Ανάκληση Πληροφορίας

Information Retrieval

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 10η

Αποτίμηση επίδοσης Μηχανών Αναζήτησης

Μέτρα επίδοσης μιας μηχανής αναζήτησης

- Πόσο γρήγορα εκτελεί την διαδικασία indexing;
  - Αριθμός εγγράφων/ώρα
  - (Μέσο μέγεθος εγγράφου)
- Πόσο γρήγορα εκτελεί την αναζήτηση;
  - Καθυστέρηση ως συνάρτηση του μεγέθους του index
- Εκφραστικότητα της γλώσσας υποβολής ερωτημάτων
  - Ικανότητα να εκφραστούν σύνθετες πληροφοριακές ανάγκες
  - Ταχύτητα σε σύνθετα ερωτήματα
- Είναι “φορτωμένο” το user interface;
- Είναι δωρεάν;

Μέτρα επίδοσης μιας μηχανής αναζήτησης

- Όλες οι προηγούμενες μετρικές είναι *μετρήσιμες*: μπορούμε να ποσοτικοποιήσουμε την ταχύτητα/μέγεθος
  - Μπορούμε να καταστήσουμε την εκφραστικότητα μετρίσιμη
- Η κρίσιμη μετρική: "ευχαρίστηση του χρήστη"
  - Η ταχύτητα της απόκρισης/μεγέθους του index είναι παράγοντες
  - Αλλά ακόμη και εάν αποκρίνεται "με την ταχύτητα του φωτός", εάν παρέχει άχρηστες απαντήσεις, δεν συνδράμει στην "ευχαρίστηση του χρήστη"
- Χρειαζόμαστε έναν τρόπο για να ποσοτικοποιήσουμε την "ευχαρίστηση του χρήστη"

Μετρώντας την "ευχαρίστηση του χρήστη"

- Ποιος είναι ο χρήστης τον οποίο προσπαθούμε να ευχαριστήσουμε; Εξαρτάται από το περιβάλλον της μελέτης
- Web engine:
  - Ο χρήστης βρίσκει αυτό που αναζητά και επιστρέφει στην μηχανή αναζήτησης
    - Μπορούμε να μετρήσουμε τον ρυθμό επιστροφής των χρηστών
  - Οι χρήστες ολοκληρώνουν τις δραστηριότητές τους – η αναζήτηση ως μέσο και όχι ως αυτοσκοπός
  - Δείτε Russell http://dmrussell.googlepages.com/JCDL-talk-June-2007-short.pdf
- eCommerce site:
  - Ο χρήστης βρίσκει αυτό που ψάχνει και κάνει κάποια αγορά
  - Άρα, μετράμε την ευχαρίστηση του τελικού χρήστη ή του eCommerce site;
  - Μετράμε τον χρόνο για να εκτελεστεί αγορά, ή το ποσοστό των αναζητήσεων που οδηγούν σε αγορά προϊόντος;
- Enterprise (company/govt/academic): Ενδιαφέρον για την "παραγωγικότητα του χρήστη"
  - Πόσο χρόνο σώζουν οι χρήστες όταν αναζητούν πληροφορίες;
  - Κι άλλα κριτήρια σχετιζόμενα με το εύρος προσπέλασης, ασφαλή προσπέλαση, κ.τ.λ.

Ευχαρίστηση: Μετρήσιμη

- Πιο κοινή προσέγγιση: *σχετικότητα (relevance)* των αποτελεσμάτων αναζήτησης
- Αλλά, πώς μετράμε την relevance;
- Θα παρουσιάσουμε μια μεθοδολογία, και κατόπιν θα εξετάσουμε τα σχετιζόμενα ζητήματα
- Η μέτρηση της relevance απαιτεί 3 στοιχεία:
  1. Μια benchmark συλλογή εγγράφων
  2. Μια benchmark συλλογή ερωτημάτων
  3. Μια συνήθη δυαδική αποτίμηση είτε Relevant είτε Nonrelevant για κάθε ερώτημα και κάθε έγγραφο

Αποτίμηση ενός IR συστήματος

- Σημείωση: η **πληροφοριακή ανάγκη** μεταφράζεται σε ένα **ερώτημα**
- Η relevance αποτιμάται ως προς την **πληροφοριακή ανάγκη** και *όχι* ως προς το **ερώτημα**
  - Πληροφοριακή ανάγκη: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine*
  - Ερώτημα: ***wine red white heart attack effective***
- Αποτιμάται κατά πόσο το έγγραφο που ανακτήθηκε απαντά στην πληροφοριακή ανάγκη, και όχι κατά πόσο περιέχει τις προαναφερθείσες λέξεις

Τυπικά relevance benchmarks

- TREC – Το National Institute of Standards and Technology (NIST) υποστηρίζει ένα μεγάλο IR test bed για πολλά χρόνια
- Reuters, και άλλες benchmark συλλογές εγγράφων
- Καθορίζονται "Retrieval tasks"
  - Μερικές φορές ως ερωτήματα
- Άνθρωποι-experts σημειώνουν, για κάθε ερώτημα και κάθε έγγραφο, Relevant ή Nonrelevant
  - Ή τουλάχιστον για ένα υποσύνολο των εγγράφων που κάποιο σύστημα επέστρεψε για το ερώτημα αυτό

Αποτίμηση μη διαβαθμισμένης ανάκτησης: Precision και Recall

- **Precision**: ποσοστό ανακτηθέντων εγγράφων τα οποία είναι σχετικά = P(relevant | retrieved)
- **Recall**: ποσοστό σχετικών εγγράφων που έχουν ανακτηθεί = P(retrieved | relevant)

	Relevant	Nonrelevant
Retrieved	**tp**	**fp**
Not Retrieved	**fn**	**tn**

- Precision: P = tp/(tp + fp)
- Recall: R = tp/(tp + fn)

Μήπως θα έπρεπε να χρησιμοποιήσουμε ως μετρική αποτίμησης την *accuracy*;

- Δεδομένου ενός ερωτήματος, μια μηχανή κατηγοριοποιεί κάθε έγγραφο ως “Relevant” ή “Nonrelevant”
- Η **accuracy** μιας μηχανής: το ποσοστό των ορθών κατηγοριοποιήσεων
  - **(tp + tn) / (tp + fp + fn + tn)**
- Η **accuracy** χρησιμοποιείται συχνά ως μετρική αποτίμησης στις κατηγοριοποιήσεις με machine learning
- Γιατί λοιπόν δεν είναι χρήσιμη στην IR;

Γιατί όχι την accuracy;

- Πώς να χτίσουμε μια φθηνή μηχανή αναζήτησης με 99.9999% accuracy …

- Οι χρήστες που κάνουν IR *επιθυμούν να βρούν κάτι*, και έχουν μια συγκεκριμένη ανοχή στα σκουπίδια

Precision/Recall

- Μπορούμε να έχουμε υψηλή recall (αλλά, χαμηλή precision) ανακτώντας όλα τα έγγραφα για όλα τα ερωτήματα!
- Η precision είναι μια φθίνουσα συνάρτηση του αριθμού των ανακτηθέντων εγγράφων

- Σε ένα καλό σύστημα, η precision ελαττώνεται καθώς αυξάνει ο αριθμός των ανακτηθέντων εγγράφων είτε καθώς αυξάνει η recall
  - Δεν είναι θεώρημα, αλλά μια εμπειρική παρατήρηση

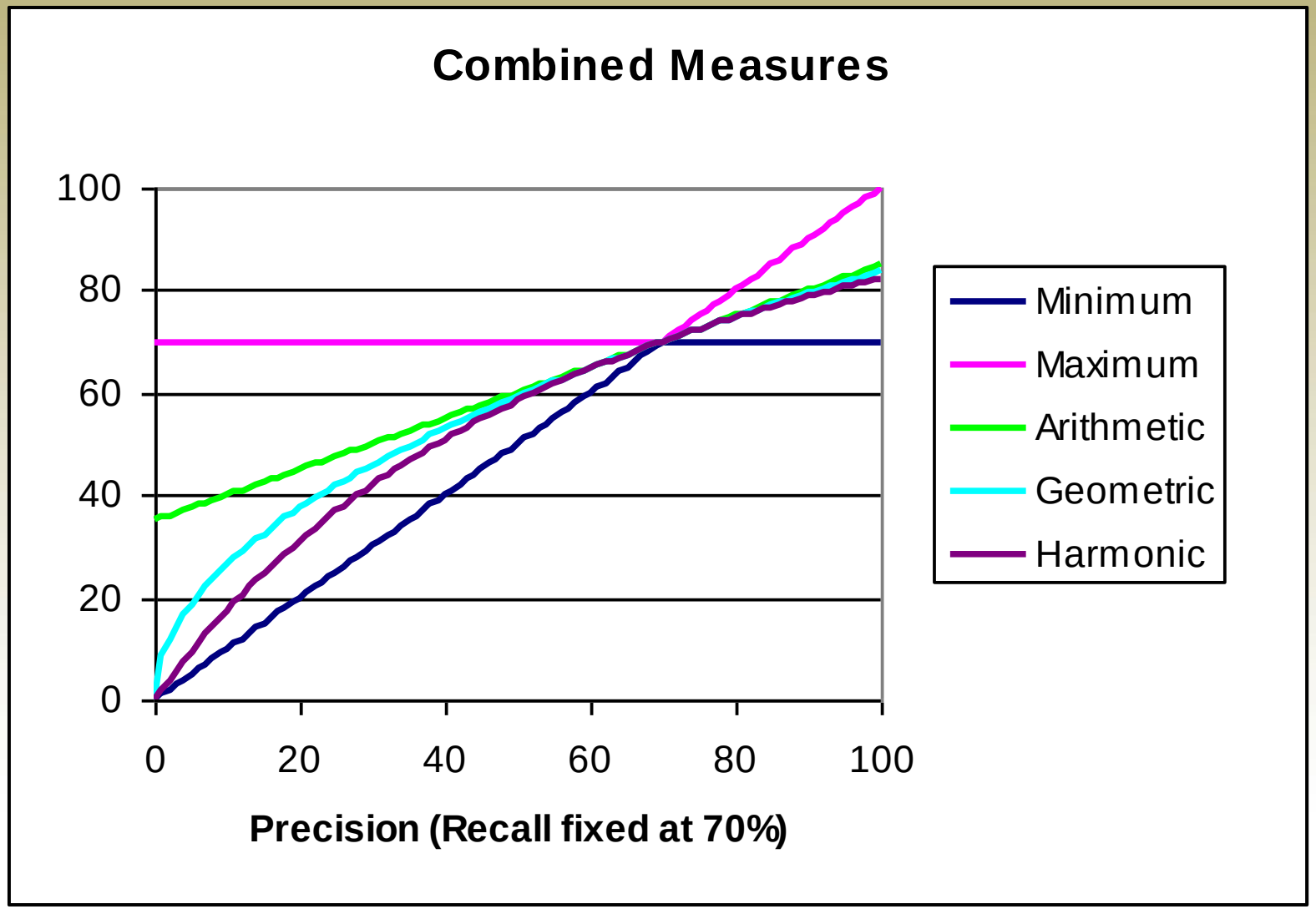
Συνδυαστική μετρική: *F*

- Μια συνδυαστική μετρική που αποτιμά το precision/recall tradeoff είναι το **F measure** (weighted harmonic mean):

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

- Συνήθως χρησιμοποιείται το balanced F_1
  - Δηλαδή, με $\beta = 1$ ή $\alpha = ½$
- Ο harmonic mean είναι συντηρητικός μέσος όρος
  - Δείτε: C.J. van Rijsbergen, *Information Retrieval*

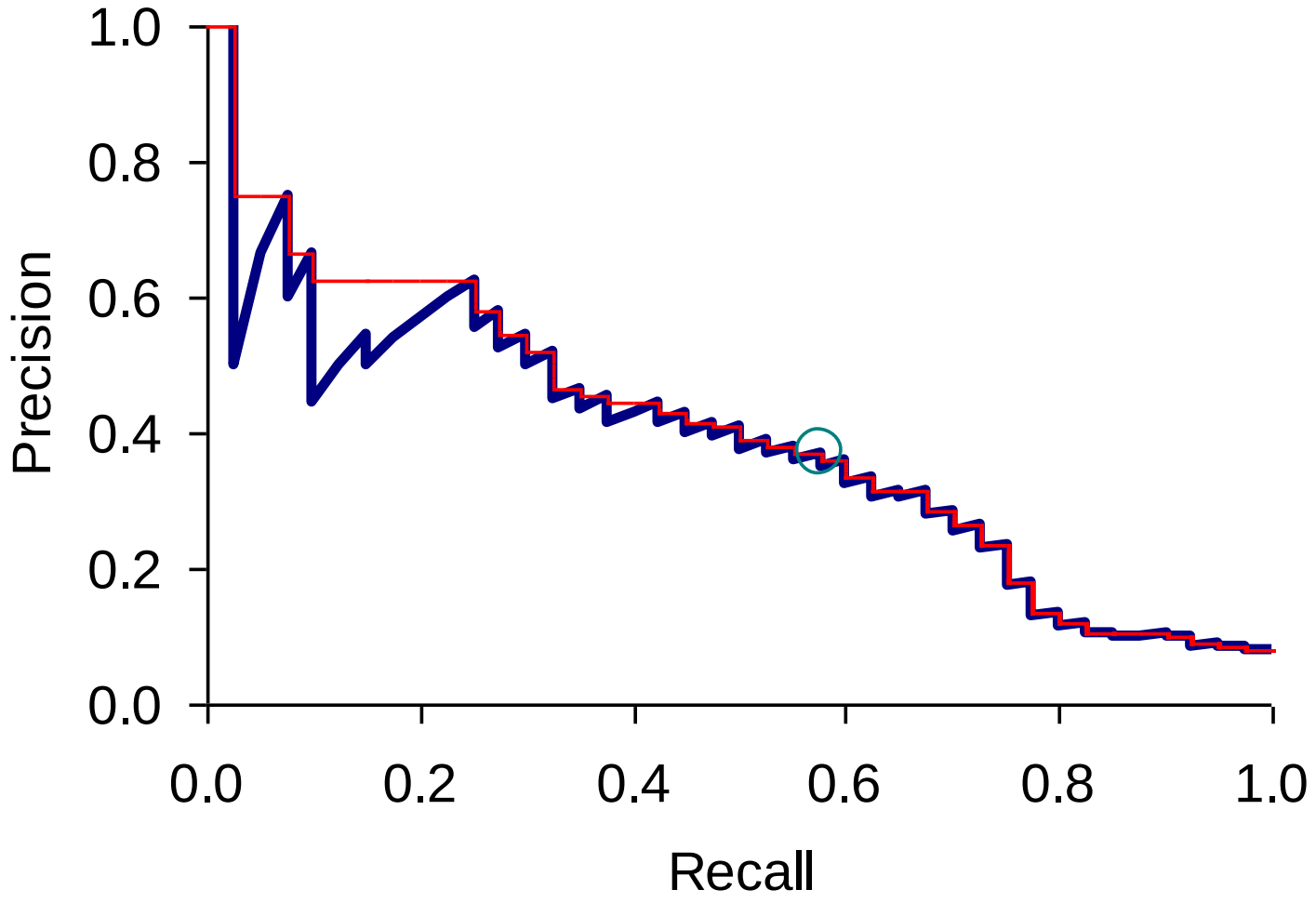
F_1 και άλλες μετρικές



Αποτίμηση διαβαθμισμένων αποτελεσμάτων

- Αποτίμηση διαβαθμισμένων αποτελεσμάτων:
  - Το σύστημα μπορεί να επιστρέψει οποιοδήποτε αριθμό αποτελεσμάτων
  - Εξετάζοντας διάφορους αριθμούς των top returned εγγράφων (δηλ., επίπεδα του recall), ο αποτιμητής μπορεί να φτιάξει την καμπύλη *precision-recall*

Καμπύλη precision-recall

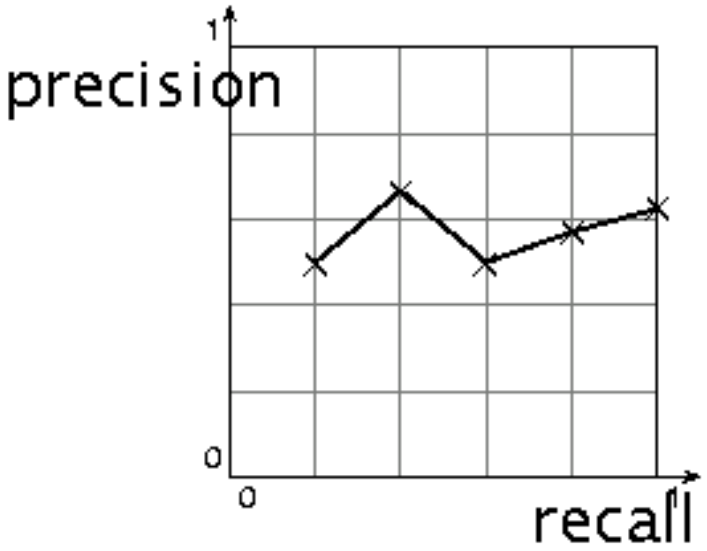


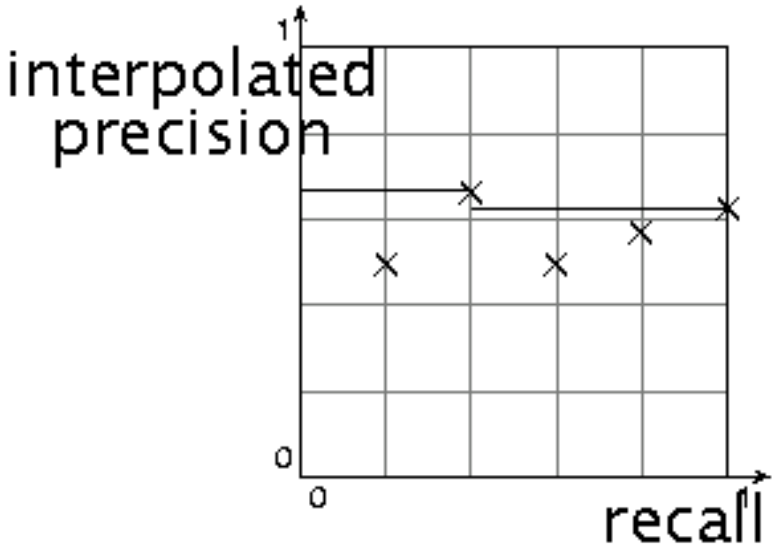
Μέσος όρος από πολλά ερωτήματα

- Το γράφημα precision-recall για ένα μόνο ερώτημα δεν είναι πολύ αξιόπιστο
- Πρέπει να δούμε τους μέσους όρους για πολλά ερωτήματα
- Αλλά, εδώ κρύβεται ένα τεχνικό ζήτημα:
  - Οι υπολογισμοί precision-recall τοποθετούν σημεία στο γράφημα
  - Πώς προσδιορίζουμε μια τιμή (interpolate) μεταξύ σημείων;

Interpolated precision

- Ιδέα: Εάν η τοπική precision αυξάνει καθώς αυξάνει η recall, τότε πρέπει να το λάβουμε υπόψην …
- Έτσι, παίρνουμε το max των precisions στα δεξιά

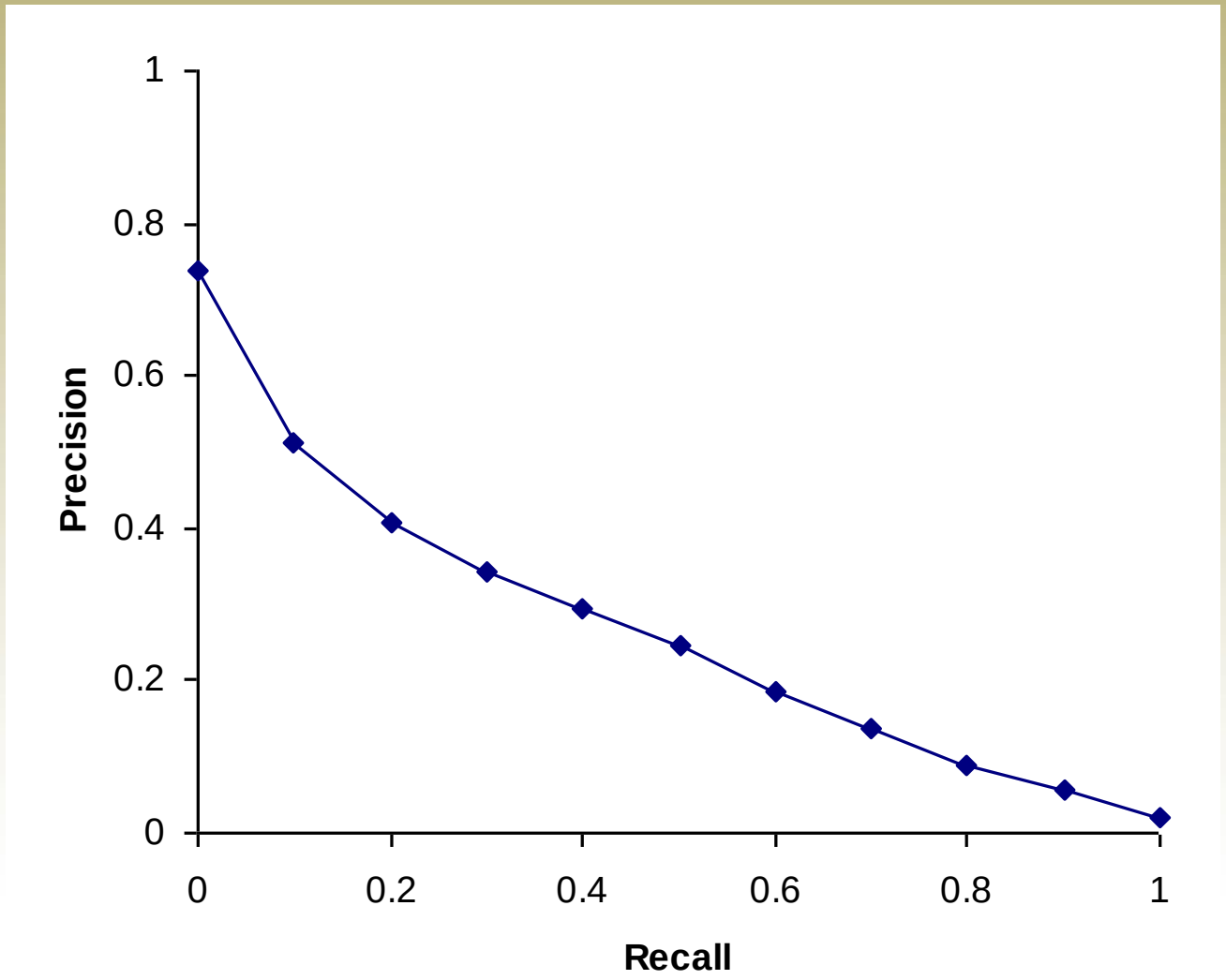




Αποτίμηση: Αθροιστικές μετρικές

- Precision σε fixed retrieval level
  - Precision-at-*k*: Precision στα top *k* αποτελέσματα
  - Ίσως καλή για τις περισσότερες Web αναζητήσεις: όλοι μας επιθυμούμε τα πιο σχετικά αποτελέσματα στις μια-δυο πρώτες σελίδες.
    - Ακραίως: 100%-at-1, το "*I am feeling lucky*" κουμπί
  - Αλλά: δεν είναι πολύ καλό και απαιτεί παράμετρο *k*
- 11-point interpolated average precision
  - Παίρνουμε την precision στα 11 επίπεδα recall μεταβάλλοντας από 0 σε 1 ανά δεκάδα εγγράφων, με χρήση interpolation (η τιμή για το 0 είναι πάντα interpolated!), και υπολογίζουμε τον μέσο όρο τους
  - Αποτιμά την επίδοση σε όλα τα επίπεδα recall

Τυπική (καλή) precision 11 σημείων



Ακόμη περισσότερες μετρικές επίδοσης …

- Mean average precision (MAP)
  - Average της precision για τα top k έγγραφα, κάθε φορά που ανακτάται ένα σχετικό έγγραφο
  - Αποφεύγει την interpolation, χρήση fixed recall επιπέδων
  - Η MAP για συλλογές ερωτημάτων είναι ο αριθμητικός μέσος όρος:

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

- R-precision
  - Έχοντας γνώση (ίσως μερική) ενός συνόλου σχετικών εγγράφων μεγέθους $|Rel|$, τότε υπολογίζουμε την precision των top Rel εγγράφων που επιστράφηκαν
  - Έστω ότι βρέθηκαν r σχετικά έγγραφα
  - Άρα: prec = rec = r/$|Rel|$

Διασπορά

- Για μια συλλογή ελέγχου, είναι σύνηθες ότι ένα σύστημα αποδίδει άσχημα για κάποιες πληροφοριακές ανάγκες (π.χ., MAP = 0.1) και εξαιρετικά για άλλες (π.χ., MAP = 0.7)
- Πράγματι, είναι συνήθης η περίπτωση ότι η διασπορά στην επίδοση του ίδιου συστήματος για διάφορα ερωτήματα είναι μεγαλύτερη από την διασπορά διαφορετικών συστημάτων για το ίδιο ερώτημα
- Δηλαδή, υπάρχουν εύκολες και δύσκολες πληροφοριακές ανάγκες