Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2005-2006 Εαρινό Εξάμηνο

1η Σειρά Ασκήσεων (*Αξιολόγηση Αποτελεσματικότητας Ανάκτησης*)

Άσκηση 1 (4 βαθμοί)

Θεωρείστε μια συλλογή αξιολόγησης που αποτελείται από 40 έγγραφα βάσει της οποίας θέλουμε να αξιολογήσουμε την αποτελεσματικότητα τριών συστημάτων S1, S2 και S3. Για το λόγο αυτό υποβάλλουμε σε κάθε σύστημα μια επερώτηση q και λαμβάνουμε τις εξής απαντήσεις:

Ans(S1,q) = <R N N N R R N N N N R N>
Ans(S2,q) = <N N N N R N N N N R R R>
Ans(S3,q) = <R R R N N N >

Το αριστερότερο στοιχείο της κάθε απάντησης παριστάνει το υψηλότερα διαβαθμισμένο έγγραφο, αυτό που το σύστημα υπολόγισε ως το πιο συναφές με την επερώτηση q. Συμβουλευόμενοι την συλλογή αξιολόγησης διακρίνουμε τα στοιχεία των απαντήσεων σε συναφή (R) και μη συναφή (N). Έστω ότι ξέρουμε ότι το σύνολο των εγγράφων της συλλογής που είναι συναφή με την επερώτηση q είναι 5.
Συγκρίνεται τα τρία αυτά συστήματα ως προς τα εξής μέτρα:
(α) Ακρίβεια (Precision)
(β) Ανάκληση (Recall)
(γ) F-Measure
(δ) R-Ακρίβεια (R-Precision)
(ε) Fallout
Για κάθε μέτρο σχολιάστε το αποτέλεσμα της σύγκρισης.

Λύση

Η συλλογή αξιολόγησής μας αποτελείται από 40 έγγραφα. Επιπλέον γνωρίζουμε ότι το σύνολο των εγγράφων της συλλογής που είναι συναφή με την επερώτηση q είναι 5. Άρα έχουμε ως προς:

α) Ακρίβεια (Precision)

Η ακρίβεια ορίζεται σαν το πηλίκο των ευρεθέντων συναφών εγγράφων προς τα ευρεθέντα έγγραφα. Άρα για τα τρία συστήματα έχουμε τις εξής τιμές:

S_1 Σύστημα: Το σύνολο των ευρεθέντων εγγράφων είναι 12, από τα οποία 4 έγγραφα είναι συναφή. Άρα **$P(S_1) = 4/12 = 0.333$**

S_2 Σύστημα: Το σύνολο των ευρεθέντων εγγράφων είναι 12, από τα οποία 4 έγγραφα είναι συναφή. Άρα **$P(S_2) = 4/12 = 0.333$**

S_3 Σύστημα: Το σύνολο των ευρεθέντων εγγράφων είναι 6, από τα οποία 3 έγγραφα είναι συναφή. Άρα **$P(S_3) = 3/6 = 0.5$**

Από τα αποτελέσματα βλέπουμε ότι τα **S_1** και **S_2** συστήματα, αν και δίνουν διαφορετικές απαντήσεις για την επερώτηση q, έχουν την ίδια ακρίβεια. Αντίθετα το **S_3** αν και έχει λιγότερα συναφή έγγραφα στα αποτελέσματά του σε σχέση με τα

δύο προηγούμενα συστήματα, έχει μεγαλύτερη ακρίβεια αφού μας επιστρέφει στην απάντησή του λιγότερα μη συναφή έγγραφα.

β) Ανάκληση (Retrieval)

Η ανάκληση ορίζεται σαν το πηλίκο των ευρεθέντων συναφών εγγράφων προς τα συναφή έγγραφα. Άρα για τα τρία συστήματα έχουμε τις εξής τιμές:

S_1 Σύστημα: Το σύνολο των ευρεθέντων συναφών εγγράφων είναι 4, ενώ όλα τα συναφή έγγραφα είναι 5. Άρα **$R(S_1) = 4/5 = 0.8$**

S_2 Σύστημα: Το σύνολο των ευρεθέντων συναφών εγγράφων είναι 4, ενώ όλα τα συναφή έγγραφα είναι 5. Άρα **$R(S_2) = 4/5 = 0.8$**

S_3 Σύστημα: Το σύνολο των ευρεθέντων συναφών εγγράφων είναι 4, ενώ όλα τα συναφή έγγραφα είναι 5. Άρα **$R(S_3) = 3/5 = 0.6$**

Από τα αποτελέσματα βλέπουμε ότι τα **S_1** και **S_2** συστήματα, αν και δίνουν διαφορετικές απαντήσεις για την επερώτηση q, έχουν την ίδια ανάκληση, η οποία είναι υψηλότερη του **S_3**. Ο λόγος είναι ότι τα δύο πρώτα συστήματα επιστρέφουν περισσότερα συναφή έγγραφα από ότι το τελευταίο.

γ) F-Measure

Το F-Measure είναι ένα μέτρο που λαμβάνει υπόψη την ακρίβεια και την ανάκληση και ορίζεται σαν το αρμονικό μέσο (harmonic mean) της ανάκλησης και της ακρίβειας σύμφωνα με τον τύπο 2*P*R/(P+R). Χρησιμοποιώντας τα αποτελέσματα από τα α) και β) έχουμε:

S_1 Σύστημα: Έχουμε **$P(S_1) = 0.333$** και **$R(S_1) = 0.8$.**
Άρα **$F(S_1)$** = 2 * 0.333 * 0.8 / (0.333 + 0.8) = **0.470.**

S_2 Σύστημα: Έχουμε **$P(S_2) = 0.333$** και **$R(S_1) = 0.8$.**
Άρα **$F(S_2)$** = 2 * 0.333 * 0.8 / (0.333 + 0.8) = **0.470.**

S_3 Σύστημα: Έχουμε **$P(S_3) = 0.5$** και **$R(S_3) = 0.6$.**
Άρα **$F(S_3)$** = 2 * 0.5 * 0.6 / (0.5 + 0.6) = **0.545.**

Από τα αποτελέσματα βλέπουμε ότι το τρίτο σύστημα έχει μεγαλύτερο F-Measure, λόγω του ότι χρειαζόμαστε παράλληλα υψηλό P και υψηλό R για να πάρουμε μεγάλη τιμή. Άρα το **S_3** είναι καλύτερο.

δ) R-Precision

Το R-Precision μέτρο είναι η ακρίβεια στην R θέση της διάταξης της απάντησης μιας επερώτησης όπου R ο αριθμός των συναφών εγγράφων της συλλογής μας. Άρα με βάση τη διάταξη που έχουμε από την εκφώνηση και γνωρίζοντας ότι ο αριθμός των συναφών εγγράφων της συλλογής μας είναι 5 έχουμε:

S_1 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_1** βλέπουμε ότι μέχρι και την $5^η$ θέση έχουμε βρεί 2 συναφή έγγραφα. Άρα **R-Precision(S_1) = 2/5 = 0.4**

S_2 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_2** βλέπουμε ότι μέχρι και την $5^η$ θέση έχουμε βρεί 1 συναφές έγγραφο. Άρα **R-Precision(S_2) = 1/5 = 0.2**

S_1 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_3** βλέπουμε ότι μέχρι και την $5^η$ θέση έχουμε βρεί 3 συναφή έγγραφα**.** Άρα **R-Precision(S_3) = 3/5 = 0.6**

Από τα παραπάνω βγάζουμε το συμπέρασμα ότι το σύστημα **S_3** έχει υψηλότερο R-Precision, πράγμα που ουσιαστικά σημαίνει ότι μας δίνει στις πρώτες θέσεις πολλά συναφή έγγραφα. Αντίθετα τα άλλα δύο συστήματα δίνουν στην αρχή και πολλά μη συναφή έγγραφα.

ε) Fallout

Το Fallout μέτρο είναι ο αριθμός των μη συναφών εγγράφων τα οποία ανακλήθηκαν από ένα σύστημα προς τον συνολικό αριθμό των μη συναφών εγγράφων της συλλογής μας.

S_1 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_1** βλέπουμε ότι έχουμε βρεί 8 μη συναφή έγγραφα, από το 35 μη συναφή έγγραφα της συλλογής μας. Άρα **Fallout(S_1) = 8/35 = 0.228**

S_2 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_2** βλέπουμε ότι έχουμε βρεί 8 μη συναφή έγγραφα, από το 35 μη συναφή έγγραφα της συλλογής μας. Άρα **Fallout(S_2) = 8/35 = 0.228**

S_3 Σύστημα: Από τη διάταξη της απάντησης του συστήματος **S_3** βλέπουμε ότι έχουμε βρεί 8 μη συναφή έγγραφα, από το 35 μη συναφή έγγραφα της συλλογής μας. Άρα **Fallout(S_3) = 3/35 = 0.085**

Από τα παραπάνω βγάζουμε το συμπέρασμα ότι το σύστημα **S_3** έχει το μικρότερο Fallout, πράγμα πολύ λογικό αφού μας δίνει το μικρότερο αριθμό μη συναφών εγγράφων.

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;

<u>Λύση</u>

Σε αυτή την άσκηση μας ζητείται να σχεδιάσουμε τις καμπύλες ακρίβειας ανάκλησης (P/R curves) των συστημάτων της προηγούμενης άσκησης. Συγκεκριμένα μας ζητούνται δύο γραφήματα, ένα με βάση τις απαντήσεις των συστημάτων και ένα χρησιμοποιώντας κανονικοποιημένα επίπεδα ανάκλησης (standard recall levels).

Για το **S_1 Σύστημα** έχουμε:

1^o συναφές έγγραφο	**$P(S_1) = 1/1 = 1$**	**$R(S_1) = 1/5 = 0.2$**
2^o συναφές έγγραφο	**$P(S_1) = 2/5 = 0.4$**	**$R(S_1) = 2/5 = 0.4$**
3^o συναφές έγγραφο	**$P(S_1) = 3/6 = 0.5$**	**$R(S_1) = 3/5 = 0.6$**
4^o συναφές έγγραφο	**$P(S_1) = 4/11 = 0.363$**	**$R(S_1) = 4/5 = 0.8$**

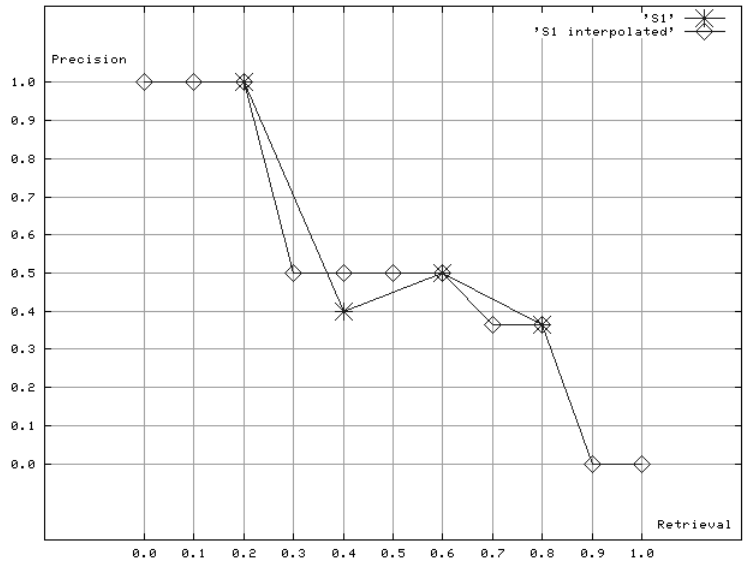
Για τις interpolated τιμές έχουμε:

$r_0 = 0.0$, $r_1 = 0.1$, $r_2 = 0.2$, $r_3 = 0.3$, $r_4 = 0.4$, $r_5 = 0.5$, $r_6 = 0.6$,
$r_7 = 0.7$, $r_8 = 0.8$, $r_9 = 0.9$, $r_{10} = 1.0$

και

$P(r_0) = 1.0$, $P(r_1) = 1.0$, $P(r_2) = 1.0$, $P(r_3) = 0.5$, $P(r_4) = 0.5$, $P(r_5) = 0.5$, $P(r_6) = 0.5$,
$P(r_7) = 0.363$, $P(r_8) = 0.363$, $P(r_9) = 0.0$, $P(r_{10}) = 0.0$

Οι δύο καμπύλες ακρίβειας και ανάκλησης με βάση τα στοιχεία της απάντησης του **S_1** συστήματος αλλά και οι interpolated τιμές δίνονται παρακάτω:



Για το **S_2 Σύστημα** έχουμε:

1^o συναφές έγγραφο	**$P(S_2) = 1/5 = 0.2$**	**$R(S_2) = 1/5 = 0.2$**
2^o συναφές έγγραφο	**$P(S_2) = 2/10 = 0.2$**	**$R(S_2) = 2/5 = 0.4$**
3^o συναφές έγγραφο	**$P(S_2) = 3/11 = 0.272$**	**$R(S_2) = 3/5 = 0.6$**
4^o συναφές έγγραφο	**$P(S_2) = 4/12 = 0.333$**	**$R(S_2) = 4/5 = 0.8$**

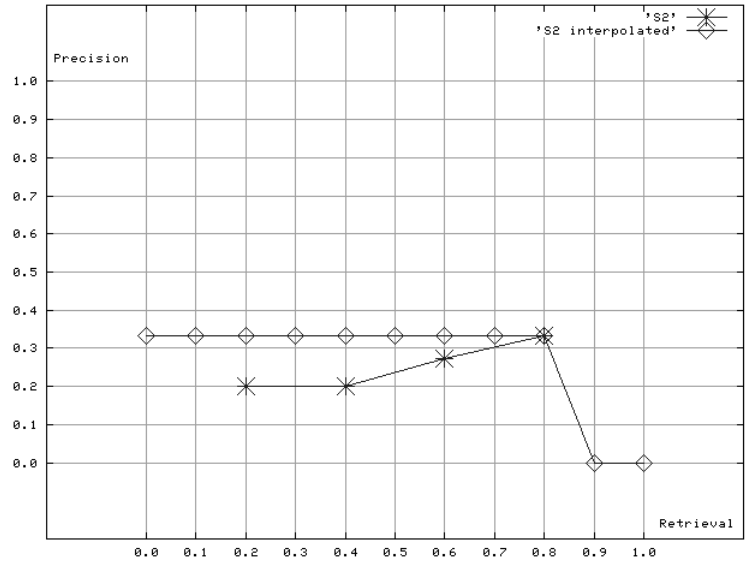
Για τις interpolated τιμές έχουμε:

$r_0 = 0.0$, $r_1 = 0.1$, $r_2 = 0.2$, $r_3 = 0.3$, $r_4 = 0.4$, $r_5 = 0.5$, $r_6 = 0.6$,
$r_7 = 0.7$, $r_8 = 0.8$, $r_9 = 0.9$, $r_{10} = 1.0$

και

$P(r_0) = 0.333$, $P(r_1) = 0.333$, $P(r_2) = 0.333$, $P(r_3) = 0.333$, $P(r_4) = 0.333$, $P(r_5) = 0.333$,
$P(r_6) = 0.5$,$P(r_7) = 0.333$, $P(r_8) = 0.333$, $P(r_9) = 0.0$, $P(r_{10}) = 0.0$

Οι δύο καμπύλες ακρίβειας και ανάκλησης με βάση τα στοιχεία της απάντησης του **S_2** συστήματος αλλά και οι interpolated τιμές δίνονται παρακάτω:



Για το **S_3 Σύστημα** έχουμε:

1ο συναφές έγγραφο	**$P(S_3) = 1/1 = 1$**	**$R(S_3) = 1/5 = 0.2$**
2ο συναφές έγγραφο	**$P(S_3) = 2/2 = 1$**	**$R(S_3) = 2/5 = 0.4$**
3ο συναφές έγγραφο	**$P(S_3) = 3/3 = 1$**	**$R(S_3) = 3/5 = 0.6$**

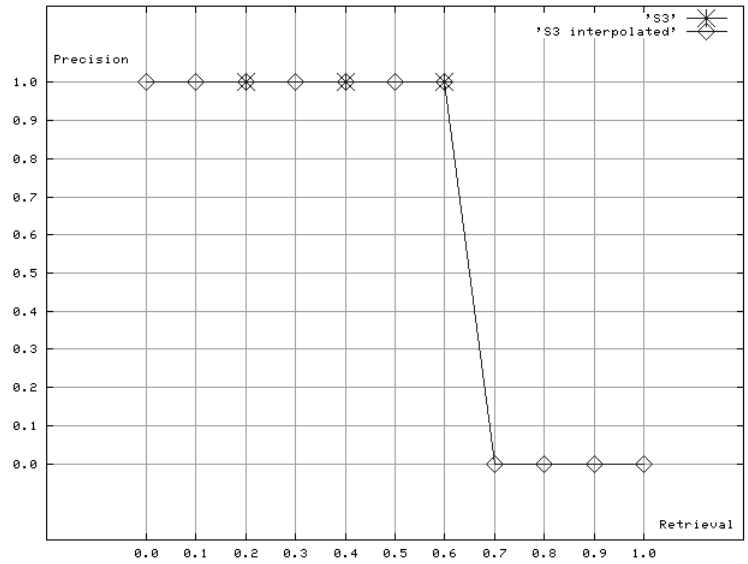
Για τις interpolated τιμές έχουμε:

$r_0 = 0.0$, $r_1 = 0.1$, $r_2 = 0.2$, $r_3 = 0.3$, $r_4 = 0.4$, $r_5 = 0.5$, $r_6 = 0.6$,
$r_7 = 0.7$, $r_8 = 0.8$, $r_9 = 0.9$, $r_{10} = 1.0$

και

$P(r_0) = 1.0$, $P(r_1) = 1.0$, $P(r_2) = 1.0$, $P(r_3) = 1.0$, $P(r_4) = 1.0$, $P(r_5) = 1.0$,
$P(r_6) = 1.0$,$P(r_7) = 0.0$, $P(r_8) = 0.0$, $P(r_9) = 0.0$, $P(r_{10}) = 0.0$

Οι δύο καμπύλες ακρίβειας και ανάκλησης με βάση τα στοιχεία της απάντησης του **S_3** συστήματος αλλά και οι interpolated τιμές δίνονται παρακάτω:



Από τις παραπάνω γραφικές παραστάσεις μπορούμε να βγάλουμε συμπεράσματα για τα τρία συστήματά μας δίχως να ξέρουμε τις απαντήσεις που μας έχει δώσει το καθένα.

Χρησιμοποιώντας τις καμπύλες Precision/Recall, βλέπουμε ότι το σύστημα $\mathbf{S_3}$ είναι το καλύτερο από όλα για μικρές και μεσαίες τιμές ανάκλησης, μιας και βλέπουμε ότι η ακρίβειά του βρίσκεται συνεχώς στο 1, πράγμα που σημαίνει ότι μας δίνει συναφής απαντήσεις, χωρίς να μας δώσει ούτε μία ασυναφή. Για μεγάλες τιμές ανάκλησης όμως βλέπουμε ότι η απόδοση του πέφτει,αφού αδυνατεί στη συνέχεια να μας δώσει κάποια συναφή απάντηση. Δεύτερο έρχεται το σύστημα $\mathbf{S_1}$ αφού βλέπουμε ότι μας δίνει ικανοποιητικές τιμές για χαμηλές τιμές ανάκλησεις και υπερτερεί του $\mathbf{S_3}$ στο γεγονός ότι έχει καλύτερη απόδοση για μεγαλύτερες τιμές ανάκλησης. Τρίτο έρχεται το $\mathbf{S_2}$ το οποίο έχει πολύ χειρότερη συμπεριφορά από το $\mathbf{S_1}$, για μικρές τιμές ανάκλησης και παρόμοια συμπεριφορά με αυτό για υψηλές τιμές ανάκλησης. Άλλωστε γενικά καλύτερο σύστημα είναι αυτό που προσεγγίζει σε μεγαλύτερο βαθμό την πάνω δεξιά γωνία του γραφήματος, η οποία ισχύει όμως μόνο σε ένα ιδανικό σύστημα.

Αντίστοιχα αποτελέσματα βγάζουμε και από τις interpolated τιμές. Εδώ θα μπορούσαμε να χρησιμοποιήσουμε το γεγονός ότι καλύτερο σύστημα είναι αυτό το οποίο έχει το μέγιστο εμβαδό στα συγκεκριμένα γραφήματα το οποίο βρίσκεται όσο γίνεται πιο αριστερά στο γράφημα, από την άποψη ότι δίνει συναφή έγγραφα στις πρώτες απαντήσεις του. Και πάλι η σειρά θα ήταν $\mathbf{S_3}$, $\mathbf{S_1}$, $\mathbf{S_2}$.

Άσκηση 3 (2 βαθμοί)

Έστω ότι έχουμε μια συλλογή N εγγράφων και K συστήματα ανάκτησης πληροφοριών. Θέλουμε να αξιολογήσουμε την αποτελεσματικότητα των συστημάτων αυτών, ώστε να επιλέξουμε το καλύτερο, αλλά δυστυχώς δεν υπάρχει καμία συλλογή αξιολόγησης. Επίσης δεν μπορούμε να κάνουμε οι ίδιοι μια άτυπη αξιολόγηση (ήτοι να υποβάλουμε σε κάθε σύστημα ένα σύνολο επερωτήσεων και να κρίνουμε τις

αποκρίσεις τους ως προς την ακρίβεια τους) διότι είτε δεν έχουμε τον απαιτούμενο χρόνο για κάτι τέτοιο (π.χ. φανταστείτε την περίπτωση που K=1000), ή διότι δεν μπορούμε να το κάνουμε (π.χ. τα έγγραφα είναι γραμμένα στην κινεζική γλώσσα).
(α) Προτείνετε τρόπους αντιμετώπισης αυτού του προβλήματος και δικαιολογείστε τις απαντήσεις σας (συγκεκριμένα τις υποθέσεις υπό τις οποίες αυτό που προτείνετε θα είχε νόημα).
(β) Έστω ότι έχετε X Ευρώ στη διάθεση σας και ότι υπάρχει ένας … κινέζος ο οποίος με 1 Ευρώ μπορεί να σας απαντήσει αν ένα έγγραφο d είναι συναφές ή όχι με μια επερώτηση q. Πως θα τον χρησιμοποιούσατε για την αξιολόγηση των συστημάτων;

<u>Λύση</u>

α)
Αντικειμενικά δεν υπάρχει τρόπος να αποφανθούμε για το ποιο σύστημα είναι καλύτερο από τα υπόλοιπα. Μπορούμε όμως θεωρήσουμε ως καλύτερο εκείνο το σύστημα του οποίου η λειτουργία είναι πιο κοντά στην λειτουργία όλων των συστημάτων. Μια τέτοια υπόθεση δεν είναι αβάσιμη υπό την έννοια ότι σε πολλές περιπτώσεις της καθημερινής μας ζωής, έτσι ορίσουμε το "αντικειμενικό" (δηλαδή πλειοψηφικά).

Η μέθοδος που μπορούμε να ακολουθήσουμε για αυτό το σκοπό είναι η εξής:
1. Επιλέγω τυχαία ένα έγγραφο.
2. Το στέλνω ως επερώτηση σε κάθε σύστημα.
3. Κατόπιν ενοποιώ τις διατάξεις που έλαβα από όλα τα συστήματα και ορίζω την συνισταμένη διάταξη. Ο τρόπος με τον οποίο θα ενοποιήσουμε τις διατάξεις είναι κρίσιμος για το αποτέλεσμα της αξιολόγησης των συστημάτων. Μια μέθοδος ενοποίησης διατάξεων που λαμβάνει υπόψιν τη σειρά του κάθε εγγράφου σε κάθε διάταξη είναι η ενοποίηση διατάξεων κατά Borda (Διάλεξη 15):

Αν κάθε πηγή S_i επιστρέφει ένα διατεταγμένο υποσύνολο O_i του συνόλου όλων των εγγράφων, τότε
- αν o_j ανήκει στο O_i, τότε $r_i(o_j)$ = θέση του o_j στο O_i
- αλλιώς, $r_i(o_j) = F+1$, όπου $F = \max\{|O_1|,\ldots,|O_k|\}$

4. Κατόπιν βαθμολογώ κάθε σύστημα ανάλογα με την απόσταση της απάντησης του από την συνισταμένη. Δεδομένης της χρήσης της μεθόδου του Borda, μπορούμε να ορίσουμε την απόσταση μεταξύ των δύο διατάξεων (δηλ. της απάντησης του συστήματος από την ενοποιημένη διάταξη) ως εξής:
$dist(i,j) = \Sigma_{o \text{ in } Obj} |r_i(o) - r_j(o)|$

Μπορούμε επίσης να τροποποιήσουμε τον παραπάνω τύπο, ώστε να πάρουμε το άθροισμα των διαφορών των θέσεων των documents σε κάθε μια από τις δύο διατάξεις μόνο για τα documents που βρίσκονται στις πρώτες M (M<N) θέσεις της ενοποιημένης διάταξης, αφού κυρίως μας ενδιαφέρει τι συμβαίνει στις κορυφαίες θέσεις της απάντησης ενός συστήματος.

Εναλλακτικά θα μπορούσαμε να θεωρήσουμε τα M πρώτα έγγραφα της ενοποιημένης διάταξης ως το σύνολο (δηλ. να αγνοήσουμε τη διάταξή τους) των συναφών εγγράφων και κατόπιν να αξιολογήσουμε τα συστήματα βάσει των μέτρων αξιολόγησης αποτελεσματικότητας (Διάλεξη 2).

5. Μπορώ να επαναλάβω τη διαδικασία αυτή για πολλά έγγραφα ή για όλα τα έγγραφα της συλλογής.

β)
Ο Κινέζος μπορεί να κρίνει το πολύ Χ έγγραφα (και συνήθως το Χ είναι μικρό). Το κρίσιμο ερώτημα είναι ποια έγγραφα μας συμφέρει να του δώσουμε να κρίνει. Μια απάντηση σε αυτό το ερώτημα θα μπορούσε να δώσει η λύση του (α) ερωτήματος, δηλαδή του δίνουμε έγγραφα που εμφανίζονται ψηλά στην συνισταμένη διάταξη. Συγκεκριμένα, του δίνω τα Χ πρώτα στοιχεία της συνισταμένης διάταξης (που προέκυψε από τις απαντήσεις των συστημάτων σε μια επερώτηση). Έστω Υ εκείνα τα οποία κατά τη γνώμη του είναι συναφή. Κρίνω τα συστήματα βάσει του αν περιέχουν (και μάλιστα ψηλά στην απάντησή τους) τα Υ έγγραφα.

Αν βέβαια το Χ είναι μεγαλύτερο από το Ν, τότε μπορούμε να κάνουμε παραπάνω από μία επερωτήσεις (στην ακραία περίπτωση για κάθε έγγραφο της συλλογής). Όμως ακόμα και αν τα χρήματα δε φτάνουν (συνήθως το Ν είναι πολύ μεγάλο) μπορούμε να επιλέξουμε τα πρώτα Χ/Μ έγγραφα της ενοποιημένης διάταξης και να ρωτήσουμε τον Κινέζο ποια από αυτά είναι συναφή με κάθε μια από τις Μ επερωτήσεις.