεσματικότητας;

- Υπάρχουν πολλά μοντέλα υπολογισμού του βαθμού συνάφειας, πολλοί αλγόριθμοι και ακόμα περισσότερα συστήματα. Ποιο είναι το καλύτερο;
- Ποιος είναι ο καλύτερος τρόπος για:
  - Επιλογή των όρων του ευρετηρίου (stopword removal, stemming…)
  - Προσδιορισμό των βαρών των όρων (Term weighting) (TF, TF-IDF,…)
  - Υπολογισμό του βαθμού συνάφειας (dot-product, cosine, …) βάσει του οποίου θα γίνει η κατάταξη των εγγράφων;
- Πόσα έγγραφα της απόκρισης ενός συστήματος πρέπει να εξετάσει ο χρήστης προκειμένου να βρει μερικά/όλα τα συναφή έγγραφα;

Οι Δυσκολίες της Αξιολόγησης

- Η αποτελεσματικότητα εξαρτάται από τη ***συνάφεια*** των ανακτημένων εγγράφων
  - Δεν υπάρχει τυπικός ορισμός της συνάφειας
- Στην ουσία η συνάφεια δεν είναι δυαδική αλλά συνεχής
- Ακόμα και αν ήταν δυαδική, η κρίση της μπορεί να μην είναι εύκολη
- Από την πλευρά του χρήστη η συνάφεια είναι:
  - **υποκειμενική**: διαφορετική από χρήστη σε χρήστη
  - **περιστασιακή** (situational): σχετίζεται με τις τρέχουσες ανάγκες του χρήστη
  - **γνωστική** (cognitive): εξαρτάται από την αντίληψη/συμπεριφορά του χρήστη
  - **δυναμική**: μεταβάλλεται με το χρόνο (δεν είναι αναλλοίωτη)

Αξιολόγηση μέσω Χειρονακτικά Μαρκαρισμένων Συλλογών

Αξιολόγηση βάσει Χειρονακτικά Μαρκαρισμένων Συλλογών (Human Labeled Corpora)

Τρόπος:

1) Επέλεξε ένα συγκεκριμένο σύνολο εγγράφων C (κατά προτίμηση του ίδιου γνωστικού πεδίου).
2) Διατύπωσε ένα σύνολο επερωτήσεων για αυτά Q
3) Βρες έναν ή περισσότερους ειδικούς (experts) του γνωστικού πεδίου, και βάλε τους να μαρκάρουν τα συναφή έγγραφα για κάθε ερώτηση

   Συνήθως, οι κρίσεις τους είναι (Συναφές, Μη-Συναφές)

   Αρα το αποτέλεσμα της διαδικασίας αυτής είναι ένα σύνολο από πλειάδες της μορφής: (c,q,Relevant) ή (c,q,Irrelevant), όπου $c \in C$, $q \in Q$.
4) Χρησιμοποίησε αυτή τη συλλογή για την αξιολόγηση της αποτελεσματικότητας ενός ΣΑΠ.

   Βάζουμε το ΣΑΠ να ευρετηριάσει τη συλλογή C, κατόπιν του στέλνουμε επερωτήσεις από το Q και αξιολογούμε τις αποκρίσεις του βάσει των κρίσεων που έχουν κάνει ήδη οι ειδικοί.

Δυσκολίες:

Η παραπάνω μέθοδος απαιτεί μεγάλη ανθρώπινη προσπάθεια για μεγάλες συλλογές εγγράφων/επερωτήσεων.

Αξιολόγηση βάσει Χειρονακτικά Μαρκαρισμένων Συλλογών (Human Labeled Corpora)

4) Χρησιμοποίησε αυτή τη συλλογή για την αξιολόγηση της αποτελεσματικότητας ενός ΣΑΠ.

   Βάζουμε το ΣΑΠ να ευρετηριάσει τη συλλογή C, κατόπιν του στέλνουμε επερωτήσεις από το Q και αξιολογούμε τις αποκρίσεις του βάσει των κρίσεων που έχουν κάνει ήδη οι ειδικοί.

Πως;

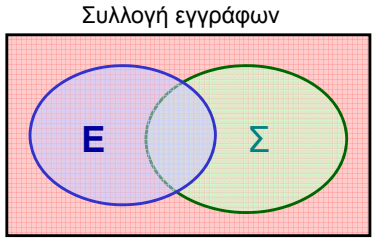
Μέτρα αξιολόγησης αποτελεσματικότητας

Μέτρα αξιολόγησης αποτελεσματικότητας: Ακρίβεια (Precision) και Ανάκληση(Recall)

- **Ακρίβεια (Precision)**:
  Διαισθητικά: Η ικανότητα ανάκτησης μόνο συναφών εγγράφων
- **Ανάκληση (Recall):**
  Διαισθητικά: Η ικανότητα εύρεσης όλων των συναφών εγγράφων της συλλογής

Ακρίβεια (Precision) και Ανάκληση(Recall)

Έστω ένα ερώτημα q



Σ: Συναφή (με το ερώτημα q)
(μας τα έχουν δώσει οι ειδικοί)

E: Ευρεθέντα (από το ΣΑΠ)

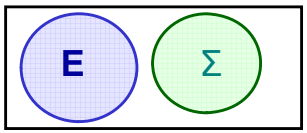
$$\text{Ακρίβεια} = \frac{|E \cap \Sigma|}{|E|}$$

P(recision)

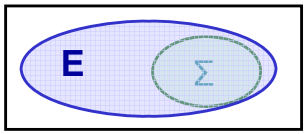
$$\text{Ανάκληση} = \frac{|E \cap \Sigma|}{|\Sigma|}$$

R(ecall)

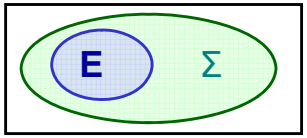
Περιπτώσεις



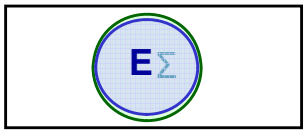
P=0, R=0 (χειρότερη περίπτωση)



P=low, R=1 (η επίτευξη R=1 είναι ευκολότατη)

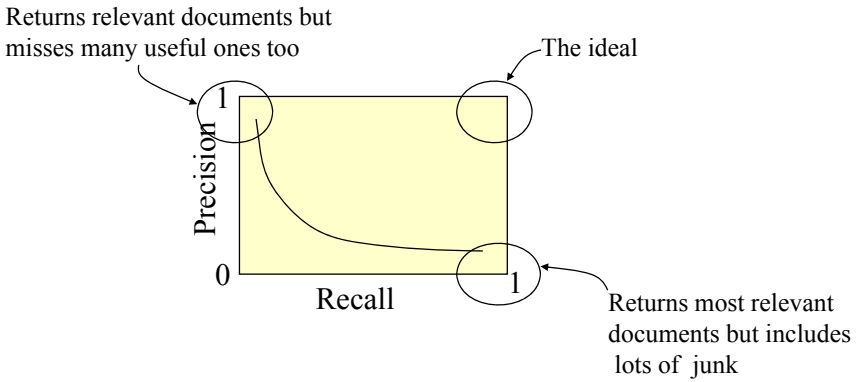


P=1, R:low



P=1, R:1 (ιδανική περίπτωση)

Αντιπαραβολή (trade-off) μεταξύ βαθμού ανάκλησης και βαθμού ακρίβειας



Ο Προσδιορισμός της Ανάκλησης είναι καμιά φορά δύσκολος (δυσκολότερος της Ακρίβειας)

Ο συνολικός αριθμός των εγγράφων που είναι συναφή με μια επερώτηση μπορεί να είναι άγνωστος

- Π.χ. Αυτό συμβαίνει με τον Ιστό

Τρόποι Αντιμετώπισης αυτού του Προβλήματος

- Δειγματοληψία (sampling)
  - Sample across the database and perform relevance judgment only on these items.
- Pooling
  - Apply different retrieval algorithms to the same database for the same query. Then the aggregate of relevant items is taken as the total relevant set.

[Τρόπους συνάθροισης διατάξεων (rank aggregation) θα δούμε στο μάθημα περί μετα-μηχανών αναζήτησης]

Θα μπορούσαμε με έναν μόνο αριθμό να χαρακτηρίσουμε την αποτελεσματικότητα ενός συστήματος;

F-Measure

Μέτρα αξιολόγησης αποτελεσματικότητας:

F-Measure

- Μέτρο που λαμβάνει υπόψη την Ακρίβεια **και** την Ανάκληση
- Είναι το αρμονικό μέσο (harmonic mean) της ανάκλησης και ακρίβειας:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Ερώτηση: Γιατί αρμονικό μέσο και όχι αριθμητικό;
- Απάντηση: Για να πάρουμε υψηλή τιμή αρμονικού μέσου χρειαζόμαστε υψηλό P **και** υψηλό R.

E-Measure

Μέτρα αξιολόγησης αποτελεσματικότητας:

E Measure (παραμετρικό F Measure)

- Παραλλαγή του F measure που μας επιτρέπει να δώσουμε περισσότερη έμφαση (βάρος) στην ακρίβεια:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

- Η τιμή του β ρυθμίζει το trade-off:
  - $\beta = 1$: Equally weight precision and recall (E-measure = F-measure).
  - $\beta > 1$: Weights precision more.
  - $\beta < 1$: Weights recall more.

Fallout

Μέτρα αξιολόγησης αποτελεσματικότητας:

Fallout Rate

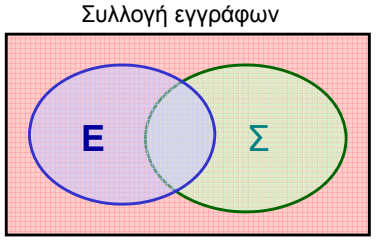
Προβλήματα της Ακρίβειας και Ανάκλησης:

- Ο αριθμός των μη-συναφών εγγράφων δεν λαμβάνεται υπόψη
- Η Ανάκληση δεν ορίζεται αν η συλλογή δεν έχει κανένα συναφές έγγραφο.
- Η Ακρίβεια δεν ορίζεται αν δεν ανακληθεί κανένα έγγραφο

$$Fallout = \frac{no.\ of\ \ nonrelevant\ items\ retrieved}{total\ no.\ of\ \ nonrelevant\ items\ in\ the\ collection}$$

Fallout

Έστω ένα ερώτημα q



Σ: Συναφή (με το ερώτημα q)
Σ^c: Μη-Συναφή

E: Ευρεθέντα (από το ΣΑΠ)

$$\text{Fallout} = \frac{|\ E \cap \Sigma^c\ |}{|\ \Sigma^c\ |}$$

Καμπύλες Ακρίβειας/Ανάκλησης (Precision/Recall Curves)

Μέτρα αξιολόγησης αποτελεσματικότητας:

Σημεία και Καμπύλες Ανάκλησης/Ακρίβειας

Κίνητρο:

Ο χρήστης δεν «καταναλώνει» όλη την απάντηση μονομιάς.
Αντίθετα αρχίζει από την κορυφή της λίστας των αποτελεσμάτων

Αυτό δεν λαμβάνεται υπόψη από τα μέτρα Recall και Precision

Θεωρείστε την περίπτωση που:

Answer(System1,q) = <N N N N N N N R R R>

Answer(System2,q) = <R R R N N N N N N N>

N: συμβολίζει ένα non-relevant έγγραφο
R: συμβολίζει ένα relevant έγγραφο

Η Ακρίβεια και η Ανάκληση των δυο συστημάτων είναι η ίδια! :(

Σημεία και Καμπύλες Ανάκλησης/Ακρίβειας (II)
(Recall/Precision Points and Curves)

Αντιμετώπιση Προβλήματος: Χρήση Recall/Precision Curves

Τρόπος υπολογισμού:

1) Για δοθείσα επερώτηση, παίρνουμε τη διατεταγμένη λίστα από το ΣΑΠ

Σημείωση: αν δεν πάρουμε όλη την απάντηση αλλά ένα τμήμα της, τότε το σύνολο των Ευρεθέντων αλλάζει, και άρα θα πάρουμε διαφορετικές recall/precision μετρήσεις

2) Σημειώνουμε κάθε έγγραφο της λίστας που είναι συναφές (βάσει της χειρονακτικά μαρκαρισμένης συλλογής)

3) Υπολογίζουμε ένα ζεύγος τιμών Ανάκλησης/Ακρίβειας **για κάθε θέση** της διατεταγμένης λίστας που περιέχει ένα συναφές έγγραφο.

Recall/Precision Points and Curves: Παράδειγμα

Έστω |Συναφή|=6

n	doc #	relevant	Recall	Precision
1	588	x	R=1/6=0.167;	P=1/1=1
2	589	x	R=2/6=0.333;	P=2/2=1
3	576			
4	590	x	R=3/6=0.5;	P=3/4=0.75
5	986			
6	592	x	R=4/6=0.667;	P=4/6=0.667
7	984			
8	988			
9	578			
10	985			
11	103			
12	591			
13	772	x	R=5/6=0.833;	P=5/13=0.38
14	990			

Missing one relevant document. Never reach 100% recall

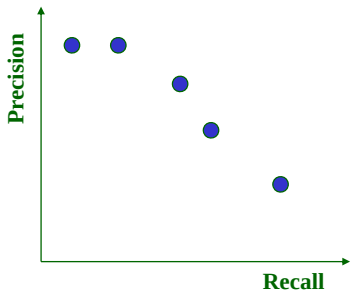
Plotting the Recall/Precision Points

R=1/6=0.167; P=1/1=1
R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

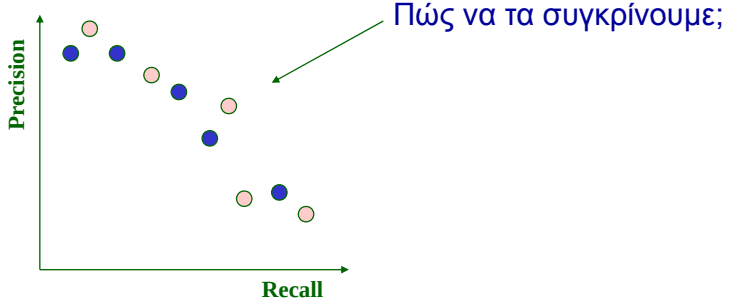
R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; P=5/13=0.38



Σύγκριση δύο συστημάτων

- Σύστημα 1
- Σύστημα 2



Interpolating a Recall/Precision Curve

Σκοπός: Δυνατότητα σύγκρισης διαφορετικών συστημάτων

Τρόπος:

Χρήση *κανονικοποιημένων επιπέδων ανάκλησης* (standard recall levels)

Παράδειγμα καθιερωμένων επιπέδων ανάκλησης (πλήθος επιπέδων: 11):

$r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

$r_0 = 0.0, r_1 = 0.1, \ldots, r_{10}=1.0$

Υπολογίζουμε μιας τιμή ακρίβειας για κάθε *standard recall level*:

Συγκεκριμένα, ως ακρίβεια στο *j* επίπεδο ανάκλησης ορίζουμε τη μέγιστη ακρίβεια που εμφανίζεται μεταξύ των βαθμών ανάκλησης *j* kai *j+1*

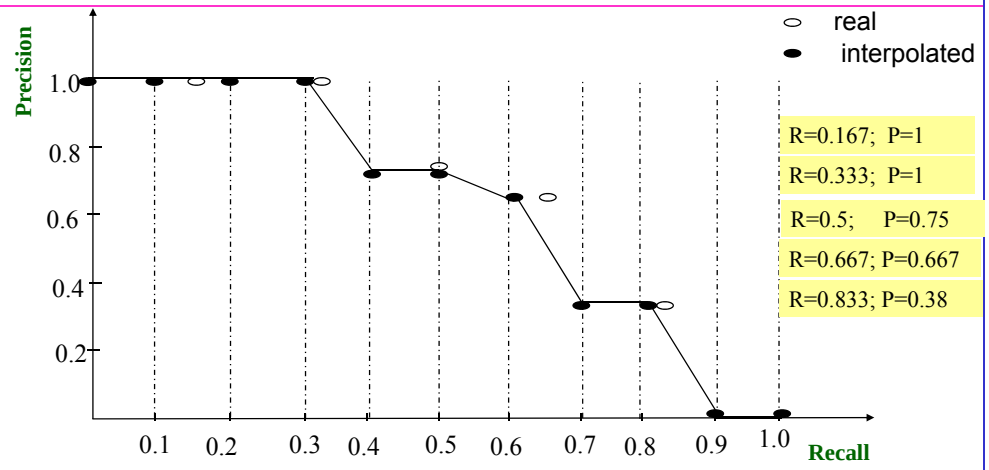
$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

Interpolating a Recall/Precision Curve (II)

- Αυτό στηρίζεται στην παρατήρηση ότι όσο η ανάκληση μεγαλώνει τόσο η ακρίβεια μειώνεται
- Για αυτό είναι λογικό να στοχεύουμε προς μια καμπύλη παρεμβολής (interpolation) που δίδει μια μονότονα φθίνουσα συνάρτηση

$$P(r_j) = \max_{r_j \le r \le r_{11}} P(r)$$

Interpolating a Recall/Precision Curve: Παράδειγμα



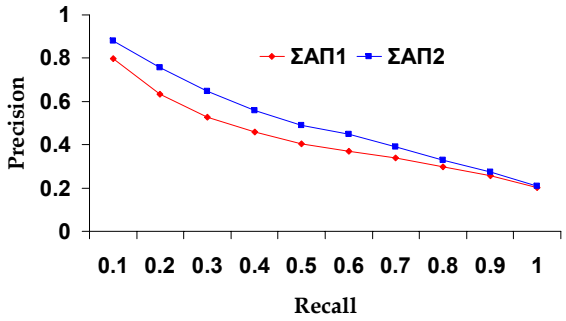
Σημείωση: Από τα 5 ζεύγη (P,R) που είχαμε πήγαμε στα 11

Τι κάνουμε αν έχουμε πολλά ερωτήματα στη συλλογή αξιολόγησης;

Μέση Καμπύλη Ανάκλησης/Ακρίβειας

- Προκύπτει αξιολογώντας την αποτελεσματικότητα του συστήματος με ένα μεγάλο πλήθος επερωτήσεων
- Υπολογίζουμε μέση ακρίβεια σε κάθε standard recall level για όλες τις επερωτήσεις
- Σχεδιάζουμε τη μέση precision/recall καμπύλη η οποία εκφράζει την επίδοση του συστήματος στη συλλογή

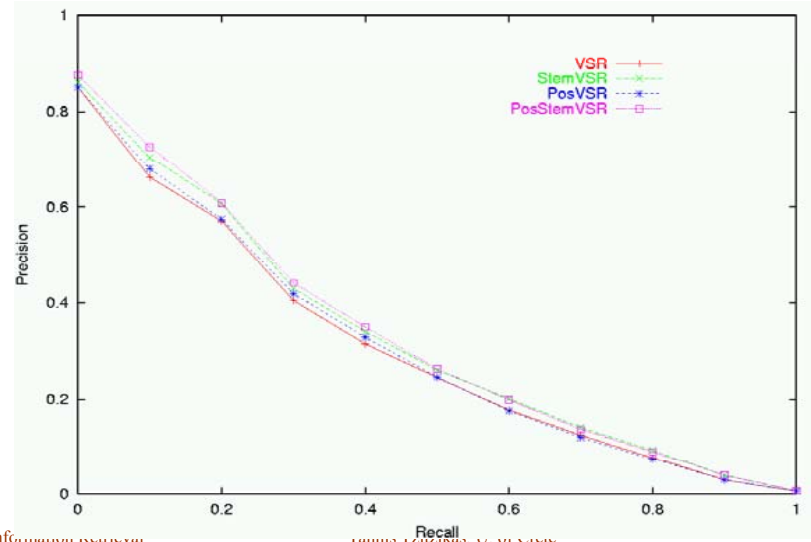
Σύγκριση Συστημάτων



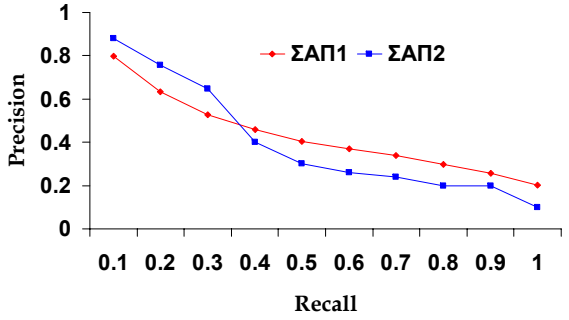
- Η καμπύλη που είναι πιο κοντά στην πάνω-δεξιά γωνία του γραφήματος υποδηλώνει καλύτερη επίδοση

Το ΣΑΠ2 είναι καλύτερο από το ΣΑΠ1

Σύγκριση Συστημάτων (II)



Σύγκριση Συστημάτων (III)



Το ΣΑΠ2 έχει καλύτερη ακρίβεια στα χαμηλά επίπεδα ανάκλησης
Το ΣΑΠ1 έχει καλύτερη ακρίβεια στα υψηλά επίπεδα ανάκλησης

Recall-Fallout graphs

- S. Robertson, ECIR'2007

R-Precision & Precision Histograms

Μέτρα αξιολόγησης αποτελεσματικότητας:

R- Precision

Ερώτημα: Μπορούμε να αξιολογήσουμε ένα σύστημα με ένα μόνο αριθμό;

(ο οποίος να λαμβάνει υπόψη τη σειρά των εγγράφων στην απάντηση;)

- **R-Precision**: Η ακρίβεια στην R θέση της διάταξης της απάντησης μιας επερώτησης που έχει R (στο πλήθος) συναφή έγγραφα

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

Μέτρα αξιολόγησης αποτελεσματικότητας:

R- Precision (II)

- Ερωτήματα:
  - Αν έχουμε πολλές επερωτήσεις αξιολόγησης, πώς υπολογίζεται αυτό το μέτρο;
  - Πως μπορούμε να συγκρίνουμε 2 συστήματα βάσει του R-Precision ;
- Απάντηση:
  - Χρησιμοποιώντας πολλές επερωτήσεις αξιολόγησης μπορούμε να σχεδιάσουμε το Ιστόγραμμα Ακρίβειας (Precision Histogram).

Σύγκριση Συστημάτων
Ιστογράμματα Ακρίβειας (Precision Histograms)

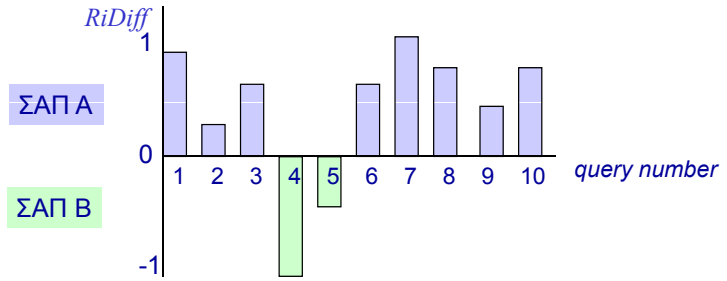
Έστω 2 συστήματα Α και Β και κ επερωτήσεις αξιολόγησης q1..qk

Τρόπος υπολογισμού του ιστογράμματος ακρίβειας:

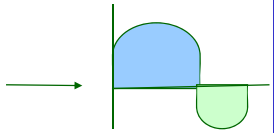
- Για κάθε i =1 έως k
  - Ri := το πλήθος των συναφών εγγράφων της επερώτησης qi
  - RiPA := Το Ri-precision του συστήματος Α για την qi
  - RiPB := Το Ri-precision του συστήματος Β για την qi
  - Ορίζουμε τη διαφορά ως εξής: RiDiff := RiPA - RiPB
- Κάνουμε την γραφική παράσταση των (i,RiDiff) (για i =1..k)

Σύγκριση Συστημάτων
Ιστογράμματα Ακρίβειας (II)

Παράδειγμα με 10 επερωτήσεις:



Μπορούμε κατόπιν να ταξινομήσουμε ως προς RiDiff ώστε να πάρουμε ένα πιο παραστατικό διάγραμμα



Expected Search Length

Expected Search Length
Αναμενόμενο μήκος αναζήτησης [Cooper 68]

- Ορισμός
  - Το **μέσο** πλήθος εγγράφων που πρέπει να εξεταστούν προκειμένου να ανακτήσουμε ένα **συγκεκριμένο** πλήθος συναφών εγγράφων.
- Παρατηρήσεις
  - Δεν είναι ένας αριθμός αλλά συνάρτηση του αριθμού των συναφών εγγράφων που επιθυμούμε
  - Μπορεί να παρασταθεί με πίνακα ή με γράφημα

Expected Search Length (II)

– Μπορεί να παρασταθεί με πίνακα ή με γράφημα

• Πίνακας

Rel Docs	Search Length
1	2.0
2	4.2
3	5.4
4	6.6
5	7.8

(στα πρώτα δυο στοιχεία της απάντησης υπάρχει ένα συναφές)

Στα 4.2 πρώτα έγγραφα υπάρχουν δύο συναφή.
Το 4.2. είναι είτε μέσος όρος (που προέκυψε κάνοντας πολλές μετρήσεις) ή/και λόγω εγγράφων που έλαβαν ίδιο βαθμό συνάφειας.

- Μπορούμε όμως να υπολογίσουμε και έναν «μέσο όρο»:
  – (2/1 + 4.2/2 + 5.4/3 + 6.6/4 + 7.8/5) / 5 =
  – (2 + 2.1 + 1.8 + 1.65 + 1.56)/5 = 9.11/5=1.82
  – Χονδρικά, ο χρήστης χρειάζεται να εξετάζει 82% παραπάνω έγγραφα από τα επιδιωκόμενα συναφή (π.χ. αν θέλει 20 θα χρειαστεί να εξετάσει τα πρώτα 1.82*20=36 έγγραφα)

Η έννοια του expected search length μας είναι επίσης χρήσιμη προκειμένου να κάνουμε ακριβείς μετρήσεις στην περίπτωση που οι απαντήσεις του συστήματος δεν είναι μια γραμμική ακολουθία εγγράφων, αλλά μια γραμμική ακολουθία συνόλων εγγράφων.

Παράδειγμα

- Answer(System1, q) = < d8, d2, {d3, d4}, d1>
  – Αυτό σημαίνει ότι τα d3 και d4 έλαβαν τον ίδιο βαθμό συνάφειας (άρα βρίσκονται και τα δύο στην 3η θέση της κατάταξης)
- Answer(System2, q) = < d1, {d2, d3}, d8 >

- Ερώτηση: *Αν ξέρουμε οτι η q έχει δύο συναφή έγγραφα, συγκεκριμένα τα d1 και d3, ποιά είναι η R-Precision του System1 και ποια του System2 ?*

User-Oriented Measures

Πιο υποκειμενικά μέτρα Συνάφειας

- *Novelty Ratio (ποσοστό ''καινοτομίας'')*:
  Το ποσοστό των ανακτημένων και συναφών εγγράφων (Ε ∩Σ) των οποίων την ύπαρξη ο χρήστης αγνοούσε (πριν την αναζήτηση).
  – Μετράει την ικανότητα εύρεσης **νέας** πληροφορίας σε ένα θέμα.

- *Coverage Ratio (ποσοστό κάλυψης)*:
  Το ποσοστό των ανακτημένων και συναφών εγγράφων (Ε ∩Σ) σε σχέση με το σύνολο των συναφών εγγράφων τα οποία είναι γνωστά στο χρήστη πριν την αναζήτηση.
  – Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000).

Άλλοι παράγοντες αξιολόγησης

- *Ανθρώπινη προσπάθεια(User effort)*:
  Work required from the user in formulating queries, conducting the search, and screening the output.
- *Χρόνος απόκρισης (Response time)*:
  Time interval between receipt of a user query and the presentation of system responses.
- *Μορφή παρουσίασης (Form of presentation)*:
  Influence of search output format on the user's ability to utilize the retrieved materials.
- *Κάλυψη συλλογής (Collection coverage)*:
  Extent to which any/all relevant items are included in the document corpus.

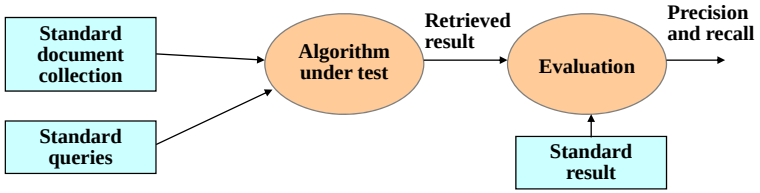
Δοκιμασίες Αποτελεσματικότητας-Συλλογές Αναφοράς (TREC)

Δοκιμασίες επιδόσεων (Benchmarking)

- Η *αναλυτική αξιολόγηση* επίδοσης είναι δύσκολη στα ΣΑΠ διότι πολλά χαρακτηριστικά (συνάφεια, κατανομή λέξεων, κλπ) δύσκολα προσδιορίζονται με μαθηματική ακρίβεια
- Η επιδόσεις συνήθως μετρώνται με ***Δοκιμασίες Επιδόσεων (benchmarking)***. Η αξιολόγηση της αποτελεσματικότητας αξιολογείται σε συγκεκριμένες συλλογές εγγράφων, επερωτήσεων και κρίσεις συνάφειας
- Τα αποτελέσματα είναι έγκυρα μόνο στο περιβάλλον που έγινε η αξιολόγηση.

Δοκιμασίες Επιδόσεων

- A benchmark collection contains:
  - A set of standard documents and queries/topics.
  - A list of relevant documents for each query.
- Standard collections for traditional IR:
  - **Smart collection**: ftp://ftp.cs.cornell.edu/pub/smart
  - **TREC**: http://trec.nist.gov/



Τα προβλήματα του Benchmarking

- Τα αποτελέσματα της αξιολόγησης είναι έγκυρα μόνο για τη συγκεκριμένη δοκιμασία αξιολόγησης
- Ο κατασκευή ενός benchmark είναι δύσκολη και χρονοβόρα
- Αφορούν κυρίως κείμενα στα ΑΓΓΛΙΚΑ

Early Test Collections

- Previous experiments were based on the SMART collection which is fairly small. (ftp://ftp.cs.cornell.edu/pub/smart)

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

- Different researchers used different test collections and evaluation techniques.

TREC
http://trec.nist.gov/

The TREC Benchmark

- TREC: Text REtrieval Conference (**http://trec.nist.gov/**)
  Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.
- Participants are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing.
- Participants submit the P/R values for the final document and query corpus and present their results at the conference.

Οι στόχοι του TREC

- Provide a common ground for comparing different IR techniques.
  - Same set of documents and queries, and same evaluation method.
- Sharing of resources and experiences in developing the benchmark.
  - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

Τα πλεονεκτήματα του TREC

- Large scale (compared to a few MB in the SMART Collection).
- Relevance judgments provided.
- Under continuous development with support from the U.S. Government.
- Wide participation:
  - TREC 1: 28 papers 360 pages.
  - TREC 4: 37 papers 560 pages.
  - TREC 7: 61 papers 600 pages.
  - TREC 8: 74 papers.

TREC Tasks

- Ad hoc: New questions are being asked on a static set of data.
- Routing: Same questions are being asked, but new information is being searched. (news clipping, library profiling).
- New tasks added after TREC 5 - Interactive, multilingual, natural language, multiple database merging, filtering, very large corpus (20 GB, 7.5 million documents), question answering.

TREC Tracks

- Cross-Language Track
  - the ability of retrieval systems to find documents that pertain to a topic **regardless of the language** in which the document is written.
  - Also studied in CLEF (Cross-Language Evaluation Forum), and the NTCIR workshops.
- Filtering Track
  - user's information need is **stable** (and some relevant documents are known) but there is a **stream of new documents**. For each document, the system must make a binary decision as to whether the document should be retrieved (as opposed to forming a ranked list).
- Genomics Track
  - study retrieval tasks in a specific domain, where the domain of interest is **genomics data** (broadly construed to include not just gene sequences but also supporting documentation such as research papers, lab reports, etc.)

TREC Tracks (II)

- HARD Track
  - achieve **High Accuracy** Retrieval from Documents by leveraging additional information about the searcher and/or the search **context**, through techniques such as ***passage retrieval*** and using very targeted interaction with the searcher.
- Interactive Track
  - A track studying user **interaction** with text retrieval systems. Participating groups develop a consensus experimental protocol and carry out studies with real users using a common collection and set of user queries.
- Novelty Track
  - ability to locate **new** (i.e., non-redundant) information.
- Question Answering Track
  - a step closer to information retrieval rather than document retrieval. Focus on definition, list, and factoid questions.

TREC Tracks (III)

- Terabyte Track
  - investigate whether/how the IR community can scale traditional IR test-collection-based evaluation to significantly **larger document collections** than those currently used in TREC. The retrieval task will be an ad hoc task using a static collection of approximately **1 terabyte of spidered web pages** (probably from the .GOV domain).
- Video Track
  - research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, the track became an independent evaluation (TRECVID).
- Web Track
  - A track featuring search tasks on a document set that is a snapshot of the World Wide Web.

Χαρακτηριστικά της συλλογής TREC

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).
- Test documents consist of:

WSJ	Wall Street Journal articles (1986-1992)	550 M
AP	Associate Press Newswire (1989)	514 M
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 M
FR	Federal Register	469 M
DOE	Abstracts from Department of Energy reports	190 M

More Details on Document Collections

- Volume 1 (Mar 1994) - Wall Street Journal (1987, 1988, 1989), Federal Register (1989), Associated Press (1989), Department of Energy abstracts, and Information from the Computer Select disks (1989, 1990)
- Volume 2 (Mar 1994) - Wall Street Journal (1990, 1991, 1992), the Federal Register (1988), Associated Press (1988) and Information from the Computer Select disks (1989, 1990)
- Volume 3 (Mar 1994) - San Jose Mercury News (1991), the Associated Press (1990), U.S. Patents (1983-1991), and Information from the Computer Select disks (1991, 1992)
- Volume 4 (May 1996) - Financial Times Limited (1991, 1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), and the Federal Register (1994).
- Volume 5 (Apr 1997) - Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990).

TREC Disk 4,5

TREC Disk 4	Congressional Record of the 103rd Congress approx. 30,000 documents approx. 235 MB
	Federal Register (1994) approx. 55,000 documents approx. 395 MB
	Financial Times (1992-1994) approx. 210,000 documents approx. 565 MB
TREC Disk 5	Data provided from the Foreign Broadcast Information Service approx. 130,000 documents approx. 470 MB
	Los Angeles Times (randomly selected articles from 1989 & 1990) approx. 130,000 document approx. 475 MB

Δείγμα Εγγράφου (σε SGML)

```
<DOC>
<DOCNO> WSJ870324-0001 </DOCNO>
<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>
<DD> 03/24/87</DD>
<SO> WALL STREET JOURNAL (J) </SO>
<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM) MARKETING, ADVERTISING (MKT)
    TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
   John Blair &amp;amp; Co. is close to an agreement to sell its TV station advertising representation operation
    and program production unit to an investor group led by James  H. Rosenfield, a former CBS Inc.
    executive, industry sources said. Industry sources put the value of the proposed acquisition at more
    than $100 million. ...
</TEXT>
</DOC>
```


Δείγμα επερώτησης (with SGML)

```
<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language processing technology which is
    being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution developing or marketing a
    natural language processing technology, identify the technology, and identify one of more features
    of the company's product.
<con> Concept(s):  1. natural language processing ;2. translation, language, dictionary
<fac> Factor(s):
<nat> Nationality: U.S.</nat>
</fac>
<def> Definitions(s):
</top>
```


TREC Properties

- Both documents and queries contain many different kinds of information (fields).
- Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system.
  - A system may be very good at querying and ranking, but if it generates poor queries from the topic, its final P/R would be poor.

Two more TREC Document Examples

ZIFF Communications Company	San Jose Mercury News
<DOC> <DOCNO> ZF109-706-077 </DOCNO> <DOCID>09 706 077.&O;</DOCID> <JOURNAL>Business Week Dec 31 1990 n3194 p93(12).&M; </JOURNAL> <TITLE>Fujitsu means business for America. (Special Advertising Section by Fujitsu Ltd.) (includes related articles on the company's business relationships with Pepsi-Cola, Convex Computer, Greenville EMS, and Sequent Computer Systems)&M; </TITLE> <TEXT> <ABSTRACT>In establishing itself as a major manufacturer in the computer hardware market, Fujitsu Ltd boasts a long list of corporate customers.&P; The company's client base includes: MCI Telecommunications Corp., Page Composition, Johns Hopkins Hospital, Tiara Computer Systems Inc., Pepsi-Cola, Convex Computer, Greenville EMS, and Sequent Computer Systems Inc. The company stresses its good customer relations and product development aspects, as well as its telecommunications products.&O; </ABSTRACT> </TEXT> <DESCRIPT> Company: Fujitsu Ltd. (Marketing).&O; Topic: Marketing Strategy Customer Relations photograph.&M; </DESCRIPT> </DOC>	<DOC> <DOCNO> SJMN91-06364024 </DOCNO> <ACCESS> 06364024 </ACCESS> <CAPTION> Photo; PHOTO: Associated Press; ANOTHER TURNOVER – Kansas City's Leonard Griffin (98) closes in on Raiders quarterback Todd Marinovich, who fumbled on the play. Marinovich also threw four interceptions. </CAPTION> <DESCRIPT> PROFESSIONAL; FOOTBALL; PLAYOFF; GAME; RESULT; BRIEF </DESCRIPT> <LEADPARA> Too much excitement on top of too much cold medication may have caused the rapid heartbeat that forced Kansas City linebacker Derrick Thomas out of the ... reliable place-kicker, kicked an 18-yard field goal at 10:26 of the fourth quarter, but he missed two field goals in the first half, from 33 and 47 yards. ... </TEXT> <FEATURE> PHOTO </FEATURE> <STATE> CA </STATE> <WORD.CT> 539 </WORD.CT> <DATELINE> Sunday, December 29, 1991 06364024.SJ1 </DATELINE> <COPYRGHT> Copyright 1991, San Jose Mercury News </COPYRGHT> <LANGUAGE> ENG </LANGUAGE> </DOC>

Another Example of TREC Topic/Query

<top>
<head> Tipster Topic Description
<num> Number: 101
<dom> Domain: Science and Technology
<title> Topic: Design of the "Star Wars" Anti-missile Defense System
<desc> Description:
Document will provide information on the proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system.
<narr> Narrative:
proposed configuration, components, and technology of the U.S.'s "star wars" anti-missile defense system. The design and technology to be used in the anti-missile defense system advocated by the Reagan administration, the Strategic Defense Initiative (SDI), also known as "star wars." Changes of constituent technologies, are also relevant documents.
<con> Concept(s):
1. Strategic Defense Initiative, SDI, star wars, peace shield
2. kinetic energy weapon, kinetic kill, directed energy weapon, laser, particle beam, ERIS (exoatmospheric reentry-vehicle interceptor system), phased-array radar, microwave
3. anti-satellite (ASAT) weapon, spaced-based technology, strategic defense technologies
<fac> Factor(s):
<nat> Nationality: U.S.
</nat>
<def> Definition(s):
</top>

Αποτελέσματα Αξιολόγησης ενός ΣΑΠ βάσει του TREC

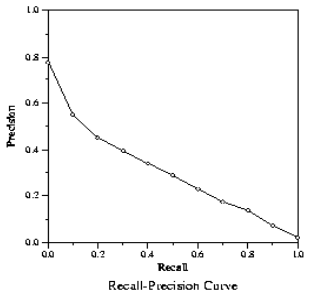
- Summary table statistics: Number of topics, number of documents retrieved, number of relevant documents.
- Recall-precision average: Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- Document level average: Average precision when 5, 10, .., 100, … 1000 documents are retrieved.
- Average precision histogram: Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

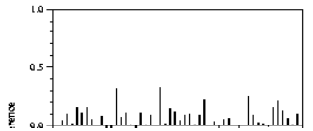
1 hoc results — Fujitsu Laboratories. Ltd.

Summary Statistics	
Run Number	Flab8atd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

Recall Level Precision Averages	
Recall	Precision
0.00	0.7796
0.10	0.5490
0.20	0.4517
0.30	0.3954
0.40	0.3397
0.50	0.2863
0.60	0.2291
0.70	0.1745
0.80	0.1381
0.90	0.0720
1.00	0.0224
Average precision over all relevant docs	
non interpolated	0.2930

Document Level Averages	
	Precision
At 5 docs	0.5480
At 10 docs	0.4880
At 15 docs	0.4587
At 20 docs	0.4200
At 30 docs	0.3887
At 100 docs	0.2490
At 200 docs	0.1777
At 500 docs	0.1011
At 1000 docs	0.0598
R Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3203





TREC 2008
http://trec.nist.gov/call08.html

- **Blog Track**
  - The purpose of the blog track is to explore information seeking behavior in the blogosphere.
- **Enterprise Track**
  - The purpose of the enterprise track is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task.
- **Legal Track**
  - The goal of the legal track is to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections.
- **Million Query Track**
  - The goal of the "million query" track is to test the hypothesis that a test collection built from very many very incompletely judged topics is a better tool than a collection built using traditional TREC pooling.
- **Relevance Feedback Track**
  - The goal of the relevance feedback track is to provide a framework for exploring the effects of different factors on the success of relevance feedback.