HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Προχωρημένες Λειτουργίες Επερώτησης
Advanced Query Operations

Γιάννης Τζίτζικας

Διάλεξη : 11
Ημερομηνία : 31-3-2006

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, Rochio Method, Ide Method, DeHi Method
  - 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)
- Γενετικοί Αλγόριθμοι

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)

Τεχνικές Βελτίωσης της Αρχικής Επερώτησης

Κατηγορίες:

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

(β) τεχνικές που **δεν απαιτούν** είσοδο

(β1) που στηρίζονται στα **κορυφαία έγγραφα** που ανακτήθηκαν

(β2) που στηρίζονται σε **όλα τα έγγραφα** της συλλογής

Ανάδραση Συνάφειας (Relevance Feedback): Η βασική ιδέα

Βήματα:

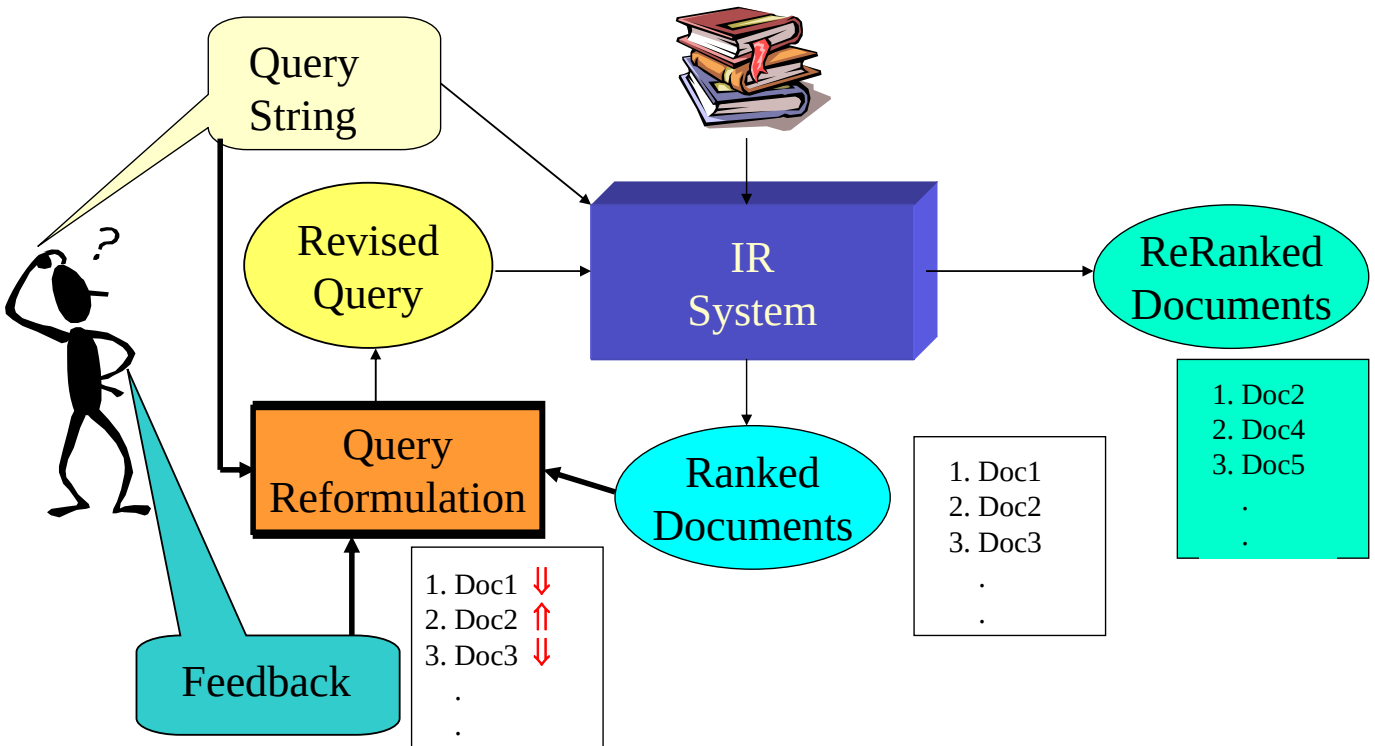
1/ Μετά την παρουσίαση των αποτελεσμάτων, επιτρέπουμε στο χρήστη **να κρίνει την συνάφεια** ενός ή περισσότερων εγγράφων της απάντησης

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

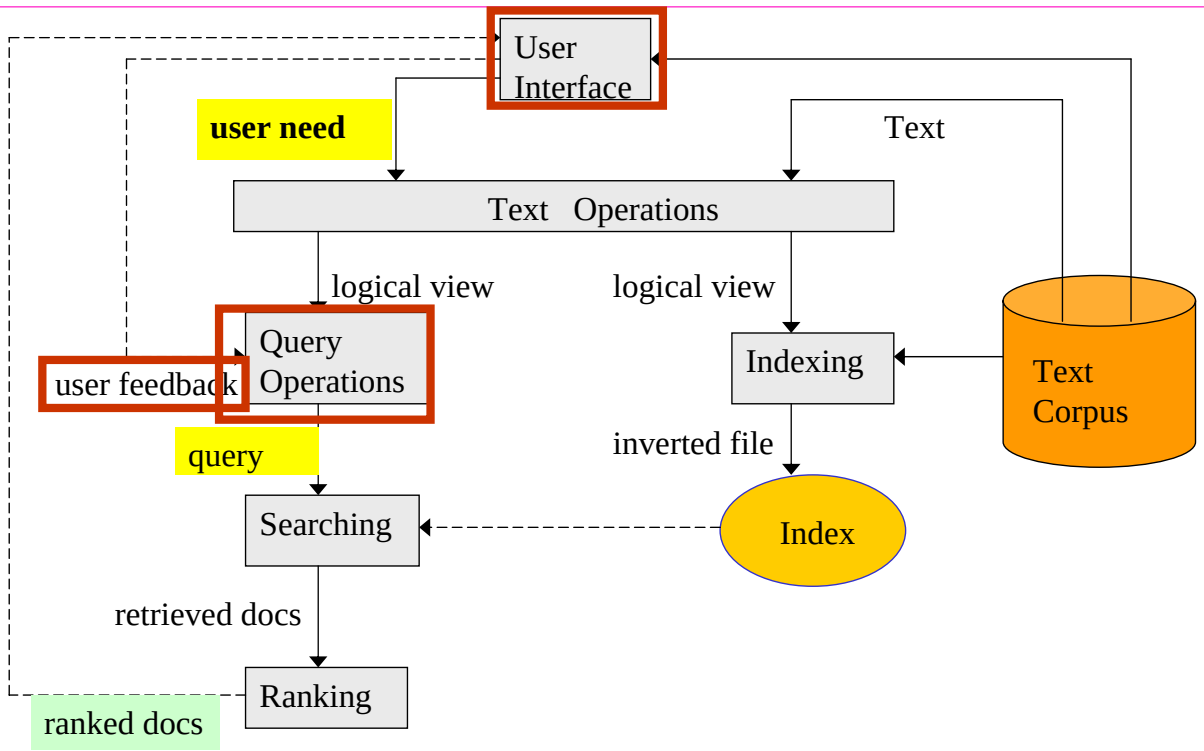
3/ Κατόπιν δίδουμε στο χρήστη την απάντηση της αναδιατυπωμένης επερώτησης

4/ Πήγαινε στο βήμα 1/

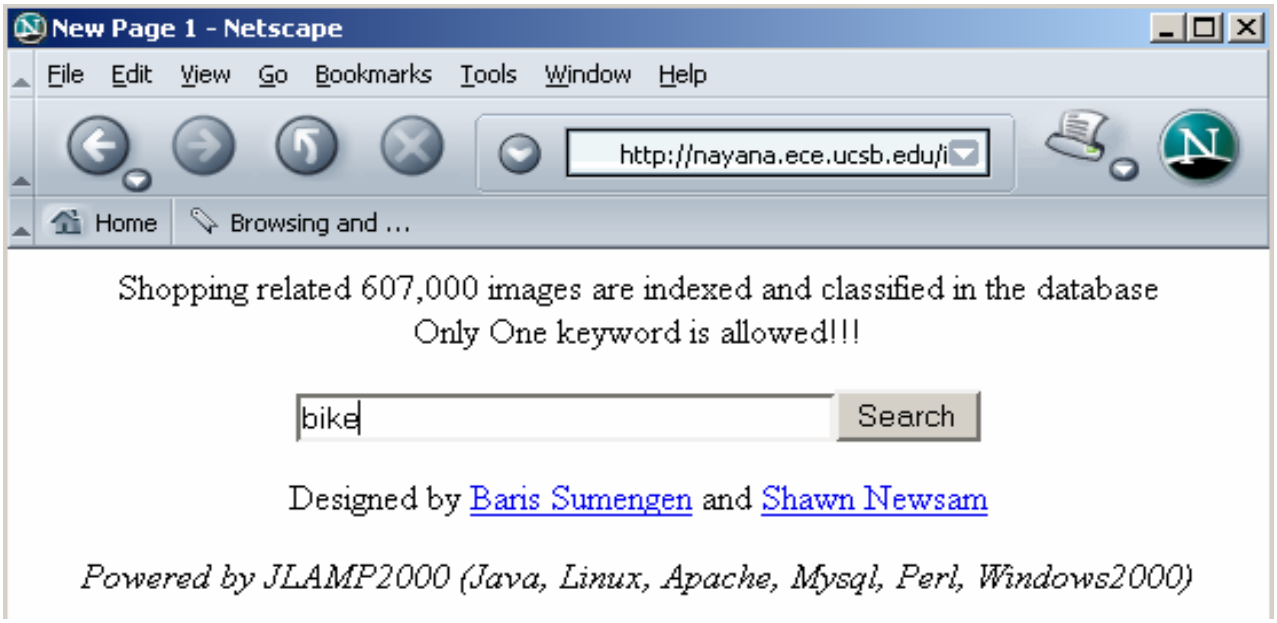
Αρχιτεκτονική για Ανάδραση Συνάφειας



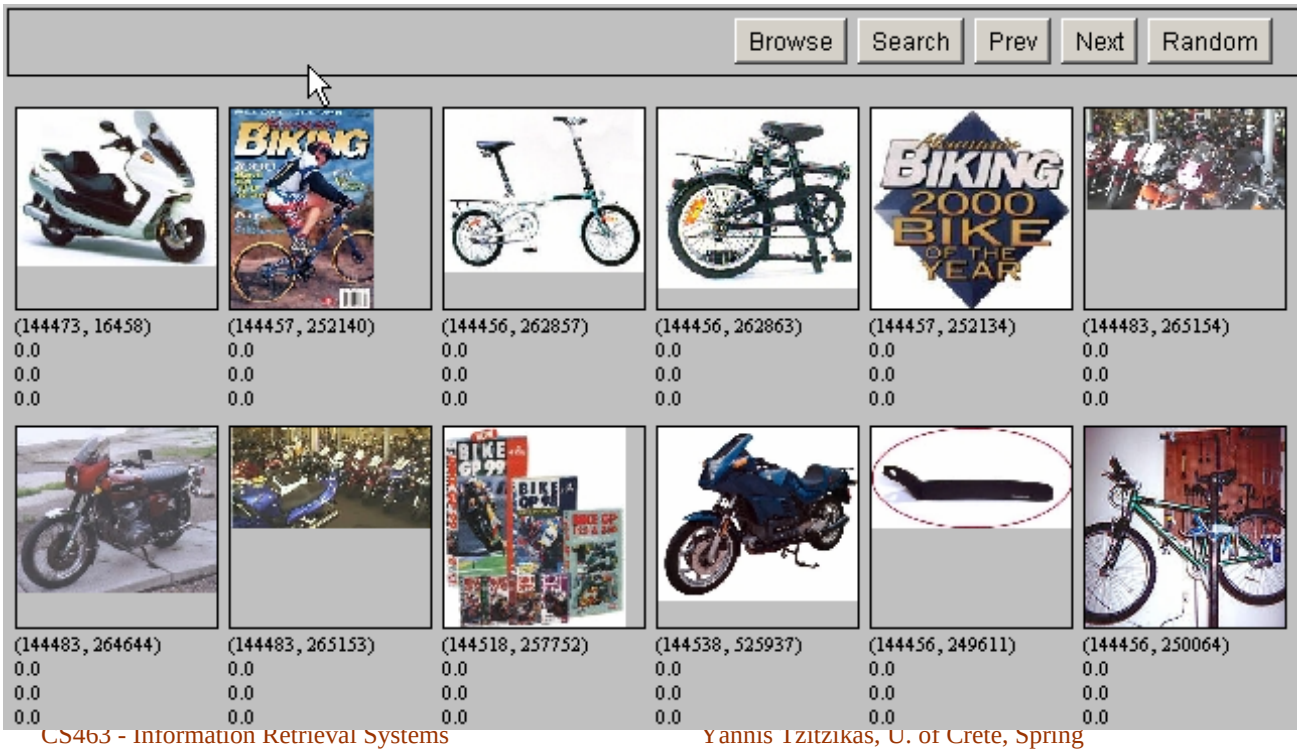
Τμήματα της Αρχιτεκτονικής που Εμπλέκονται



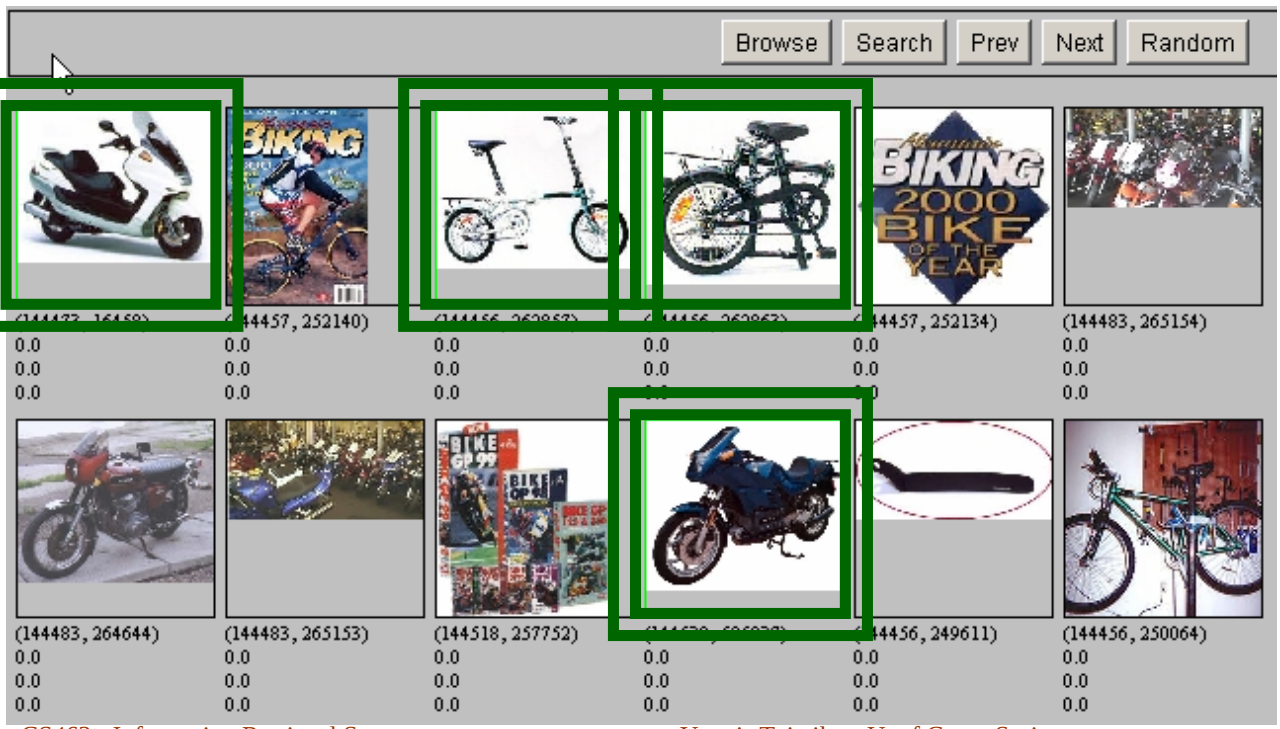
http://nayana.ece.ucsb.edu/imsearch/imsearch.html
q=bike



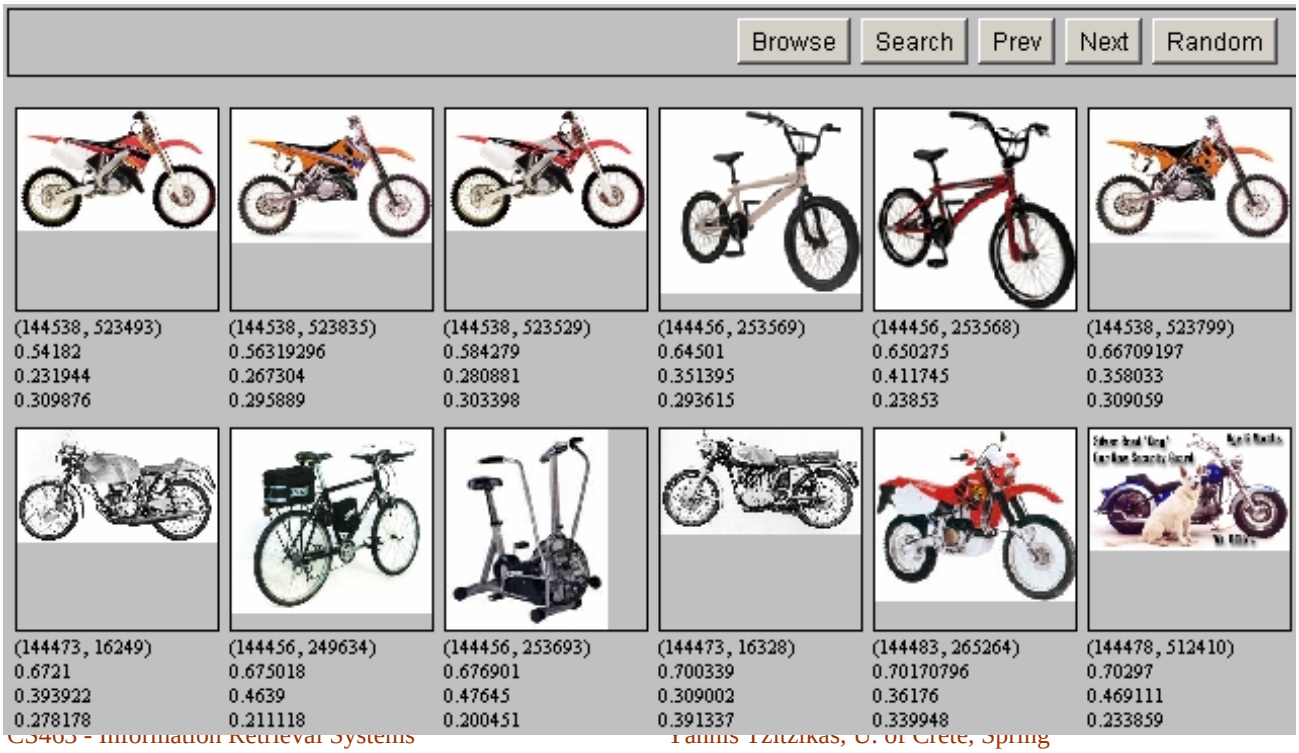
Αποτελέσματα
Ans(«bike»)



Μαρκάρισμα των Συναφών



Απάντηση Αναδιατυπωμένης Επερώτησης



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

Τρόποι αναδιατύπωσης της επερώτησης μετά την ανάδραση:

- **Αναβάρυνση των Όρων (Term Reweighting)**:
  - Αύξηση των βαρών των όρων που εμφανίζονται στα συναφή έγγραφα και μείωση των βαρών των όρων που εμφανίζονται στα μη-συναφή έγγραφα.
- **Επέκταση επερώτησης (Query Expansion):**
  - **Προσθήκη νέων όρων** στην επερώτηση (π.χ. από γνωστά συναφή έγγραφα)
- Υπάρχουν πολλοί αλγόριθμοι για επαναδιατύπωση επερώτησης

Αναδιατύπωση επερώτησης στο **Διανυσματικό Μοντέλο**

Η βέλτιστη επερώτηση (Optimal Query)

- Ας υποθέσουμε ότι γνωρίζουμε το σύνολο C_r **όλων** των συναφών (με την πληροφοριακή ανάγκη του χρήστη) εγγράφων.
- Η «καλύτερη επερώτηση» (αυτή που κατατάσσει στην κορυφή όλα τα συναφή έγγραφα και μόνο αυτά) θα ήταν:

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

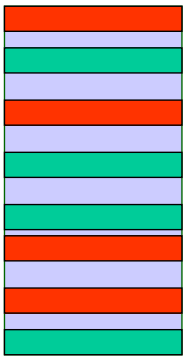
Where N is the total number of documents.

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

Αναδιατύπωση επερώτησης στο **Διανυσματικό Μοντέλο**

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

answer(q):



Τεχνικές
(I) Rochio Method
(II) Ide Method
(III) DeHi Method

(I) Standard Rochio Method

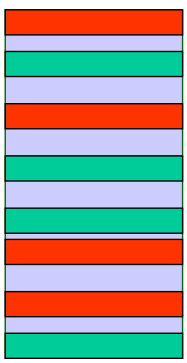
Αφού το σύνολο όλων των συναφών είναι άγνωστο,

χρησιμοποίησε τα **γνωστά** συναφή (D_r) και **γνωστά** μη-συναφή (D_n) 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 γ < β (the relevant docs are more important)
If γ=0 then we have positive feedback only

(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.

(III) Ide "Dec Hi" Method

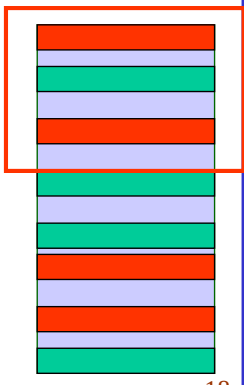
Τάση για απόρριψη **μόνο** των μη-συναφών εγγράφων που έχουν υψηλό σκορ

(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.

answer(q):



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

Αξιολόγηση Αποτελεσματικότητας Τεχνικών Ανάδρασης Συνάφειας

Remarks

- By construction, reformulated query will rank **explicitly-marked relevant** documents higher and **explicitly-marked irrelevant** documents lower.
- 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.

Fair Evaluation of Relevance Feedback

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

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						
		+49%	+44%	+92%	+32%	+79%	+59%
Rocchio (standard $\beta = .75$, $\alpha = .25$ expand by all terms	Rank	2	39	8	14	17	16
	Precision	.2552	.1404	.2955	.1821	.5630	
	Improvement	+75%	+19%	+156%	+33%	+68%	+70%
expand by most common terms	Rank	3	12	12	10	24	
	Precision	.2491	.1623	.2534	.1861	.5279	
	Improvement	+71%	+37%	+119%	+36%	+55%	+64%
Probabilistic (adjusted revised							

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

Rutgers INQUERY

Reset All | UNDO LAST RUN QUERY | Show Search Topic Text | Show Tutorial | Exit RU INQUERY

Enter (next) query term below and hit <RETURN>

Current Query Has 4 term(s):

automobil* manufactur*
car*
defect*
recal*

Run Query

Clear All Marks | You marked 0 documents

1. GM Plans to Recall 62,000 1988-89 Cars With Quad 4 Engines
2. GM, Ford Recall Vehicles to Repair Defective Parts ---- By Neal Templin S
3. Isuzu Motors, Honda Commence Car Recalls ---- A Wall Street Journal News I
4. Ford and GM Recall Series Of Pickup Trucks, Coupes
5. General Motors Corp. Recalls 196,000 Cars For Defective Brakes

Total of 6747 documents retrieved | Jump to rank:

Document # 1 of 6747

GM Plans to Recall
62,000 1988-89 Cars
With Quad 4 Engines

WSJ900413-0013
04/13/90 WALL STREET JOURNAL (J), PAGE B2

DETROIT -- General Motors Corp. said it is recalling 62,000 1988-89 model cars equipped with its high-tech Quad 4 engine to fix defective fuel lines linked to 24 engine fires.

GM said the 1988-89 Pontiac Grand Am, Oldsmobile Cutlass Calais and Buick Skylark cars equipped with the 16-valve, four-cylinder Quad 4 engine have fuel lines that could crack or separate from the engines. Although GM has received reports of 24 fires caused by leaks attributable to the faulty fuel lines, a spokesman says the company knows of no injuries resulting from the incidents. GM sold about 312,000 cars equipped with Quad 4 engines in the 1988-89 model years.

In another action, GM said it is recalling about 3,200 of its 1990 Oldsmobile Cutlass Calais and Buick Skylark models to fix fuel-line defects on three engines: the Quad 4, 3.3-liter V-6, and 2.5-liter four cylinder. GM isn't aware of any fires or injuries related to the fuel line problems in this group of cars, the spokesman said.

All repairs will be done free of charge to owners, the company said.

Separately, the U.S. sales arm of Volkswagen AG's Audi subsidiary said it is recalling 1,600 1990-model Audi 80, 90 and Coupe Quattro luxury cars to replace a defective bolt in the assembly that locks the steering when the car is parked. The defective bolt could break, causing the steering wheel to

Credit: Marti Hearst

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

Criterion: p@30 (precision at 30 documents)

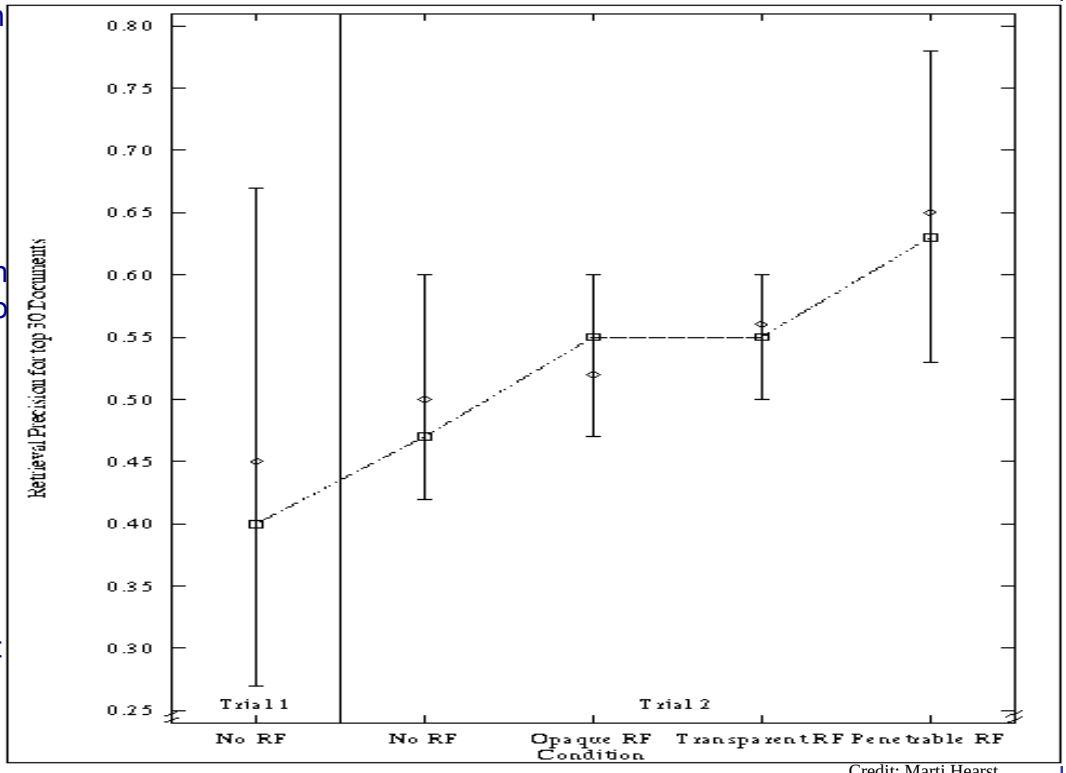
Compare:

- 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***.

- However:
  - User does not always have sufficient initial knowledge.
  - Examples: Misspellings, Mismatch of searcher's vocabulary vs collection vocabulary.

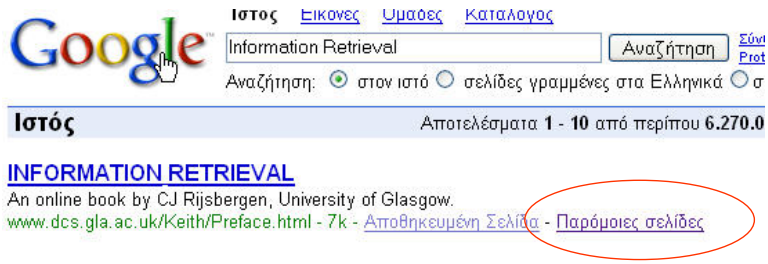
A2: ***Relevance prototypes are "well-behaved".***

- Either: All relevant documents are similar to a single prototype.
- Or: There are different prototypes, but they have significant vocabulary overlap.
- However:
  - There are several relevance prototypes.

Γιατί η ανάδραση συνάφειας δεν χρησιμοποιείται ευρέως;

- Οι χρήστες συχνά διστάζουν να δώσουν είσοδο
- Η ανάδραση έχει ως αποτέλεσμα μεγάλες επερωτήσεις των οποίων ο υπολογισμός απαιτεί περισσότερο χρόνο
  - σε σύγκριση με τις συνηθισμένες επερωτήσεις που διατυπώνουν οι χρήστες οι οποίες αποτελούνται από 2-3 λέξεις
  - (search engines process lots of queries and allow little time for each one)
- Μερικές φορές η νέα απάντηση περιέχει έγγραφα τα οποία δεν μπορούμε να καταλάβουμε πως προέκυψαν

Ανάδραση Συνάφειας στον Παγκόσμιο Ιστό



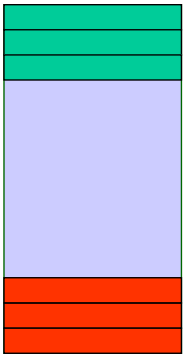
- Some search engines offer a *similar/related pages* feature (simplest form of relevance feedback)
  - Google (link-based)
- But some don't because it's hard to explain to average user.
  - "Excite" initially had true relevance feedback, but abandoned it due to lack of use.

Ψευδοανάδραση Συνάφειας

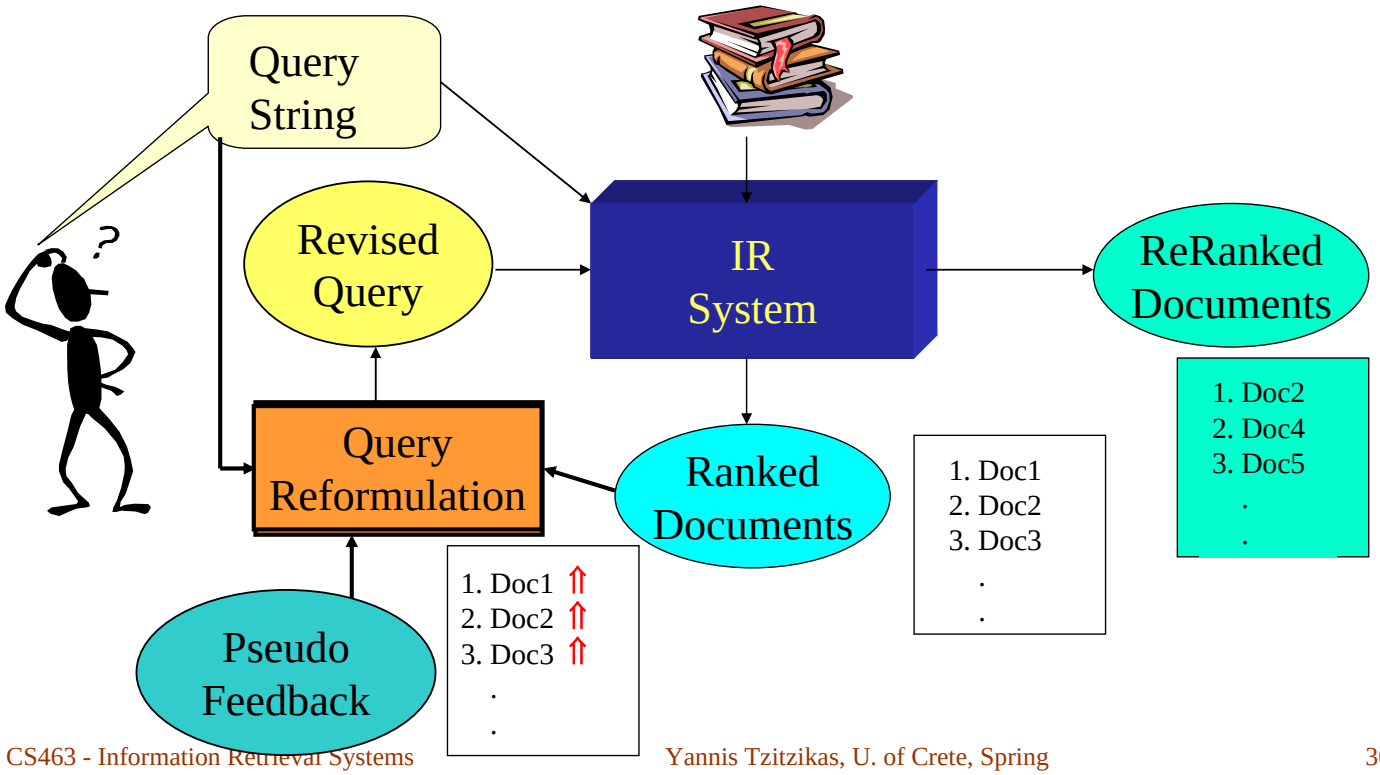
Ψευδοανάδραση Συνάφειας
Pseudo Relevance Feedback

- 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



Αξιολόγηση Ψευδοανάδρασης

- Βρέθηκε να βελτιώνει την απόδοση στο διαγωνισμό του TREC (ad-hoc retrieval task)
- 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)

Επέκταση Επερώτησης
(Query Expansion)

In *relevance feedback*, users give additional input (relevant/non-relevant) on documents.

In *query expansion*, users give additional input (good/bad search term) on words or phrases

(HIDE)
Επέκταση Επερώτησης
(Query Expansion)

- In *relevance feedback*, users give additional input (relevant/non-relevant) on documents.
- In *query expansion*, users give additional input (good/bad search term) on words or phrases

Τεχνικές Επέκτασης Επερώτησης

- **Local Analysis**:
  - Analysis of documents in result set
- **Global Analysis**: Thesaurus-based
  - Controlled vocabulary
    - Maintained by editors (e.g., medline)
  - Automatically derived thesaurus
    - (co-occurrence statistics)
- Refinements based on query log mining
  - Common on the web

Επέκταση Επερώτησης
(Query Expansion)

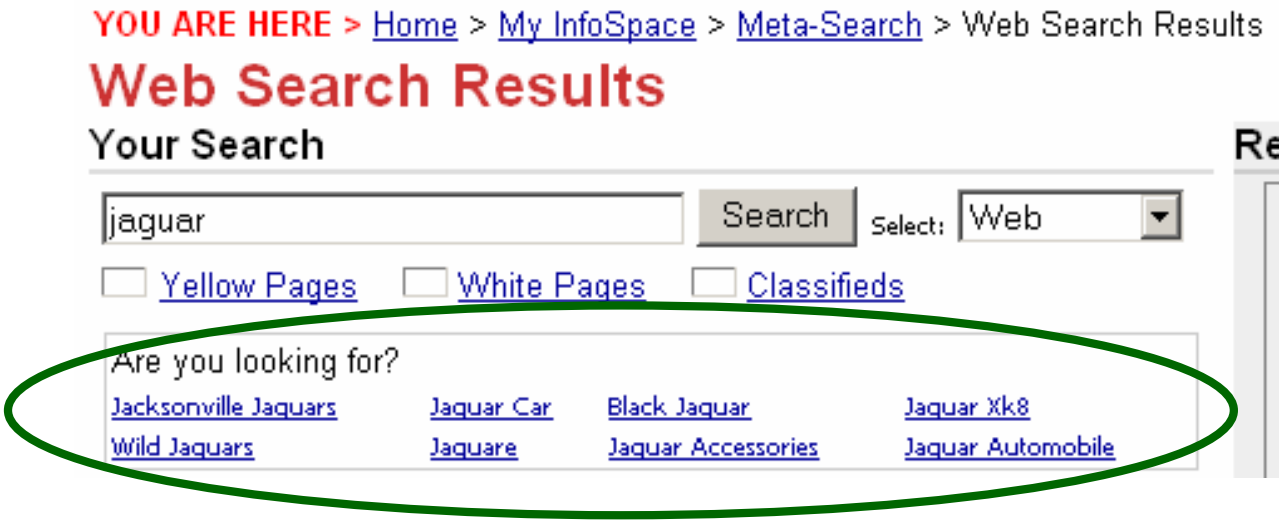
- Τοπική Ανάλυση
  - Αναλύουμε τα (κορυφαία) έγγραφα της απάντησης
- Καθολική Ανάλυση
  - Αναλύουμε όλα τα έγγραφα της συλλογής

Επέκταση Επερώτησης (Query Expansion)
Τοπική Ανάλυση (Local Analysis)

Αυτόματη Τοπική (Επιτόπια) Ανάλυση
Automatic Local Analysis

- Μετά την διατύπωση της επερώτησης, ανέλυσε (στατιστικά) τις λέξεις που εμφανίζονται **μόνο** στα κορυφαία ανακτημένα έγγραφα
  - π.χ. επιλέγουμε τις 10 πιο συχνά εμφανιζόμενες λέξεις των κορυφαίων 5 εγγράφων
- Το σύστημα παρουσιάζει στο χρήστη τις σχετικές λέξεις και ο αυτός επιλέγει εκείνες που θέλει να προστεθούν στην επερώτηση
  - εναλλακτικά η επιλογή μπορεί να γίνει αυτόματα (χωρίς την παρέμβαση του χρήστη)
- Αποτέλεσμα
  - Οι ασαφείς (ή αμφίσημες) λέξεις δημιουργούν λιγότερα προβλήματα (απ' ότι στην καθολική ανάλυση)
  - Παράδειγμα
    - "Apple computer" → "Apple computer Powerbook laptop"

Example of Query Expansion



Also: see altavista, teoma

Αυτόματη Τοπική (Επιτόπια) Ανάλυση

Τεχνικές αυτόματης τοπικής ανάλυσης

- Association Matrix
  - based on the co-occurrence of terms in documents
- Metric Correlation Matrix
  - based on the co-occurrence and proximity of terms in documents
- //Scalar Clusters
- //Local context analysis

(a) Association Matrix and Normalized Association Matrix

	w_1	w_2	w_3	.....	w_n
w_1	c_{11}	c_{12}	c_{13}	.....	c_{1n}
w_2	c_{21}				
w_3	c_{31}				
.	.				
.	.				
w_n	c_{n1}				

c_{ij}: Correlation factor between term i and term j:

$$c_{ij} = \sum_{d_k \in D} f_{ik} \times f_{jk}$$

f_{ik} : frequency of term i in document k

- Normalized Association Matrix
  - Frequency based correlation factor favors more frequent terms.
  - Normalize association scores: $$s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}$$
  - Normalized score is 1 if two terms have the same frequency in all documents.

Από αυτόν τον πίνακα μπορούμε να βρούμε τους όρους που είναι πιο κοντά σε αυτούς της επερώτησης

(b) Metric Correlation Matrix

- Association correlation does not account for the **proximity** of terms in documents, just co-occurrence frequencies within documents.
- **Metric correlations** account for term proximity.

$$c_{ij} = \sum_{k_u \in V_i} \sum_{k_v \in V_j} \frac{1}{r(k_u, k_v)}$$

V_i: Set of all occurrences of term i in any document.
$r(k_u,k_v)$: Distance in words between word occurrences k_u and k_v (∞ if k_u and k_v are occurrences in different documents).

- Normalized Metric Correlation Matrix
  - to account for term frequencies: $$s_{ij} = \frac{c_{ij}}{|V_i| \times |V_j|}$$

Query Expansion with Correlation Matrix

- For each term *i* in query, expand query with *n* terms, those with the highest value of c_{ij}.
- This adds semantically related terms in the "neighborhood" of the query terms.

Αυτόματη Καθολική Ανάλυση
(Automatic Global Analysis)

Αυτόματη Καθολική Ανάλυση
Automatic Global Analysis

- Προσδιορισμός βαθμού ομοιότητας μεταξύ των όρων βάσει στατιστικής ανάλυσης ολόκληρης της συλλογής
- Υπολογισμός πινάκων συσχέτισης (association matrices) που ποσοτικοποιούν την ομοιότητα μεταξύ των όρων ανάλογα με το πόσο συχνά συνεμφανίζονται
- Επέκταση επερώτησης με του πιο όμοιους όρους.
- Τρόποι
  - Query Expansion Based on a Similarity Thesaurus

Προβλήματα Καθολικής Ανάλυσης

- Term ambiguity may introduce irrelevant statistically correlated terms.
  - "Apple computer" → "Apple red fruit computer"
- Since terms are highly correlated anyway, expansion may not retrieve many additional documents.

Query Expansion Based on a **Similarity Thesaurus**

- Keypoint
  - Οι όροι που προστίθενται καθορίζονται με βάση την απόσταση τους **από ολόκληρη την επερώτηση** (και όχι βάσει της απόστασής τους από κάθε όρο της επερώτησης ξεχωριστά)
- Στην αντίθετη περίπτωση θα είχαμε:
  - "Apple computer" → "Apple red fruit computer"
- Ενώ τώρα
  - "fruit" not added to "Apple computer" since it is far from "computer."
  - "fruit" added to "apple pie" since "fruit" close to both "apple" and "pie."

Query Expansion Based on a Similarity Thesaurus

- Τρόπος
  - Έστω N έγγραφα, t όροι K={k1,…,kt}
  - Παριστάνουμε **κάθε όρο** με ένα διάνυσμα στον χώρο των N διαστάσεων
    - (είναι σαν να έχουμε αντιστρέψει το ρόλο των όρων και των εγγράφων)

$$\vec{k}_i = (w_{i1},...,w_{iN})$$

$$w_{ij} = \frac{(0.5+0.5\frac{f_{ij}}{\max_j(f_{ij})})itf_j}{\sqrt{\sum_{l=1}^{N}(0.5+0.5\frac{f_{il}}{\max_l(f_{il})})^2 itf_j^2}}$$

Query Expansion Based on a Similarity Thesaurus (II)

- Η σχέση μεταξύ δυο όρων $c_{u,v} = \vec{k}_u \cdot \vec{k}_v$
- Query Expansion
  - (1) Represent query in the concept space $\vec{q} = \sum_{ki \in q} w_{iq} \vec{k}_i$
  - (2) Compute sim(q,ku) for each ku $sim(q, k_u) = \vec{q} \cdot \vec{k}_u$
  - (3) Expand q with the top r ranked terms
- Results
  - 20% improved retrieval performance $w_{uq'} = \frac{sim(q, k_u)}{\sum_{k_i \in q} w_{iq}}$

Καθολική vs. Επιτόπια Ανάλυση

- Η καθολική ανάλυση έχει μεγάλο υπολογιστικό κόστος αλλά μόνο στην αρχή
  - υποθέτοντας ότι τα έγγραφα της συλλογής είναι σταθερά
- Η τοπική ανάλυση έχει αρκετό υπολογιστικό κόστος για κάθε επερώτηση
  - (παρόλο που το πλήθος των όρων και των εγγράφων είναι μικρότερο αυτού της καθολικής)
- Η τοπική ανάλυση δίδει καλύτερα αποτελέσματα

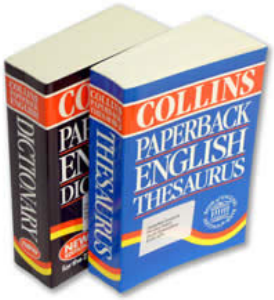
Επέκταση επερωτήσεων: Συμπεράσματα

- Η επέκταση των επερωτήσεων με σχετιζόμενους όρους μπορεί να βελτιώσει την αποτελεσματικότητα της ανάκτησης, ιδιαίτερα την ανάκληση (recall).
- Η αλόγιστη επιλογή σχετιζόμενων όρων μπορεί να μειώσει την ακρίβεια (precision).

Θησαυροί Όρων και Καθολική Ανάλυση

Θησαυροί Όρων

- Ένας θησαυρός παρέχει πληροφορίες για συνώνυμα κα σημασιολογικά κοντινές λέξεις και φράσεις [see also Sec 7.2.5]



- Παράδειγμα:

```
physician
  syn: ||croaker, doc, doctor, MD, medical, mediciner,
medico, ||sawbones
  rel: medic, general practitioner, surgeon,
```

- Online-θησαυροί:
  - Roget's thesaurus
  - INSPEC thesaurus
  - WordNet (http://wordnet.princeton.edu/)
  - The free dictionary http://www.thefreedictionary.com/

Χρήσεις Θησαυρού

- Ευρετηρίαση κειμένων/βιβλίων με επιλογή όρου από θησαυρό
- Αναζήτηση χρησιμοποιώντας όρους του θησαυρού
  - (αυτόματη ή ύστερα από επιλογή του χρήστη)
- Για βελτίωση της ανάκτησης
  - Αν η απάντηση μιας επερώτησης είναι μικρή, μπορούμε να προσθέσουμε όρους βάσει των σχέσεων του θησαυρού (συνώνυμα, ..)
  - Αν απάντηση είναι πολύ μεγάλη, μπορούμε να συμβουλευτούμε το θησαυρό και να αντικαταστήσουμε κάποιους όρους της επερώτησης με πιο ειδικούς.

Διάκριση Θησαυρών

- Γλωσσικοί Θησαυροί
  - Πχ Roget's thesaurus. Designed to assist the writer in creatively selecting covabulary
- Θησαυροί κατάλληλοι για Information Retrieval
  - for coordinating the basic processes of indexing and retrieval
  - designed for speific subject areas and are therefore domain dependent
  - Examples
    - INSPEC

INSPEC thesaurus (for IR)

- Domain: physics, electrical engineering, electronics, computers
- Example:
  - **computer-aided instruction**
    - **see also education**
    - **UF teaching machines (UF: Used For, converse: USE)**
    - **BT educational computing (BT: Broader Term)**
    - **TT computer applications (TT: Top Node, I.e. root of the hierarchy)**
    - **RT education , teaching (RT: Related Term)**

WordNet (http://wordnet.princeton.edu/)

- A more detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144,000 English words.Nouns, adjectives, verbs, and adverbs grouped into about 109,000 synonym sets called *synsets*.

Synset Relationships

- Antonym: front → back
- Attribute: benevolence → good (noun to adjective)
- Pertainym: alphabetical → alphabet (adjective to noun)
- Similar: unquestioning → absolute
- Cause: kill → die
- Entailment: breathe → inhale
- Holonym: chapter → text (part-of)
- Meronym: computer → cpu (whole-of)
- Hyponym: tree → plant (specialization)
- Hypernym: fruit → apple (generalization)

AAT (Art and Architecture Thesaurus)

- Controlled vocabulary for describing and retrieving information: fine art, architecture, decorative art, and material culture.
- Almost 120,000 terms for objects, textual materials, images, architecture and culture from all periods and all cultures.
- Used by archives, museums, and libraries to describe items in their collections.
- Used to search for materials.
- Used by computer programs, for information retrieval, and natural language processing.

Χαρακτηριστικά Θησαυρών

- Coordination Level (βαθμός συντονισμού)
  - refers to the construction of phrases from individual terms
  - ***precoordination***: the thesaurus contain phrases
    - + the vocabulary is very precise
    - - the user has to be aware of the phrase construction rules, large size
  - ***postcoordination***: the thesaurus does not contain phrases. They are constructed while indexing/searching
    - + user does not worry about the order of the words
    - - precision may fall
- Term Relationships
  - equivalence relations (e.g. synonymy)
  - hierarchical relations (e.g. dogs BT animals,)
  - nonhierarchical relations (e.g. RT)

Χαρακτηριστικά Θησαυρών (2)

- Number of Entries per Term
  - preferably: a single entry for each thesaurus term
  - however homonyms does not make this possible
    - parenthetical qualifiers:
      - bonds(chemical), bonds(adhesive) // χημικός δεσμός / υλικό συγκόλλησης
- Specificity of Vocabulary
  - high specificity -> large vocabulary size
- Control of Term Frequency of Class Members (for statistical thesauri)
  - the terms of a thesaurus should have roughly equal frequencies
  - the total frequency in each class (of terms) should be equal
- Normalization of Vocabulary
  - terms should be in noun form
  - other rules related to singularity of terms, spelling, capitalization, abbreviations, initials, acronyms, punctuation

Τρόποι Κατασκευής Θησαυρών

[A] Χειροποίητη Δημιουργία

[B] Αυτόματη Κατασκευή

[B.1] από **συλλογή κειμένων**

Προϋπόθεση: Να υπάρχει μια μεγάλη και αντιπροσωπευτική συλλογή κειμένων

[B.2] από **συγχώνευση** άλλων θησαυρών

Προϋπόθεση: Να υπάρχουν >2 διαθέσιμοι θησαυροί για την περιοχή που μας ενδιαφέρει

[A] Χειροποίητοι Θησαυροί

- Define subject boundaries
- partition into divisions and subject areas
- collection of terms
  - sources: encyclopedias, handbooks, textbooks, journal titles, catalogues, other thesauri, subject experts, potential users
- analysis of terms (synonyms, hierarchical structure, definitions, scope notes)
- reviewing phase

- delivery in both hierarchical and in alphabetical arrangement
- maintenance (new terms, etc)

Very long, laborious and costly process

Faceted Taxonomies + CTCA

- A new flexible and fast approach (to be discussed in another lecture)

Επέκταση επερωτήσεων βάσει Θησαυρού
Thesaurus-based Query Expansion

- Τρόπος:
  - Γιά κάθε όρο t της επερώτησης, πρόσθεσε στην επερώτηση τα συνώνυμα και τις σχετικές λέξεις (related terms) του t
  - Τα βάρη των νέων λέξεων μπορεί να είναι χαμηλότερα των βαρών των λέξεων της αρχικής επερώτησης
  - E.g. of a WordNet-based Query Expansion
    - Add synonyms in the same synset.
    - Add hyponyms to add specialized terms.
    - Add hypernyms to generalize a query.
    - Add other related terms to expand query.
- Αποτέλεσμα
  - Αυξάνει την ανάκληση (recall.)
  - Μπορεί να μειώσει την ακρίβεια (precision), ιδιαίτερα όταν η επερώτηση περιέχει αμφίσημες λέξεις
    - "interest rate" → "interest rate fascinate evaluate"

[B1] Αυτόματη Κατασκευή Θησαυρών από Κείμενα

- Η κατασκευή (από ανθρώπους) ενός θησαυρού είναι πολύ **χρονοβόρα** και δεν υπάρχουν θησαυροί για όλες τις γλώσσες
- Οι πληροφορίες που μπορούμε να χρησιμοποιήσουμε από έναν θησαυρό περιορίζονται στις σχέσεις που υποστηρίζει ο θησαυρός
- **Ιδέα: Μπορούμε να ανακαλύψουμε σημασιολογικές σχέσεις μεταξύ λέξεων αναλύοντας στατιστικά μια μεγάλη συλλογή κειμένων**
- Στάδια
  - ***1/ Κατασκευή λεξιλογίου***
  - ***2/ Υπολογισμός ομοιότητας μεταξύ όρων***
  - ***3/ Οργάνωση (συνήθως ιεραρχική) του λεξιλογίου***

Αυτόματη Κατασκευή Θησαυρών από Κείμενα (II)

- 1/ Κατασκευή Λεξιλογίου
  - Decision: Desired specificity
    - if high then emphasis will be given on identifying precise phrases
  - Terms can be selected from titles, abstracts, or even the full text
  - Normalization: stemming, stoplists
  - Criteria for selecting a term:
    - frequency of occurrence (divide words to 3 categoris: low, medium, high, select terms with medium frequency)
    - discrimination value ~ idf
  - Phrase construction (if desired, recall coordination level)
- 2/ Υπολογισμός Ομοιότητας
  - Παραδείγματα μετρικών: Cosine, Dice

Αυτόματη Κατασκευή Θησαυρών από Κείμενα (III)

3/ Οργάνωση (συνήθως ιεραρχική) του λεξιλογίου

- Οποιοσδήποτε αλγόριθμος clustering μπορεί να χρησιμοποιηθεί

Ένας Αλγόριθμος

1/ Identify a set of frequency ranges

2/ Group the vocabulary terms into different classes based on their frequencies and the ranges selected in Step 1. There will be one term class for each frequency range

3/ The highest frequency class is assigned level 0, the next level 1, and so on

4/ Parent-child links: The parent(s) of a term at level i is the most similar term in level i-1 (a term is allowed to have multiple parent)

5/ Continue until reaching level 1

Παράδειγμα με 3 κλάσεις συχνοτήτων