Introduction to

Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 8α: Αξιολόγηση στην Ανάκτηση Πληροφοριών.

Τι θα δούμε σήμερα;

- Πως ξέρουμε αν τα αποτελέσματα είναι καλά
  - Αξιολόγηση μηχανών αναζήτησης: μεθοδολογία και μέτρα

Αξιολόγηση συστήματος

Αποδοτικότητα (Performance)

- Πόσο γρήγορη είναι η κατασκευή του ευρετηρίου;
  - αριθμός εγγράφων την ώρα
  - Μέγεθος ευρετηρίου
- Πόσο γρήγορη είναι η αναζήτηση;
  - π.χ., latency ως συνάρτηση των ερωτημάτων ανά δευτερόλεπτο ή του μεγέθους του ευρετηρίου

Εκφραστικότητα της γλώσσας ερωτημάτων

επιτρέπει τη διατύπωση περίπλοκων αναγκών πληροφόρησης;

Ποιο είναι το κόστος ανά ερώτημα;

- Π.χ., σε δολάρια

Μέτρα για μηχανές αναζήτησης

- Όλα αυτά τα κριτήρια είναι μετρήσιμα (measurable): μπορούμε να ποσοτικοποιήσουμε την ταχύτητα/μέγεθος/χρήματα και να κάνουμε την εκφραστικότητα συγκεκριμένη
- Ωστόσο μια βασική μέτρηση για μια μηχανή αναζήτησης είναι η **ικανοποίηση των χρηστών (user happiness)**
- *Τι κάνει ένα χρήστη χαρούμενο;* Οι παράγοντες περιλαμβάνουν:
  - Ταχύτητα απόκρισης (Speed of response)
  - Μέγεθος/κάλυψη ευρετηρίου
  - Εύχρηστη διεπαφή (Uncluttered UI)
  - Χωρίς κόστος (free)
- *Κανένα από αυτά δεν αρκεί*: εξαιρετικά γρήγορες αλλά άχρηστες απαντήσεις δεν ικανοποιούν ένα χρήστη **(συνάφεια-relevance)**
- *Πως μπορούμε να το μετρήσουμε;*

Ποιοι είναι οι χρήστες

Ποιος είναι ο χρήστης που προσπαθούμε να ικανοποιήσουμε;

Εξαρτάται από την εφαρμογή

- *Μηχανές αναζήτησης στο Web: searcher.* Επιτυχία: Ο χρήστης βρίσκει αυτό που ψάχνει. Μέτρο: ρυθμός επιστροφής στη συγκεκριμένη μηχανή αναζήτησης
- *Μηχανές αναζήτησης στο Web: διαφημιστής*. Επιτυχία: Searcher «κλικάρει» στη διαφήμιση. Μέτρο: clickthrough rate
- *Ecommerce: Αγοραστής.* Επιτυχία: Ο αγοραστής αγοράζει κάτι. Μέτρο: χρόνος για την αγορά, ποσοστό των searchers που γίνονται αγοραστές
- *Ecommerce: Πωλητής.* Επιτυχία: Ο πωλητής πουλάει κάτι. Μέτρο: κέρδος ανά πώληση.
- *Επιχείρηση: CEO.* Επιτυχία: Οι εργαζόμενοι γίνονται πιο αποδοτικοί (λόγω αποτελεσματικής αναζήτησης). Μέτρο: κέρδος της εταιρείας.

5

Συνήθης ορισμός: Συνάφεια

Η ικανοποίηση του χρήστη συνήθως εξισώνεται με τη **συνάφεια (relevance)** των αποτελεσμάτων της αναζήτησης με το ερώτημα

Μα πως θα μετρήσουμε τη συνάφεια;

Η καθιερωμένη μεθοδολογία στην Ανάκτηση Πληροφορίας αποτελείται από τρία στοιχεία:

1. Μία πρότυπη συλλογή εγγράφων (benchmark document collection)
2. Μια πρότυπη ομάδα ερωτημάτων (benchmark suite of queries)
3. Ένα σύνολο αποτίμησης της συνάφειας κάθε ζεύγους ερωτήματος-εγγράφου (συνήθως δυαδικές: συναφής-μη συναφής) - gold standard/ground truth

6

Συνάφεια και Ανάγκη Πληροφόρησης

- Συνάφεια ως προς τι;

Συνάφεια ως προς την ερώτηση

- Ανάγκη Πληροφόρησης (Information need *i*) : «Ψάχνω για πληροφορία σχετικά με το αν το κόκκινο κρασί είναι πιο αποτελεσματικό από το λευκό κρασί για τη μείωση του ρίσκου για καρδιακή προσβολή»

Μεταφράζεται σε ερώτημα:

- Ερώτημα *q*: [red wine white wine heart attack]

Έστω το έγγραφο *d'*: At heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.

  - *d'* άριστο ταίριασμα στο ερώτημα *q*
  - *d'* δεν είναι συναφές με την ανάγκη πληροφόρησης *i*

Συνάφεια και Ανάγκη Πληροφόρησης

- Η ικανοποίηση του χρήστη μπορεί να μετρηθεί μόνο με τη συνάφεια ως *προς την ανάγκη πληροφόρησης* και όχι ως προς το ερώτημα

- Το ακριβές είναι *συνάφεια έγγραφου-ανάγκης πληροφόρησης* αν και χρησιμοποιούμε συνάφεια έγγραφου-ερωτήματος.

Ακρίβεια και Ανάκληση

- **Precision (P) – Ακρίβεια** είναι το ποσοστό των ανακτημένων εγγράφων που είναι συναφή

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

- **Recall (R) – Ανάκληση** είναι το ποσοστό των συναφών εγγράφων που ανακτώνται

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

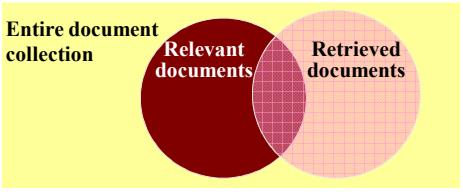
Ακρίβεια και Ανάκληση

Πίνακας Ενδεχόμενων (Incidence Matrix)

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$P = TP / (TP + FP)$

$R = TP / (TP + FN)$



Ακρίβεια vs Ανάκληση

- Η ανάκληση μπορεί να αυξηθεί με το να επιστρέψουμε *περισσότερα έγγραφα*
- Η ανάκτηση είναι μια μη-φθίνουσα συνάρτηση των εγγράφων που ανακτώνται.
  - Ένα σύστημα που επιστρέφει όλα τα έγγραφα έχει ποσοστό ανάκτησης 100%!
- Το αντίστροφο ισχύει επίσης (συνήθως): *Είναι εύκολο να πετύχεις μεγάλη ακρίβεια με πολύ μικρή ανάκληση*
  - Έστω ότι το έγγραφο με το μεγαλύτερο βαθμό είναι συναφές. Πως μπορούμε να μεγιστοποιήσουμε την ακρίβεια;
- Σε ένα καλό σύστημα η ακρίβεια ελαττώνεται όσο περισσότερα έγγραφα ανακτούμε ή με την αύξηση της ανάκλησης

Αρμονικό Μέσο

Πως θα συνδυάσουμε το P και R;

Π.χ., το *αριθμητικό μέσο* (arithmetic mean)

❖ Το απλό αριθμητικό μέσο μιας μηχανής αναζήτησης που επιστρέφει τα πάντα είναι 50%, που είναι πολύ υψηλό

Γεωμετρικό μέσο (geometric mean) γινόμενο

Θα θέλαμε με κάποιο τρόπο να τιμωρήσουμε *την πολύ κακή συμπεριφορά* σε οποιοδήποτε από τα δύο μέτρα.

Αυτό επιτυγχάνεται παίρνοντας το *ελάχιστο*

Αλλά το ελάχιστο είναι λιγότερο ομαλό (smooth) και είναι δύσκολο να σταθμιστεί

Το F (αρμονικό μέσο) είναι ένα είδος ομαλού ελάχιστου

Ένα συνδυαστικό μέτρο *F*

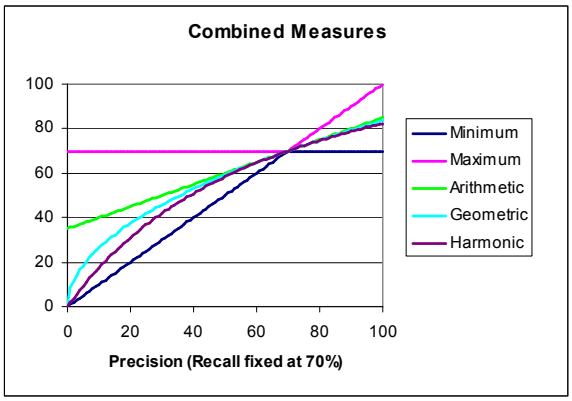
Συνήθως ισορροπημένο (balanced) F_1

- Αρμονικό μέσο των P και R

$$F1 = 1/ [(1/2)1/P + (1/2)1/R] = 2PR/P+R$$

✓ Πιο κοντά στη μικρότερη από δύο τιμές

Αρμονικό Μέσο



Τιμές στο 0-1, αλλά συνήθως σε ποσοστά

Ένα συνδυαστικό μέτρο *F*

Το μέτρο ***F*** επιτρέπει μια αντιστάθμιση (trade off) της ακρίβεια και της ανάκλησης.

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

όπου $\beta^2 = \frac{1-\alpha}{\alpha}$ α ϵ [0, 1] and thus b^2 ϵ [0,∞]

Συνήθως ισορροπημένο (balanced) F_1 με α = 0.5 και b = 1

- Αυτό είναι το αρμονικό μέσο των P και R $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$

- Για ποια περιοχή τιμών του β η ανάκληση σταθμίζεται περισσότερο από την ακρίβεια;

15

Παράδειγμα

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

P = 20/(20 + 40) = 1/3

R = 20/(20 + 60) = 1/4

$$F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

16

Ορθότητα (Accuracy)

- Γιατί να χρησιμοποιούμε περίπλοκα μέτρα όπως ακρίβεια, ανάκληση και F?
- Γιατί όχι κάτι πιο απλό;

Ορθότητα (Accuracy) είναι το ποσοστό των αποφάσεων (συναφή/μη συναφή) που είναι σωστές.

Με βάση τον πίνακα ενδεχομένων:

accuracy = (TP + TN)/(TP + FP + FN + TN).

Γιατί αυτό δεν είναι χρήσιμο στην ΑΠ;

Ορθότητα

Παράδειγμα

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

Ορθότητα

Η μηχανή αναζήτησης snoogle επιστρέφει πάντα 0 αποτελέσματα ("0 matching results found"), ανεξάρτητα από το ερώτημα. Τι μας λέει όπως το accuracy

Ορθότητα

- Απλό κόλπο για τη μεγιστοποίηση της ορθότητας στην ΑΠ: πες πάντα όχι και μην επιστρέφεις κανένα έγγραφο
- Αυτό έχει ως αποτέλεσμα 99.99% ορθότητα στα περισσότερα ερωτήματα

Searchers στο web (και γενικά στην ΑΠ) θέλουν να βρουν κάτι και έχουν κάποια ανεκτικότητα στα «σκουπίδια»

Καλύτερα να επιστρέφεις κάποια κακά hits αρκεί να επιστέφεις κάτι

→ Για την αποτίμηση, χρησιμοποιούμε την ακρίβεια, ανάκληση και F

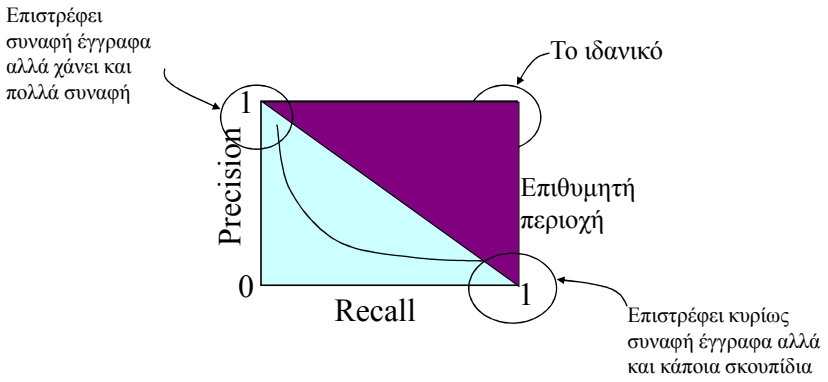
Δυσκολίες στη χρήση P/R

- Πρέπει να υπολογιστούν μέσοι όροι για μεγάλες ομάδες συλλογών εγγράφων/ερωτημάτων
- Χρειάζονται εκτιμήσεις συνάφειας από ανθρώπους
  - Οι χρήστες γενικά δεν είναι αξιόπιστοι αξιολογητές
- Οι εκτιμήσεις πρέπει να είναι δυαδικές
  - Ενδιάμεσες αξιολογήσεις;
- Εξαρτώνται από τη συλλογή/συγγραφή
  - Τα αποτελέσματα μπορεί να διαφέρουν από το ένα πεδίο στο άλλο

Μη γνωστή ανάκληση

- Ο συνολικός αριθμός των συναφών εγγράφων δεν είναι πάντα γνωστός:
  - Δειγματοληψία – πάρε έγγραφα από τη συλλογή και αξιολόγησε τη συνάφεια τους.
  - Εφάρμοσε διαφορετικούς αλγόριθμους για την ίδια συλλογή και την ίδια ερώτηση και χρησιμοποίησε το άθροισμα των συναφών εγγράφων

Ακρίβεια και Ανάκληση



Τι γίνεται όταν υπάρχει διάταξη των αποτελεσμάτων;

Αξιολόγηση Καταταγμένης Ανάκτησης

Ο χρήστης δε βλέπει όλη την απάντηση, αντίθετα αρχίζει από την κορυφή της λίστας των αποτελεσμάτων

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>

✓ *Η ακρίβεια, ανάκληση και το F είναι μέτρα για μη καταταγμένα (unranked) σύνολα .*

Πως μπορούμε να τα τροποποιήσουμε τα μέτρα για λίστες με διάταξη;

Καμπύλη Ακρίβειας/Ανάκλησης

Πως μπορούμε να τα τροποποιήσουμε τα μέτρα για λίστες με διάταξη;

- Απλώς υπολόγισε το μέτρο συνόλου για κάθε πρόθεμα: το κορυφαίο 1, κορυφαία 2, κορυφαία 3, κορυφαία 4 κλπ αποτελέσματα

Με αυτόν τον τρόπο παίρνουμε μια καμπύλη ακρίβειας-ανάκλησης (precision-recall curve).

Παράδειγμα I

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	

total # of relevant docs = 6
Check each new recall point:

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

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

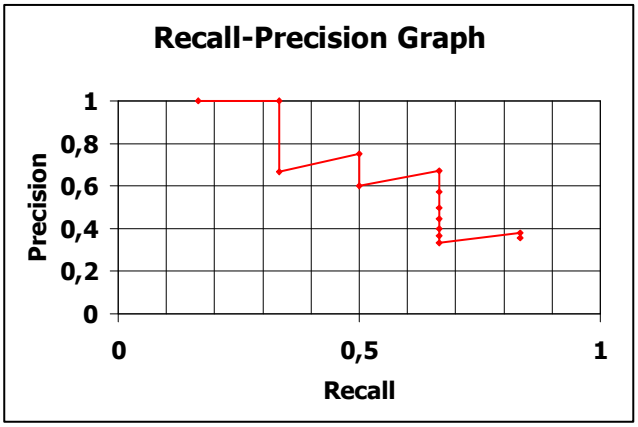
R=3/6=0.5; P=3/4=0.75

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

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

Missing one relevant document. Never reach 100% recall

Παράδειγμα I (συνέχεια)

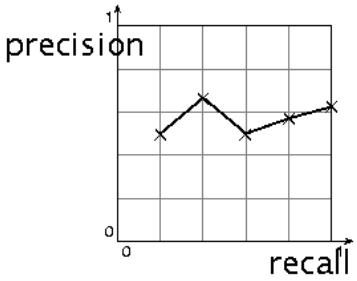


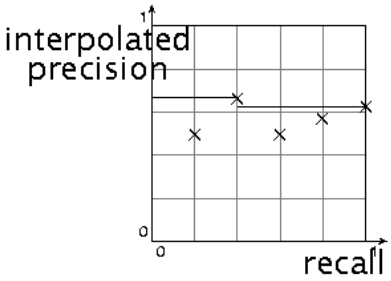
Πριονωτή – το precision ελαττώνεται για το ίδιο recall μέχρι να βρεθεί το επόμενο συναφές έγγραφο

Ακρίβεια εκ παρεμβολής (Interpolated precision)

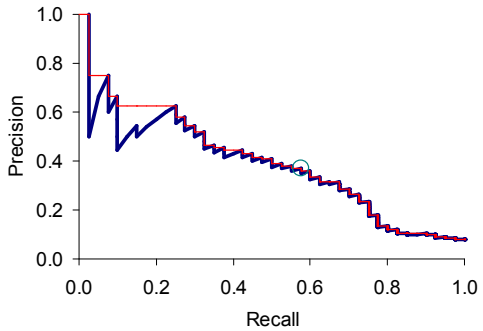
- Αν η ακρίβεια αλλάζει τοπικά με την αύξηση της ανάκλησης, το λαμβάνουμε υπ' όψιν – *ο χρήστης θέλει να δει και άλλα έγγραφα αν αυξάνεται και η ακρίβεια και η ανάκληση*
- Παίρνουμε τη μέγιστη τιμή της ακρίβειας στα δεξιά της τιμής

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





Καμπύλη Ακρίβειας/Ανάκλησης



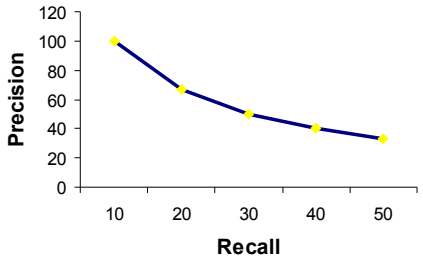
Κάθε σημείο αντιστοιχεί σε ένα αποτέλεσμα για τα κορυφαία *k* έγγραφα (*k* = 1, 2, 3, 4, . . .).

Παρεμβολή (με κόκκινο): μέγιστο των μελλοντικών σημείων

30

Παράδειγμα II

$$\text{Relevant} = \left\{ \begin{matrix} d_3, d_5, d_9, d_{25}, d_{39}, \\ d_{44}, d_{56}, d_{71}, d_{89}, d_{123} \end{matrix} \right\}$$



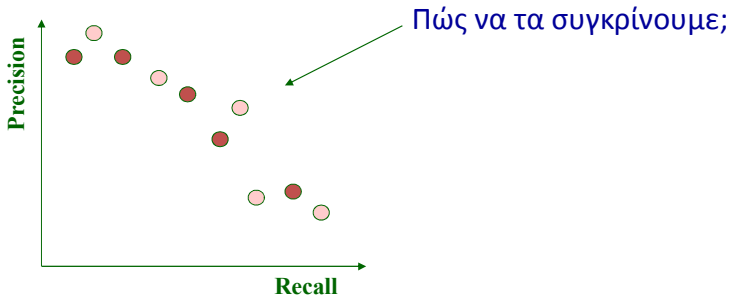
Rank	Doc	Rel	R_{ecall}	$P_{recision}$
0			0 %	0 %
1	d_{123}	✔	10 %	100 %
2	d_{84}		10 %	50 %
3	d_{56}	✔	20 %	67 %
4	d_{6}		20 %	50 %
5	d_{84}		20 %	40 %
6	d_{9}	✔	30 %	50 %
7	d_{511}		30 %	43 %
8	d_{129}		30 %	38 %
9	d_{187}		30 %	33 %
10	d_{25}	✔	40 %	40 %
11	d_{38}		40 %	36 %
12	d_{48}		40 %	33 %
13	d_{250}		40 %	31 %
14	d_{113}		40 %	29 %
15	d_{3}	✔	50 %	33 %

Μέσοι όροι από πολλά ερωτήματα

- Το γράφημα για ένα ερώτημα δεν αρκεί
- Χρειαζόμαστε τη μέση απόδοση σε αρκετά ερωτήματα.
- Αλλά:
  - Οι υπολογισμοί ακρίβειας-ανάκλησης τοποθετούν κάποια σημεία στο γράφημα
  - Πως καθορίζουμε μια τιμή ανάμεσα στα σημεία;

Σύγκριση Συστημάτων

- Σύστημα 1
- Σύστημα 2



Σύγκριση Συστημάτων

Σκοπός: Δυνατότητα σύγκρισης διαφορετικών συστημάτων

Πως; Χρήση *κανονικοποιημένων επιπέδων ανάκλησης* (standard recall levels)

Παράδειγμα καθιερωμένων επιπέδων ανάκλησης (πλήθος επιπέδων: 11):

Standard Recall levels at 0%, 10%, 20%, …, 100%

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

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

Μέση ακρίβεια 11-σημείων με παρεμβολή (11-point interpolated average precision)

- Υπολόγισε την ακρίβεια με παρεμβολή στα επίπεδα ανάκτησης 0.0, 0.1, 0.2, . . .
- Επανέλαβε το για όλα τα ερωτήματα στο evaluation benchmark και πάρε το μέσο όρο
- Αυτό το μέτρο μετρά την απόδοση σε όλα τα επίπεδα ανάκλησης (at all recall levels).

Μέση ακρίβεια 11-σημείων με παρεμβολή (11-point interpolated average precision)

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

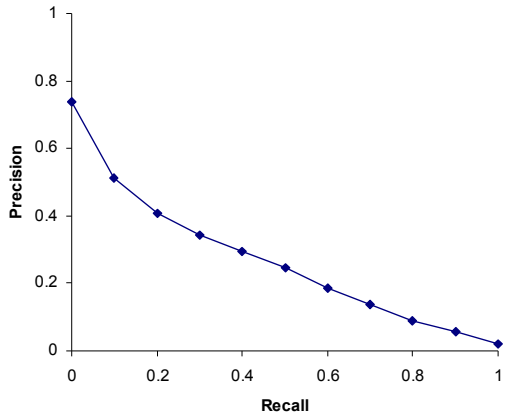
11-point average: ≈ 0.425

Γενικά υπολόγισε το μέσο precision recall για ένα σύνολο από ερωτήματα
N_q – πλήθος ερωτημάτων
$P_i(r)$ - precision at recall level r for i^{th} query

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

Τυπική (καλή;) ακρίβεια 11-σημείων

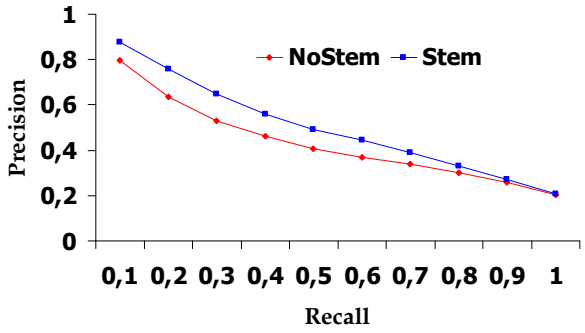
- SabIR/Cornell 8A1 11pt precision from TREC 8 (1999)



37

Σύγκριση Συστημάτων

- Η καμπύλη που είναι πιο κοντά στη πάνω δεξιά γωνία του γραφήματος υποδηλώνει και καλύτερη απόδοση



Και άλλα μέτρα

- Τα γραφήματα είναι καλά, αλλά οι χρήστες θέλουν περιληπτικά μέτρα
- Ακρίβεια σε προκαθορισμένα επίπεδα ανάκτησης
- Είδαμε το μέσο με τα 11-σημεία επίσης ->
  - ***Και άλλα μέτρα …***

ΜΑΠ

Μέση αντιπροσωπευτική ακρίβεια - Mean average precision (MAP)

- Μέσος όρος της τιμής της ακρίβειας των κορυφαίων *k* εγγράφων, κάθε φορά που επιστρέφεται ένα σχετικό έγγραφο
- Αποφεύγει την παρεμβολή και τη χρήση προκαθορισμένων επιπέδων ανάκλησης
- MAP για μια *συλλογή ερωτημάτων* είναι το αριθμητικό μέσο.
  - Macro-averaging: κάθε ερώτημα μετράει το ίδιο

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

Q σύνολο ερωτημάτων, $\mathbf{q_j}$ ένα από τα ερωτήματα, $\{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_{mj}}\}$ είναι τα συναφή έγγραφα και $\mathbf{R_{jk}}$ είναι ο αριθμός των εγγράφων στο αποτέλεσμα μέχρι να φτάσουμε στο d_{jk} (0 αν το d_{jk} δεν ανήκει στο αποτέλεσμα)

Διασπορά (Variance)

- Για μια συλλογή ελέγχου, συχνά ένα σύστημα έχει κακή απόδοση σε κάποιες πληροφοριακές ανάγκες (π.χ., MAP = 0.1) και άριστη σε άλλες (π.χ., MAP = 0.7)
- Συχνά, η διασπορά στην απόδοση είναι πιο μεγάλη για διαφορετικά ερωτήματα του ίδιου συστήματος παρά η διασπορά στην απόδοση διαφορετικών συστημάτων στην ίδια ερώτηση
- Δηλαδή, υπάρχουν εύκολες ανάγκες πληροφόρηση και δύσκολες ανάγκες πληροφόρησης!

Ακρίβεια στα *k*

- ***Ακρίβεια-στα-k (Precision-at-k):*** *Η ακρίβεια των κορυφαίων k* αποτελεσμάτων

Πχ ακρίβεια-στα-10

  - Πιθανόν κατάλληλο για τις περισσότερες αναζητήσεις στο web: όλοι οι χρήστες θέλουν καλά αποτελέσματα στις πρώτες μία ή δύο σελίδες
  - Αλλά: μη αντιπροσωπευτικό μέσο όρο

R-ακρίβεια

R-ακρίβεια

- Αν έχουμε ένα γνωστό (πιθανών μη πλήρες) σύνολο από συναφή έγγραφα μεγέθους *Rel,* τότε υπολογίζουμε την ακρίβεια των κορυφαίων *Rel* εγγράφων που επιστρέφει το σύστημα
- Το τέλειο σύστημα μπορεί να πετύχει βαθμό 1.0

Αν υπάρχουν r, τότε r/Rel

R-Ακρίβεια

- Ακρίβεια στην R-οστή θέση στη κατάταξη των αποτελεσμάτων για ένα ερώτημα που έχει R συναφή αποτελέσματα.

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

Μη δυαδικές εκτιμήσεις συνάφειας

Normalized discounted cumulative gain

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log_2(1+m)},$$

Υπολογίζεται για τα *k* πρώτα έγγραφα
Q σύνολο ερωτημάτων, q_j ένα από τα ερωτήματα,
R(j, d) είναι ο βαθμός που έδωσαν οι κριτές στο έγγραφο *d* του q_j
Z συντελεστής κανονικοποίησης ώστε για μια τέλεια βαθμολογία το NDCG να είναι 1

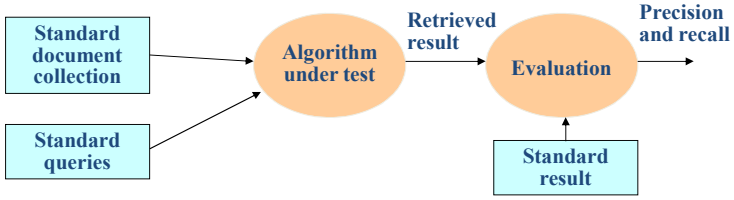
Μεθοδολογία – πρότυπες συλλογές (benchmarks)

Απαιτήσεις από ένα πρότυπο (benchmark)

1. Ένα σύνολο από έγγραφα
   - Τα έγγραφα πρέπει να είναι αντιπροσωπευτικά των πραγματικών εγγράφων
2. Μια συλλογή από ανάγκες πληροφόρησης
   - (ή, καταχρηστικά ερωτημάτων)
   - Να σχετίζονται με τα διαθέσιμα έγγραφα
   - Οι ανάγκες πληροφόρησης πρέπει να είναι αντιπροσωπευτικές των πραγματικών - τυχαίοι όροι δεν είναι καλή ιδέα
   - Συχνά από ειδικούς της περιοχής
3. Εκτιμήσεις συνάφειας από χρήστες (Human relevance assessments)
   - Χρειάζεται να προσλάβουμε/πληρώσουμε κριτές ή αξιολογητές.
   - Ακριβό χρονοβόρο
   - Οι κριτές πρέπει να είναι αντιπροσωπευτικοί των πραγματικών χρηστών
   - Mechanical truck

47

Benchmarks



Standard benchmarks συνάφειας

TREC - National Institute of Standards and Technology (NIST) τρέχει ένα μεγάλο IR test bed εδώ και πολλά χρόνια

- Χρησιμοποιεί το Reuters και άλλες πρότυπες συλλογές εγγράφων
- Καθορισμένα "Retrieval tasks"
  - Μερικές φορές ως ερωτήματα
- Ειδικοί (Human experts) βαθμολογούν κάθε ζεύγος ερωτήματος, εγγράφου ως Συναφές Relevant ή μη Συναφές Nonrelevant
  - Ή τουλάχιστον ένα υποσύνολο των εγγράφων που επιστρέφονται για κάθε ερώτημα

Standard benchmarks συνάφειας

Cranfield

Πρωτοπόρο: το πρώτο testbed που επέτρεπε ακριβή ποσοτικοποιημένα μέτρα της αποτελεσματικότητας της ανάκτησης

Στα τέλη του 1950, UK

- 1398 abstracts από άρθρα περιοδικών αεροδυναμικής, ένα σύνολο από 225 ερωτήματα, εξαντλητική κρίση συνάφειας όλων των ζευγών
- Πολύ μικρό, μη τυπικό για τα σημερινά δεδομένα της ΑΠ

TREC

TREC Ad Hoc task από τα πρώτα 8 TRECs είναι ένα standard task, μεταξύ του 1992-1999

- 1.89 εκατομμύρια έγγραφα, κυρίως newswire άρθρα
- 50 λεπτομερείς ανάγκες πληροφόρησης το χρόνο (σύνολο 450)
- Επιστρέφετε η αξιολόγηση χρηστών σε pooled αποτελέσματα (δηλαδή όχι εξαντλητική αξιολόγηση όλων των ζευγών)
- Πρόσφατα και Web track

A TREC query (TREC 5)

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>

Άλλα benchmarks

- GOV2
  - Ακόμα μια TREC/NIST συλλογή
  - 25 εκατομμύρια web σελίδες
  - Αλλά ακόμα τουλάχιστον 3 τάξης μεγέθους μικρότερη από το ευρετήριο της Google/Yahoo/MSN
- NTCIR
  - Ανάκτηση πληροφορίας για τις γλώσσες της Ανατολικής Ασίας και cross-language ανάκτηση
- Cross Language Evaluation Forum (CLEF)
  - Το ίδιο για Ευρωπαϊκές γλώσσες

Συλλογές ελέγχου

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Αξιοπιστία των αξιολογήσεων των κριτών

- Οι αξιολογήσεις συνάφειας είναι χρήσιμες αν είναι συνεπής consistent.
- Πως μπορούμε να μετρήσουμε τη συνέπεια ή τη συμφωνία ανάμεσα στους κριτές

Μέτρο Kappa της διαφωνίας (συμφωνίας) (dis-agreement) μεταξύ των κριτών

Μέτρο Kappa

- Συμφωνία μεταξύ των κριτών
- Αφορά κατηγορική κρίση
- Λαμβάνει υπό όψιν την συμφωνία από τύχη

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A): ποσοστό των περιπτώσεων που οι κριτές συμφωνούν

P(E): τι συμφωνία θα είχαμε από τύχη

κ = 1 Για πλήρη συμφωνία, 0 για τυχαία συμφωνία, αρνητική για μικρότερη της τυχαίας

Kappa: παράδειγμα

Number of docs	ΚΡΙΤΗΣ 1	ΚΡΙΤΗΣ 2
300	Relevant	Relevant
70	Nonrelevant	Nonrelevant
20	Relevant	Nonrelevant
10	Nonrelevant	Relevant

P(A) = 370/400 = 0.925

P(nonrelevant) = (70+10+70+20)/800 = 0.2125

P(relevant) = (300+20+300+10)/800 = 0.7878

P(E) = 0.2125^2 + 0.7878^2 = 0.665

Kappa = (0.925 – 0.665)/(1-0.665) = 0.776

Kappa

- Kappa > 0.8 = καλή συμφωνία

0.67 < Kappa < 0.8 -> “tentative conclusions” (Carletta ’96)

- Εξαρτάται από το στόχο της μελέτης
- Για >2 κριτές: μέσοι όροι ανά-δύο κ

Kappa: παράδειγμα

Information need	number of docs judged	disagreements
51	211	6
62	400	157
67	400	68
95	400	110
127	400	106

Συμφωνία κριτών στο TREC

Επίπτωση της Διαφωνίας

- Επηρεάζει την απόλυτη (absolute) μέτρηση απόδοσης αλλά όχι τη σχετική απόδοση ανάμεσα σε συστήματα

Επίπτωση της Διαφωνίας

- Μπορούμε να αποφύγουμε τις κρίσεις από χρήστες
  - Όχι
- Αλλά μπορούμε να τα επαναχρησιμοποιήσουμε

Crowdsourcing

- Το Mechanical Truck της Amazon

Αξιολόγηση σε μεγάλες μηχανές αναζήτησης

- Οι μηχανές αναζήτησης διαθέτουν συλλογές ελέγχου ερωτημάτων και αποτελέσματα καταταγμένα με το χέρι (hand-ranked)
- Στο web είναι δύσκολο να υπολογίσουμε την ανάκληση

  Συνήθως οι μηχανές αναζήτησης χρησιμοποιούν την ακρίβεια στα κορυφαία k π.χ., $k = 10$
- Υπάρχουν επίσης μέτρα που σταθμίζουν περισσότερο την επιτυχία στην ανάκτηση του $1^{ου}$ π.χ, από το $10^{ο}$ αποτέλεσμα.
  - NDCG (Normalized Cumulative Discounted Gain)

Αξιολόγηση σε μεγάλες μηχανές αναζήτησης

Οι μηχανές αναζήτησης χρησιμοποιούν επίσης και άλλα μέτρα εκτός της συνάφειας

- Clickthrough on first result
  - Όχι πολύ αξιόπιστο όταν ένα clickthrough (μπορεί απλώς η περίληψη να φάνηκε χρήσιμη αλλά όχι το ίδιο το έγγραφο) αλλά αρκετά αξιόπιστα συναθροιστικά
- Μετρήσεις σε εργαστήριο
- Έλεγχος A/B

A/B testing

Στόχος: έλεγχος μιας νέας ιδέας (a single innovation)
Προϋπόθεση: Υπάρχει μια μεγάλη μηχανή αναζήτησης σε λειτουργία

- Οι πιο πολλοί χρήστες χρησιμοποιούν τα παλιό σύστημα
- Παράκαμψε ένα μικρό ποσοστό της κυκλοφορίας (π.χ., 1%) στο νέο σύστημα που χρησιμοποιεί την καινούργια
- Αξιολόγησε με ένα αυτόματο μέτρο όπως το clickthrough τα πρώτα αποτελέσματα

Κριτική της Συνάφειας

- Οριακή Συνάφεια (Marginal Relevance)
- Και άλλα κριτήρια όπως
  - Novelty
  - Coverage

ΤΕΛΟΣ α’ μέρους 8ου Μαθήματος

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό από:
✓Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)
✓Hinrich Schütze and Christina Lioma, Stuttgart IIR class
✓ διαφάνειες του καθ. Γιάννη Τζίτζικα (Παν. Κρήτης)