



HY463 - Συστήματα Ανάκτησης Πληροφοριών Information Retrieval (IR) Systems

Προχωρημένες Λειτουργίες Επερώτησης Advanced Query Operations

Γιάννης Τζίτζικας

Διάλεξη : 11

Ημερομηνία : 2-5-2007



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
 - 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)
- **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



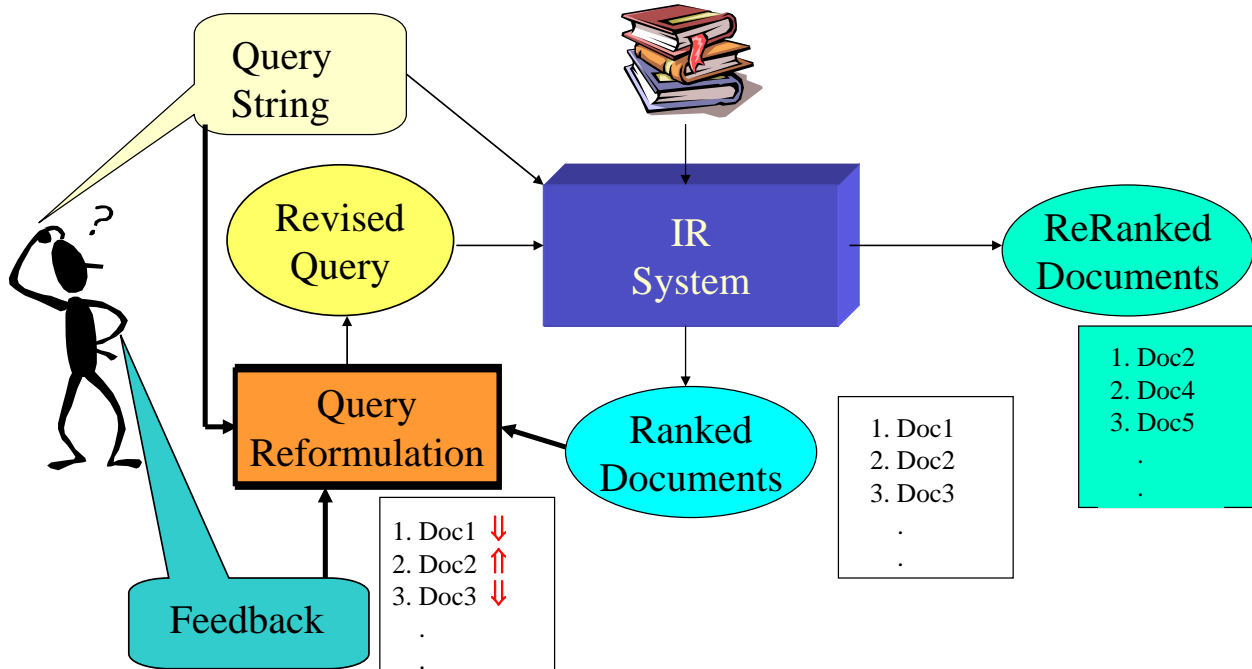
Ανάδραση Συνάφειας (Relevance Feedback): Η βασική ιδέα

Βήματα:

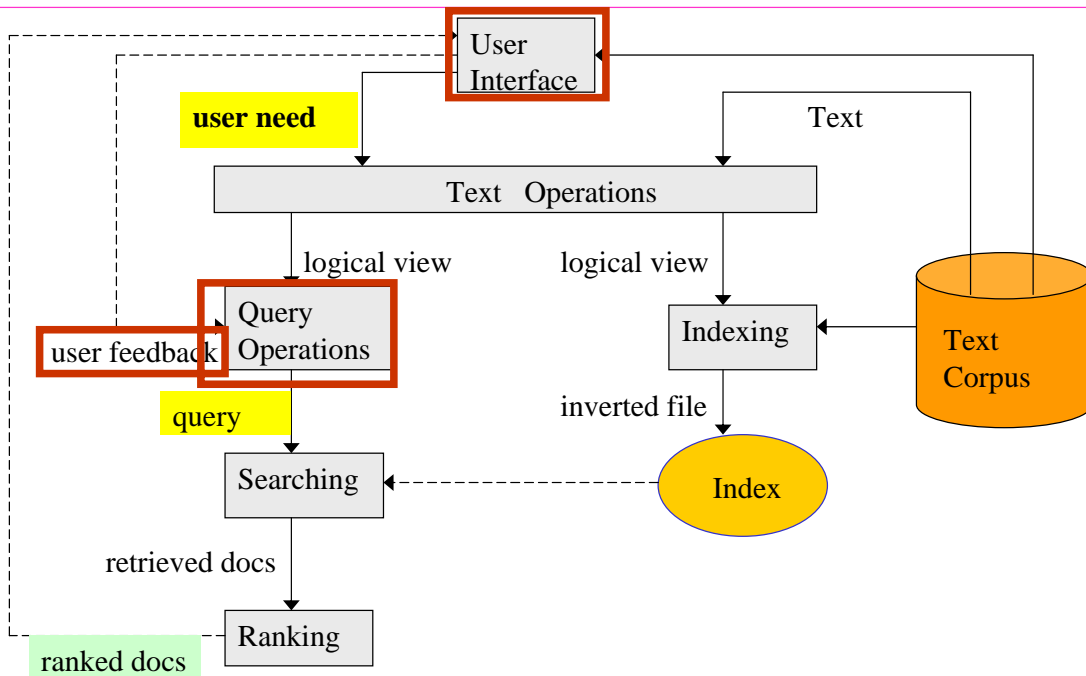
- 1/ Μετά την παρουσίαση των αποτελεσμάτων, επιτρέπουμε στο χρήστη **να κρίνει την συνάφεια** ενός ή περισσότερων εγγράφων της απάντησης
- 2/ Αξιοποιούμε αυτήν την πληροφορία για να **αναδιατυπώσουμε** την επερώτηση
- 3/ Κατόπιν δίδουμε στο χρήστη την απάντηση της αναδιατυπωμένης επερώτησης
- 4/ Πήγαινε στο βήμα 1/



Αρχιτεκτονική για Ανάδραση Συνάφειας

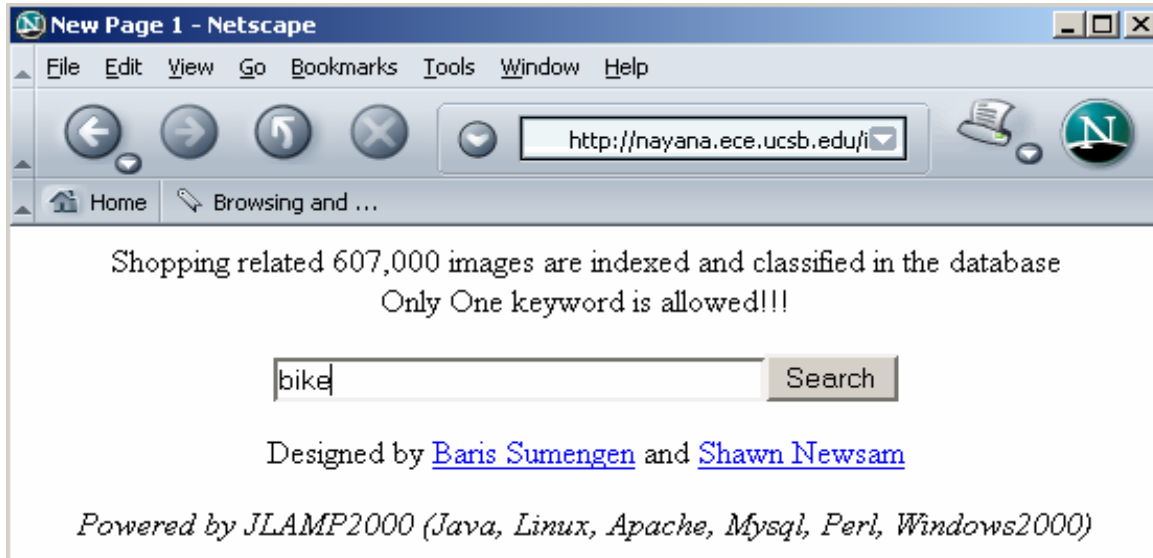


Τμήματα της Αρχιτεκτονικής που Εμπλέκονται

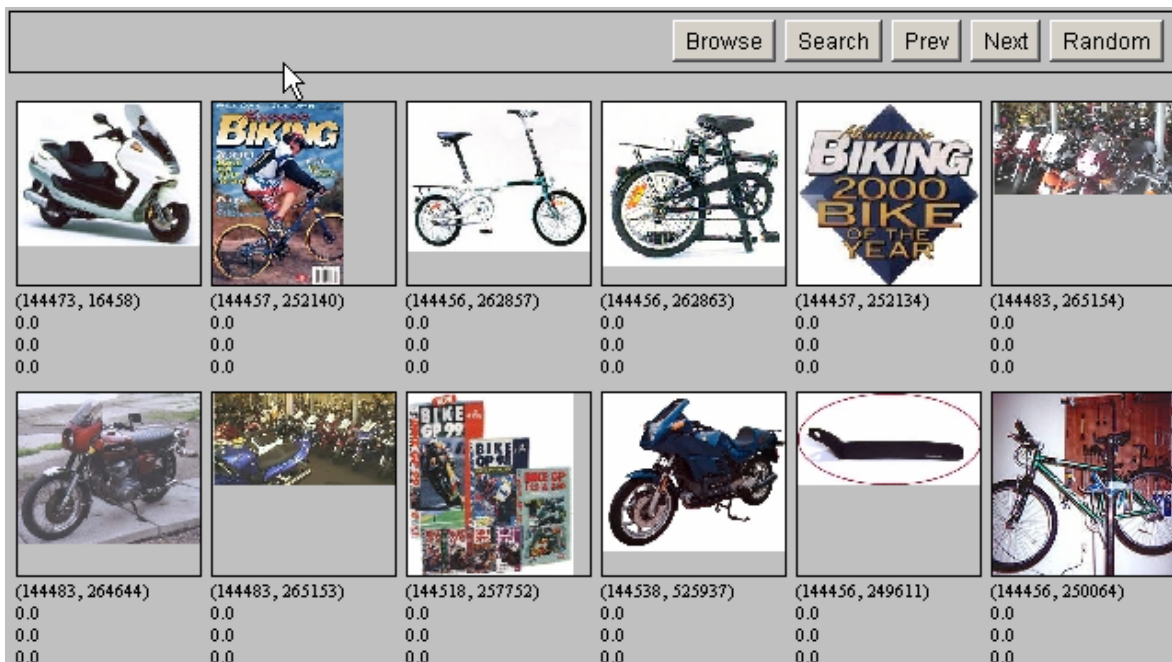




<http://nayana.ece.ucsb.edu/imsearch/imsearch.html>
q=bike















Αποτελέσματα
Ans(«bike»)





Μαρκάρισμα των Συναφών από τον Χρήστη













Browse Search Prev Next Random

					
(144473, 16458)	(144457, 252140)	(144456, 263952)	(144456, 263962)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144473, 16328)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0



Απάντηση Αναδιατυπωμένης Επερώτησης

Browse Search Prev Next Random

					
(144458, 523493)	(144458, 523835)	(144458, 523529)	(144456, 253569)	(144456, 253568)	(144458, 523799)
0.54182	0.56319296	0.584279	0.64501	0.650275	0.66709197
0.231944	0.267304	0.280881	0.351395	0.411745	0.358033
0.309876	0.295889	0.303398	0.293615	0.23853	0.309059
					
(144473, 16249)	(144456, 249634)	(144456, 253693)	(144473, 16328)	(144483, 265264)	(144478, 512410)
0.6721	0.675018	0.676901	0.700339	0.70170796	0.70297
0.393922	0.4639	0.47645	0.309002	0.36176	0.469111
0.278178	0.211118	0.200451	0.391337	0.339948	0.233859



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) θα ήταν:

$$q_{opt} = \frac{1}{|C_r|} \sum_{\forall d_j \in C_r} d_j^{\rho} - \frac{1}{N - |C_r|} \sum_{\forall d_j \notin C_r} d_j^{\rho}$$

Where N is the total number of documents.

(θα αναλύσουμε περισσότερο αυτό το ζήτημα αργότερα)

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):



Κόκκινα: ο χρήστης έδωσε αρνητική ανάδραση

Πράσινα: ο χρήστης έδωσε θετική ανάδραση

Μπλε: ο χρήστης δεν έδωσε ανάδραση



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



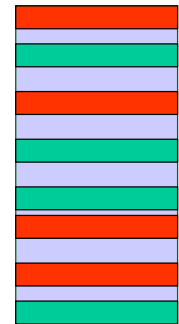
(I) Standard Rochio Method

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:

$$q_m^p = \alpha q^p + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} d_j^p - \frac{\gamma}{|D_n|} \sum_{\forall d_j \in D_n} d_j^p$$

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



(II) Ide Regular Method

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

$$q_m^p = \alpha q^p + \beta \sum_{\forall d_j \in D_r} d_j^p - \gamma \sum_{\forall d_j \in D_n} d_j^p$$

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



(III) Ide “Dec Hi” Method

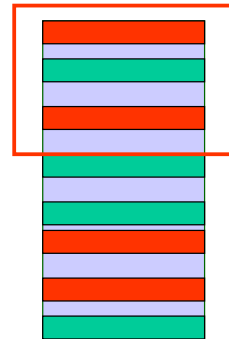
Τάση για απόρριψη **μόνο** των μη-συναφών εγγράφων που έχουν υψηλό σκορ

(Bias towards rejecting **just** the highest ranked of the irrelevant documents:)

$$q_m^p = \alpha q^p + \beta \sum_{\forall d_j \in D_r} d_j^p - \gamma \max_{non-relevant} (d_j^p)$$

- 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)

$$q_m^p = \alpha q^p + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} d_j^p - \frac{\gamma}{|D_n|} \sum_{\forall d_j \in D_n} d_j^p$$

$$q_m^p = \alpha q^p + \beta \sum_{\forall d_j \in D_r} d_j^p - \gamma \sum_{\forall d_j \in D_n} d_j^p$$

$$q_m^p = \alpha q^p + \beta \sum_{\forall d_j \in D_r} d_j^p - \gamma \max_{non-relevant} (d_j^p)$$

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



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)							
Idle (dec hi)		.1459	.1184	.1156	.1368	.3346	
expand by	Rank						
Rocchio (standard $\beta = .75, \alpha = .25$)		+49%	+44%	+92%	+32%	+79%	+59%
expand by all terms	Rank	2	39	8	14	17	16
	Precision	.2552	.1404	.2955	.1821	.5630	
expand by most common terms	Improvement	+75%	+19%	+156%	+33%	+68%	+70%
	Rank	3	12	12	10	24	
	Precision	.2491	.1623	.2534	.1861	.5279	
Probabilistic (adjusted revised derivation)	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



The screenshot shows the Rutgers INQUERY interface. At the top, there are buttons for "Reset All", "UNDO LAST RUN QUERY", "Show Search Topic Text", "Show Tutorial", and "Exit RU INQUERY". Below these is a text input field for the query and a "Run Query" button. The current query is "automobil* manufactur* car* defect* recal*". To the right, a list of search results is displayed, each with a checkbox and a title. The first result is selected. Below the list, the full text of the selected document is shown, including the title "GM Plans to Recall 62,000 1988-89 Cars With Quad 4 Engines" and the start of the article text.

Enter (next) query term below and hit <RETURN> Clear All Marks You marked 0 documents

Current Query Has 4 term(s):
automobil* manufactur*
car*
defect*
recal*

Run Query

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 remain locked even after the driver starts the car and begins



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

Criterion: $p@30$ (precision at 30 documents)

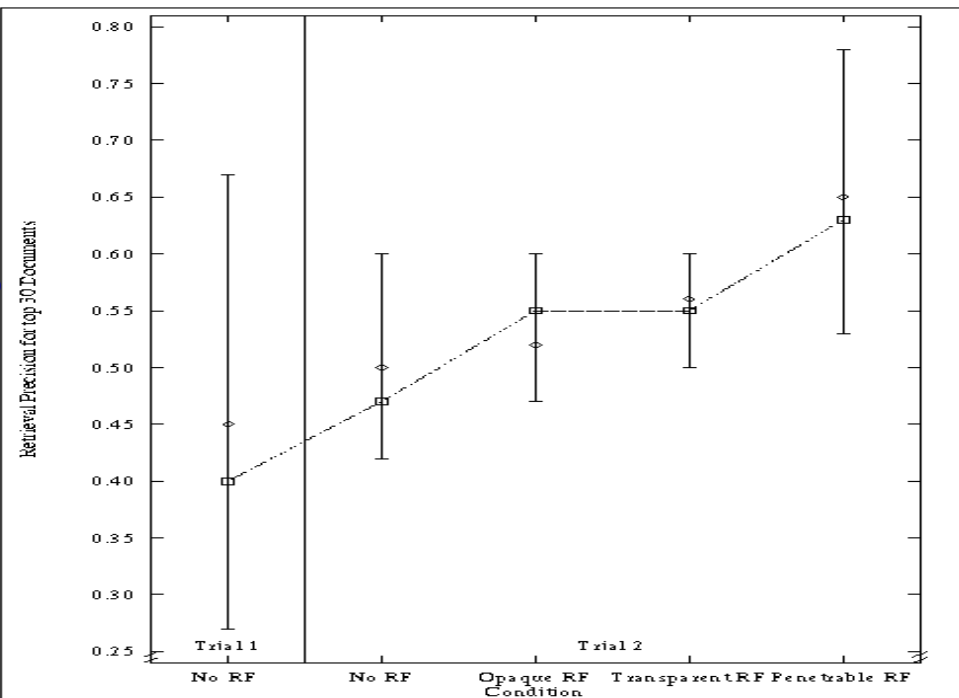
Compare:

- $p@30$ 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



CS463 - Information Retrieval Systems

Yannis Tzitzikas, U. of Crete, Spring 2007

Credit: Marti Hearst

25



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

The screenshot shows a Google search interface. The search bar contains the text 'Information Retrieval'. Below the search bar, there are radio buttons for search preferences: 'στον ιστό' (selected), 'σελίδες γραμμένες στα Ελληνικά', and 'σ'. Below the search results, the first result is titled 'INFORMATION RETRIEVAL' and is described as 'An online book by CJ Rijsbergen, University of Glasgow'. At the bottom of the result, there is a link labeled 'Παρόμοιες σελίδες' (Similar/related pages), which is circled in red in the original image.

- 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

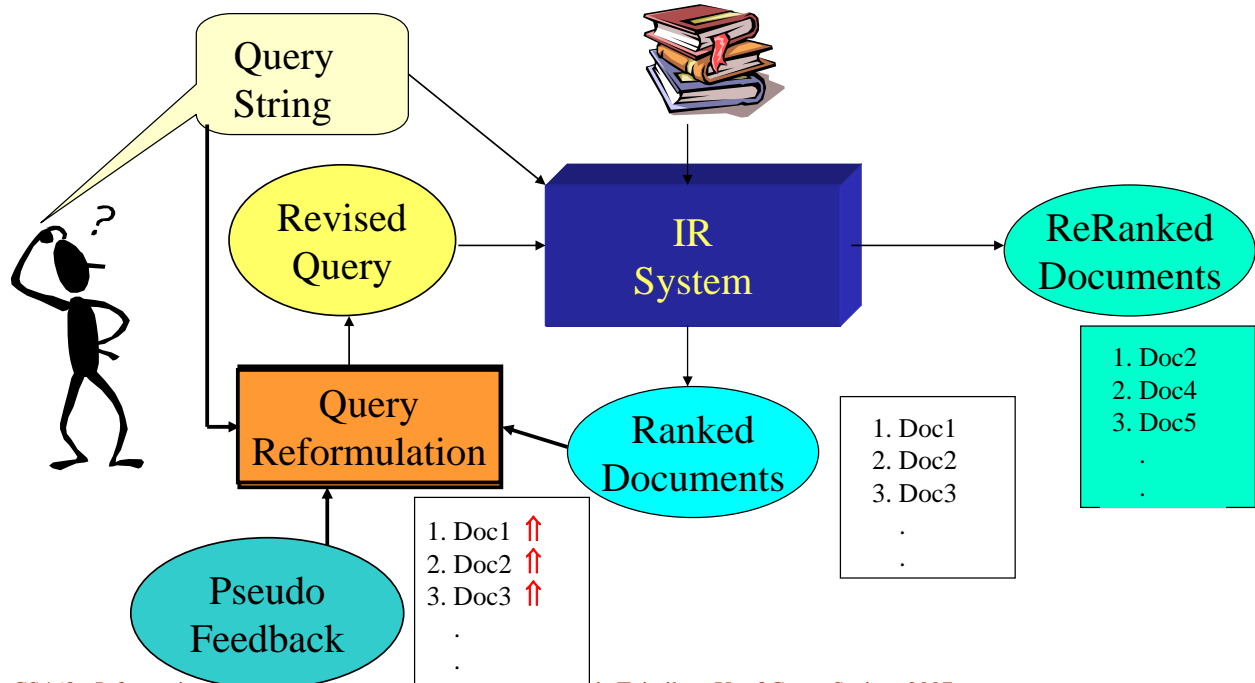
- Χρήση μεθόδων ανάδρασης αλλά **χωρίς είσοδο από το χρήστη**
- **Υπόθεση** ότι τα **κορυφαία m** από τα ανακτημένα έγγραφα είναι συναφή (και χρήση αυτών για ανάδραση)
 - Μπορούμε επίσης να χρησιμοποιήσουμε τα τελευταία έγγραφα για αρνητική ανάδραση
- Επιτρέπει την επέκταση της επερώτησης με όρους που 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)



Αναλύοντας περισσότερο την έννοια της βέλτιστης επερώτησης

Πηγή:

Yannis Tzitzikas and Yannis Theoharis, **Naming Functions for the Vector Space Model**, *29th European Conference on Information Retrieval, Rome 2-5 April 2007*



The Naming Problem

We can view an IR system as a function from set of Queries to set of Answers

$$S: \text{Queries} \longrightarrow \text{Answers}$$

Classically IR systems are good at solving the equation $S(q)=A$ wrt A, i.e.:

$$S(q) = ?$$

The **naming problem** is the problem of solving the equation wrt q, i.e. :

$$S(?) = A$$

We can distinguish two formulations of the naming problem:

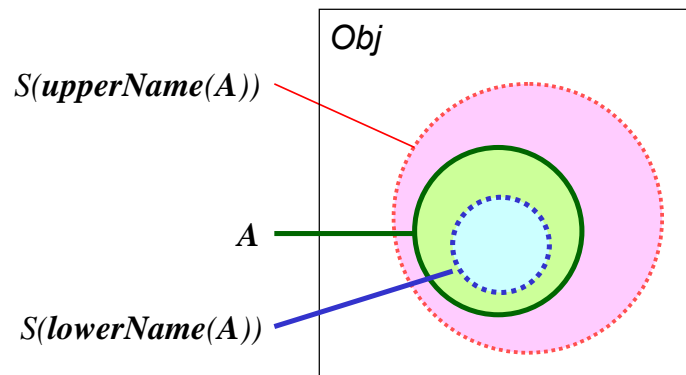
- For **unordered** answers (here A is a subset of Obj)
 - For **ordered** answers (here A is an ordered subset of Obj).
- where Obj is the set of stored objects (e.g. documents)



The Naming Problem for **Unordered** Sets

Basic Notions

- **Exact** name
- **Upper** name
 - **Best Upper** Name
- **Lower** name
 - **Best Lower** Name
- **Relaxed** name



The Naming Problem for **Ordered** Sets

Basic Notions

- **Exact** name
- **Upper** name
 - **Best Upper** Name
- **Lower** name
 - **Best Lower** Name
- **Relaxed** name

$$A = \langle \mathbf{d1}, \mathbf{d2}, \mathbf{d3} \rangle$$

$$S(\mathit{exactName}(A)) = \langle \mathbf{d1}, \mathbf{d2}, \mathbf{d3}, d8, d9, \dots \rangle$$

$$S(\mathit{lowerName}(A)) = \langle \mathbf{d1}, \mathbf{d2}, d5, d7, \dots \rangle$$

$$S(\mathit{upperName}(A)) = \langle \mathbf{d1}, d9, \mathbf{d2}, d5, \mathbf{d3}, d8, \dots \rangle$$



Notations

- $A(k)$: the **ordered** set comprising the **first k** elements of A
- $A\{k\}$: the **set** of elements that appear in $A(k)$
- $A|F$: the **restriction** of A on the set F , i.e. the **ordered set** obtained if we **exclude** from A those elements that do not belong to F ,

Example

if $A = \langle d1, d2, d3 \rangle$ then

- $A(2) = \langle d1, d2 \rangle$
- $A\{2\} = \{d1, d2\}$
- $\{A\} = \{d1, d2, d3\}$
- if $F = \{d1, d3\}$, then $A|_F = \langle d1, d3 \rangle$



Defining Formally Relaxed/Upper/Lower/Exact Names

A query is a **relaxed name** of an answer A iff :

(A is a set) : $|S(q)\{m\} \cap A| = j$ where $m \geq j > 0$.

(A is an ordered set) : $S(q)(m)_{\{A(j)\}} = A(j)$ where $m \geq j > 0$.

- If $m=j=|A|$ then q is an **exact** name
- If $m>j=|A|$ then q is an **upper** name (the **best upper** name if m is the least possible)
- If $m=j<|A|$ then q is a **lower** name (the **best lower** name if m is the greatest possible)

- So each query can be characterized by a pair (m,j) . We can now define an ordering over these pairs.

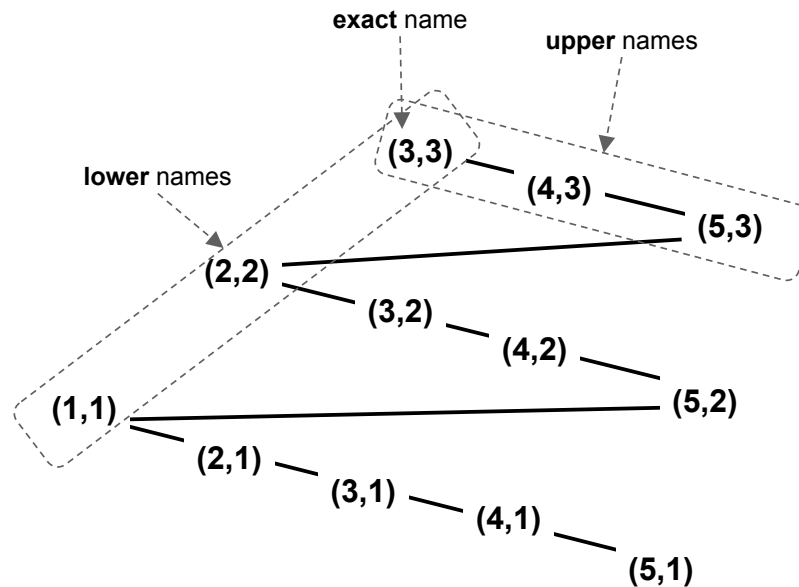


Defining a total ordering of queries

$(m,j) > (m',j')$ iff: $(j > j')$ or $(j = j'$ and $m < m')$

Let $|\text{Obj}|=5$ and $|A|=3$

Let $|\text{Obj}|=5$ and $|A|=3$.
The ordering of the
corresponding pairs are:



Investigating possible approaches for solving the Naming Problem for **Unordered Sets**

Let $A = \{d_1, d_2, d_3, \dots, d_n\}$

Possible name queries

- $q_a = \frac{1}{2}(d_i, d_j)$ where $(d_i, d_j) = \arg \max \{ \text{dist}(d, d') \mid d, d' \in A \}$
- $q_b = 1/|A| \sum \{ d \mid d \in A \}$
- $q_c = 1/|A| \sum \{ a \mid d \in A \} - 1/|\text{Obj}-A| \sum \{ d \mid d \notin A \}$

Notes

- q_a : minimizes the maximum dist from the elements of A
- q_b : minimizes the average dist from the elements of A
- q_c : is the Rocchio method (avg A , - avg $\text{Obj}-A$)

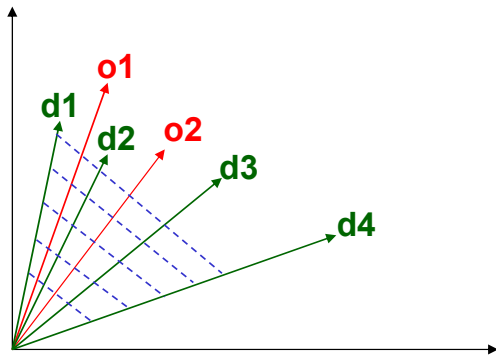
Important remarks:

- **None** is guaranteed to be the exact name (they are all **relaxed** names)
- However, if an exact name does not exist, then q_a is the **best upper** name
- Moreover, the evaluation of q_a (i.e. the computation of its answer) is **faster** than the evaluation of q_b, q_c .



The Naming Problem for **Unordered** Sets (III)

Method (a): Let $A = \{d1, d2, d3, d4\}$



1st step:

Find the pair of most distant documents

Here: $(d1, d4)$

2nd step:

Find whether other documents lie in the area specified by $(d1, d4)$

Here: $\{o1, o2\} \neq \emptyset$

Thus, $qa = \frac{1}{2}(d1, d4)$ is an upper name of A

- Cost
 - 1st step: $O(|A|^2)$ computations of similarity
 - 2nd step: $O(|Obj|)$ or compute the answer of qa and check if $S(qa)\{A\}=A$

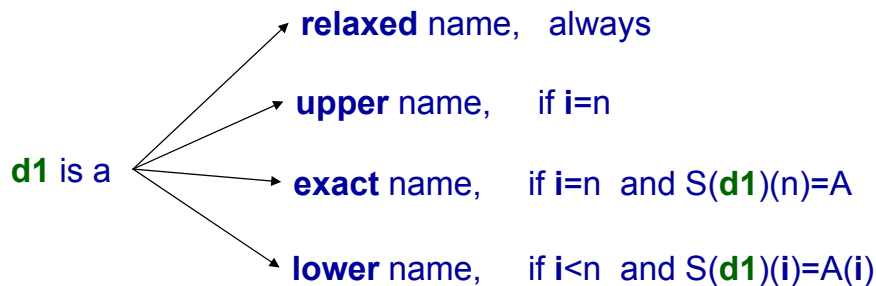


The Naming Problem for **Ordered** Sets

Let $A = \langle d1, d2, d3, \dots, dn \rangle$

Let i the max integer for which it holds:

$$\text{sim}(d1, d2) > \text{sim}(d1, d3) > \dots > \text{sim}(d1, di)$$



Computational cost:

- $O(n)$ computations of similarity to find i .
- One query evaluation in order to decide whether $d1$ is lower/exact name.



Experimental Evaluation

- Experiments conducted over the experimental web search engine GRooble
 - <http://grooble.csd.uoc.gr:8080/google>
- The set A was selected randomly, the average times are listed in the Table below.

Collection			Naming Functions (in sec)					
			Unordered			Ordered		
$ O $	$ T $	$ \{A\} $	t_a	$t_b(\text{query terms})$	t_c	t_a	$t_b(\text{query terms})$	t_c
1K	40K	10	0.015	1.566 (5) - 3.174 (10)	0.001	0	1.878 (5) - 3.575 (10)	0.008
1K	40K	100	0.328	1.56 (5) - 3.176 (10)	0.005	0	1.624 (5) - 3.192 (10)	0.004
5K	255K	10	0.015	112.1 (5) - 262.5 (10)	0.001	0	131.2 (5) - 264.7 (10)	0.048
5K	255K	100	0.328	116.0 (5) - 251.1 (10)	0.006	0	153.4 (5) - 271.1 (10)	0.152

t_a : time to find the query (fast, depends on $|A|$, not the size of the db)

t_b : time to evaluate the query (this is the only expensive task)

t_c : time to decide what kind of name it is (fast)



Concluding Remarks and Further Research

- The naming problem has several applications (e.g. for relevance feedback or for providing a flexible interaction scheme between the system and the users)
- We expressed formally several variations of this problem and related optimality criteria (to best of our knowledge this analysis is novel)
- We provided optimal solutions
 - E.g. we provided a method that certainly returns the best upper name for unordered sets (this is not true for the Rocchio method)
- We described (polynomial) algorithms for solving these problems
- Future research
 - Extend the problem statement with an additional parameter: the maximum number of words that a name could have, e.g. "find the best upper name with no more than 3 words".
 - Indexing structures for efficient computation of names
- For more see
 - http://www.ics.forth.gr/~tzitzik/publications/2007_TzitzikasTheoharisNaming.pdf