Introduction to

Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 8: Ανάδραση στην Ανάκτηση Πληροφορίας.

Τι είδαμε στο προηγούμενο μάθημα

- Πως θα αξιολογήσουμε ένα Σύστημα Ανάκτησης Πληροφορίας

Τι θα δούμε σήμερα;

- Πως μπορούμε να βελτιώσουμε τα αποτελέσματα
  - Για βελτίωση της ανάκλησης, π.χ., αναζήτηση για *aircraft* δεν ταιριάζει με το *plane,* ή το *thermodynamic* με *heat*

- Ανάδραση (ανατροφοδότηση) Συνάφειας (Relevance Feedback)

1. Αναδιατύπωση Ερωτήσεων (Query Reformulation)

   Ψευδο-ανάδραση συνάφειας (Pseudo relevance feedback)

2. Επέκταση Επερωτήσεων (Query Expansion)

Γιατί τη χρειαζόμαστε;

- Έχει παρατηρηθεί ότι οι χρήστες των ΣΑΠ δαπανούν πολύ χρόνο *αναδιατυπώνοντας* την αρχική τους ερώτηση προκειμένου να βρουν ικανοποιητικά έγγραφα
- Πιθανές αιτίες
  - ο χρήστης *δε γνωρίζει το περιεχόμενο* των υποκείμενων εγγράφων
  - το *λεξιλόγιο* του χρήστη μπορεί να *διαφέρει* από αυτό της συλλογής
  - η αρχική ερώτηση μπορεί να είναι *πιο γενική ή πιο ειδική* από αυτή που θα έπρεπε (καταλήγοντας είτε σε πάρα πολλά ή σε πολύ λίγα έγγραφα)
- Η αρχική ερώτηση μπορεί να θεωρηθεί ως η πρώτη προσπάθεια έκφρασης της πληροφοριακής ανάγκης του χρήστη
- Ανάγκη για τεχνικές αντιμετώπισης αυτού του προβλήματος

Πως;

(1) Βελτίωση της αρχικής ερώτησης

(2) Χρήση Προφίλ Χρήστη (personalization)

(3) Βελτίωση αναπαράστασης κειμένων

(4) Βελτίωση αλγορίθμου (μοντέλου) ανάκτησης

- Τα (2), (3), (4) έχουν πιο μόνιμο αποτέλεσμα (επηρεάζουν την απάντηση και των επόμενων ερωτήσεων)
- Εδώ θα εστιάσουμε στο (1)

Αναδιατύπωση ερώτησης

Αναβάρυνση των Όρων (Term Reweighting):

Αύξηση των βαρών των όρων που εμφανίζονται στα συναφή/επιθυμητά έγγραφα και μείωση των βαρών των όρων που εμφανίζονται στα μη-συναφή/επιθυμητά έγγραφα.

Επέκταση ερώτησης (Query Expansion):

Προσθήκη νέων όρων στην ερώτηση (π.χ. από γνωστά συναφή έγγραφα)

Τεχνικές βελτίωσης της ερώτησης

(α) τεχνικές που *απαιτούν είσοδο* από τον χρήστη

(β) τεχνικές *που δεν απαιτούν είσοδο*

(β1) που βασίζονται στα κορυφαία έγγραφα που ανακτήθηκαν

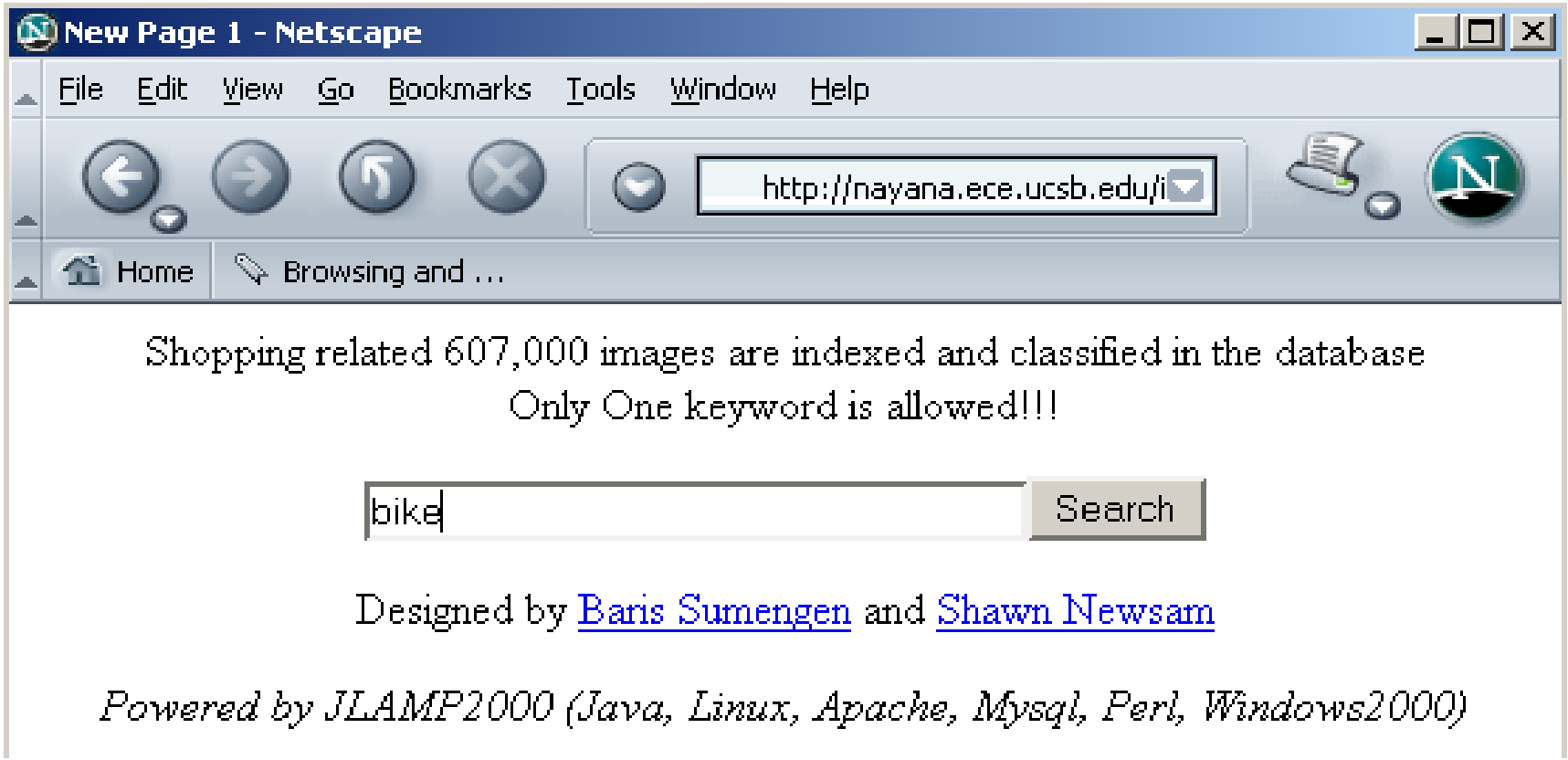
(β2) που βασίζονται σε όλα τα έγγραφα της συλλογής

Ανατροφοδότηση Συνάφειας (Relevance Feedback)

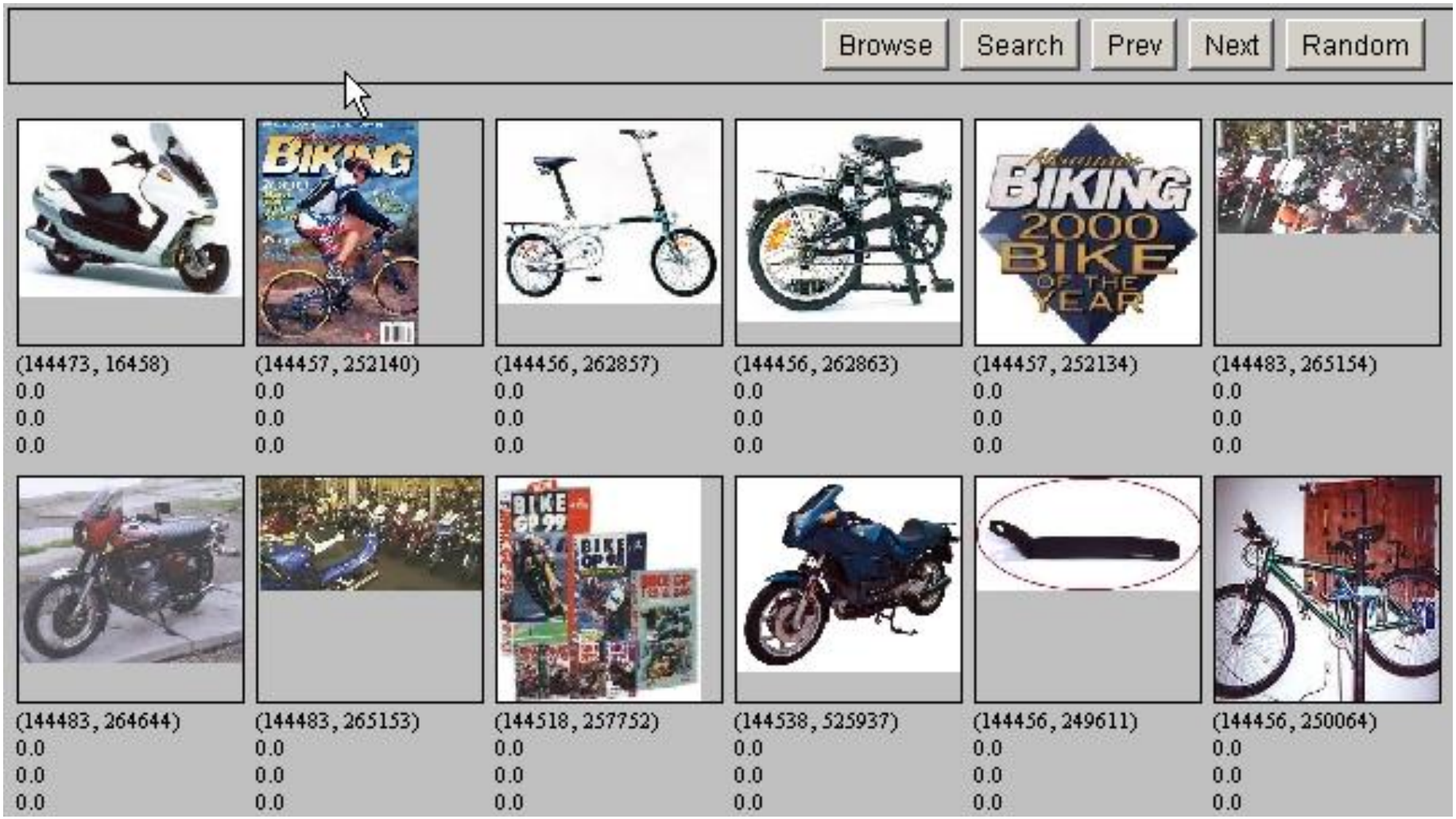
- Ιδέα: Μπορεί να είναι δύσκολο να διατυπωθεί μια καλή ερώτηση αν δεν είναι επαρκώς γνωστή η συλλογή
- Ανατροφοδότηση συνάφειας (relevance feedback): ο χρήστης παρέχει ανατροφοδότηση (feedback) σχετικά με τη συνάφεια των αρχικών αποτελεσμάτων
  - Ο χρήστης υποβάλει μια (σύντομη, απλή) ερώτηση
  - Ο χρήστης χαρακτηρίζει τα έγγραφα που ανακτώνται ως συναφή ή μη συναφή
  - Το σύστημα υπολογίζει μια καλύτερη αναπαράσταση της ανάγκης πληροφορίας βασισμένη στην ανατροφοδότηση.
  - Η ανατροφοδότηση συνάφειας μπορεί να επαναληφθεί μία ή περισσότερες φορές (iterations).

Παράδειγμα

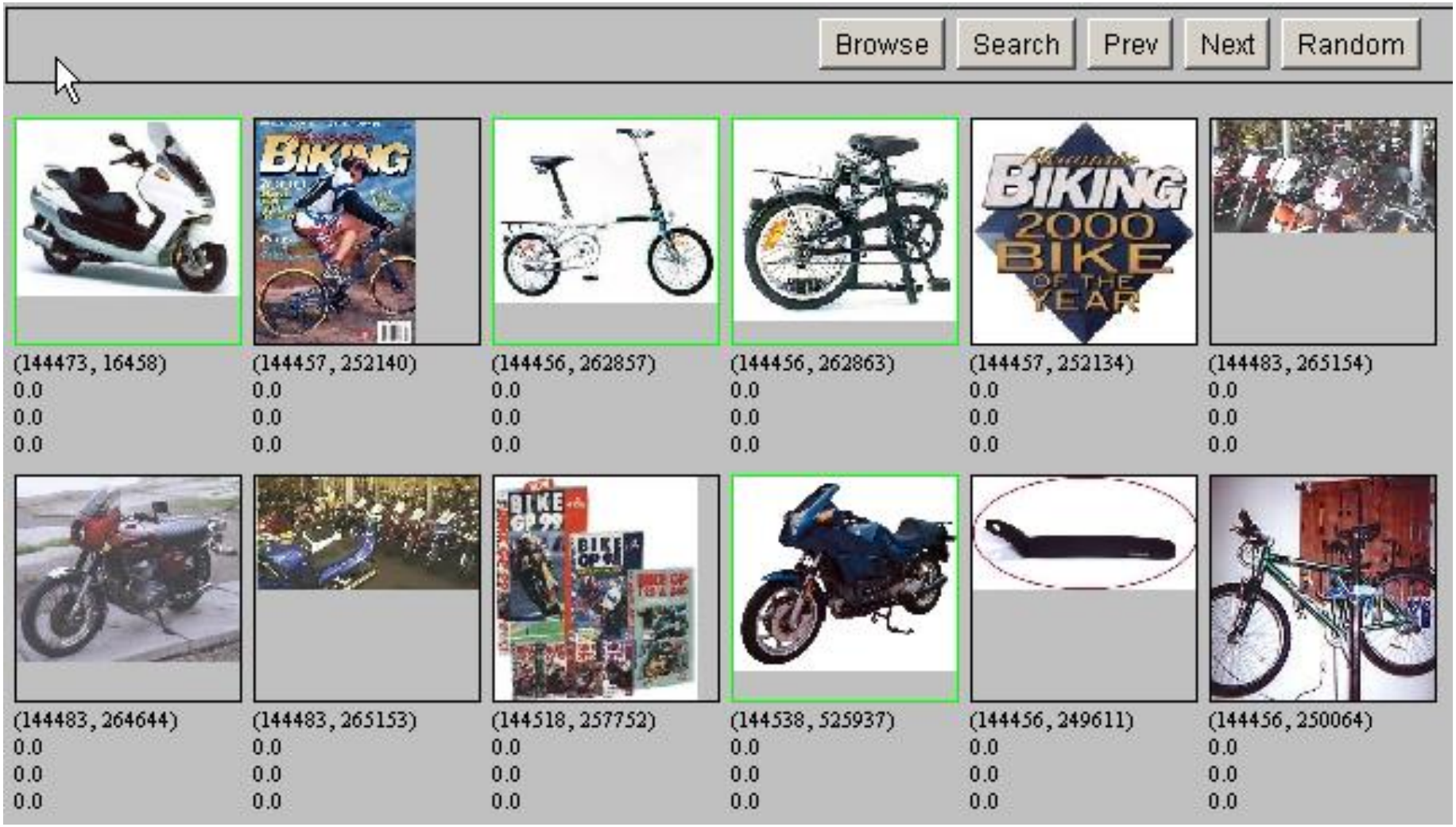
- Μηχανή αναζήτησης εικόνων
- http://nayana.ece.ucsb.edu/imsearch/imsearch.html



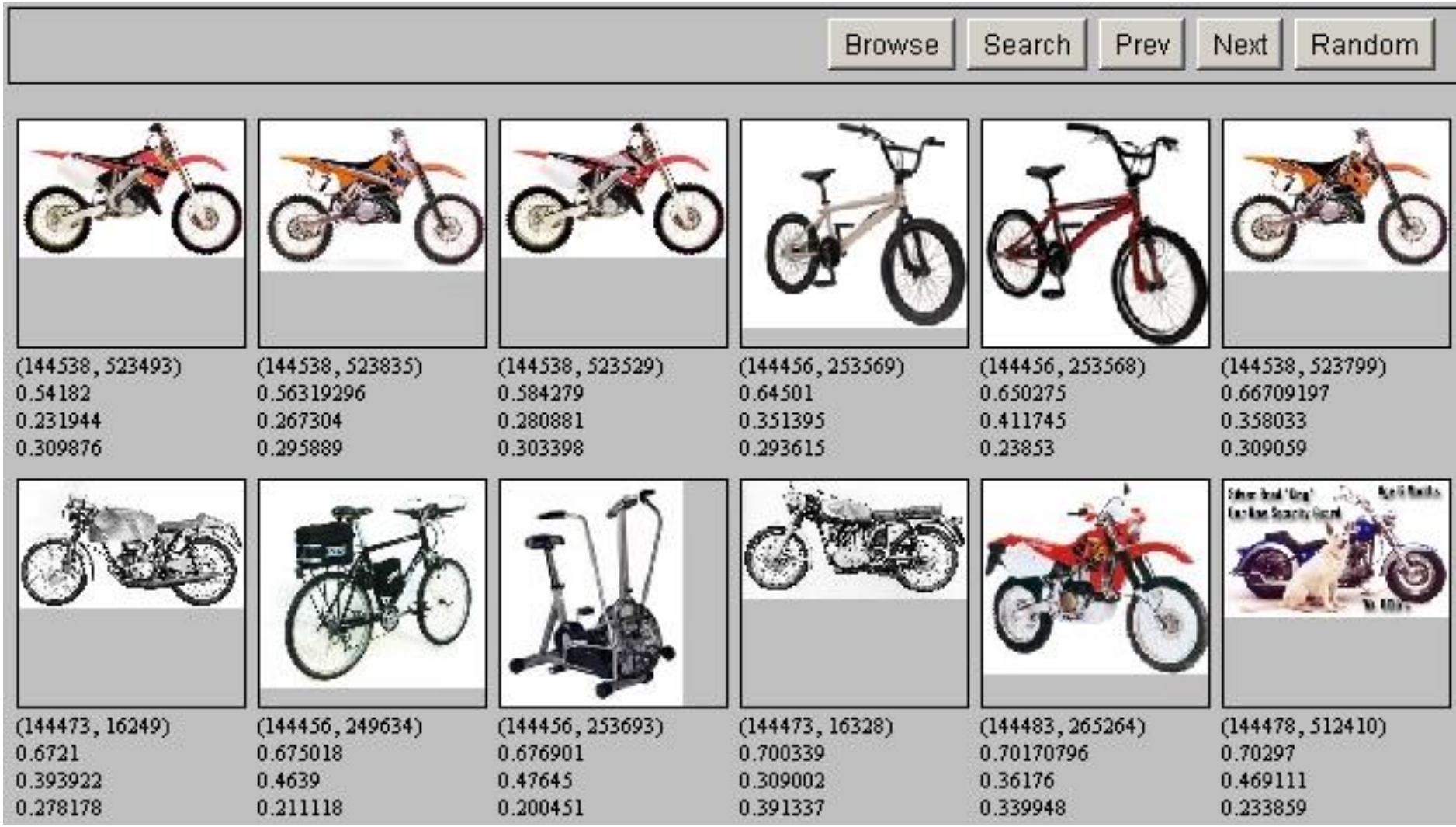
Αποτελέσματα της αρχικής ερώτησης

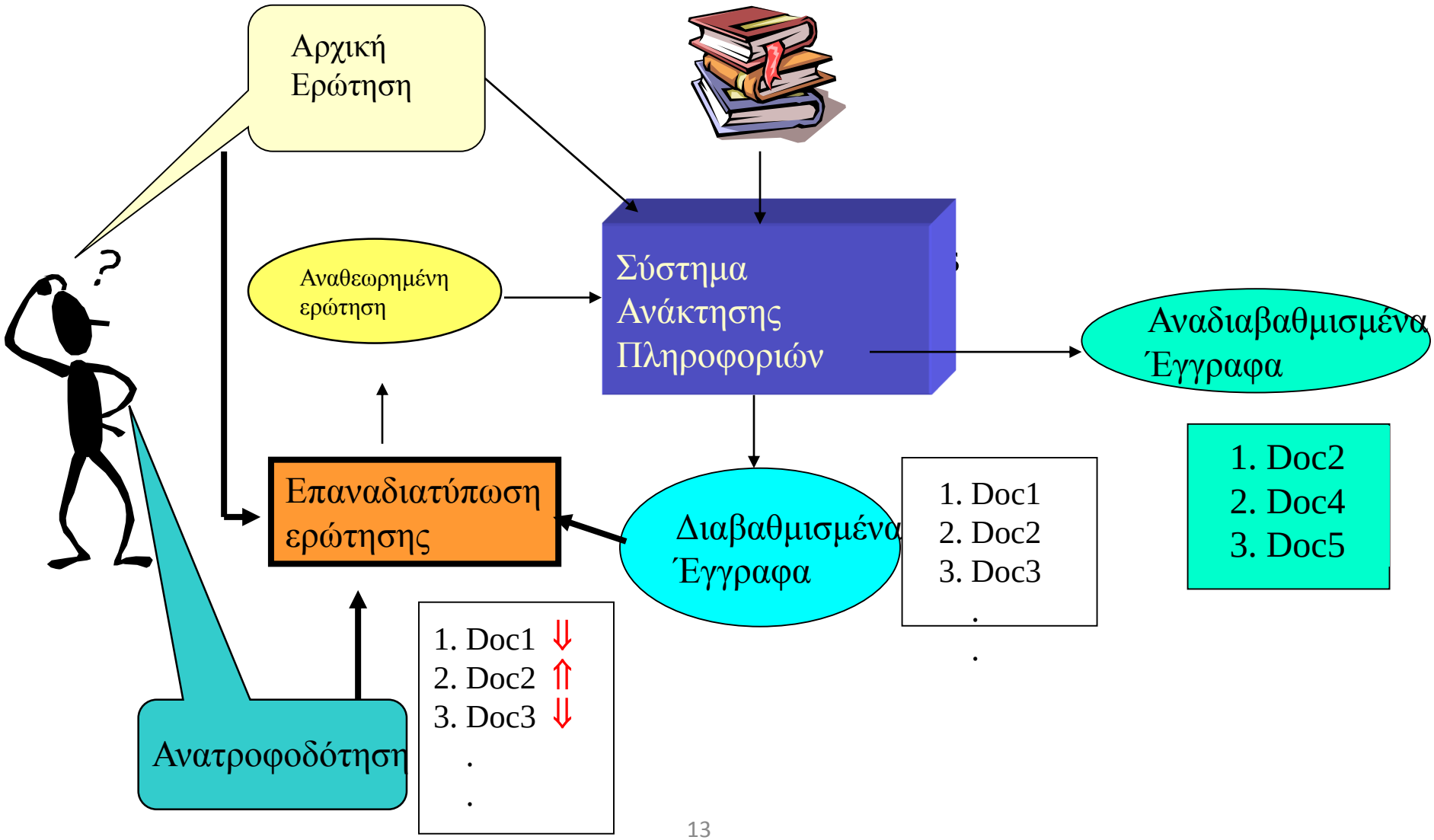


Ανατροφοδότηση συνάφειας



Αποτελέσματα μετά την ανατροφοδότηση

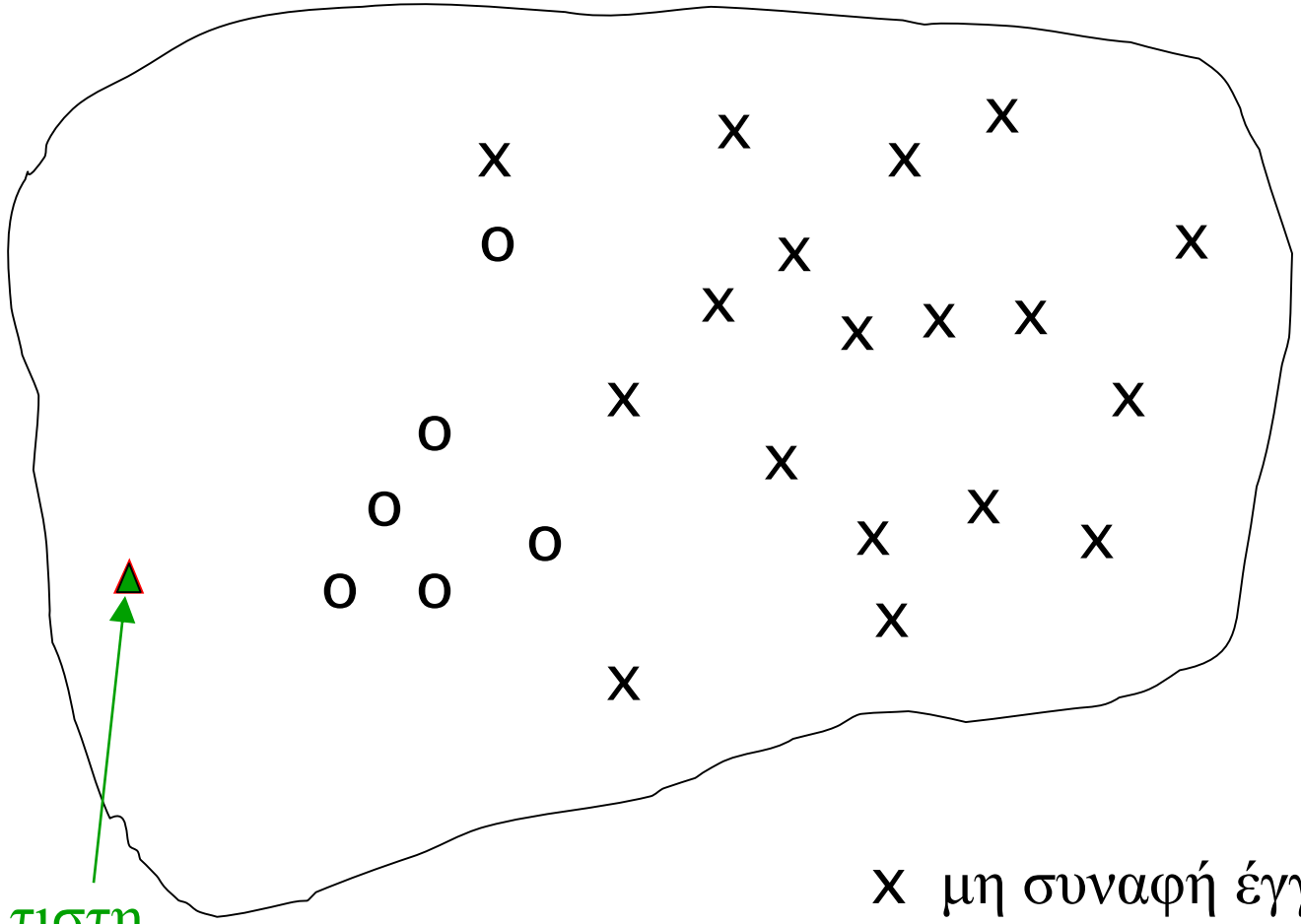



Αρχική Ερώτηση
Αναθεωρημένη ερώτηση
Σύστημα Ανάκτησης Πληροφοριών
Αναδιαβαθμισμένα Έγγραφα
1. Doc2
2. Doc4
3. Doc5
Επαναδιατύπωση ερώτησης
Διαβαθμισμένα Έγγραφα
1. Doc1
2. Doc2
3. Doc3
.
.
Ανατροφοδότηση
1. Doc1
2. Doc2
3. Doc3
.
.

Ο Αλγόριθμος Rocchio

- Ο αλγόριθμος Rocchio ενσωματώνει πληροφορία ανατροφοδότησης συνάφειας στο διανυσματικό μοντέλο.
- Θέλουμε να μεγιστοποιήσουμε το $sim\ (Q, C_r)\ -\ sim\ (Q, C_{nr})$,
  όπου Q = query, Cr = relevant documents Cnr = not relevant documents

Η θεωρητικά καλύτερη ερώτηση

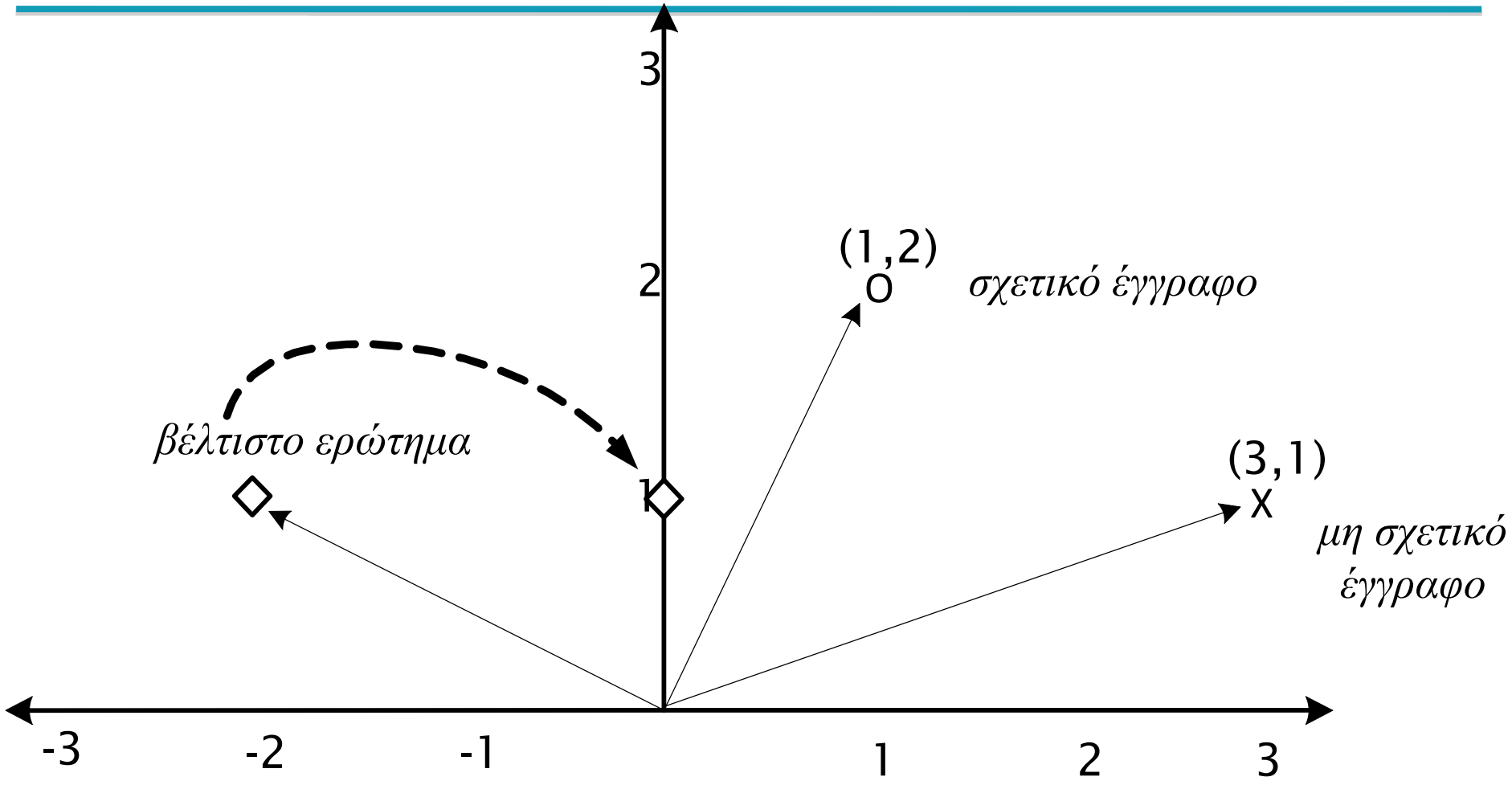


Ανάδραση στο Διανυσματικό Μοντέλο

Έστω ότι έχουμε ένα μόνο σχετικό έγγραφο (έστω *r*) και ένα μόνο μη σχετικό έγγραφο (έστω *nr*). Για να μπορέσουμε να διαχωρίσουμε το ένα από το άλλο, το διάνυσμα του βέλτιστου ερωτήματος Q_{opt} θα είναι (για συνημίτονο):

$$\text{vec}(Q_{opt}) = \text{vec}(r) - \text{vec}(nr)$$

Ανάδραση στο Διανυσματικό Μοντέλο



Ο Αλγόριθμος Rocchio

- Το βέλτιστο διάνυσμα ερώτησης για να ξεχωρίσουμε τα συναφή από τα μη συναφή έγγραφα (με ομοιότητα συνημίτονου):

$$\vec{Q}_{opt} = \frac{1}{|C_r|}\sum_{\vec{d}_j \in C_r}\vec{d}_j - \frac{1}{N-|C_r|}\sum_{\vec{d}_j \notin C_r}\vec{d}_j$$

Q_{opt} = optimal query; C_r = set of rel. doc vectors; N = collection size

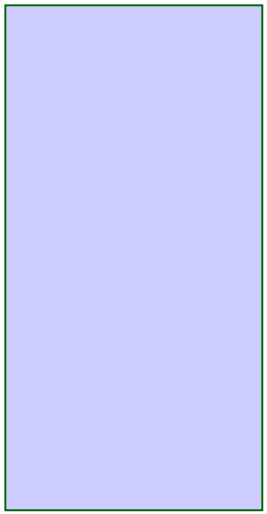
διαφορά ανάμεσα στα centroids των συναφών και μη συναφών

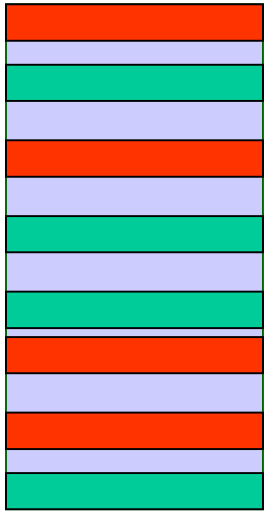
- Μη ρεαλιστικό γιατί δε ξέρουμε τα συναφή έγγραφα.

Χρήση Ανατροφοδότησης

Answer(q) =

Answer (q) + user feedback =





Κόκκινα: ο χρήστης έδωσε αρνητική ανατροφοδότηση

Πράσινα: ο χρήστης έδωσε θετική ανατροφοδότηση

Μπλε: ο χρήστης δεν έδωσε ανάδραση

Rocchio 1971 Algorithm (SMART)

Ο Rocchio 1971 Αλγόριθμος (SMART)

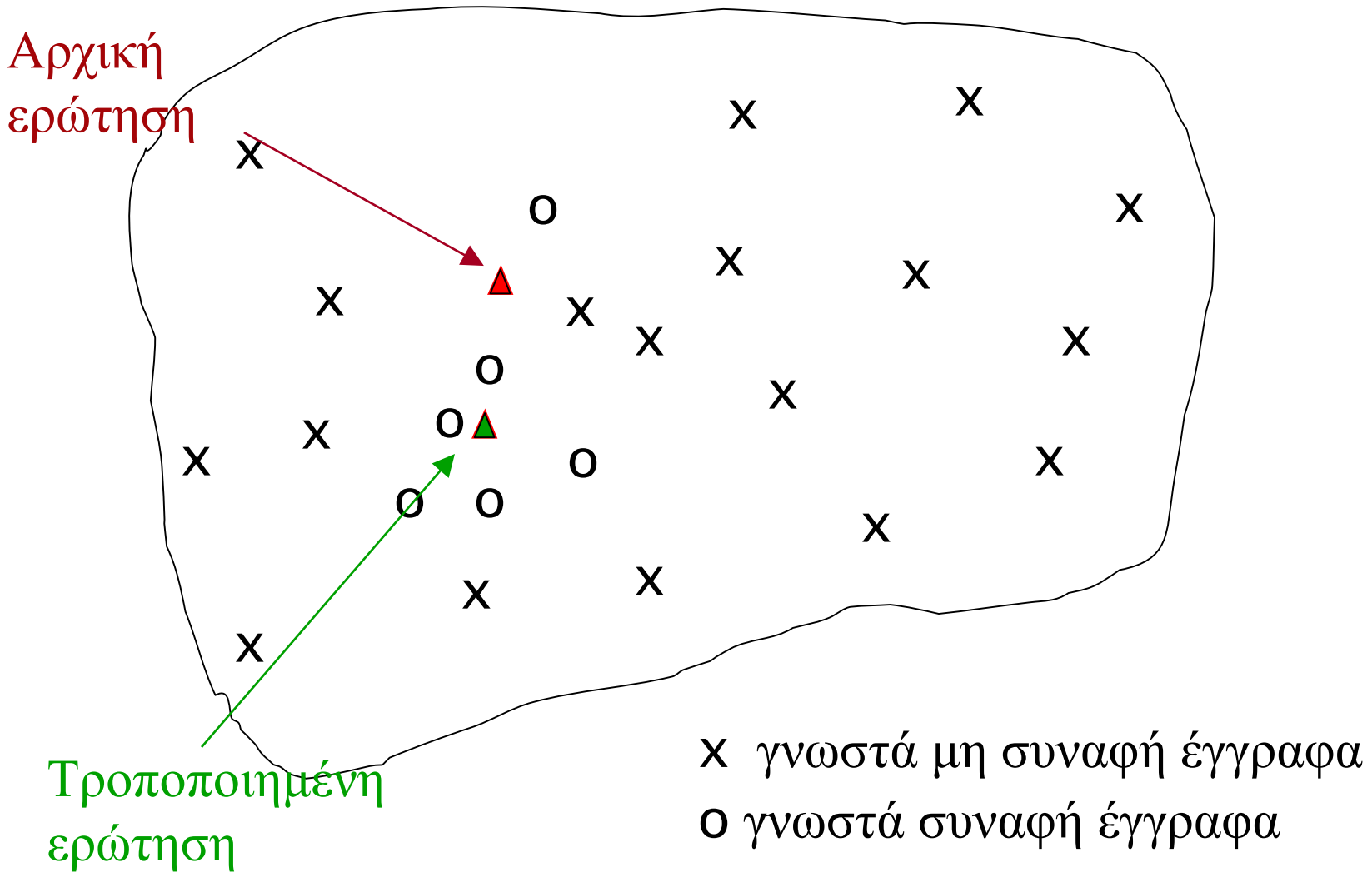
- Στην πράξη:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

q_m = τροποποιημένο διάνυσμα ερώτησης, q_0 = αρχικό διάνυσμα ερώτησης, α,β,γ: βάρη (επιλεγμένα εμπειρικά με το χέρι) D_r = *σύνολο διανυσμάτων γνωστών συναφών εγγράφων*, D_{nr} = *σύνολο διανυσμάτων γνωστών μη συναφών εγγράφων*

- Η νέα ερώτηση μετακινείται προς τα συναφή έγγραφα και μακριά από τα μη συναφή
- Tradeoff α vs. β/γ : Αν έχουμε πολλές αξιολογήσεις εγγράφων, θέλουμε μεγαλύτερο β/γ.
- Μπορεί να προκύψουν αρνητικά βάρη για κάποιους όρους
  - Τα αγνοούμε (τίθενται ίσα με 0)

Ανάδραση συνάφειας στην αρχική ερώτηση



Θετική vs Αρνητική Ανάδραση

- Η θετική ανάδραση είναι πιο χρήσιμη από την αρνητική (οπότε, θέτουμε $\gamma < \beta$; Π.χ., $\gamma = 0.25$, $\beta = 0.75$).
  - Γενικά πιο δύσκολο για τους χρήστες να αξιολογήσουν αρνητικά
  - Επίσης, πιο δύσκολο να χρησιμοποιηθεί γιατί τα συναφή συχνά σχηματίζουν συστάδες
- Πολλά συστήματα χρησιμοποιούν μόνο θετική ανάδραση (γ=0).

Η ανάδραση σημαντική κυρίως όταν έχει σημασία η ανάκληση (τότε, επίσης οι χρήστες είναι πιο πρόθυμοι να αφιερώσουν χρόνο)

Παράδειγμα

$$\text{query vector} = \alpha \cdot \text{αρχικό διάνυσμα ερωτήματος} + \beta \cdot \text{θετική ανάδραση} - \gamma \cdot \text{αρνητική ανάδραση}$$

Συνήθως, γ < β

αρχικό ερώτημα	0	4	0	8	0	0	$\alpha = 1.0$	0	4	0	8	0	0
													(+)
θετική ανάδραση	2	4	8	0	0	2	$\beta = 0.5$	1	2	4	0	0	1
													(-)
αρνητική ανάδραση	8	0	4	4	0	16	$\gamma = 0.25$	2	0	1	1	0	4
νέο ερώτημα								-1	6	3	7	0	-3

Ανάδραση συνάφειας: Υποθέσεις

- Υ1: Ο χρήστης έχει επαρκή γνώση για την αρχική ερώτηση
- Υ2: Το συναφή πρότυπα συμπεριφέρονται καλά
  - Η κατανομή των όρων στα συναφή έγγραφα είναι παρόμοια
  - Η κατανομή των όρων στα μη συναφή έγγραφα είναι διαφορετική από αυτήν στα συναφή
    - Είτε: όλα τα συναφή έγγραφα είναι ομαδοποιημένα γύρω από ένα πρότυπο (σχηματίζουν συστάδα).
    - Είτε: Υπάρχουν διαφορετικά πρότυπα με σημαντική επικάλυψη στο λεξιλόγιο τους.
    - Μικρές ομοιότητες μεταξύ συναφών και μη συναφών εγγράφων

Πότε δεν ισχύει η Υ1

- Ο χρήστης δεν έχει επαρκή αρχική γνώση
- Παραδείγματα:
  - Τυπογραφικά (ορθογραφικά) λάθη (Brittany Speers).
  - Ανάκτηση πληροφορίας από πολλές γλώσσες Cross-language (hígado).
  - Μη ταίριασμα (mismatch) ανάμεσα στο λεξιλόγιο του χρήστη και το λεξιλόγιο της συλλογής
    - Cosmonaut/astronaut

Πότε δεν ισχύει η A2

- Υπάρχουν πολλά διαφορετικά συναφή πρότυπα
- Παραδείγματα:
  - Υποσύνολα εγγράφων με διαφορετικό λεξιλόγιο (Burma/Myanmar)
  - Contradictory government policies
  - Ερωτήσεις των οποίων το αποτέλεσμα είναι διαζευκτικό (ανεξάρτητο) Pop stars that worked at Burger King
- Παραδείγματα γενικών εννοιών (general concepts) που εμφανίζονται σε διάζευξη ειδικών όρων
- Βοηθά να υπάρχει καλό editorial content
  - Πχ Report on contradictory government policies

Ανάδραση συνάφειας: Προβλήματα

- Γιατί οι περισσότερες μηχανές αναζήτησης δε χρησιμοποιούν ανάδραση;

Προβλήματα

- Οι μεγάλες ερωτήσεις είναι συνήθως *μη αποδοτικές*
  - Μεγάλοι χρόνοι απόκρισης
  - Μεγάλο κόστος για το σύστημα ανάκτησης
  - Μερική αντιμετώπισης:
    - Στάθμιση μόνο συγκεκριμένων σημαντικών όρων
      - Πιθανών των 20 πιο συχνών
- Οι χρήστες συνήθως *απρόθυμοι* να παρέχουν άμεσο feedback
- Συχνά είναι *δύσκολο να καταλάβει* κανείς γιατί ένα συγκεκριμένο έγγραφο εμφανίζεται στην απάντηση μετά την εφαρμογή της ανατροφοδότησης συνάφειας

Παράδειγμα: αρχική ερώτηση και κορυφαία 8 έγγραφα

Query: New space satellite applications

Θέλουμε μεγάλη ανάκληση

+ 1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
+ 2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget
5. 0.525, 07/24/90, Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6. 0.524, 08/22/90, Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7. 0.516, 04/13/87, Arianespace Receives Satellite Launch Pact From Telesat Canada
+ 8. 0.509, 12/02/87, Telecommunications Tale of Two Companies

Επέκταση ερώτησης (νέοι όροι και βάρη)

2.074 new 15.106 space

30.816 satellite 5.660 application

5.991 nasa 5.196 eos

4.196 launch 3.972 aster

3.516 instrument 3.446 arianespace

3.004 bundespost 2.806 ss

2.790 rocket 2.053 scientist

2.003 broadcast 1.172 earth

0.836 oil 0.646 measure

Κορυφαία 8 αποτελέσματα μετά την επαναδιατύπωση της ερώτησης

* 1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
* 2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm‘ Superconductors For Fast Circuit
* 5. 0.492, 12/02/87, Telecommunications Tale of Two Companies
6. 0.491, 07/09/91, Soviets May Adapt Parts of SS-20 Missile For Commercial Use
7. 0.490, 07/12/88, Gaping Gap: Pentagon Lags in Race To Match the Soviets In Rocket Launchers
8. 0.490, 06/14/90, Rescue of Satellite By Space Agency To Cost $90 Million

Στρατηγικές αξιολόγησης

- Χρησιμοποίησε την q_0 και υπολόγισε την ανάκληση και την ακρίβεια
- Χρησιμοποίησε την q_m και υπολόγισε την ανάκληση και την ακρίβεια
  - Χρήση όλων των εγγράφων της συλλογής
    - Γιατί δεν είναι «τίμιο»;
  - Χρήση των υπόλοιπων εγγράφων (residual collection) (όλα τα έγγραφα - αυτά που έχουν αξιολογηθεί)
    - Πρόβλημα, ειδικά αν τα συναφή είναι λίγα
  - Χρήση διαφορετικών συλλογών για την ανατροφοδότηση και την αξιολόγηση

- Εμπειρικά ένας γύρος ανατροφοδότησης είναι χρήσιμος, μερικές φορές ένας δεύτερος είναι οριακά χρήσιμος

Ανάδραση Συνάφειας στον Παγκόσμιο Ιστό

- Some search engines offer a *similar/related pages* feature (simplest form of relevance feedback)
  - **Πολλές φορές ο υπολογισμός αυτών των όμοιων/σχετικών σελίδων δεν γίνεται βάσει του περιεχομένου αλλά βάσει της δομής του γράφου (ανάλυση συνδέσμων). Ο υπολογισμός είναι αρκετά πιο γρήγορος.**

Excite Relevance Feedback

Spink et al. 2000

- Only about 4% of query sessions from a user used relevance feedback option
  - Expressed as "More like this" link next to each result
- But about 70% of users only looked at first page of results and didn't pursue things further
  - So 4% is about 1/8 of people extending search
- Relevance feedback improved results about 2/3 of the time

Ψευδοανάδραση Συνάφειας (Pseudo Relevance Feedback)

- Ανάδραση *χωρίς είσοδο από το χρήση*.
- Υπόθεση ότι τα κορυφαία *m* από τα ανακτημένα έγγραφα είναι συναφή (και χρήση αυτών για ανάδραση)
  - Μπορούμε επίσης να χρησιμοποιήσουμε τα τελευταία έγγραφα για αρνητική ανάδραση
- Επιτρέπει την επέκταση της ερώτησης με όρους που σχετίζονται με τους όρους της ερώτησης

Ψευδοανάδραση Συνάφειας

- Γνωστή και ως blind ή ad-hoc, χρησιμοποιεί τεχνικές RF για να *βελτιώσει αυτόματα* την κατάταξη *πριν παρουσιαστούν* τα έγγραφα στο χρήστη.
- Η ψευδοανάδραση λειτουργεί ικανοποιητικά για *«καλά» αρχικά ερωτήματα* αλλά είναι αναποτελεσματική για «κακά» αρχικά ερωτήματα.
- Δουλεύει ακόμα καλύτερα αν τα κορυφαία έγγραφα πρέπει να *ικανοποιούν και μια boolean έκφραση* προκειμένου να χρησιμοποιηθούν για ανάδραση

  (π.χ. να περιέχουν όλους του όρους της επερώτησης)

Ψευδοανάδραση Συνάφειας

- Βρέθηκε να *βελτιώνει* την απόδοση στο διαγωνισμό του TREC
- Κίνδυνος query drift

Indirect (έμμεσο) relevance feedback

- On the web, DirectHit introduced a form of indirect relevance feedback.
- DirectHit ranked documents higher that users look at more often.
  - Clicked on links are assumed likely to be relevant
    - Assuming the displayed summaries are good, etc.
- Globally: Not user or query specific.
- This is the general area of clickstream mining

Δυναμική Αναζήτηση

- Κατά ένα μεγάλο μέρος υπάρχει η παραδοχή ότι η πληροφορία που αναζητεί ο χρήστης δε μεταβάλλεται κατά τη διάρκεια της αναζήτησης.
- Αν αυτό δεν συμβαίνει, τότε τα έγγραφα που χαρακτηρίστηκαν σχετικά στην αρχή της αναζήτησης μπορεί να μην είναι καλά παραδείγματα για το τι θεωρεί ο χρήστης σχετικό μετά.
- Με χρήση *ageing component* μπορεί να μειώνεται το βάρος ενός όρου όσο περνάει ο χρόνος ώστε να έχει μικρότερη επίδραση στην εύρεση εγγράφων.

Επέκταση Ερώτησης (Query Expansion)

- Στην ανατροφοδότηση συνάφειας, οι χρήστες δίνουν επιπρόσθετη πληροφορία (συναφή/μη συναφή) στα έγγραφα, η οποία χρειάζεται για να επαναπροσδιοριστούν τα βάρη των όρων στα έγγραφα
- Στην επέκταση της ερώτησης, οι χρήστες δίνουν επιπρόσθετη πληροφορία (καλός/κακός όρος αναζήτησης) σε λέξεις ή φράσεις.

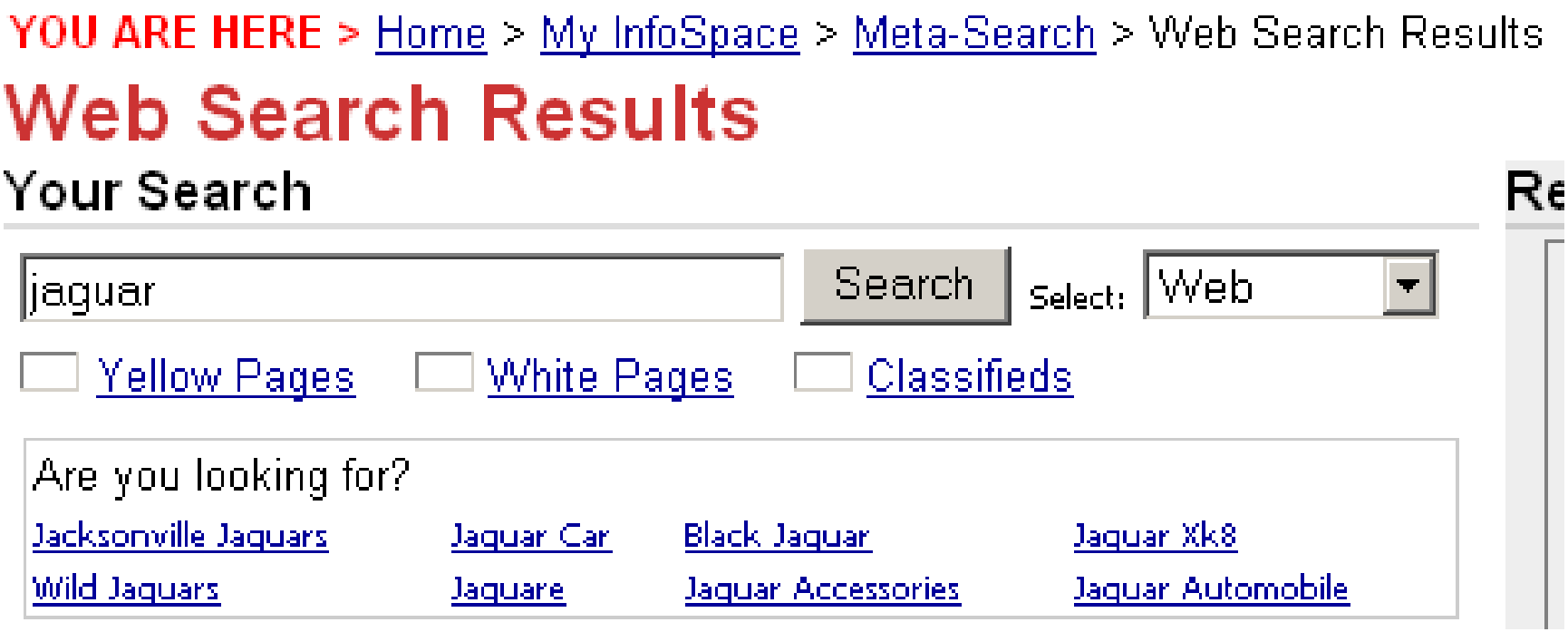
Πως παράγεται η εναλλακτική ή επέκταση της ερώτησης

- Συνήθως με *ολική ανάλυση* (global analysis) χρησιμοποιώντας μια μορφή *θησαυρού* (thesaurus)
- Για κάθε όρο t στην ερώτηση, η ερώτηση επεκτείνεται αυτόματα με *συνώνυμα και σχετικές λέξεις* με το t από το θησαυρό
- Η χρήση του θησαυρού μπορεί να συνδυαστεί με κάποιο είδος *στάθμιση των όρων* (π.χ., μικρότερα βάρη στους νέους όρους από ότι στους αρχικούς)

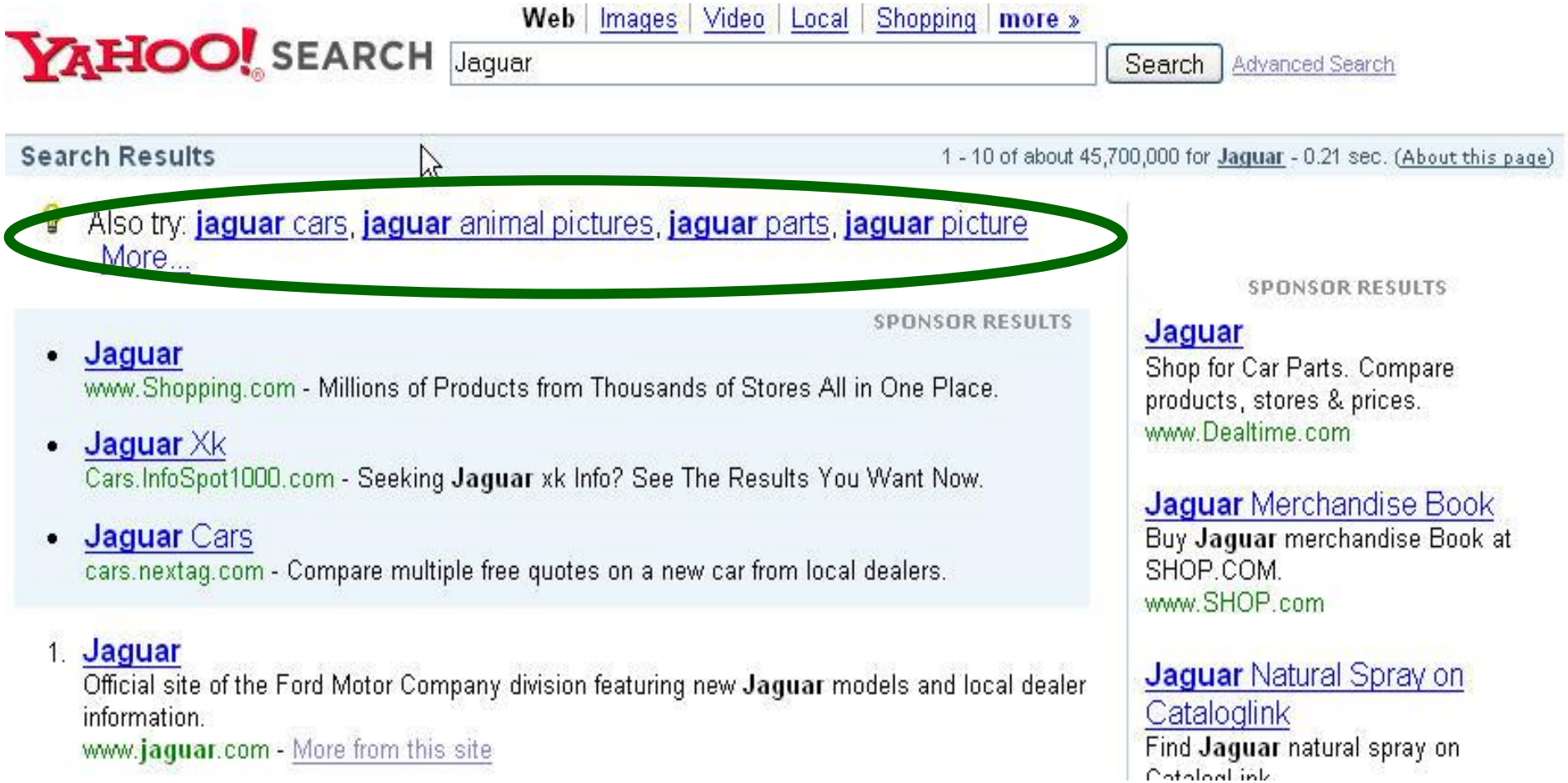
Είδη Επέκτασης της Ερώτησης

- Ολική ανάλυση: (στατική, χρήση όλων των εγγράφων στη συλλογή)
- Τοπική ανάλυση: (δυναμική)
  - Ανάλυση των (κορυφαίων) εγγράφων στο σύνολο της απάντησης

Παράδειγμα



Επίσης: www.altavista.com, www.teoma.com


YAHOO! SEARCH
Web | Images | Video | Local | Shopping | more »
Jaguar
Search
Advanced Search
Search Results
1 - 10 of about 45,700,000 for Jaguar - 0.21 sec. (About this page)
Also try: jaguar cars, jaguar animal pictures, jaguar parts, jaguar picture
More...
SPONSOR RESULTS
Jaguar
www.Shopping.com - Millions of Products from Thousands of Stores All in One Place.
Jaguar Xk
Cars.InfoSpot1000.com - Seeking Jaguar xk Info? See The Results You Want Now.
Jaguar Cars
cars.nextag.com - Compare multiple free quotes on a new car from local dealers.
1. Jaguar
Official site of the Ford Motor Company division featuring new Jaguar models and local dealer information.
www.jaguar.com - More from this site
SPONSOR RESULTS
Jaguar
Shop for Car Parts. Compare products, stores & prices.
www.Dealtime.com
Jaguar Merchandise Book
Buy Jaguar merchandise Book at SHOP.COM.
www.SHOP.com
Jaguar Natural Spray on Cataloglink
Find Jaguar natural spray on

MyStuff | Settings

Ask.com

Web | Images | Video | More ▸

jaguar

Advanced Search

Narrow

Jaguar Cars
Black **Jaguar**
Cat **Jaguar**
Jaguar Big Cats
Jaguars Habitat
What Do Jaguars Eat
Panthera Onca
Where Do Jaguars Live
More ▸

Expand

Cheetah
Ferrari
More ▸

Related Names

Ford
Wolf
More ▸

jaguar

Showing results **1-10** of 10,190,000

Source

Jaguar | Save

Kingdom: Animalia **Phylum:** Chordata **Class:** Mammalia **Order:** Carnivora **Family:** Felidae **Genus:** Panthera **Species:** Panthera onca

The biggest and most powerful North American cat, the Jaguar is the only one that roars. It moves over a large home range with a diameter of 3 to 15 miles (5-25 km) where prey is abundant, larger where prey is scarce. This cat hunts... More »

Key Facts | Images | Encyclopedia

Jaguar
Gama actual, concesionarios, historia, noticias, anuncios y servicios financieros.
· www.jaguar.com/

Jaguar (Panthera onca)
Jaguar (Panthera onca) facts, photos and videos. ... The **Jaguar** is the largest cat in the Western Hemisphere and the third largest cat in ...
· www.thebigzoo.com/Animals/Jaguar.asp · Cached

Jaguar
The **jaguar** measures five to six feet from its nose to the tip of its tail and weighs 140 to 220 pounds (females are slightly smaller).
· www.kidsplanet.org/factsheets/jaguar.html · Cached

Jaguar

Images



More »

Dictionary

Definitions of 'jaguar'

(jăgwär, jăgyū-är) - 1 definition
The American Heritage® Dictiona

jaguar (n.) A large feline mammal (Panthera onca) of Central and South America, closely related to the leopard a having a tawny coat spotted with black rosettes.

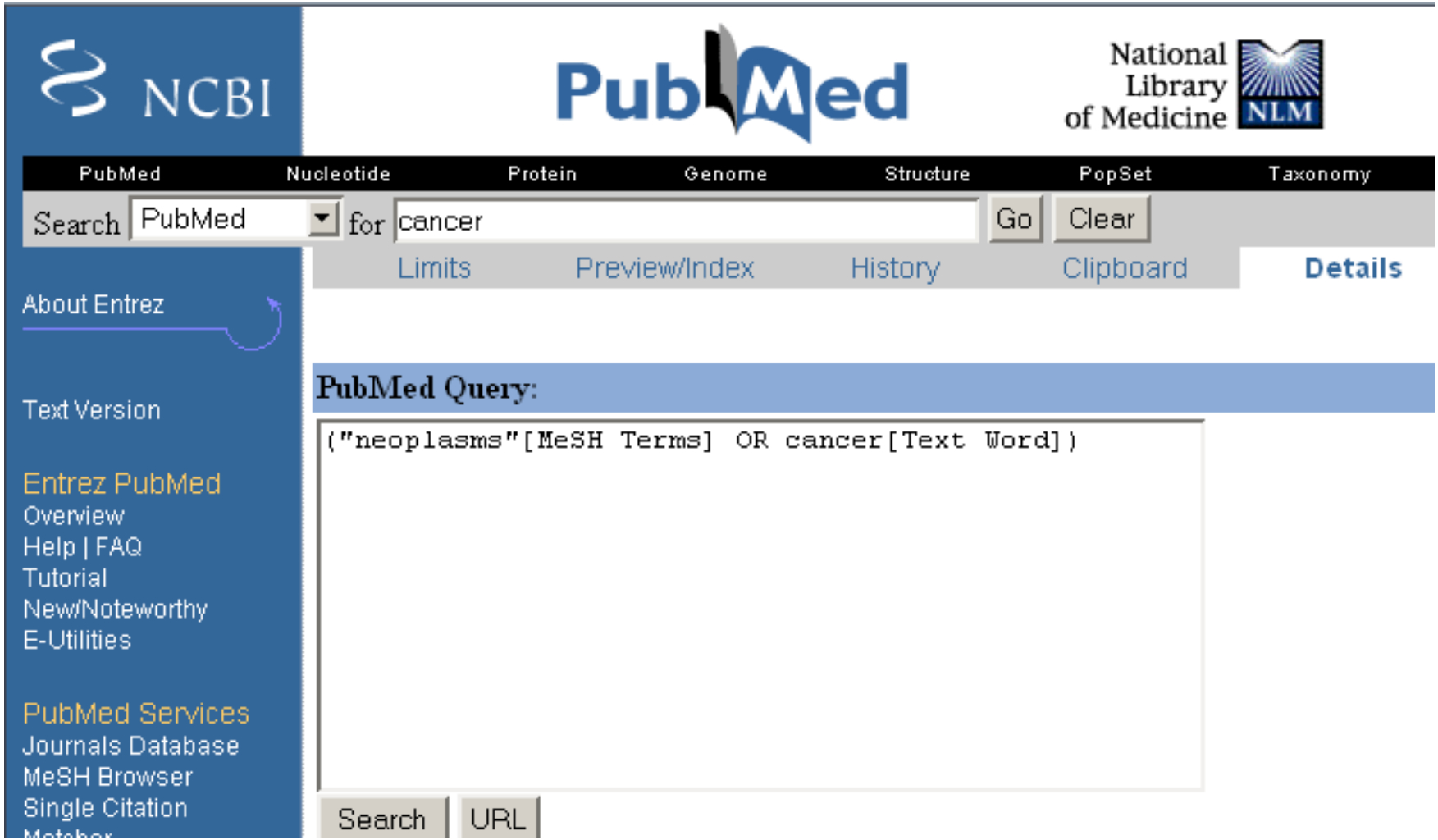
All Music Guide

Jaguar
By: Fred Small
Whether an artist is conservative, centrist, liberal or downright radical, there nothing wrong with getting on a

Είδη λεξιλογίου

- Controlled vocabulary
  - Maintained by editors (e.g., medline)
- Manual thesaurus
  - E.g. MedLine: physician, syn: doc, doctor, MD, medico
- Automatically derived thesaurus
  - (co-occurrence statistics)
- Refinements based on query log mining
  - Συνηθισμένο στο web, χρησιμοποιούμε τις λέξεις που χρησιμοποίησαν προηγούμενη χρήστες για να επαναπροσδιορίσουν τις ερωτήσεις τους

Controlled Vocabulary

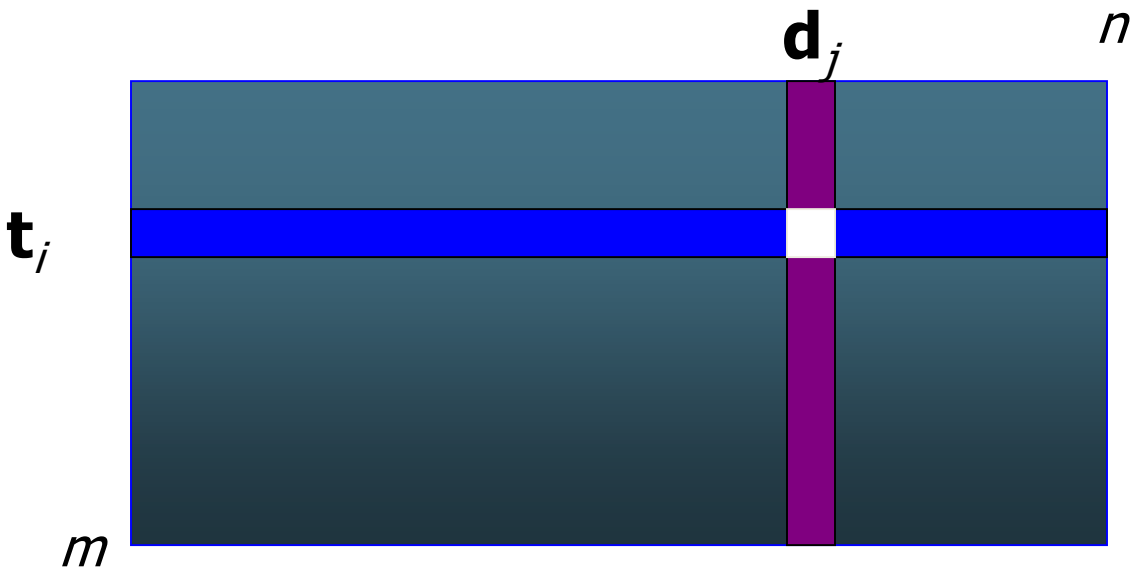


Automatic Thesaurus Generation

- Attempt to generate a thesaurus automatically by analyzing the collection of documents
- Two main approaches
  - Co-occurrence based (co-occurring words are more likely to be similar)
  - Shallow analysis of grammatical relations
    - Entities that are grown, cooked, eaten, and digested are more likely to be food items.
- Co-occurrence based is more robust, grammatical relations are more accurate.

Co-occurrence Thesaurus

- Simplest way to compute one is based on term-term similarities in C = AA^T where A is term-document matrix.
- $w_{i,j}$ = (normalized) weighted count (t_i, $\mathbf{d}_j$)



With integer counts – what do you get for a boolean cooccurrence matrix?

Automatic Thesaurus Generation Example

word	ten nearest neighbors
absolutely	absurd whatsoever totally exactly nothing
bottomed	dip copper drops topped slide trimmed slig
captivating	shimmer stunningly superbly plucky witty
doghouse	dog porch crawling beside downstairs gazec
Makeup	repellent lotion glossy sunscreen Skin gel p
mediating	reconciliation negotiate cease conciliation p
keeping	hoping bring wiping could some would othe
lithographs	drawings Picasso Dali sculptures Gauguin
pathogens	toxins bacteria organisms bacterial parasite
senses	grasp psyche truly clumsy naive innate awl

Automatic Thesaurus Generation Discussion

- Quality of associations is usually a problem.
- Term ambiguity may introduce irrelevant statistically correlated terms.
  - "Apple computer" → "Apple red fruit computer"
- Problems:
  - False positives: Words deemed similar that are not
  - False negatives: Words deemed dissimilar that are similar
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

Thesaurus-based Query Expansion

- This doesn't require user input
- For each term, *t*, in a query, expand the query with synonyms and related words of *t* from the thesaurus
  - feline → feline cat
- May weight added terms less than original query terms.
- Generally increases recall.
- Widely used in many science/engineering fields
- May significantly decrease precision, particularly with ambiguous terms.
  - "interest rate" → "interest rate fascinate evaluate"
- There is a high cost of manually producing a thesaurus
  - And for updating it for scientific changes

ΤΕΛΟΣ 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)*