Προχωρημένες Λειτουργίες Επερώτησης
Advanced Query Operations
ΜΕΡΟΣ Ι

Κεφάλαιο 5

Διάρθρωση Διάλεξης

- Κίνητρο
- **Ανάδραση Συνάφειας (Relevance Feedback)**
- **Αναδιατύπωση Επερωτήσεων (Query Reformulation)**
  - Αναβάρυνση Όρων (Term Reweighting)
  - Επέκταση (Διαστολή) Επερώτησης (Query Expansion),
  - Αναδιατύπωση Επερωτήσεων για το Διανυσματικό Μοντέλο
    - Optimal Query, Rocchio Method, Ide Method, DeHi Method
  - Η έννοια του Optimal (or Best) Query
  - Αξιολόγηση
- **Ψευδο-ανάδραση συνάφειας (Pseudo relevance feedback)**
- **Επέκταση Επερωτήσεων**
  - Αυτόματη Τοπική (Επιτόπια) Ανάλυση (Automatic Local Analysis)
  - Καθολική Ανάλυση
  - Επέκταση Επερώτησης βάσει Θησαυρού (Thesaurus-based Query Expansion)
  - Αυτόματη Καθολική Ανάλυση (Automatic Global Analysis)
  - Στατιστικοί Θησαυροί (Statistical Thesaurus)
  - Κατασκευή Θησαυρών

Κίνητρο

- Έχει παρατηρηθεί ότι οι χρήστες των ΣΑΠ δαπανούν πολύ χρόνο *αναδιατυπώνοντας* την αρχική τους επερώτηση προκειμένου να βρουν ικανοποιητικά έγγραφα
- Πιθανές αιτίες
  - ο χρήστης δεν γνωρίζει το περιεχόμενο των υποκείμενων εγγράφων
  - το λεξιλόγιο του χρήστη μπορεί να διαφέρει από αυτό της συλλογής
  - η αρχική επερώτηση 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 επερώτηση μπορεί να θεωρηθεί ως η πρώτη προσπάθεια έκφρασης της πληροφοριακής ανάγκης του χρήστη
- Ανάγκη για τεχνικές αντιμετώπισης αυτού του προβλήματος

Τρόποι Αντιμετώπισης

(1) Βελτίωση της αρχικής επερώτησης
(2) Χρήση Προφίλ Χρήστη
(3) Βελτίωση παράστασης κειμένων
(4) Βελτίωση αλγορίθμου (μοντέλου) ανάκτησης

Παρατηρήσεις

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

Αναδιατύπωση επερώτησης βάσει Ανάδρασης Συνάφειας (Relevance Feedback: Query Reformulation)

Τρόποι αναδιατύπωσης της επερώτησης:

1. 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.

Αναδιατύπωση επερώτησης βάσει Ανάδρασης Συνάφειας (Relevance Feedback: Query Reformulation)

2. Επέκταση επερώτησης (Query Expansion):

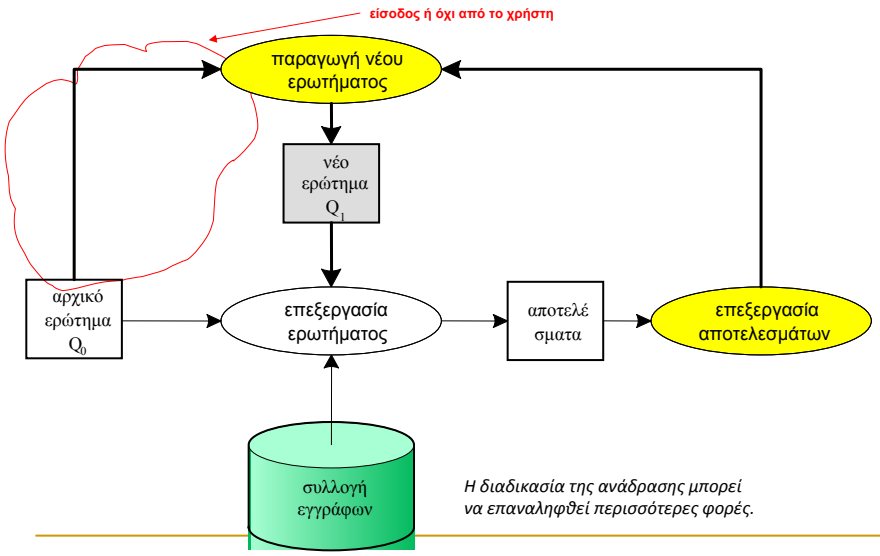
Προσθήκη νέων όρων στην επερώτηση (π.χ. από γνωστά συναφή έγγραφα)

Υπάρχουν πολλοί αλγόριθμοι για επαναδιατύπωση επερώτησης

Τεχνικές Βελτίωσης της Αρχικής Επερώτησης

Κατηγορίες:
(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

Η Διαδικασία της Ανάδρασης



Η διαδικασία της ανάδρασης μπορεί να επαναληφθεί περισσότερες φορές.

Ανάδραση Συνάφειας (Relevance Feedback): Η βασική ιδέα

Με είσοδο από το χρήστη

Βήματα:

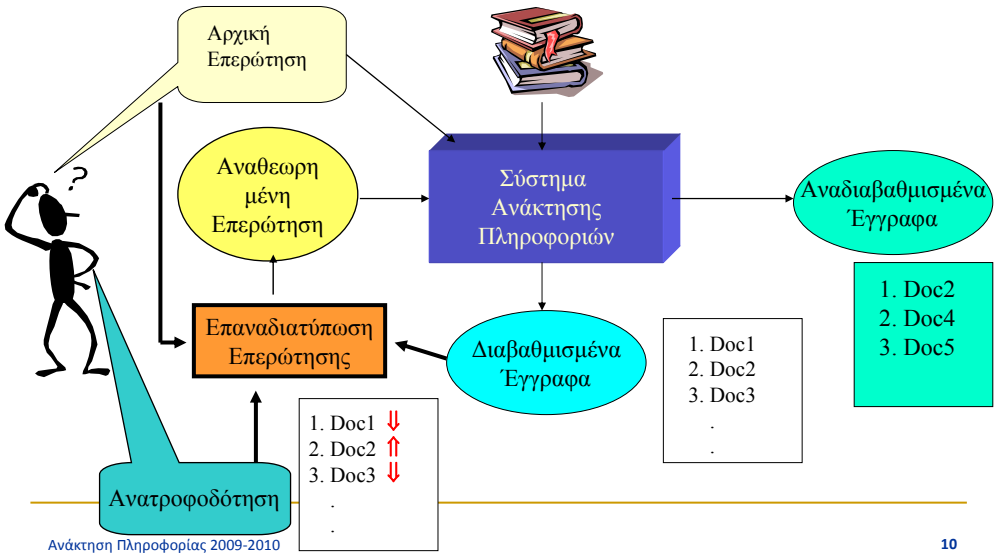
1/ Μετά την παρουσίαση των αποτελεσμάτων, επιτρέπουμε στο χρήστη **να κρίνει (θετικά ή αρνητικά) την συνάφεια** ενός ή περισσότερων εγγράφων της απάντησης

2/ Αξιοποιούμε αυτήν την πληροφορία για να **αναδιατυπώσουμε** την επερώτηση

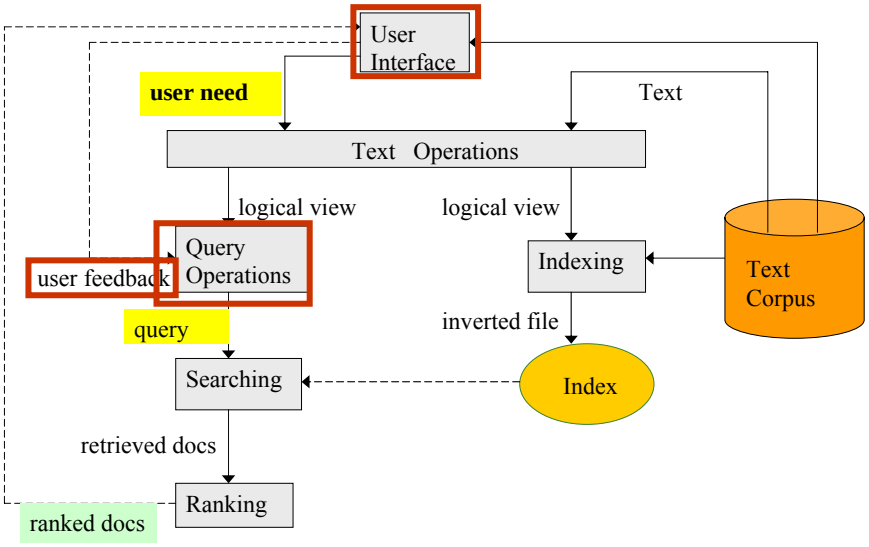
3/ Κατόπιν δίνουμε στο χρήστη την απάντηση της αναδιατυπωμένης επερώτησης

4/ Πήγαινε στο βήμα 1/

Ανάδραση Συνάφειας (Relevance Feedback): Η βασική ιδέα

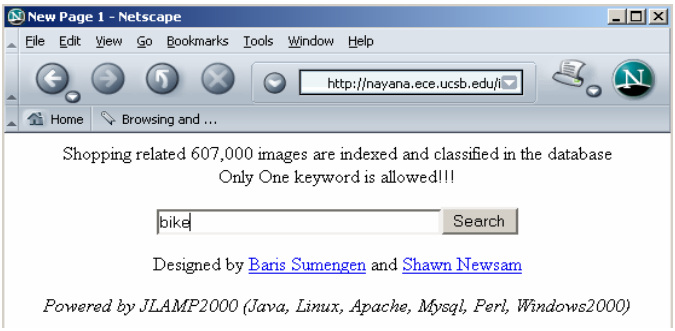


Τμήματα της Αρχιτεκτονικής που Εμπλέκονται



Παράδειγμα ανατροφοδότησης συνάφειας σε σύστημα ανάκτησης εικόνων

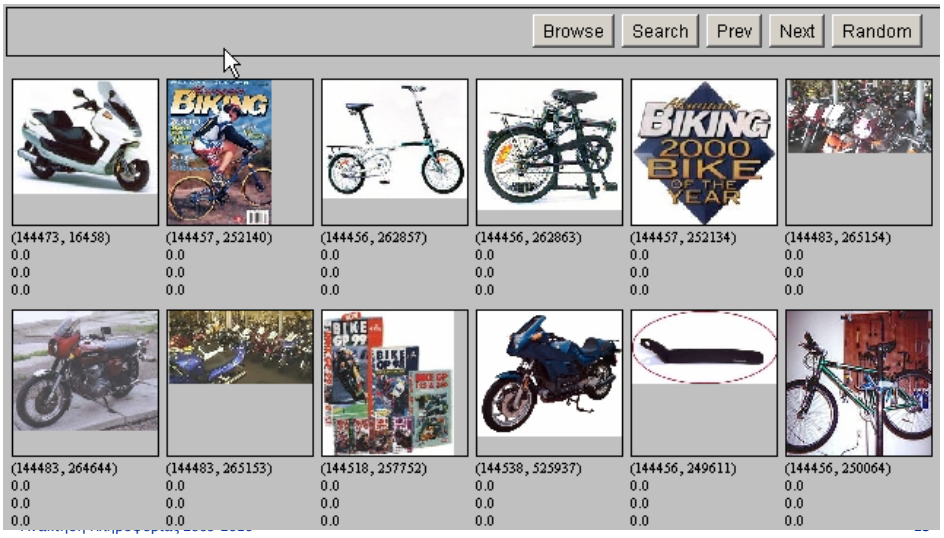
q = bike



(http://nayana.ece.ucsb.edu/imsearch/imsearch.html)

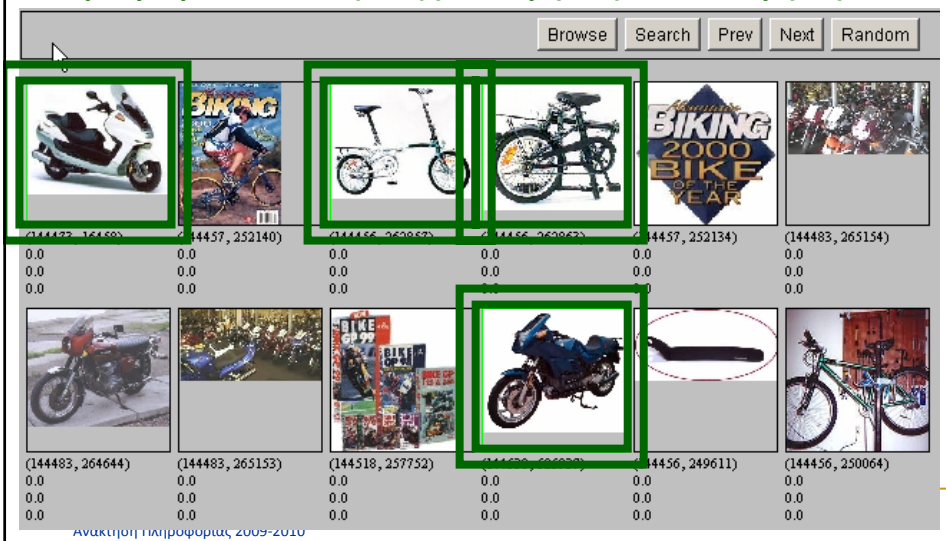
Παράδειγμα ανατροφοδότησης συνάφειας σε σύστημα ανάκτησης εικόνων

Answer("bike")=



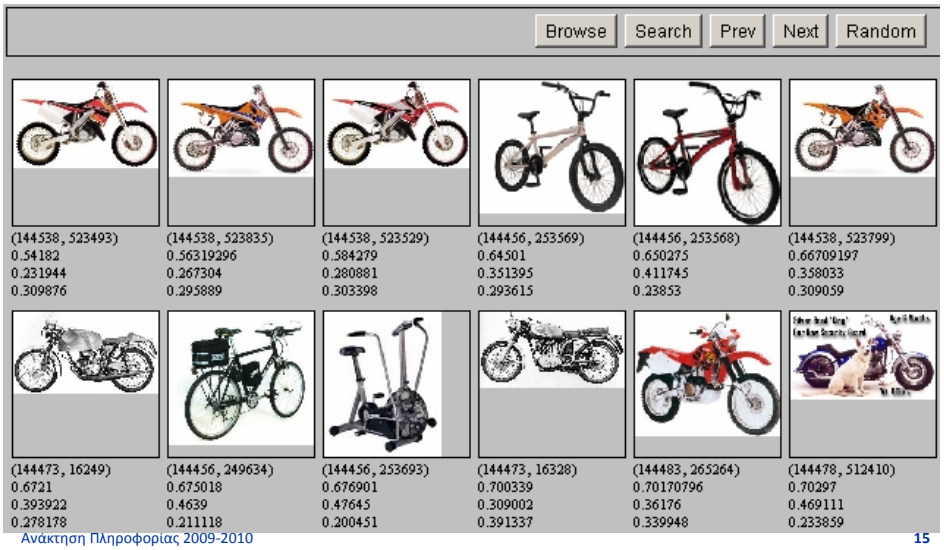
Παράδειγμα ανατροφοδότησης συνάφειας σε σύστημα ανάκτησης εικόνων

Μαρκάρισμα των Συναφών (η Επιθυμητών) από τον Χρήστη



Παράδειγμα ανατροφοδότησης συνάφειας σε σύστημα ανάκτησης εικόνων

Απάντηση της αναδιατυπωμένης απάντησης =



Ανάδραση στο Διανυσματικό Μοντέλο

Βέλτιστη ερώτηση: Η πιο γνωστή μέθοδος ανάδρασης στο Διανυσματικό μοντέλο είναι η μέθοδος του **Rocchio**.

A query vector, q_{opt} that

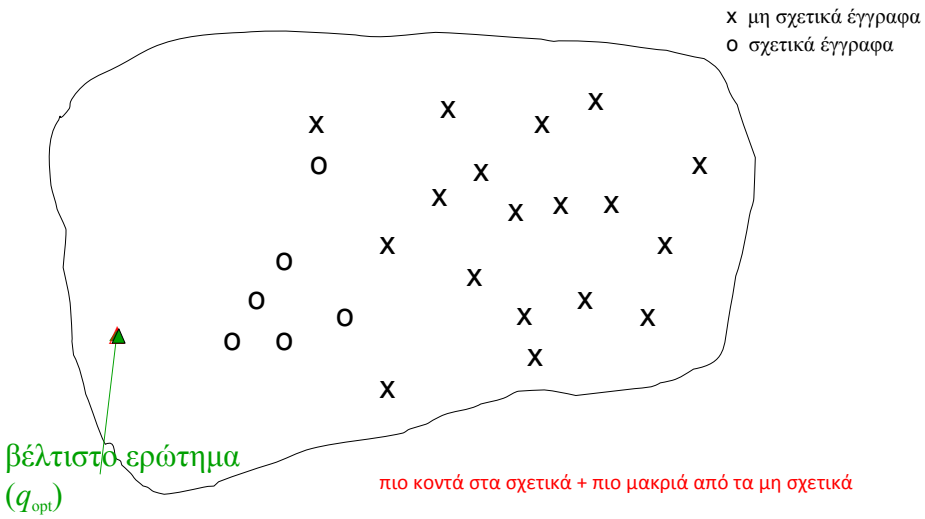
- maximizes similarity with relevant documents
- minimizes similarity with nonrelevant documents.

$$\vec{q}_{opt} = \arg\max_{\vec{q}}[\text{sim}(\vec{q}, C_r) - \text{sim}(\vec{q}, C_{nr})],$$

C: συλλογή εγγράφων, *Cr*: σύνολο σχετικών, *Cnr*: σύνολο μη σχετικών

Υπενθύμιση: $$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|},$$

Ανάδραση στο Διανυσματικό Μοντέλο

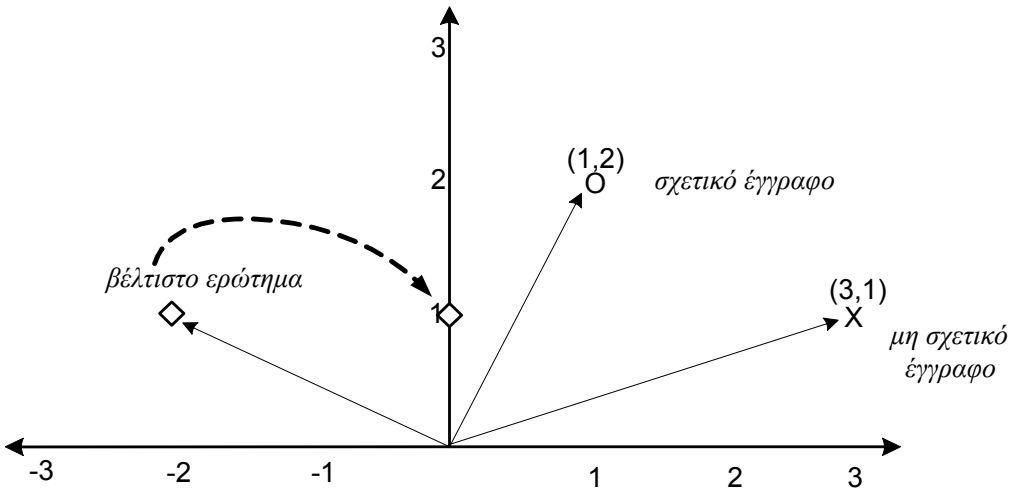


Ανάδραση στο Διανυσματικό Μοντέλο

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



Ανάδραση στο Διανυσματικό Μοντέλο

Το βέλτιστο ερώτημα που πρέπει να διατυπώσουμε ώστε να διαχωριστούν τα σχετικά έγγραφα από τα μη σχετικά είναι (όταν χρησιμοποιείται ομοιότητα συνημιτόνου):

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

C: συλλογή εγγράφων, R: σύνολο σχετικών, N-R: σύνολο μη σχετικών

The optimal query is the vector difference between the centroids of the relevant and nonrelevant documents

Ανάδραση στο Διανυσματικό Μοντέλο

Ομαδοποίηση (Clustering)

Relevant documents have similarities among themselves -> Ιδανική ερώτηση στο κέντρο (centroid) της συστάδα τους

Irrelevant documents have term-weight vectors which are dissimilar from the ones for the relevant documents

Ανάδραση στο Διανυσματικό Μοντέλο

Πρόβλημα: το βέλτιστο ερώτημα δεν μπορεί να βρεθεί στην πράξη.

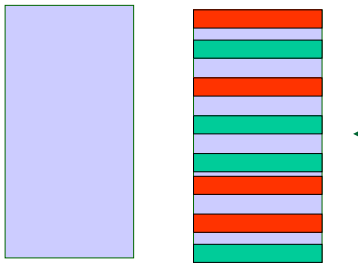
Γιατί?

Διότι δε γνωρίζουμε εκ των προτέρων το σύνολο των σχετικών εγγράφων. Αν τα γνωρίζαμε ποιος ο λόγος να εκτελέσουμε το ερώτημα?

Αναδιατύπωση επερώτησης στο Διανυσματικό Χώρο

Αφού όμως δεν γνωρίζουμε το σύνολο C_r, θα λάβουμε υπόψη την αρχική επερώτηση και την είσοδο του χρήστη.

Answer(q)= Answer (q) + user feedback =



Κόκκινα: ο χρήστης έδωσε αρνητική ανάδραση
Πράσινα: ο χρήστης έδωσε θετική ανάδραση
Μπλε: ο χρήστης δεν έδωσε ανάδραση

Rocchio 1971 Algorithm (SMART)

(I) Standard **Rocchio** Method

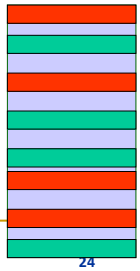
Αφού το σύνολο όλων των συναφών είναι άγνωστο,

χρησιμοποίησε τα **γνωστά** συναφή (*Dr*) και **γνωστά** μη-συναφή (*Dn*) έγγραφα (από την απάντηση της αρχικής επερώτησης και βάσει της εισόδου από τον χρήστη) και επίσης συμπεριέλαβε την αρχική επερώτηση *q*.

Αναδιατυπωμένη επερώτηση

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

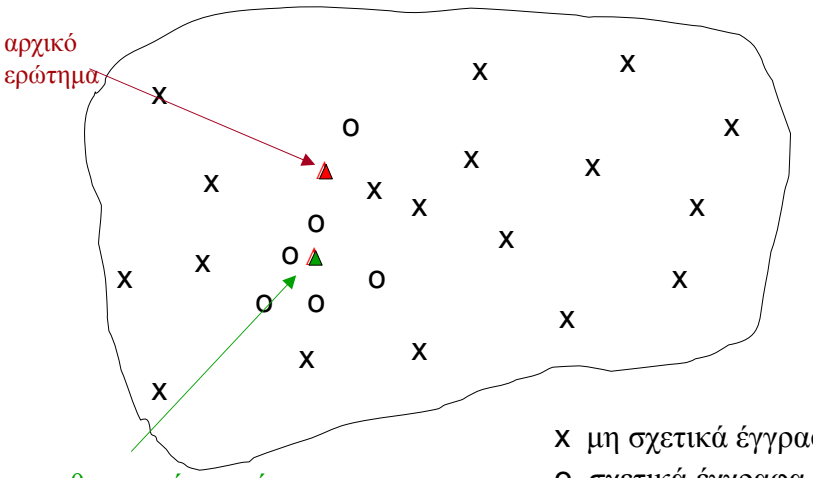
answer(q):



α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for irrelevant documents.

Usually $\gamma < \beta$ (the relevant docs are more important)
If $\gamma = 0$ then we have positive feedback only

Ανάδραση στο Διανυσματικό Μοντέλο



Ανάδραση στο Διανυσματικό Μοντέλο

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

Συνήθως, $\gamma < \beta$

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

(+)

(-)

Term weight can go negative

Negative term weights are ignored (set to 0)

Αναδιατύπωση επερώτησης στο Διανυσματικό Χώρο

Τρόποι αξιοποίησης της ανατροφοδότησης του χρήστη
(I) **Rocchio** Method
(II) **Ide** Method
(III) **DeHi** Method

(II) IDE Regular Method

Περισσότερη ανάδραση => μεγαλύτερος βαθμός αναδιατύπωσης.
Για αυτό, κατά την IDE Regular μέθοδο **δεν** κάνουμε κανονικοποίηση (βάσει του ποσού ανάδρασης)

$$\vec{q}_m = \alpha\vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for irrelevant documents.

Αρνητική Ανάδραση

- Η αρνητική ανάδραση είναι μια μορφή ανάδρασης που χρησιμοποιεί πληροφορία από έγγραφα που έχουν χαρακτηριστεί ως **μη-σχετικά** από το χρήστη.
- Η αρνητική ανάδραση θεωρείται προβληματική:
  - Πότε ένας χρήστης θα πρέπει να χαρακτηρίζει ένα έγγραφο ως μη-σχετικό; Είναι πιο δύσκολο από το χαρακτηρισμό ως σχετικό.
  - Το να περιμένει κανείς από τους χρήστες να επιλέξουν έγγραφα ως μη-σχετικά στην αναζήτηση μπορεί να είναι δύσκολο στην υλοποίηση πρακτικά.

(III) IDE "Dec Hi" Method

Τάση για απόρριψη **μόνο** των μη-συναφών εγγράφων που έχουν υψηλό σκορ
(Bias towards rejecting **just** the highest ranked of the irrelevant documents:)

$$\vec{q}_m = \alpha \vec{q} + \beta \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \max_{non-relevant}(\vec{d}_j)$$

α: Tunable weight for initial query.
β: Tunable weight for relevant documents.
γ: Tunable weight for irrelevant document.

Σύγκριση μεθόδων (I) (II) (III)

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\forall\vec{d}_j\in D_r}\vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall\vec{d}_j\in D_n}\vec{d}_j$$

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j\in D_r}\vec{d}_j - \gamma\sum_{\forall\vec{d}_j\in D_n}\vec{d}_j$$

$$\vec{q}_m = \alpha\vec{q} + \beta\sum_{\forall\vec{d}_j\in D_r}\vec{d}_j - \gamma\max_{non-relevant}(\vec{d}_j)$$

- Γενικά, τα πειραματικά δεδομένα δεν δίνουν καθαρό προβάδισμα σε κάποια τεχνική.
- Όλες οι τεχνικές βελτιώνουν την απόδοση (recall & precision)
- Συνήθως α = β = γ = 1. Αν γ = 0, μόνο θετική ανάδραση

Relevance Feedback Example: Initial Query and Top 8 Results

Note: want high recall

Query: New space satellite applications

+ 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

Relevance Feedback Example: Expanded Query

Νέα ερώτηση με 18 όρους και νέα 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

Top 8 Results After Relevance Feedback

+ 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

Αξιολόγηση Αποτελεσματικότητας Τεχνικών Ανάδρασης Συνάφειας

Remarks

- Relevance feedback is most useful for increasing *recall* in situations where recall is important
  - Users can be expected to review results and to take time to iterate

Incremental Refinement

Empirically, one round of relevance feedback is often very useful. Two rounds is sometimes marginally useful.

Overall time spent (user satisfaction)

Αξιολόγηση Αποτελεσματικότητας Τεχνικών Ανάδρασης Συνάφειας

Αξιολόγηση:

Χρήση αρχικής q_0 και υπολογισμός precision and recall graph

Χρήση βελτιωμένης q_m και υπολογισμός precision and recall graph

Remarks

- By construction, reformulated query will rank **explicitly-marked relevant** documents higher and **explicitly-marked irrelevant** documents lower.
- When evaluating such methods, a method should not get credit for improvement on ***these*** documents, since it was told their relevance.

  In machine learning, this error is called "testing on the training data."
- Evaluation should focus on generalizing to **other** un-rated documents.

Συμπέρασμα: θα αγνοήσουμε τα έγγραφα για τα οποία έχουμε ήδη τη γνώμη του χρήστη;

Αξιολόγηση Αποτελεσματικότητας Τεχνικών Ανάδρασης Συνάφειας

Fair Process for Evaluating the Effectiveness of Relevance Feedback

- **Remove** from the corpus any document for which feedback was provided.
- Measure recall/precision performance after relevance feedback on the remaining **residual collection**.
- Relative performance on the residual collection provides **fair** data on the effectiveness of relevance feedback
- But, compared to complete corpus, specific recall/precision numbers **may decrease** since relevant documents were removed.

Relevance Feedback Evaluation

TABLE 4. Evaluation of typical relevance feedback methods for five collections (weighted documents, weighted queries).

Relevance Feedback Method	Rank of Method and Avg Precision	CACM 3204 docs 64 queries	CISI 1460 docs 112 docs	CRAN 1397 docs 225 queries	INSPEC 12684 docs 84 queries	MED 1033 docs 30 queries	Average
Initial Run (reduced collection)		.1459	.1184	.1156	.1368	.3346	
Ide (dec hi) expand by	Rank						
Rocchio (standard $\beta = .75$, $\alpha = .25$ expand by all terms		+49%	+44%	+92%	+32%	.5980 +79%	+59%
expand by most common terms	Rank	2	39	8	14	17	16
	Precision	.2552	.1404	.2955	.1821	.5630	
	Improvement	+75%	+19%	+156%	+33%	+68%	+70%
Probabilistic (adjusted revised	Rank	3	12	12	10	24	
	Precision	.2491	.1623	.2534	.1861	.5279	
	Improvement	+71%	+37%	+119%	+36%	+55%	+64%

Simulated interactive retrieval consistently outperforms non-interactive retrieval (70% here).

Relevance Feedback Evaluation: Case Study

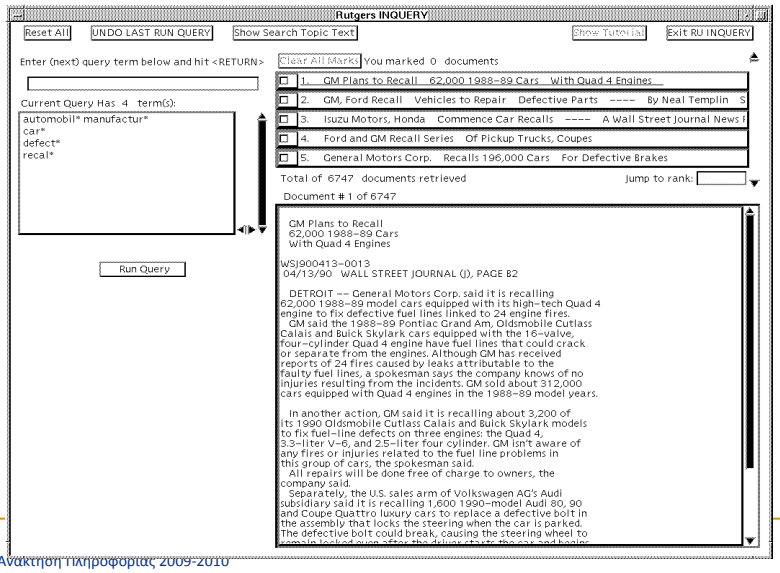
Example of evaluation of interactive information retrieval [Koenemann & Belkin 1996]

Goal of study: show that relevance feedback improves retrieval effectiveness

Details

- 64 novice searchers (43 female, 21 male, native English)
- TREC test bed (Wall Street Journal subset)
- Two search topics
  - Automobile Recalls
  - Tobacco Advertising and the Young
- Relevance judgements from TREC and experimenter
- System was INQUERY (vector space with some bells and whistles)
- Subjects had a tutorial session to learn the system
- Their goal was to keep modifying the query until they have developed one that gets high precision
- Reweighting of terms similar to but different from Rocchio

Relevance Feedback Evaluation: Case Study



Credit: Marti Hearst

Πειραματικά Αποτελέσματα

- Opaque (black box)
  - Ο χρήστης δεν μπορεί να δει τη διαδικασία ανάδρασης.
- Transparent
  - Ο χρήστης μπορεί να δει τους όρους που δημιουργήθηκαν από την ανάδραση αλλά δεν μπορεί να μεταβάλει το ερώτημα.
- Penetrable
  - Ο χρήστης μπορεί να μεταβάλει το ερώτημα.

Evaluation:Precision vs. RF condition (from Koenemann & Belkin 96)

Criterion: p@30 (precision at 30 documents)

Compare:

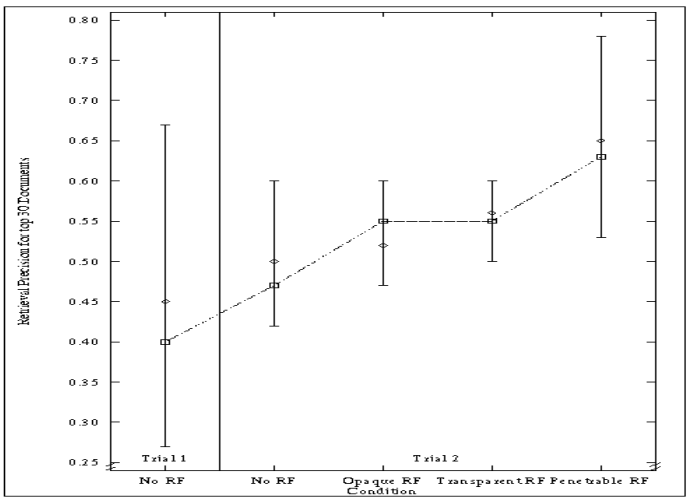
p@30 for users with relevance feedback

p@30 for users without relevance feedback

Goal: show that users with relevance feedback do better

Results:

- Subjects with relevance feedback had 17-34% better performance
- But: Difference in precision numbers not statistically significant. Search times approximately equal



Credit: Marti Hearst

Σιωπηρές Υποθέσεις της Ανάδρασης Συνάφειας

A1: User has sufficient knowledge for initial query.

Violation of A1

- User does not have sufficient initial knowledge.
- Examples:
  - Misspellings (Brittany Speers).
  - Cross-language information retrieval (hígado).
  - Mismatch of searcher's vocabulary vs. collection vocabulary
    - Cosmonaut/astronaut – laptop/notebook

Σιωπηρές Υποθέσεις της Ανάδρασης Συνάφειας

A2: Relevance prototypes are “well-behaved”.

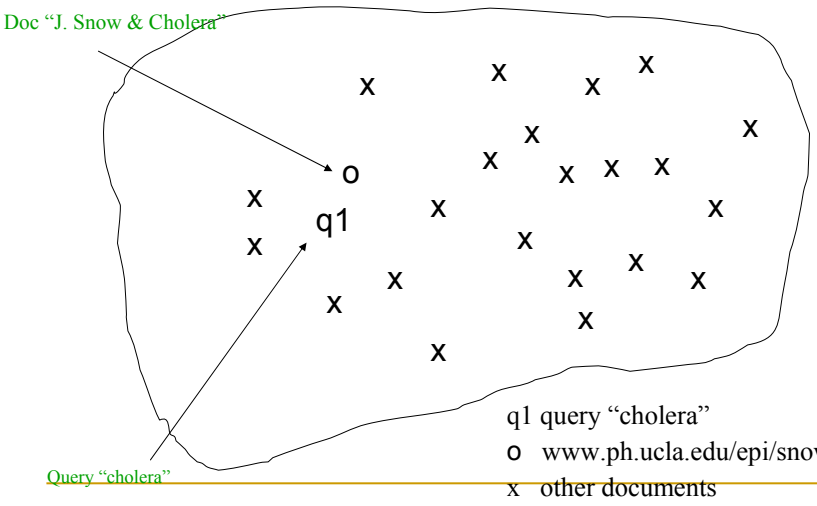
- Term distribution in relevant documents will be similar
- Term distribution in non-relevant documents will be different from those in relevant documents
  - Either: All relevant documents are tightly clustered around a single prototype.
  - Or: There are different prototypes, but they have significant vocabulary overlap.
  - Similarities between relevant and irrelevant documents are small

Violation of A2

However, there are several relevance prototypes.

- Examples:
  - (using different vocabulary) Burma/Myanmar
  - Pop stars that worked at Burger King (query whose answer set is inherently disjunctive)
  - Often: instances of a general concept
- Good editorial content can address problem
  - Report on contradictory government policies

Aside: Vector Space can be Counterintuitive.



q1 query "cholera"
o www.ph.ucla.edu/epi/snow.html
x other documents

High-dimensional Vector Spaces

- The queries "cholera" and "john snow" are far from each other in vector space.
  How can the document "John Snow and Cholera" be close to both of them?
- Our intuitions for 2- and 3-dimensional space don't work in >10,000 dimensions.
- 3 dimensions: If a document is close to many queries, then some of these queries must be close to each other.
- Doesn't hold for a high-dimensional space.

Γιατί η ανάδραση συνάφειας δεν χρησιμοποιείται ευρέως;

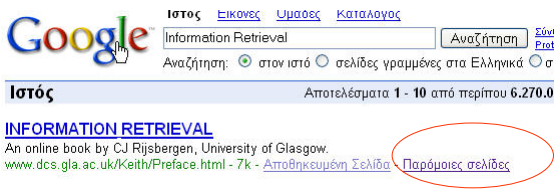
- Long queries are inefficient for typical IR engine.
  - Long response times for user.
  - High cost for retrieval system.
  - Partial solution:
    - Only reweight certain prominent terms
      - Perhaps top 20 by term frequency
- Users are often reluctant to provide explicit feedback
- It's often harder to understand why a particular document was retrieved after applying relevance feedback

Relevance Feedback on the Web

[in 2003: now less major search engines, but same general story]

- Some search engines offer a similar/related pages feature (this is a trivial form of relevance feedback)
  - Google (link-based)
  - Altavista
  - Stanford WebBase
- But some don't because it's hard to explain to average user:
  - Alltheweb
  - msn
  - Yahoo
- Excite initially had true relevance feedback, but abandoned it due to lack of use.

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

Other Uses of Relevance Feedback

- Following a changing information need (name of car models change over time)
- Maintaining an information filter (e.g., for a news feed)
- Active learning
  - [Deciding which examples it is most useful to know the class of to reduce annotation costs]

Δυναμική Αναζήτηση

- Κατά ένα μεγάλο μέρος στη δουλειά που γίνεται για το RF 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.

Ψευτοανάδραση Συνάφειας

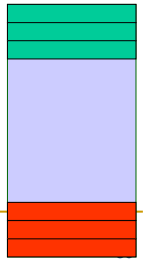
Ανάδραση χωρίς είσοδο από το χρήστη

Ψευδοανάδραση Συνάφειας Pseudo Relevance Feedback

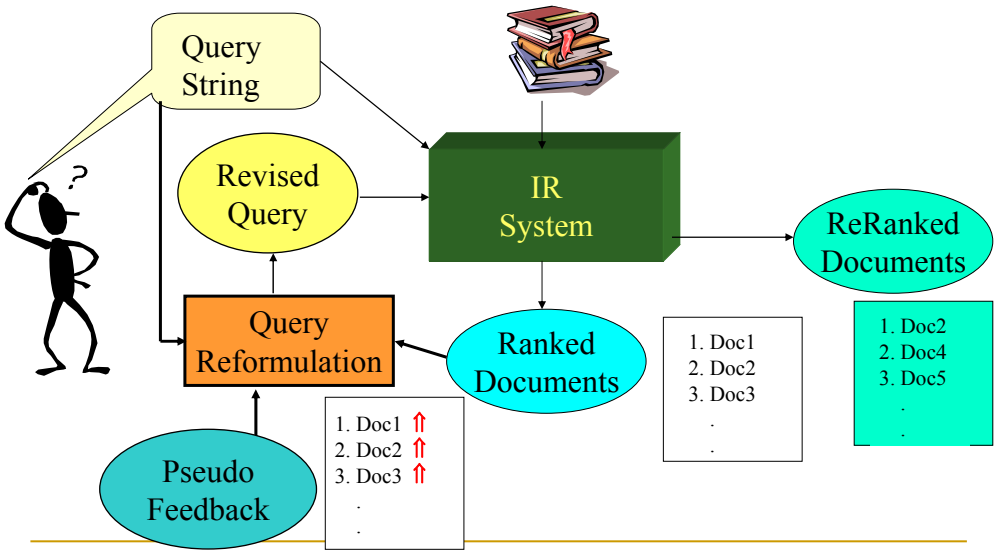
Χρήση μεθόδων ανάδρασης αλλά **χωρίς είσοδο από το χρήστη**

- **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 επερώτησης με όρους που σχετίζονται με τους όρους της επερώτησης

answer(q):



Ψευδοανάδραση Συνάφειας



Σχετικά με την Ψευδοανάδραση

- Γνωστή και ως **blind** ή **ad-hoc** RF, χρησιμοποιεί τεχνικές RF για να *βελτιώσει αυτόματα* την κατάταξη πριν παρουσιαστούν τα έγγραφα στο χρήστη.
- Η pseudo RF τεχνική λειτουργεί ικανοποιητικά για «καλά» αρχικά ερωτήματα αλλά είναι αναποτελεσματική για «κακά» αρχικά ερωτήματα.

Αξιολόγηση Ψευδοανάδρασης

- Βρέθηκε να βελτιώνει την απόδοση στο διαγωνισμό του TREC (ad-hoc retrieval task)
- Δίνει τη δυνατότητα να βρεθεί ένα καλύτερο ερώτημα **χωρίς** να απαιτείται από το χρήστη να διατυπώσει ένα νέο ερώτημα από την αρχή.
- Εύκολη υλοποίηση και εφαρμογή στα πιο δημοφιλή μοντέλα ανάκτησης (boolean, vector, probabilistic)
- 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)

Indirect Relevance Feedback

Implicit relevance feedback

clickstream mining or clickthrough

Ανάδραση στο Πιθανοκρατικό Μοντέλο

Η συνάρτηση ομοιότητας του πιθανοκρατικού μοντέλου είναι:

$$S_{prob}(q,d) = \sum_i \log \frac{p_i \cdot (1-r_i)}{r_i \cdot (1-p_i)}$$

Όπου η άθροιση αφορά στους όρους που βρίσκονται και στο ερώτημα και στο έγγραφο.

p_i πιθανότητα ο όρος t_i να εμφανίζεται σε σχετικό έγγραφο

r_i πιθανότητα ο όρος t_i να εμφανίζεται σε μη σχετικό έγγραφο

Ανάδραση στο Πιθανοκρατικό Μοντέλο

Αρχικά θέτουμε τιμές στις πιθανότητες :

$p_i = P(x_i \mid R) = c$

$r_i = P(x_i \mid \overline{R}) = n_i / N$

όπου:

c είναι μία τυχαία σταθερά (π.χ., 0.5)

n_i είναι το πλήθος των εγγράφων που περιέχουν τον i-οστό όρο

N πλήθος εγγράφων συλλογής

Ανάδραση στο Πιθανοκρατικό Μοντέλο

- Για τη βελτίωση της ποιότητας των αποτελεσμάτων οι πρώτες εφαρμογές του Πιθανοκρατικού μοντέλου χρειάζονται την παρέμβαση του χρήστη για την αναπροσαρμογή των τιμών (πχ Rocchio).

Εναλλακτικά μπορεί να χρησιμοποιηθεί και αυτοματοποιημένος τρόπος (ψευτο-ανάδραση).

Αρχικά εκτελείται το ερώτημα με τις αρχικές εκτιμήσεις. Επιλέγονται τα ***k*** καλύτερα έγγραφα.

Έστω k_i ο αριθμός των εγγράφων που περιέχουν τον i-οστό όρο. Θέτουμε:

$p_i = P(x_i \mid R) = k_i / k$

$r_i = P(x_i \mid R) = (n_i - k_i) / (N - k)$

+ 0.5 adjustment

Ανάδραση στο Πιθανοκρατικό Μοντέλο

Για μικρές τιμές του k και k_i

$$p_i = P(x_i \mid R) = (k_i + 0.5)/ (k + 1)$$

+ 0.5 adjustment

$$r_i = P(x_i \mid \bar{R}) = [(n_i - k_i) + 0.5] / [(N - k) + 1]$$

Άλλο πιθανό + k_i/N

Χρήσιμοι Σύνδεσμοι

Σύστημα ανάκτησης εικόνων με δυνατότητα ανάδρασης

http://amazon.ece.utexas.edu/~qasim/cires.htm

Survey of relevance feedback over VSM, Probabilistic and Logic Model

http://www.dcs.qmul.ac.uk/~mounia/CV/Papers/ker_ruthven_lalmas.pdf