Ευρετηρίαση κειμένων, βαθμονόμηση και αξιολόγηση αποτελεσμάτων

Ηρακλής Βαρλάμης

Περιεχόμενα

- Μέθοδοι αξιολόγησης του IR
- Ανεστραμμένα ευρετήρια

Γιατί χρειαζόμαστε αξιολόγηση

- Για να εντοπίσουμε την καλύτερη μέθοδο, αλγόριθμο, προεπεξεργασία κλπ
- Επιλογές:
  - Συνάρτηση ομοιότητας (cosine similarity, dot product, euclidian distance, …)
  - Προεπεξεργασία (stopword removal, stemming, …)
  - Βάρος όρων (tf, tf-idf, …)
- Πόσα αποτελέσματα (“κατ εκτίμηση σχετικά έγγραφα”) πρέπει να δω για να βρω όλα ή τα περισσότερα πραγματικά σχετικά έγγραφα;
- Ακόμη και αν η αξιολόγηση για ένα κείμενο είναι δυαδική (σχετικό ή όχι-σχετικό), είναι υποκειμενική. Ακόμη και οι άνθρωποι δε συμφωνούν πάντοτε.

Σύγκριση μετρικών

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

Δίνει τιμές στο [-1,1]

Ιδανικό για σύγκριση ομοιότητας περιεχομένου (και embeddings), όπου δε μας ενδιαφέρει το μέγεθος του κάθε διανύσματος αλλά η κατεύθυνση.

Αν τα διανύσματα δείχνουν στην ίδια κατεύθυνση η ομοιότητα είναι μεγάλη ανεξάρτητα από το μήκος τους (π.χ. τα αντίστοιχα βάρη μπορεί να είναι διπλάσια από το ένα στο άλλο).

$$A \cdot B = \sum A_i B_i$$

Δίνει τιμές στο (-οο,+οο)

Χρήσιμο όταν τόσο το μέγεθος όσο και η κατεύθυνση των διανυσμάτων έχουν σημασία (π.χ. σε συστήματα συστάσεων όπου το μέγεθος των προτιμήσεων των χρηστών ή των χαρακτηριστικών στοιχείων είναι σημαντικό).

Αν τα διανύσματα δείχνουν στην ίδια κατεύθυνση η ομοιότητα είναι θετική και μεγαλώνει ανάλογα με το μήκος των διανυσμάτων.

$$d(A, B) = \sqrt{\sum (A_i - B_i)^2}$$

Δίνει τιμές στο [0,+οο)

Ευαίσθητο σε διαφορές κλίμακας. απαιτεί κανονικοποιημένα δεδομένα για ισορροπημένες συγκρίσεις

Χρήση σε εργασίες χωρικής εγγύτητας (π.χ. ομαδοποίηση). Σε δεδομένα υψηλών διαστάσεων δεν αποδίδει καλά.

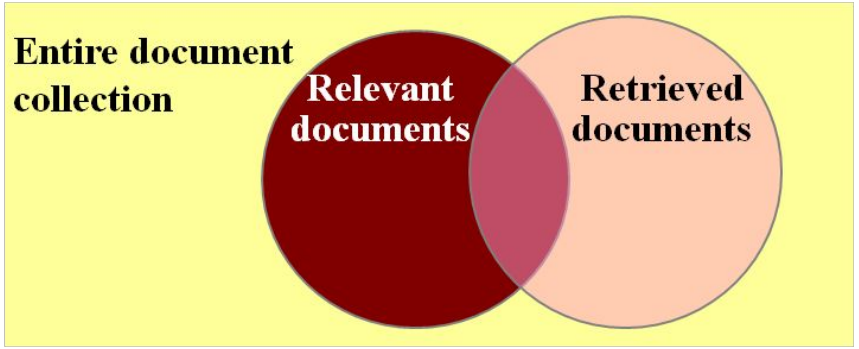
Τι χρειάζομαι - Ένα gold standard

- Ένα σύνολο από κείμενα που έχουν χαρακτηριστεί ως σχετικά με την πληροφοριακή ανάγκη (ερώτηση) του χρήστη.
- Χρειάζομαι πολλούς ειδικούς για να ετοιμάσω ένα τέτοιο σύνολο (με πολλές ερωτήσεις και πολλά κείμενα).

Ακρίβεια (Precision) και Ανάκληση (Recall)

- Precision
  - Η ικανότητα να ανακτήσω στην απάντησή μου κείμενα που είναι πολύ σχετικά.
- Recall
  - Η ικανότητα να βρω όλα τα σχετικά κείμενα στο σώμα κειμένων και να τα επιστρέψω ως απάντηση.
- *Όταν ψάχνω σε συλλογές χιλιάδων ή και περισσοτέρων κειμένων όπου τα σχετικά κείμενα σε μια ερώτηση είναι εκατοντάδες ή και περισσότερα, τότε η ποιότητα της απάντησής μου καθορίζεται από το πόσα κείμενα αποφασίζω να επιστρέψω*

Ακρίβεια (Precision) και Ανάκληση (Recall)



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

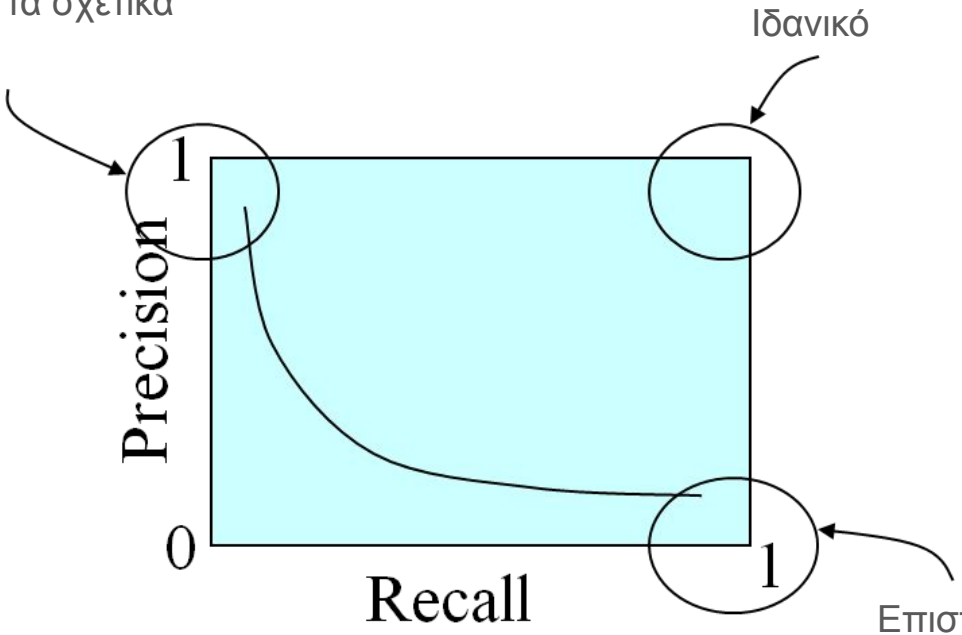
$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Δύσκολο να πετύχω ψηλό recall

- Δύσκολο και να αξιολογήσω το recall
- Δεν γνωρίζω εξ αρχής πόσα και ποια είναι όλα τα κείμενα που απαντούν σωστά την ερώτηση
  - Κάνω δειγματοληπτική αξιολόγηση για πολλαπλά ερωτήματα
  - Αξιολογώ πολλούς εναλλακτικούς αλγορίθμους και κρατώ αυτόν με την καλύτερη μέση αξιολόγηση (ακρίβεια;... ανάκληση;)

Trade-off μεταξύ ανάκλησης και ακρίβειας

Επιστρέφει σχετικά κείμενα μόνο, αλλά χάνει πολλά από τα σχετικά στο σώμα



Επιστρέφει όλα τα σχετικά κείμενα στο σώμα, αλλά περιλαμβάνει και πολλά σκουπίδια

Υπολογισμός των σημείων Recall/Precision

- Για ένα δεδομένο ερώτημα, παράγω την ταξινομημένη λίστα των σχετικών κειμένων (μπορεί να περιέχει και όλα τα κείμενα με τα οποία η ομοιότητα δεν είναι μηδενική).
- Θέτοντας ένα κατώφλι σε αυτή τη ταξινομημένη λίστα παράγω διαφορετικά σύνολα ανακτηθέντων εγγράφων και, επομένως, διαφορετικά μέτρα ανάκλησης/ακρίβειας.
- Σημειώνω στη λίστα αυτή τα σχετικά και άσχετα κείμενα σύμφωνα με το golden standard.
- Υπολογίζω ένα ζεύγος ανάκλησης/ακρίβειας για κάθε θέση στη ταξινομημένη λίστα που περιέχει ένα σχετικό έγγραφο. Εναλλακτικά για τα 10 ποσοστημόρια της λίστας.

Παράδειγμα - Μοντέλο Α

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	

Έστω ότι έχω συνολικά 6 σχετικά κείμενα
Ελέγχω κάθε νέο 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

Αν χάσω αυτό το σχετικό κείμενο δεν θα φτάσω στο 100% recall

Παράδειγμα - Μοντέλο Β

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Έστω ότι έχω συνολικά 6 σχετικά κείμενα
Ελέγχω κάθε νέο recall point:

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

R=2/6=0.333; P=2/3=0.667

R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

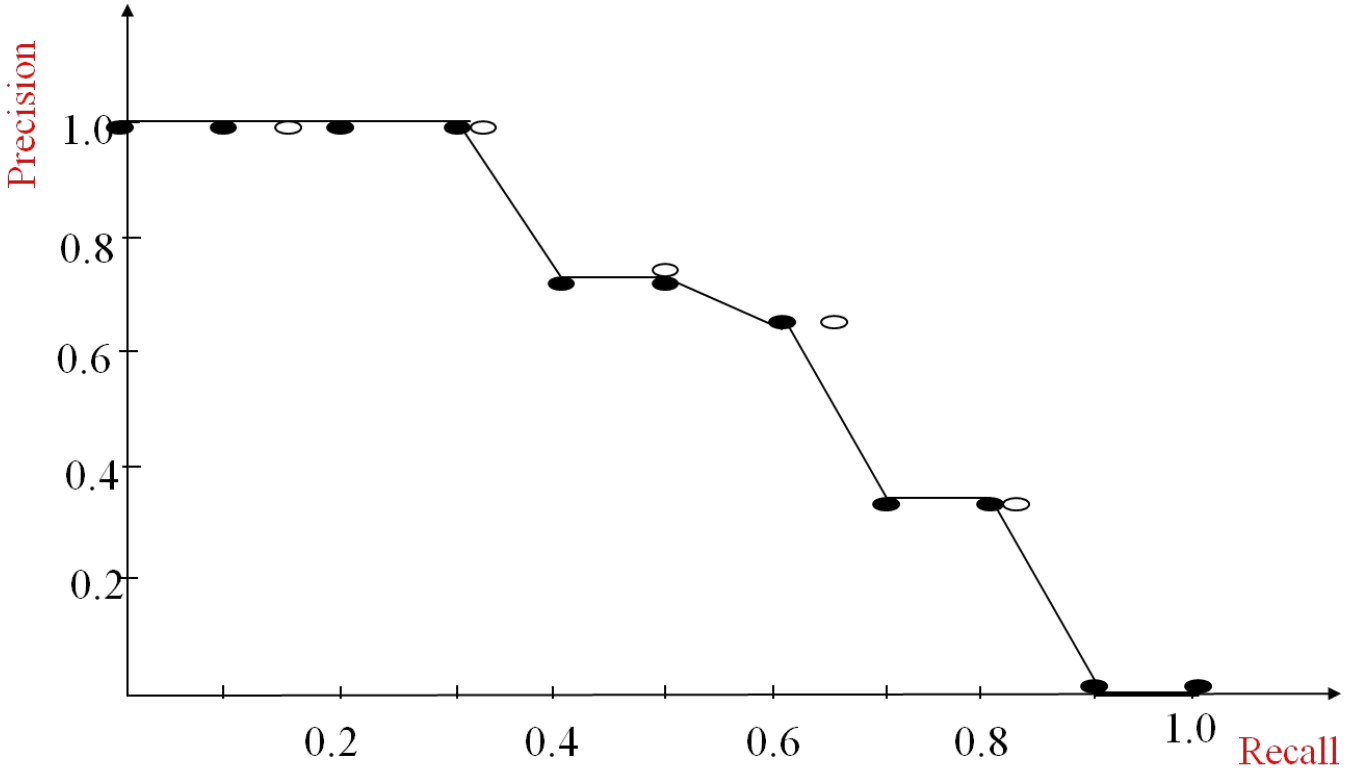
R=6/6=1.0; p=6/14=0.429

Προσαρμογή καμπύλης Ανάκλησης/Ακρίβειας

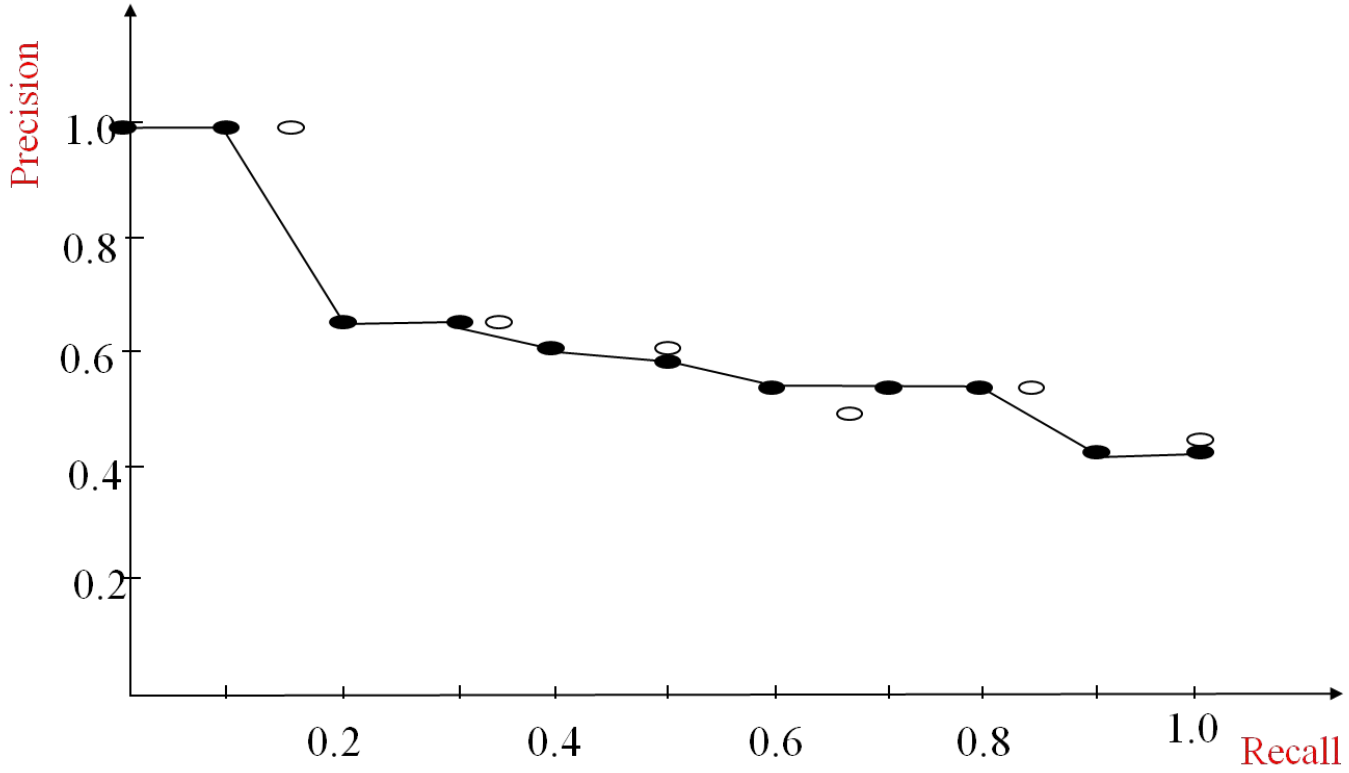
- Κάνω μια προσαρμογή (interpolation) στα διαστήματα που δεν έχω τιμές για να έχω μια τιμή ακρίβειας για κάθε επίπεδο ανάκληση
- $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- $r_0 = 0.0$, $r_1 = 0.1$, …, $r_{10}=1.0$
- Η προσαρμοσμένη ακρίβεια στο j-th standard recall level είναι η μέγιστη ακρίβεια σε οποιοδήποτε recall level ανάμεσα στο j-th και στο (j + 1)-th:

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

Μοντέλο Α με interpolation

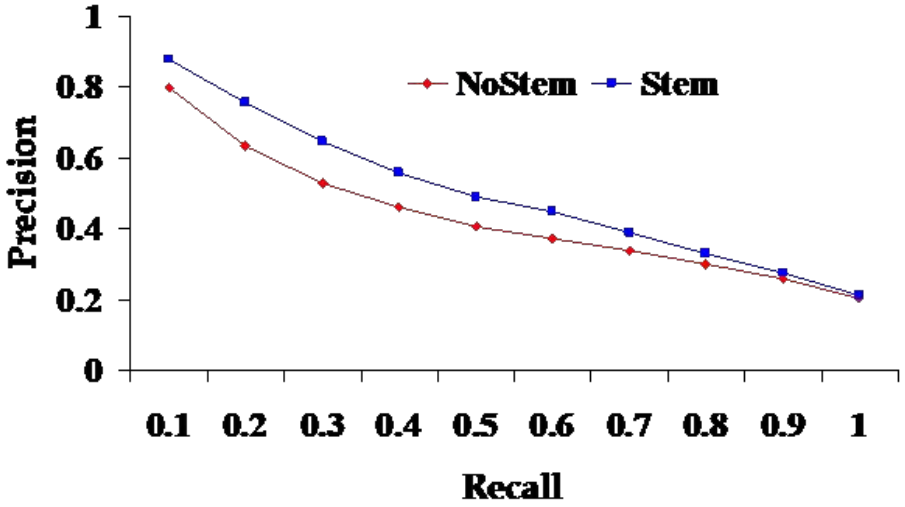


Μοντέλο Β με interpolation



Σύγκριση μοντέλων

- Συνήθως υπολογίζουμε τις μέσες τιμές ακρίβειας και ανάκλησης σε πολλά ερωτήματα
- Η καμπύλη που είναι πιο κοντά στην πάνω δεξιά γωνία του γραφήματος αντιστοιχεί στο καλύτερο μοντέλο.



Εναλλακτικά - Ακρίβεια στα top-k

- Επιλέγουμε έναν αριθμό κειμένων που θεωρούμε ότι συνήθως ελέγχουν οι χρήστες (10, 20, 30,...) και αναφέρουμε την ακρίβεια στο σημείο αυτό.

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@6 = 4/6 = 0.67

F-Measure

- Ένα μέτρο απόδοσης που λαμβάνει υπόψη ταυτόχρονα την ακρίβεια και την ανάκληση
- Είναι ο αρμονικός μέσος των δύο μεγεθών

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

- Σε σχέση με τον αριθμητικό μέσο, απαιτεί και τα δύο μεγέθη να είναι ψηλά για να δώσει υψηλή τιμή

E Measure (parameterized F Measure)

- Μια παραλλαγή του F measure που επιτρέπει να δώσουμε έμφαση στην ακρίβεια ή στην ανάκληση

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

- Το β καθορίζει την ισορροπία των δύο:
  - β=1 ⇒ E=F
  - β>1 ⇒ θεωρεί πιο σημαντική την ανάκληση
  - β<1 ⇒ θεωρεί πιο σημαντική την ακρίβεια

Mean Average Precision (MAP)

- Average Precision: Μέση τιμή όλων των τιμών ακρίβειας στα διαφορετικά σημεία ανάκλησης
  - Μοντέλο 1: (1 + 1 + 0.75 + 0.667 + 0.38 + 0)/6 = 0.633
  - Μοντέλο 2: (1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429)/6 = 0.625

$$\overline{p}(r_i) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} p_j(r_i),$$

- Mean Average Precision: Μέση τιμή του average precision για πολλά ερωτήματα

Μη δυαδική σχετικότητα

- Τα κείμενα δεν είναι σχετική ή μη-σχετικά μόνο
- Μπορούν να έχουν και ενδιάμεσες τιμές (π.χ. 5-βάθμια κλίμακα)
- Μπορεί να έχουμε πολλαπλούς κριτές (που δε συμφωνούν απόλυτα) και δίνουν ένα μέσο σκορ.
- Μπορεί να μετράμε τη σχετικότητα με βάση το αν ο χρήστης επέλεξε (κλίκαρε) το αποτέλεσμα

Inter-annotator agreement

- Kappa measure
  - Έχει σχεδιαστεί για binary αξιολογήσεις (σχετικό vs μη-σχετικό)
- Kappa = [P(A) – P(E)] / [1 – P(E)]
- P(A) – ποσοστό συμφωνίας των κριτών
- P(E) – πιθανότητα να συμφωνήσουν κατά τύχη
- P(E) = P(relevant)^2+P(non-relevant)^2
- P(relevant) - η πιθανότητα να έχω έναν χαρακτηρισμό με τιμή “σχετικό”
- Kappa = 0 για τυχαία συμφωνία, 1 για πλήρη συμφωνία.

Παράδειγμα

- P(A) = 370/400 = 0.925
- P(nonrelevant) = (10+20+70+70)/800 = 0.2125
- P(relevant) = (10+20+300+300)/800 = 0.7878
- P(E) = 0.2125^2 + 0.7878^2 = 0.665
- Kappa = (0.925 – 0.665)/(1-0.665) = 0.776

- Kappa > 0.8 ⇒ καλή συμφωνία
- 0.67 < Kappa < 0.8 ⇒ μερική συμφωνία

- For >2 judges: average pairwise kappas

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

Cumulative Gain

- Αν η σχετικότητα βαθμολογείται σε μια κλίμακα (π.χ. 0 ως 1) μπορούμε να υπολογίζουμε το κέρδος σε κάθε αλλαγή ανάκλησης.
- Cumulative Gain στη θέση ανάκλησης n:

$$CG_n = \sum_{i=1}^{n} rel_i$$

(όπου rel_i είναι η σχετικότητα του εγγράφου που ανακτήθηκε στη θέση i - και ήταν σχετικό)

n	doc #	relevance (gain)	CG_n
1	588	1,0	1,0
2	589	0,6	1,6
3	576	0,0	1,6
4	590	0,8	2,4
5	986	0,0	2,4
6	592	1,0	3,4
7	984	0,0	3,4
8	988	0,0	3,4
9	578	0,0	3,4
10	985	0,0	3,4
11	103	0,0	3,4
12	591	0,0	3,4
13	772	0,2	3,6
14	990	0,0	3,6

Discounted Cumulative Gain

- Οι χρήστες νοιάζονται περισσότερο για τα πρώτα κείμενα που τους επιστρέφουμε, άρα τα πιο κάτω αποτελέσματα έχουν μειωμένο κέρδος κατά 1/log2(rank).
- Discounted Cumulative Gain στη θέση ανάκλησης n:

$$DCG_n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$$

n	doc #	rel (gain)	CG_n	$\log_n$	DCG_n
1	588	1.0	1.0	-	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

Normalized Discounted Cumulative Gain (NDCG)

- Κανονικοποιούμε τις τιμές έτσι ώστε το ιδανικό ranking να έχει ένα NDCG=1
- Το ιδανικό ranking δεξιά θα έπαιρνε το DCG της τελευταίας στήλης

n	doc #	rel (gain)	CG_n	log_n	DCG_n
1	588	1.0	1.0	0.00	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

n	doc #	rel (gain)	CG_n	log_n	$IDCG_n$
1	588	1.0	1.0	0.00	1.00
2	592	1.0	2.0	1.00	2.00
3	590	0.8	2.8	1.58	2.50
4	589	0.6	3.4	2.00	2.80
5	772	0.2	3.6	2.32	2.89
6	576	0.0	3.6	2.58	2.89
7	986	0.0	3.6	2.81	2.89
8	984	0.0	3.6	3.00	2.89
9	988	0.0	3.6	3.17	2.89
10	578	0.0	3.6	3.32	2.89
11	985	0.0	3.6	3.46	2.89
12	103	0.0	3.6	3.58	2.89
13	591	0.0	3.6	3.70	2.89
14	990	0.0	3.6	3.81	2.89

Normalized Discounted Cumulative Gain (NDCG)

- Κανονικοποίηση του DCG με το DCG του ideal ranking:

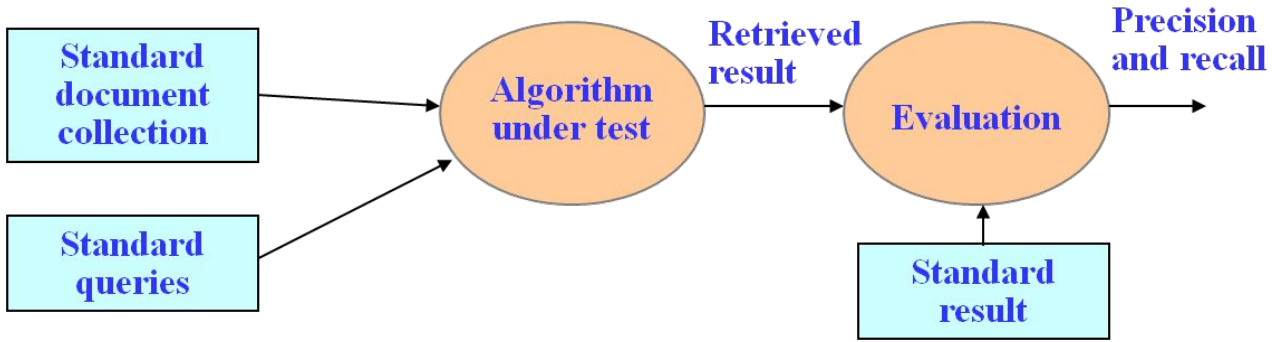
$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

- NDCG ≤ 1 σε κάθε θέση
- NDCG είναι συγκρίσιμο μεταξύ πολλαπλών ερωτήσεων

n	doc #	rel (gain)	DCG_n	$IDCG_n$	$NDCG_n$
1	588	1.0	1.00	1.00	**1.00**
2	589	0.6	1.60	2.00	**0.80**
3	576	0.0	1.60	2.50	**0.64**
4	590	0.8	2.00	2.80	**0.71**
5	986	0.0	2.00	2.89	**0.69**
6	592	1.0	2.39	2.89	**0.83**
7	984	0.0	2.39	2.89	**0.83**
8	988	0.0	2.39	2.89	**0.83**
9	578	0.0	2.39	2.89	**0.83**
10	985	0.0	2.39	2.89	**0.83**
11	103	0.0	2.39	2.89	**0.83**
12	591	0.0	2.39	2.89	**0.83**
13	772	0.2	2.44	2.89	**0.84**
14	990	0.0	2.44	2.89	**0.84**

Benchmarks

- Μια συλλογή αξιολόγησης περιέχει:
  - Ένα σύνολο από κείμενα και ερωτήσεις/θέματα.
  - Μια λίστα από σχετικά κείμενα για κάθε ερώτημα.
- Standard collections for traditional IR:
  - Smart collection: ftp://ftp.cs.cornell.edu/pub/smart
  - TREC: http://trec.nist.gov/
  - NFCorpus: https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/



Τι κάνουμε report

- Summary table statistics: Πλήθος θεμάτων, πλήθος εγγράφων που ανακτήθηκαν, πλήθος σχετικών εγγράφων
- Recall-precision average: Μέση ακρίβεια στα 11 recall levels (0 ως 1 με βήμα 0.1).
- Document level average: Μέση ακρίβεια στα πρώτα 5, 10, .., 100, … 1000 κείμενα.
- Average precision histogram: Διαφορές στο R-precision για κάθε θέμα και το average R-precision όλων το συστημάτων για το ίδιο θέμα

Περιεχόμενα

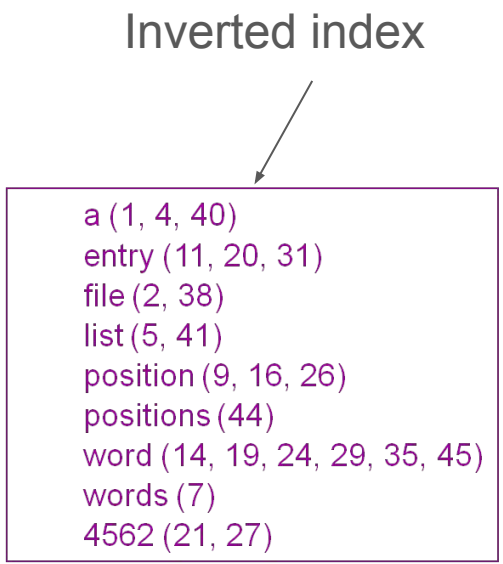
- Μέθοδοι αξιολόγησης του IR
- Ανεστραμμένα ευρετήρια

Απλοϊκή ανάκτηση

- Υπολογίζω την ομοιότητα/απόσταση του q από κάθε κείμενο d_i στο σώμα D
- Λαμβάνω υπόψη μόνο τις λέξεις με μη-μηδενικό βάρος
- Λαμβάνω υπόψη τις λέξεις που υπάρχουν ταυτόχρονα στο q και στο d_i
- Ταξινομώ τα κείμενα με φθίνουσα σειρά ομοιότητας ως προς το q
- Επιλέγω τα top-k

Ανάκτηση με χρήση ευρετηρίων

- Οργανώνω τα έγγραφα στο σώμα D σε μια δομή που να μου επιτρέπει πολύ γρήγορα να βρώ ένα μικρό υποσύνολο του D και μετά να κάνω την απλοϊκή ανάκτηση σε αυτό.
- Χρησιμοποιώ ανεστραμμένο ευρετήριο:
  - Κάθε λέξη στο λεξικό μου συνοδεύεται από μια λίστα με τα doc_ids που την περιέχουν
- Για ένα:
  - Σύνολο εγγράφων D = $\{d_1, d_2, \ldots, d_N\}$
  - Κάθε έγγραφο έχει μοναδικό ID id_j
- Ένα ανεστραμμένο ευρετήριο αποτελείται από:
  - Ένα λεξιλογιο V που περιέχει όλους τους διακριτούς όρους t_i στο D
  - Και μια ανεστραμμένη λίστα με αναφορές(postings) για κάθε t_i
- Κάθε αναφορά περιέχει το id_j *του* d_j και επιπλέον πληροφορίες, π.χ. συχνότητα, θέση του όρου στο έγγραφο κλπ



Παράδειγμα

- Έχουμε 3 έγγραφα id1, id2, id3:
  - id1: Web mining is useful.
    1 2 3 4
  - id2: Usage mining applications.
    1 2 3
  - id3: Web structure mining studies the Web hyperlink structure.
    1 2 3 4 5 6 7 8

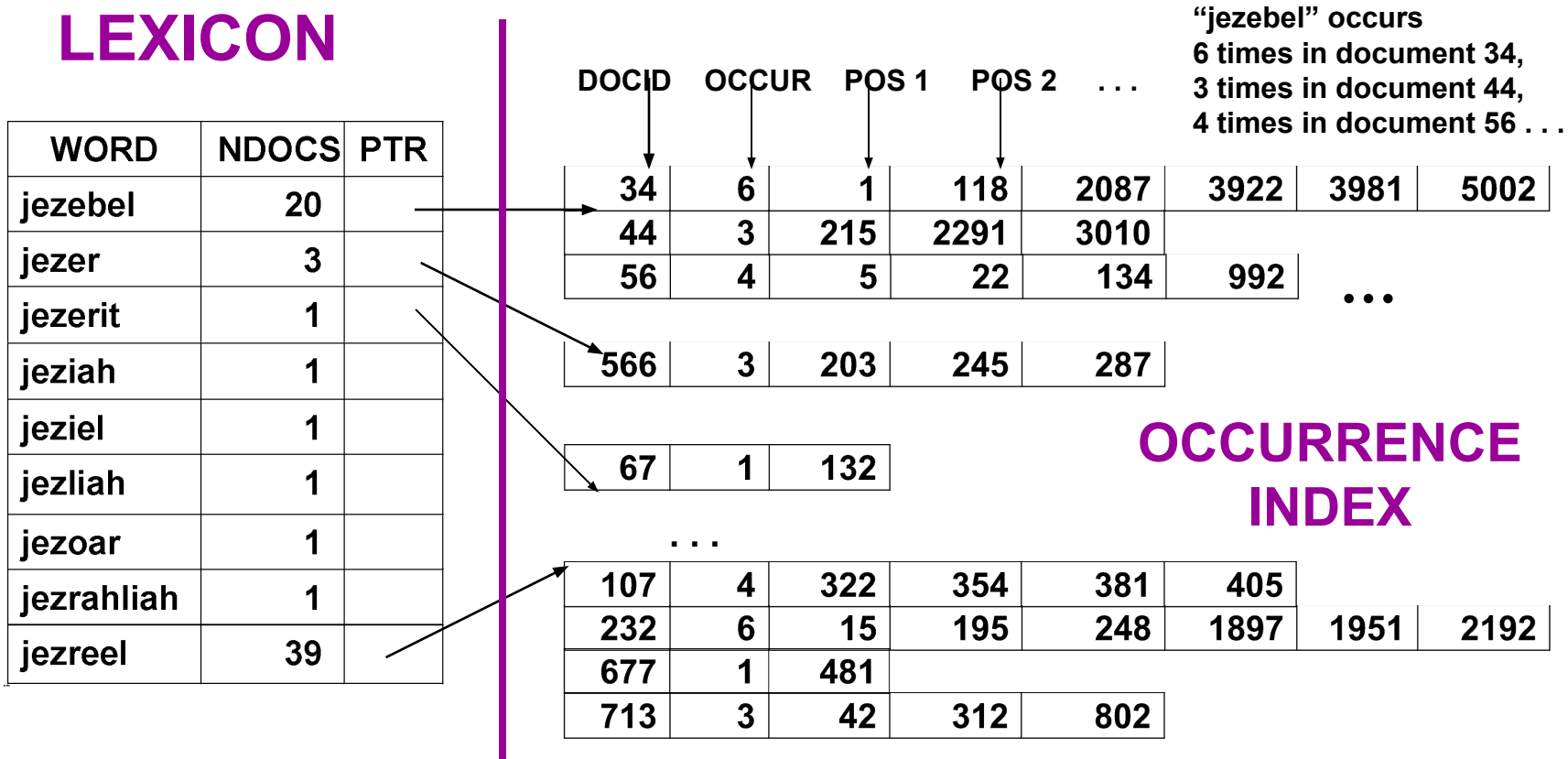
Applications: id2
Hyperlink: id3
Mining: id1,id2,id3
Structure: id3
Studies: id3
Usage:id2
Useful:id1
Web:id1,id3

Απλό ανεστραμμένο ευρετήριο

Applications: <id2,1,[3]>
Hyperlink: <id3,1,[7]>
Mining: <id1,1,[2]>, <id2,1,[2]>, <id3,1,[3]>
Structure: <id3,2,[2,8]>
Studies: <id3,1,[4]>
Usage:<id2,1,[1]>
Useful:<id1,1,[4]>
Web:<id1,1,[1]>, <id3,2,[1,6]>

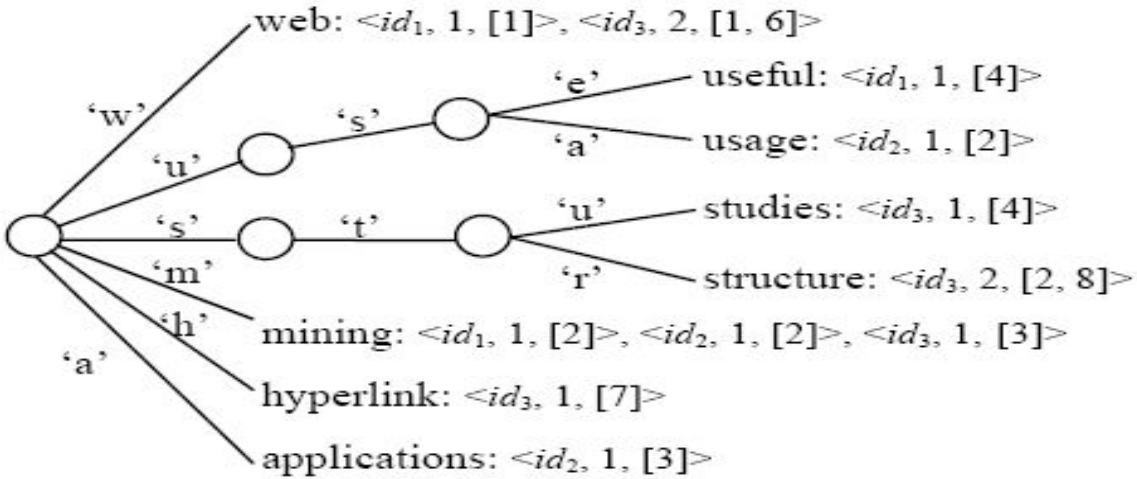
Σύνθετο ανεστραμμένο ευρετήριο

Παράδειγμα


LEXICON
WORD
NDOCS
PTR
jezebel
20
jezer
3
jezerit
1
jeziah
1
jeziel
1
jezliah
1
jezoar
1
jezrahliah
1
jezreel
39
DOCID
OCCUR
POS 1
POS 2
. . .
"jezebel" occurs
6 times in document 34,
3 times in document 44,
4 times in document 56 . . .
34 6 1 118 2087 3922 3981 5002
44 3 215 2291 3010
56 4 5 22 134 992
. . .
566 3 203 245 287
67 1 132
. . .
107 4 322 354 381 405
232 6 15 195 248 1897 1951 2192
677 1 481
713 3 42 312 802
OCCURRENCE
INDEX

Λεξικό και trie index

- Ειδική δομή (trie structure):
- id_1: Web mining is useful
- id_2: Usage mining applications
- id_3: Web structure mining studies the Web hyperlink structure



Αναζήτηση με ανεστραμμένο ευρετήριο

- Για το σύνολο των όρων ενός ερωτήματος:
  1. Αναζήτηση στο λεξιλόγιο: βρίσκω για κάθε όρο του ερωτήματος την λίστα αναφορών στο λεξιλόγιο.
     - Το λεξιλόγιο συνήθως χωρά στη μνήμη
     - Αν έχω κάποιο ευρετήριο επιταχύνω την αναζήτηση
     - Αν έχω λεξικογραφική διάταξη κάνω δυαδική αναζήτηση
     - Αν το ερώτημα έχει 1 όρο, πάω στο βήμα 3, αλλιώς πάω στο βήμα 2
  2. Συγχώνευση αποτελεσμάτων: συγχωνεύω τις λίστες κάθε όρου και κρατώ την τομή τους
     - Έτσι βρίσκω τα έγγραφα που περιέχουν όλους τους όρους του ερωτήματος
     - Ευριστικό: Ξεκινώ από τη μικρότερη λίστα
     - Ενδεχομένως να δεχτώ μερικό ταίριασμα στους όρους
  3. Υπολογισμός βαθμού σχετικότητας: για κάθε έγγραφο που περιέχει τον όρο (τους όρους) ως προς το ερώτημα
     - Χρήση συνάρτησης ομοιότητας (π.χ. cosine)
     - Μπορεί να λάβει υπόψη την εγγύτητα των όρων στο έγγραφο και το ερώτημα αντίστοιχα

Ευρετήρια μηχανών αναζήτησης

- Το ευρετήριο του Web δεν χωρά στη μνήμη
- Δημιουργούμε μερικά ευρετήρια που χωρούν στη μνήμη και τα αποθηκεύουμε στο δίσκο
- Μόλις φτιάξουμε όλα τα ευρετήρια τα συγχωνεύουμε ιεραρχικά
- Εναλλακτικά χρησιμοποιούμε ένα hash-table (ή κάποια άλλη δομή) ή συγχωνεύουμε τα ευρετήρια αυξητικά
- Συγχωνεύουμε τα ευρετήρια