Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2006-2007 Εαρινό Εξάμηνο

Μαρκετάκης Γιάννης [1772] *marketak@csd.uoc.gr*

1η Σειρά Ασκήσεων (Αξιολόγηση της Αποτελεσματικότητας της Ανάκτησης)

Άσκηση 1 (4 βαθμοί)

Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 40 έγγραφα {d_1,…,d_{40}}. Η συλλογή αξιολόγησης περιλαμβάνει μια επερώτηση q για την οποία γνωρίζουμε ότι τα έγγραφα της συλλογής που είναι συναφή με αυτήν είναι 5, συγκεκριμένα τα {d_1, d_{11}, d_{18}, d_{21}, d_{33}}. Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων S1, S2 και S3.

Για το λόγο αυτό υποβάλλουμε σε κάθε σύστημα την επερώτηση q και λαμβάνουμε τις εξής απαντήσεις:

Ans(S1,q) = <d_{11}, d_4, d_{18}, d_2, d_{21}, d_{33}, d_9, d_7, d_8, d_6, d_1, d_5>
Ans(S2,q) = < d_9, d_7, d_5, d_6, d_{11}, d_4, d_8, d_2, d_1, d_{33}, d_{18}, d_{21}>
Ans(S3,q) = < d_{18}, d_{33}, d_{11}, d_1, d_5, d_2 >

Το αριστερότερο στοιχείο της κάθε απάντησης παριστάνει το υψηλότερα διαβαθμισμένο έγγραφο, αυτό που το σύστημα υπολόγισε ως το πιο συναφές με την επερώτηση q. Συγκρίνετε τα τρία αυτά συστήματα ως προς τα εξής μέτρα: (α) Ακρίβεια (Precision), (β) Ανάκληση (Recall), (γ) F-Measure, (δ) R-Ακρίβεια (R-Precision) και (ε) Fallout. Για κάθε μέτρο σχολιάστε το αποτέλεσμα της σύγκρισης.

Λύση

α) Ακρίβεια (Precision)

Γνωρίζουμε ότι τα συναφή με την επερώτηση q έγγραφα της συλλογής είναι 5 (συγκεκριμένα {d_1, d_{11}, d_{18}, d_{21}, d_{33}}). Έτσι για κάθε σύστημα έχουμε:

S1 : Τα συναφή έγγραφα που επιστρέφει το σύστημα **S1** είναι { d_{11}, d_{18}, d_{21}, d_{33}, d_1}, άρα 5 ενώ το συνολικό πλήθος είναι 12 έγγραφα. Επομένως το σύστημα **S1** έχει ακρίβεια

$$P(S1) = \frac{5}{12} = 0.417$$

S2 : Τα συναφή έγγραφα που επιστρέφει το σύστημα **S2** είναι { d_{11}, d_1, d_{33}, d_{18}, d_{21}}, άρα 5 ενώ το συνολικό πλήθος είναι 12 έγγραφα. Επομένως το σύστημα **S2** έχει ακρίβεια

$$P(S2) = \frac{5}{12} = 0.417$$

S3 : Τα συναφή έγγραφα που επιστρέφει το σύστημα **S3** είναι { d_{18}, d_{33}, d_{11}, d_1}, άρα 4 ενώ το συνολικό πλήθος είναι 6 έγγραφα. Επομένως το σύστημα **S3** έχει ακρίβεια

$$P(S3) = \frac{4}{6} = 0.667$$

Έτσι βλέπουμε ότι τα συστήματα **S1** και **S2** έχουν την ίδια ακρίβεια αν και έχουν διαφορετικές απαντήσεις. Το σύστημα **S3** έχει μεγαλύτερη ακρίβεια αν και δίνει λιγότερα συναφή έγγραφα από τα άλλα 2 άλλα από την άλλη δίνει και λιγότερα μη-συναφή έγγραφα και είναι προτιμητέο από πλευράς ακρίβειας.

β) Ανάκληση (Recall)

Γνωρίζουμε ότι το σύνολο των συναφών εγγράφων είναι 5. Οπότε:

S1 : Επιστρέφει 5 συναφή αποτελέσματα οπότε :

$$R(S1) = \frac{5}{5} = 1$$

S2 : Επιστρέφει 5 συναφή αποτελέσματα οπότε :

$$R(S2) = \frac{5}{5} = 1$$

S3 : Επιστρέφει 4 συναφή αποτελέσματα οπότε :

$$R(S3) = \frac{4}{5} = 0.8$$

Παρατηρούμε ότι το **S1** και το **S2** έχουν την καλύτερη δυνατή ανάκληση (επιστρέφουν όλα τα συναφή έγγραφα) ενώ από την άλλη το **S3** επιστρέφει μόνο 4. Από πλευράς ανάκλησης τα 2 πρώτα είναι προτιμητέα.

γ) F-Measure

Το F-Measure είναι το αρμονικό μέσο της ανάκλησης και της ακρίβειας. Συγκεκριμένα είναι $F = \frac{2*P*R}{P+R}$. Επομένως για κάθε σύστημα έχουμε :

S1 : $F(S1) = \frac{2*0.417*1}{0.417+1} = \frac{0.834}{1.417} = 0.589$

S2 : $F(S2) = \frac{2*0.417*1}{0.417+1} = \frac{0.834}{1.417} = 0.589$

S3 : $F(S3) = \frac{2*0.667*0.8}{0.667+0.8} = \frac{1.0672}{1.467} = 0.727$

Ο λόγος που χρησιμοποιούμε τον αρμονικό μέσο της ανάκλησης και ακρίβειας είναι επειδή υψηλή τιμή F-Measure επιτυγχάνεται όταν έχουμε υψηλό R και υψηλό P. Επομένως προτιμητέο σύστημα με βάση το F-Measure είναι το τελευταίο σύστημα.

δ) R-Precision

R-Precision είναι ακρίβεια ενός συστήματος στην R θέση της διάταξης της απάντησης σε μία επερώτηση που έχει R συναφή έγγραφα. Γνωρίζουμε ότι για την επερώτηση q υπάρχουν 5 συναφή έγγραφα στην συλλογή. Άρα R=5

S1 : Για το σύστημα **S1** στις 5 πρώτες θέσεις βρίσκονται 3 συναφή έγγραφα και 2 μη-συναφή. Επομένως

$$R-\Pr ecision(S1)=\frac{3}{5}=0.6$$

S2 : Για το σύστημα **S2** στις 5 πρώτες θέσεις βρίσκεται 1 συναφές έγγραφο και 4 μη-συναφή. Επομένως

$$R-\Pr ecision(S2)=\frac{1}{5}=0.2$$

S3 : Για το σύστημα **S3** στις 5 πρώτες θέσεις βρίσκονται 4 συναφή έγγραφα και 1 μη-συναφές. Επομένως

$$R-\Pr ecision(S3)=\frac{4}{5}=0.8$$

Από το R-Precision καταλαβαίνουμε εάν ένα σύστημα επιστρέφει στις πρώτες θέσεις πολλά συναφή έγγραφα. Όσο πιο μεγάλο είναι τόσο πιο "πυκνά" είναι τα συναφή έγγραφα στις πρώτες θέσεις του αποτελέσματος. Σύμφωνα λοιπόν με το R-Measure προτιμητέο σύστημα είναι το **S3.**

ε) Fallout

Το Fallout είναι ο λόγος των μη-συναφών εγγράφων που έχουν ανακληθεί προς τον συνολικό αριθμό των μη-συναφών εγγράφων. Συνολικά η συλλογή μας περιλαμβάνει 40 έγγραφα εκ των οποίων 5 είναι συναφή. Άρα τα μη συναφή έγγραφα της συλλογής μας είναι 35.

S1 : Το σύστημα **S1** ανακτά συνολικά 12 έγγραφα εκ των οποίων 7 είναι μη συναφή οπότε

$$Fallout(S1)=\frac{7}{35}=0.2$$

S1 : Το σύστημα **S2** ανακτά συνολικά 12 έγγραφα εκ των οποίων 7 είναι μη συναφή οπότε

$$Fallout(S2)=\frac{7}{35}=0.2$$

S3 : Για Το σύστημα **S3** ανακτά συνολικά 6 έγγραφα εκ των οποίων 2 είναι μη συναφή οπότε

$$Fallout(S3) = \frac{2}{35} = 0.057$$

Όσο πιο μικρό είναι το Fallout για ένα σύστημα τόσο πιο λίγα συναφή έγγραφα επιστρέφει. Άρα από πλευράς Fallout το **S3** είναι καλύτερο μιας και επιστρέφει λιγότερα μη-συναφή έγγραφα για την επερώτηση q.

Άσκηση 2 (4 βαθμοί)

Σχεδιάστε τις καμπύλες ακρίβειας/ανάκλησης (P/R curves) των συστημάτων της προηγούμενης άσκησης. Για κάθε σύστημα δώστε 2 γραφήματα: ένα που να απεικονίζει τα P/R σημεία όπως προκύπτουν από τις απαντήσεις, και ένα χρησιμοποιώντας κανονικοποιημένα επίπεδα ανάκλησης (standard recall levels). Αν βλέπατε μόνο αυτά τα γραφήματα (και όχι τις απαντήσεις) θα μπορούσατε να επιλέξετε το καλύτερο σύστημα;

Λύση

S1:

Αρχικά πρέπει να βρούμε τα σημεία της καμπύλης ακρίβειας/ανάκλησης επομένως κοιτάζουμε σε ποιες θέσεις έχουμε συναφή έγγραφα.
Το **S1** ανακτά 5 συναφή έγγραφα οπότε:

1ο συναφές: $R(S1) = \frac{1}{5} = 0.2$ $P(S1) = \frac{1}{1} = 1$

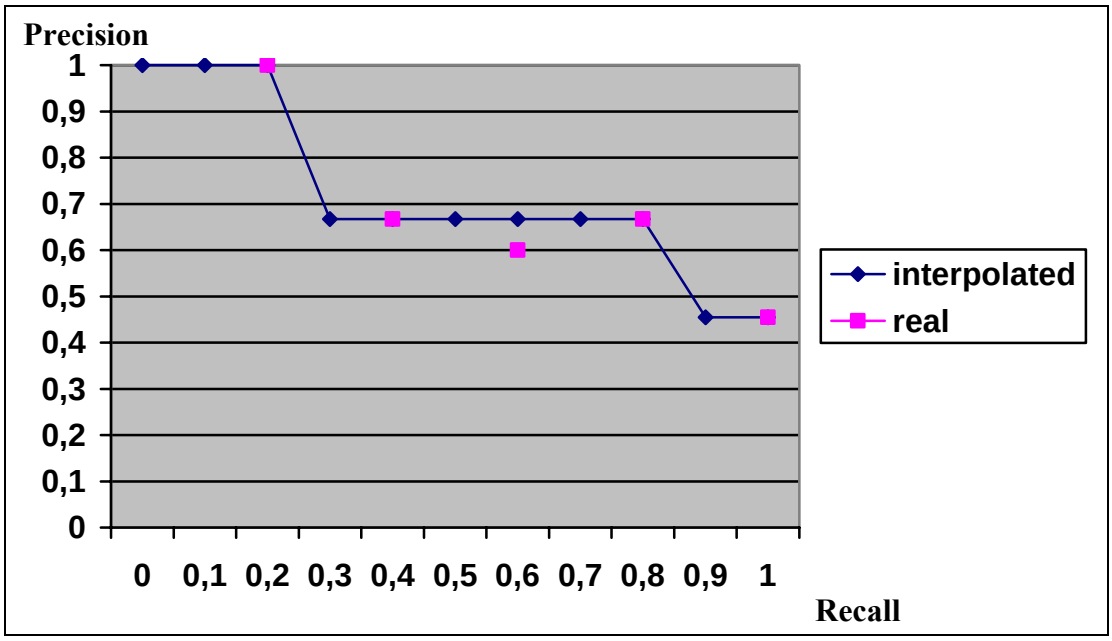
2ο συναφές: $R(S1) = \frac{2}{5} = 0.4$ $P(S1) = \frac{2}{3} = 0.667$

3ο συναφές: $R(S1) = \frac{3}{5} = 0.6$ $P(S1) = \frac{3}{5} = 0.6$

4ο συναφές: $R(S1) = \frac{4}{5} = 0.8$ $P(S1) = \frac{4}{6} = 0.667$

5ο συναφές: $R(S1) = \frac{5}{5} = 1$ $P(S1) = \frac{5}{11} = 0.455$

Και για την κανονικοποίηση χρησιμοποιούμε τα καθιερωμένα επίπεδα ανάκλησης
$r_j = \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$



S2:

Βρίσκουμε ξανά τα σημεία της καμπύλης ανάκλησης/ακρίβειας.
Το **S2** ανακτά 5 συναφή έγγραφα οπότε:

1ο συναφές: $R(S2)=\frac{1}{5}=0.2$ $P(S2)=\frac{1}{5}=0.2$

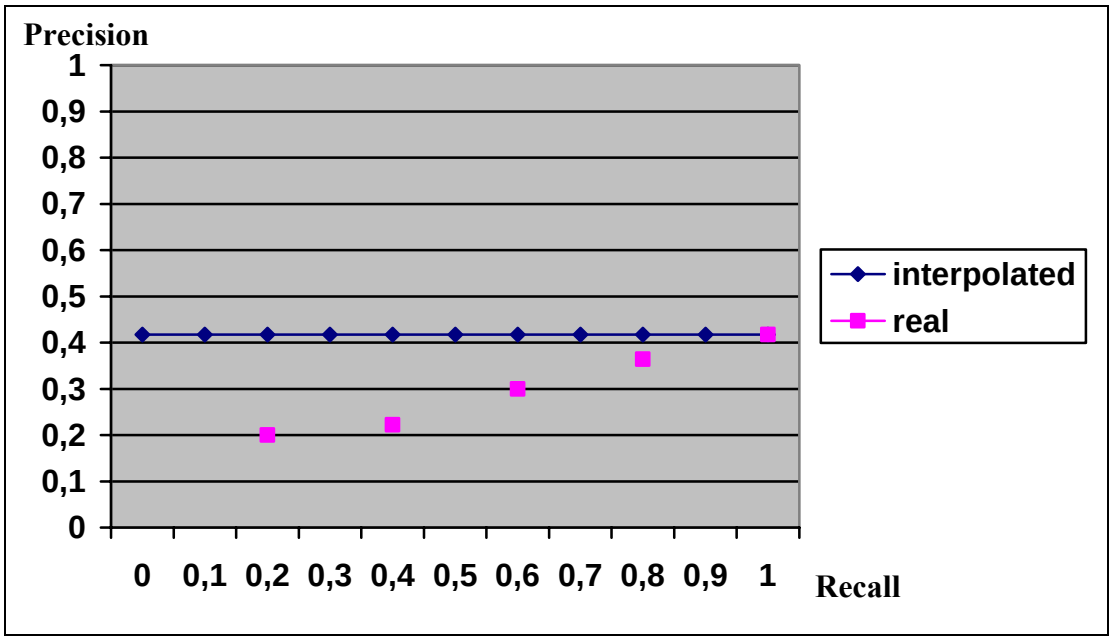
2ο συναφές: $R(S2)=\frac{2}{5}=0.4$ $P(S2)=\frac{2}{9}=0.222$

3ο συναφές: $R(S2)=\frac{3}{5}=0.6$ $P(S2)=\frac{3}{10}=0.3$

4ο συναφές: $R(S2)=\frac{4}{5}=0.8$ $P(S2)=\frac{4}{11}=0.364$

5ο συναφές: $R(S2)=\frac{5}{5}=1$ $P(S2)=\frac{5}{12}=0.417$

Και χρησιμοποιώντας τα ίδια επίπεδα κανονικοποίησης με παραπάνω προκύπτει το παρακάτω γράφημα :



S3:

Βρίσκουμε ξανά τα σημεία της καμπύλης ανάκλησης/ακρίβειας.
Το **S3** ανακτά 4 συναφή έγγραφα οπότε:

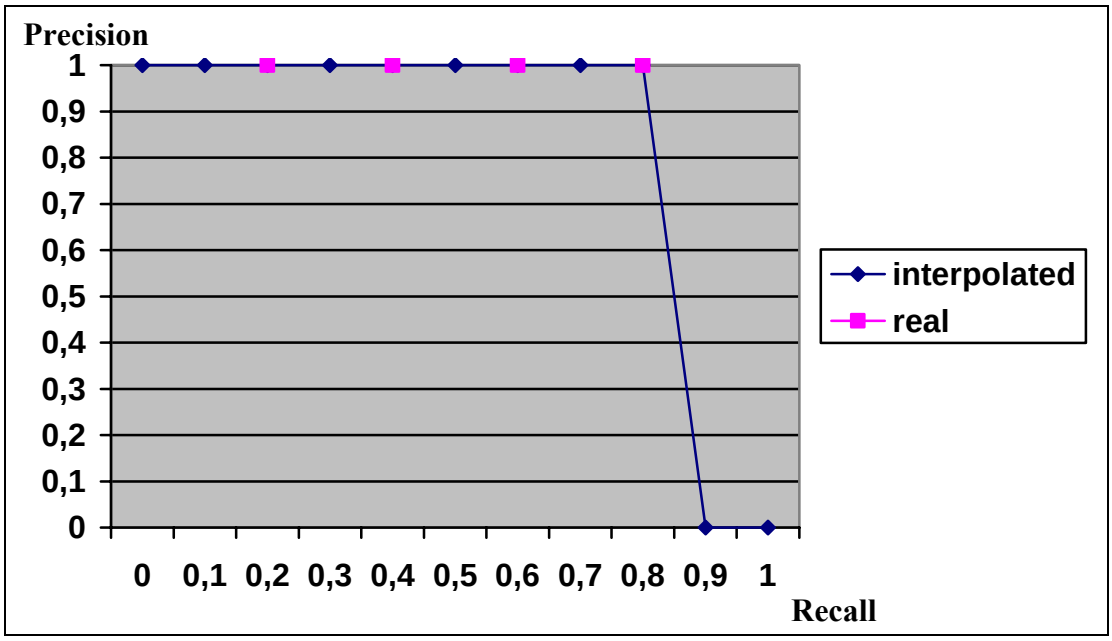
1ο συναφές: $R(S3) = \frac{1}{5} = 0.2$ $P(S3) = \frac{1}{1} = 1$

2ο συναφές: $R(S3) = \frac{2}{5} = 0.4$ $P(S3) = \frac{2}{2} = 1$

3ο συναφές: $R(S3) = \frac{3}{5} = 0.6$ $P(S3) = \frac{3}{3} = 1$

4ο συναφές: $R(S3) = \frac{4}{5} = 0.8$ $P(S3) = \frac{4}{4} = 1$

Και χρησιμοποιώντας τα ίδια επίπεδα κανονικοποίησης με παραπάνω προκύπτει το παρακάτω γράφημα :



Αυτό που παρατηρούμε είναι ότι για το σύστημα **S1** έχει υψηλές τιμές ακρίβειας για χαμηλές τιμές της ανάκλησης και οι τιμές της ακρίβειας όσο η ανάκληση ανεβαίνει προς το 1 μειώνεται μέχρι κάποιο όριο (0,455). Σε γενικές γραμμές το σύστημα αυτό έχει αρκετά καλή συμπεριφορά. Το σύστημα **S2** έχει χαμηλές τιμές ακρίβειας για σχεδόν όλα τα επίπεδα ανάκλησης. Το σύστημα αυτό μας επιστρέφει μεγάλο πλήθος μη-συναφών εγγράφων από τα πρώτα κιόλας αποτελέσματα. Το σύστημα **S3** τέλος έχει εξαιρετικές τιμές ακρίβειας για χαμηλά και μεσαία επίπεδα ανάκλησης και μετά μειώνεται στο μηδέν η ακρίβεια. Αυτό πρακτικά σημαίνει ότι τα πρώτα αποτελέσματα της αναζήτησης θα είναι συναφή έγγραφα και από ένα σημείο και μετά θα είναι μη-συναφή. Χρησιμοποιώντας τα γραφήματα με τις κανονικοποιημένες τιμές είναι πιο ευδιάκριτα τα παραπάνω. Βλέποντας τα γραφήματα λοιπόν μπορούμε να επιλέξουμε το καλύτερο αναζητώντας αυτό που προσεγγίζει την πάνω δεξιά γωνία (εκεί όπου ακρίβεια =1 και ανάκληση =1). Αυτό θα μπορούσε να γίνει μετρώντας το εμβαδό των γραφημάτων στην κάτω από την γραμμή του γραφήματος περιοχή. Έτσι εδώ το σύστημα **S3** βγαίνει ως καλύτερο.

Άσκηση 3 (2 βαθμοί)

Έστω ότι η συλλογή αξιολόγησης αποτελείται από 200 έγγραφα $\{d_1,\ldots,d_{200}\}$ και γνωρίζουμε ότι υπάρχουν 3 έγγραφα της συλλογής, συγκεκριμένα τα $\{d_1, d_2, d_3\}$, που είναι συναφή με την επερώτηση q. Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων S1, S2 και S3 τα οποία επιστρέφουν ως απάντηση έγγραφα συνοδευμένα από ένα βαθμό συνάφειας.

Υποβάλλουμε σε κάθε σύστημα την επερώτηση q και λαμβάνουμε τις εξής απαντήσεις:

Ans(S1,q) = $<d_1, \{d_2, d_{100} – d_{200}\}, d_3>$

Ans(S2,q) = $<d_1, d_2, d_3>$

Ans(S3,q) = $<\{d_1, d_8\}, d_2, d_3>$

Η απάντηση $<\{d_1, d_8\}, d_2, d_3>$ σημαίνει ότι τα d_1, d_8 ισοβαθμούν στην πρώτη θέση (άρα

έλαβαν τον μεγαλύτερο βαθμό συνάφειας). Η απάντηση <d_1, {d_2, d_{100} –d_{200}}, d_3> σημαίνει ότι το d_1 έλαβε το μεγαλύτερο βαθμό, ενώ μετά ακολουθεί μια ομάδα από 102 έγγραφα τα οποία ισοβαθμούν, και στο τέλος της κατάταξης βρίσκεται το d_3. Για κάθε ένα από τα 3 συστήματα απαντήστε τα ακόλουθα ερωτήματα:
(α) Ποια είναι η R-ακρίβεια (R-precision);
(β) Ποιο είναι το αναμενόμενο μήκος αναζήτησης για να βρούμε 2 συναφή;
(γ) Ποιο είναι το μέσο αναμενόμενο μήκος αναζήτησης;

Λύση

α) R-ακρίβεια (R-precision)

Η R-ακρίβεια για το **S1** σύστημα , η ακρίβεια δηλαδή στην R θέση της διάταξης , όπου εδώ R=3. Όμως στην 3η θέση υπάρχει ένα block από έγγραφα με ίδιο βαθμό συνάφειας και μέσα σε αυτά βρίσκεται και ένα συναφές έγγραφο το d2. Αν το d2 βρίσκεται στην 2η ή στην 3η θέση τότε $R-\Pr esicion(S1)=\frac{2}{3}=0.667$. Αν όμως αυτό βρίσκεται σε κάποια άλλη θέση τότε $R-\Pr esicion(S1)=\frac{1}{3}=0.333$.

Πρέπει λοιπόν να λάβουμε υπ' όψιν τις διαφορετικές μεταθέσεις του d2. Οι συνολικοί συνδυασμοί εγγράφων που μπορούν να βρίσκονται στις θέσεις 2 και 3 είναι 102*101= 10302.
Το d2 μπορεί να βρίσκεται στην 2η θέση με 101 συνδυασμούς από τους παραπάνω.
Το d2 μπορεί να βρίσκεται στην 3η θέση με 101 συνδυασμούς από τους παραπάνω.

Επομένως με πιθανότητα $\frac{202}{10302}=0{,}0196$ το R-Presicion(S1) = 0.667 ενώ με πιθανότητα $\frac{10100}{10302}=0{,}9804$ το R-Presicion(S1) = 0.333.

Συνολικά το R-Presicion(S1) =(0,0196 * 0,667) + (0,9804 * 0,333) = 0,0131 + 0,3265 = 0,3396

Η R-ακρίβεια για το **S2** σύστημα είναι $R-\Pr esicision(S2)=\frac{3}{3}=1$

Η R-ακρίβεια για το **S3** σύστημα είναι $R-\Pr esicision(S3)=\frac{2}{3}=0.667$

Ό λόγος είναι προφανής . Κοιτάζοντας τα 3 πρώτα αποτελέσματα που επιστρέφει κάθε σύστημα βλέπουμε ότι:

Ans(S1,q) = <**d_1**, {**d_2**, d_{100} …}, …>
Ans(S2,q) = <**d_1**, **d_2**, **d_3**>
Ans(S3,q) = <{**d_1**, d_8},**d_2**, …>

β) αναμενόμενο μήκος αναζήτησης για να βρούμε 2 συναφή αποτελέσματα

Γνωρίζουμε ότι μήκος αναζήτησης είναι το πλήθος των μη συναφών εγγράφων τα οποία πρέπει να αναζητήσουμε μέχρι να βρούμε 2 συναφή έγγραφα.

S1

Βλέπουμε ότι στην 2η θέση βρίσκονται 102 έγγραφα με ίδιο βαθμό. Αν το d2 βρίσκεται στην 2η θέση τότε θα έχουμε μήκος αναζήτησης 2. Αν βρίσκεται στην θέση 3 θα έχουμε μήκος αναζήτησης 3 κ.ο.κ. Πρέπει λοιπόν να λάβουμε υπ'όψιν όλες τις πιθανές θέσεις που μπορεί να έχει το d2. Ανάλογα με την θέση του d2 λοιπόν το μήκος αναζήτησης θα είναι 2 , 3, 4, 5, …, 103 και το μέσο μήκος αναζήτησης για να έχουμε 2 συναφή αποτελέσματα θα είναι 52.5

S2

Εδώ είναι πιο ξεκάθαρα τα πράγματα επειδή δεν έχουμε ένα έγγραφα με τον ίδιο βαθμό συνάφειας όταν βρίσκουμε το 2ο συναφές έγγραφο. Εδώ το μήκος αναζήτησης για να βρούμε 2 συναφή έγγραφα είναι 2 και το αναμενόμενο μήκος αναζήτησης είναι το ίδιο.

S3

Το μήκος αναζήτησης εδώ όταν βρούμε 2 συναφή έγγραφα είναι 3 . και το ίδιο είναι και το αναμενόμενο μήκος αναζήτησης.

γ) μέσο αναμενόμενο μήκος αναζήτησης

Χρειάζεται για κάθε σύστημα να φτιάξουμε τον πίνακα των συναφών εγγράφων και του αναμενόμενου μήκος αναζήτησης. Έχουμε λοιπόν

S1

Συναφή έγγραφα	Αν. μήκος αναζήτησης
1	1
2	52.5
3	104

Το μέσο μήκος αναζήτησης είναι :

$$\frac{\frac{1}{1}+\frac{52.5}{2}+\frac{104}{3}}{3}=\frac{1+26.25+34.667}{3}=\frac{61.917}{3}=20.639$$

Από το οποίο παρατηρούμε ότι πρέπει να ψάξουμε πολλά επιπλέον έγγραφα εκτός των συναφών για να ανακτήσουμε ένα πλήθος συναφών.

S2

Συναφή έγγραφα	Αν. μήκος αναζήτησης
1	1
2	2
3	3

Το μέσο μήκος αναζήτησης είναι :

$$\frac{\frac{1}{1}+\frac{2}{2}+\frac{3}{3}}{3}=\frac{3}{3}=1$$

Είναι το ιδανικότερο σύστημα αφού μέσο μήκος αναζήτησης 1 σημαίνει ότι δεν χρειάζεται να αναζητήσουμε καθόλου μη συναφή έγγραφα για να ανακτήσουμε συναφή.

S3

Συναφή έγγραφα	Αν. μήκος αναζήτησης
1	1,5
2	3
3	4

Το μέσο μήκος αναζήτησης είναι :

$$\frac{\frac{1.5}{1}+\frac{3}{2}+\frac{4}{3}}{3}=\frac{1.5+1.5+1.33}{3}=\frac{4.33}{3}=1.44$$

Το οποίο σημαίνει ότι χρειάζεται να αναζητήσουμε 44 % επιπλέον μη συναφή έγγραφα για να βρούμε και συναφή.