Προχωρημένες Λειτουργίες Επερώτησης
Advanced Query Operations
ΜΕΡΟΣ ΙΙ

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

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

Διαδραστική Επαύξηση Ερωτήματος

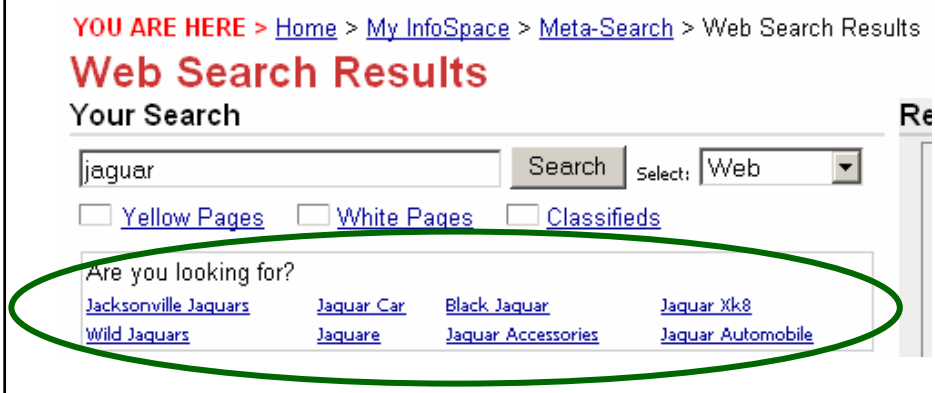
- Οι μέθοδοι για τροποποίηση ερωτήματος που έχουν περιγραφεί ως τώρα επιλέγουν *όρους από έγγραφα* και προσθέτουν κάποιους στο ερώτημα.
- Ένας εναλλακτικός τρόπος είναι οι χρήστες να επιλέγουν τους όρους που θα προστεθούν στο ερώτημα (**IQE - interactive query expansion**).
  - Το σύστημα επιλέγει τους πιο σχετικούς όρους
- Αναζητήσεις σε περιπτώσεις που έχει ανακτηθεί λίγη σχετική πληροφορία ωφελούνται από IQE.
- Είναι πιο πιθανό οι χρήστες να χρησιμοποιήσουν IQE σε μια πολύπλοκη ή δύσκολη αναζήτηση.

Επέκταση Επερώτησης (Query Expansion)

Πως θα βρούμε τους πιο σχετικούς όρους

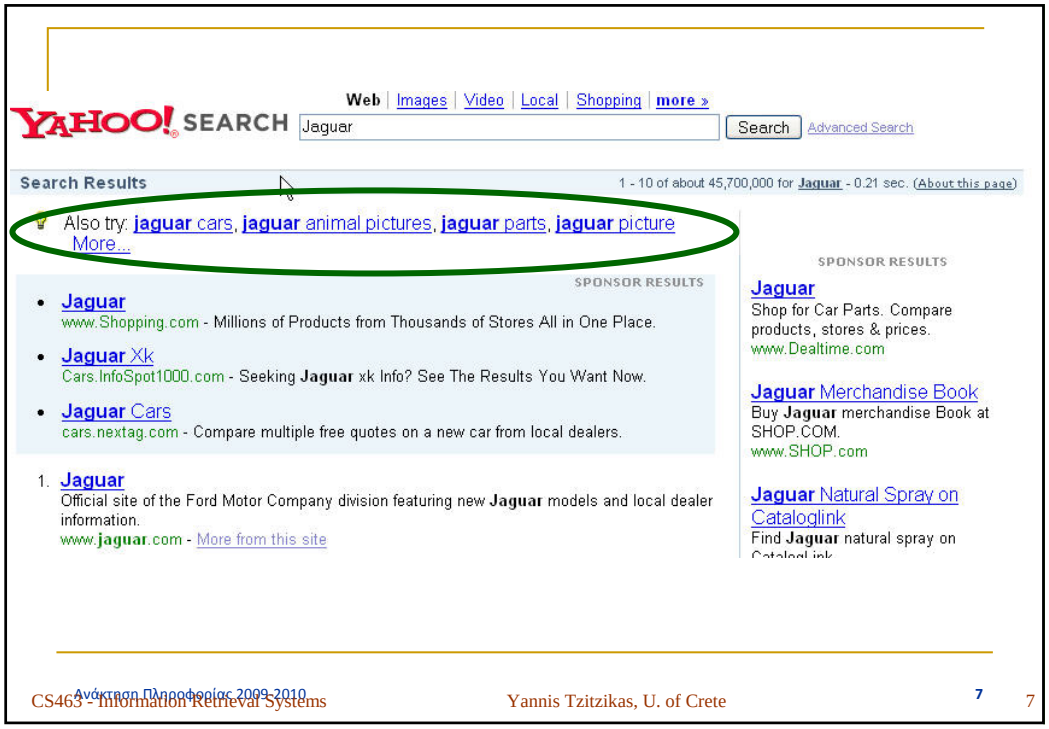
- Τοπική Ανάλυση
  - Αναλύουμε τα (κορυφαία) έγγραφα της απάντησης
- Καθολική Ανάλυση
  - Αναλύουμε όλα τα έγγραφα της συλλογής

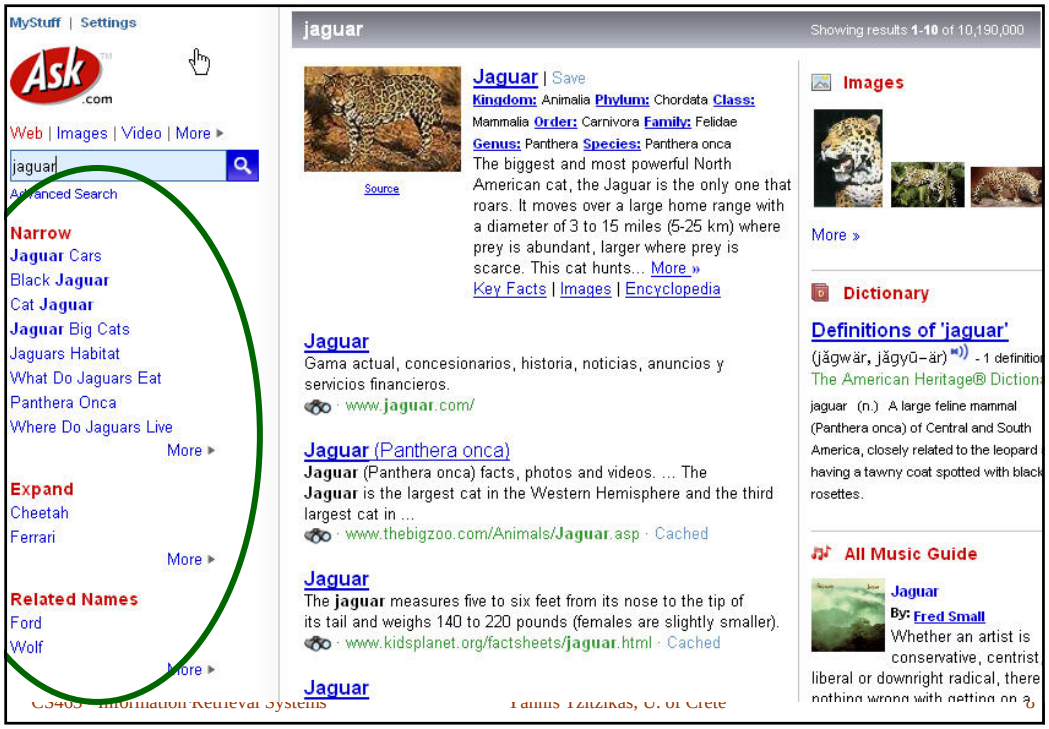
Παράδειγμα εφαρμογής



Also: see altavista, teoma

Ο χρήστης βλέπει τις κορυφαίες λέξεις (και όχι έγγραφα)


Web | Images | Video | Local | Shopping | more »
YAHOO! SEARCH
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
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
Source
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® Dictionary
jaguar (n.) A large feline mammal (Panthera onca) of Central and South America, closely related to the leopard 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

Στο groogle/mitos

A very simple technique is currently supported:

- For each term t_i that appears in the top L (by default L=5) documents returned by the Query Evaluator, we sum its term frequencies (i.e. all tf_{ij} where j in top-L documents) and we recommend to the user the S terms (by default S=5) with the highest accumulative frequency.

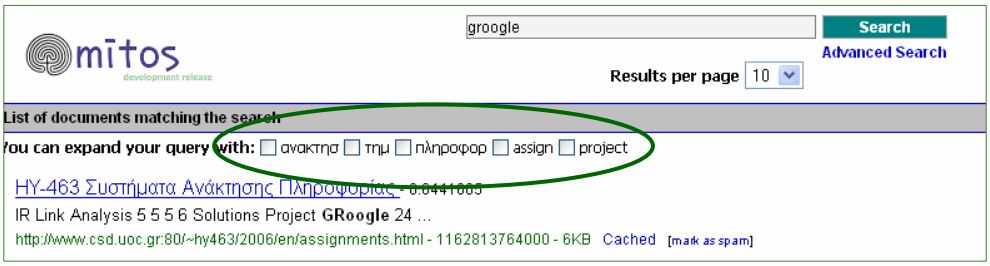


Table 9. Query Expansion Examples

	Initial Query	*Expanded Terms*				
1	retrieval	imag	medic	index	storag	system
2	web	system	servic	page	process	cours
3	user	interfac	layer	system	develop	softwar

Table 10. Query Expansion Average Times

L	*Time (sec)*
5	0.002
10	0.003
15	0.004
20	0.004

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

Αυτόματη Τοπική (Επιτόπια) Ανάλυση
Automatic Local Analysis

1. Μετά την διατύπωση της επερώτησης, ανάλυσε (στατιστικά) τις λέξεις που εμφανίζονται **μόνο** στα κορυφαία ανακτημένα έγγραφα (κοιτάμε μόνο την απάντηση)

   π.χ. επιλέγουμε τις 10 πιο συχνά εμφανιζόμενες λέξεις των κορυφαίων 5 εγγράφων

2. Το σύστημα παρουσιάζει στο χρήστη τις πιο συχνά εμφανιζόμενες λέξεις και αυτός επιλέγει εκείνες που θέλει να προστεθούν στην επερώτηση

   εναλλακτικά η επιλογή μπορεί να γίνει αυτόματα (χωρίς την παρέμβαση ή συγκατάθεση του χρήστη)

Επίδραση στην αποτελεσματικότητα της ανάκτησης

- Οι ασαφείς (ή αμφίσημες) λέξεις δημιουργούν λιγότερα προβλήματα (απ' ότι στην καθολική ανάλυση – την οποία θα αναλύσουμε παρακάτω)

  Παράδειγμα: με τοπική ανάλυση η επερώτηση "Apple computer" μπορεί να επεκταθεί στην "Apple computer Powerbook laptop"

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

D: τα έγγραφα της απάντησης και $t_1 \ldots t_n$ οι όροι που εμφανίζονται σε αυτά.

	t_1	t_2	t_3 ……..t_n
t_1	c_{11}	c_{12}	c_{13}……c_{1n}
t_2	c_{21}		
t_3	c_{31}		
.	.		
.	.		
t_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:

  Normalized score is 1 if two terms have the same frequency in all documents in D.

$$s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}$$

Από αυτόν τον πίνακα μπορούμε να βρούμε τους όρους που είναι **πιο κοντά σε αυτούς της επερώτησης** (θυμηθείτε και τον πίνακα συσχέτισης στο fuzzy model)

(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.

V_i: Set of all occurrences of term *i* in any document in D.

$r(k_u, k_v)$: Distance in words between word occurrences k_u and k_v ($=\infty$ if k_u and k_v are occurrences in different documents).

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

Normalized Metric Correlation Matrix

- to account for term frequencies:

$$s_{ij} = \frac{c_{ij}}{|V_i| \times |V_j|}$$

Query Expansion with (Association or Metric) Correlation Matrix

	t_1	t_2	t_3	…….	t_n
t_1	c_{11}	c_{12}	c_{13}	…….	c_{1n}
t_2	c_{21}				
t_3	c_{31}				
.	.				
.	.				
t_n	c_{n1}				

- For each term *i* in the query q, 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.

Query Expansion with Scalar Clusters

	t_1	t_2	t_3	…….	t_n
t_1	c_{11}	c_{12}	c_{13}	…….	c_{1n}
t_2	c_{21}				
t_3	c_{31}				
.	.				
.	.				
t_n	c_{n1}				

For each query term t_q: consider the terms that have similar correlation values with it

Inner product of the related vectors

Αυτόματη Καθολική Ανάλυση (Automatic Global Analysis)

Αυτόματη Καθολική Ανάλυση Automatic Global Analysis

1. 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
2. Επέκταση επερώτησης με τους πιο όμοιους όρους.

Επίδραση στην αποτελεσματικότητα της ανάκτησης

- Οι ασαφείς (ή αμφίσημες) λέξεις δημιουργούν περισσότερα προβλήματα (απ' ότι στην τοπική ανάλυση)
- Παράδειγμα: με καθολική ανάλυση η επερώτηση "Apple computer" μπορεί να επεκταθεί στην "Apple red fruit orange computer"

Μια λύση:

- Query Expansion Based on a Similarity Thesaurus

Query Expansion Based on a Similarity Thesaurus

Βασική ιδέα

- Οι όροι που προστίθενται στην επερώτηση 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={k_1,…,k_t}
- Παριστάνουμε **κάθε όρο** με ένα διάνυσμα στο χώρο των N διαστάσεων

  Είναι σαν να έχουμε αντιστρέψει το ρόλο των όρων και των εγγράφων: έχουμε λοιπόν μια διανυσματική παράσταση των όρων (κάθε έγγραφο αποτελεί μια διάσταση στο χώρο των διανυσμάτων). Προσαρμόζουμε το σχήμα βάρυνσης TF-IDF βάσει αυτής της θεώρησης.

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

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

itf: Inverse term frequency (το ανάλογο του idf):

$$itf_j = \frac{\text{Num of terms in the collection}}{\text{Num of distinct terms in } d_j}$$

←ανάλογο της βάρυνσης TF*IDF μόνο που εδώ χρησιμοποιούμε το inverse term frequency.

Query Expansion Based on a Similarity Thesaurus

Υπενθύμιση

idf_i = inverse document frequency of term i := $\log(N/df_i)$

$tf_{ij} = freq_{ij} / \max_k \{freq_{kj}\}$

Όπου

- $freq_{ij}$ = πλήθος εμφανίσεων του όρου i στο έγγραφο
- $\max_k \{freq_{kj}\}$ *το μεγαλύτερο πλήθος εμφανίσεων ενός όρου στο έγγραφο j*

$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log(N/df_i)$$

Query Expansion Based on a Similarity Thesaurus (II)

Υπολογισμός ομοιότητας δυο όρων

- (π.χ. με εσωτερικό γινόμενο)

$$c_{u,v} = \vec{k}_u \cdot \vec{k}_v$$

- Τα βήματα για την επέκταση της επερώτησης
  - (1) Represent query in the concept space that we used to represent terms

$$\vec{q} = \sum_{ki \in q} w_{iq} \vec{k}_i$$

  - (2) Compute sim(q,k_u) for each k_u

$$sim(q, k_u) = \vec{q} \cdot \vec{k}_u$$

  - (3) Expand q with the top r ranked terms. The weight of each added term k_u is set

$$w_{uq'} = \frac{sim(q, k_u)}{\sum_{k_i \in q} w_{iq}}$$

- Results
  - 20% improved retrieval performance

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 δίνει 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

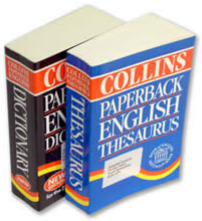
Θησαυροί Όρων

- Ένας θησαυρός παρέχει πληροφορίες για συνώνυμα και σημασιολογικά κοντινές λέξεις και φράσεις
- 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 vocabulary
- Θησαυροί κατάλληλοι για Information Retrieval
  - for coordinating the basic processes of indexing and retrieval
  - designed for specific subject areas and are therefore domain dependent
  - Examples
    - INSPEC

INSPEC thesaurus (for IR)

- Domain: physics, electrical engineering, electronics, computers
- Types of relationships between two terms
  - **UF: Used For (converse: USE)** // π.χ. USE X σημαίνει ότι ο X είναι ο δόκιμος όρος
  - **BT: Broader Term (converse NT)**
  - **TT: Top Node, I.e. root of the hierarchy**
  - **RT: Related Term**
- 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)

Επέκταση επερωτήσεων βάσει Θησαυρού
Thesaurus-based Query Expansion

- Τρόπος:
  - Για 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"

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
  - ***Pre-coordination***: the thesaurus contain phrases
    + the vocabulary is very precise
    - the user has to be aware of the phrase construction rules, large size
  - ***Post-coordination***: 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

[Α] Χειροποίητη Δημιουργία

[Β] Αυτόματη Κατασκευή

[Β.1] από **συλλογή κειμένων**

Προϋπόθεση: Να υπάρχει μια μεγάλη και αντιπροσωπευτική συλλογή κειμένων

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

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

[Β1] 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

- Απόφαση: Ποιος θέλουμε να είναι ο βαθμός εξειδίκευσης (desired specificity)
  - if high then emphasis will be given on identifying precise phrases
- Οι όροι (terms) μπορούν να επιλεγούν από τους *τίτλους,* τις *περιλήψεις* (abstracs), ή ακόμα και από το *πλήρες κείμενο* (full text)
- Normalization: stemming, stoplists
- Criteria for selecting a term:
  - frequency of occurrence (**divide words to 3 categories: low, medium, high, select terms with medium frequency**)
  - discrimination value ~ idf
- Κατασκευή φράσεων (phrase construction) αν κάτι τέτοιο είναι επιθυμητό (θυμηθείτε coordination level)

2/ Υπολογισμός Ομοιότητας μεταξύ όρων

- Παραδείγματα μετρικών: Cosine, Dice

Αυτόματη Κατασκευή Θησαυρών από Κείμενα (III)

3/ Οργάνωση (συνήθως ιεραρχική) του λεξιλογίου

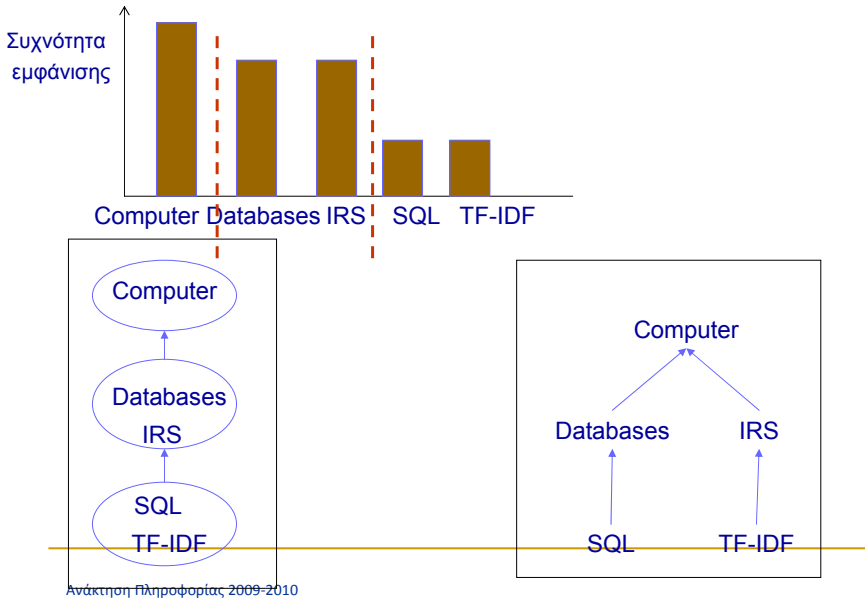
- Οποιοσδήποτε αλγόριθμος clustering μπορεί να χρησιμοποιηθεί

Ένας αλγόριθμος για ιεραρχική οργάνωση ενός λεξιλογίου:

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 parents)

5/ Continue until reaching level 1

Παράδειγμα με 3 κλάσεις συχνοτήτων



Case: grOOGLE'2007

(1) Compute the minimum and maximum frequency of the words in the lexicon (denoted by df_{mn} and df_{mx} respectively).
(2) Partition the interval $[df_{mn}, df_{mx}]$ into L successive intervals (where L is administrator-provided), i.e. $[df_{mn}, df_1], \ldots, [df_{L-1}, df_{mx}]$. We will refer to them with $lev_1, \ldots, lev_L$ respectively.
(3) Ignore the intervals corresponding to low frequencies, specifically keep only the M intervals with the highest frequencies (M is administrator-provided and it should be $M < L$), i.e. keep only $lev_{L-M-1}, \ldots, lev_L$.
(4) Assign to each of these M intervals those words whose frequency falls to that interval.
(5) For each word w_i of level z (where $z \leq L-1$) connect it with the most "correlated" word of the level $z+1$ (that word will be the "parent" of w_i).

Regarding step (5), the correlation c_{ij} between two words w_i and w_j is computed using the formula:

$$c_{ij} = \sum_{d_k \in D} tf_{ik} \times tf_{jk} \quad (1)$$

where tf_{ik} is the frequency of term i in document k.

(cont)

As an example, Table 11 describes the partitioning obtained assuming $L = 20$ (for each level the table shows the number of words that belong to that level). To construct the taxonomy we have considered only the last 5 groups (empty groups, like level 19, are considered as non existant). So the taxonomy includes 35 words in total. After creating the connections between words we realized that each word has an average of 1.4 child nodes.

	Low frequency																		High frequency
Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Num. Of Words	217142	1103	523	292	199	128	83	83	53	52	25	18	18	14	14	7	8	2	0	4

The reason for partitioning words into groups (according to their frequency) is for avoiding computing the correlation matrix between all pairs of words (which would be formidably expensive[7]). In addition, ignoring those words that occur rarely further improves efficiency (as more than 95% of the vocabulary has a very small document frequency) and does not harm the quality of the result as these words do not describe the main concepts of the document corpus, and we have not anyway adequate statistical information to connect them right in a hierarchy.

(cont)

Resulting taxonomy:

<1> http
<2> system
<3> us
<3.1> new
<3.1.1> url
<3.1.1.1> map
<3.1.1.2> network
<3.1.1.2.1> commun
<3.1.1.2.2> gener
<3.1.1.2.3> site
<3.1.1.2.4> scienc
<3.1.1.2.5> web
<3.1.1.3> page
<3.1.1.3.1> link
<3.1.1.3.2> applic
<3.1.1.4> document
<3.1.1.5> access
<3.1.1.6> univers
<3.1.1.7> data
<3.1.1.8> process
<3.1.2> time
<3.1.3> base
<4> inform

As you can see this taxonomy is not very good/useful

Possible improvements:

- Better vocabulary construction
  - The terms with high frequency are not very informative as you can see (e.g.. http, system, url, ...). Therefore we should try the **middle** levels.
  - Furthermore if we had selected words that appear only in titles/abstracts then we would avoid words like: http, url, ..
  - The user at run-time could even specify how specific/general the taxonomy should be (his/her choice would determine the visible part of the taxonomy)
- Other improvements
  - It's better to show the original words (rather than stems)
  - Use phrases instead of single words as terms

	Low frequency																		High frequency
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Of Words	217142	1103	523	292	199	128	83	83	53	52	25	18	18	14	14	7	8	2	0	4

Αυτόματη Κατασκευή Ιεραρχιών [M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In SIGIR'1999]

- Given two terms x and y from a document collection, we say that x *subsumes* y, and we write x → y if: **P(x|y) > 0.8 and P(y|x) < 1**

  P(x|y) is the probability that term x occurs in a document, given that term y does

- This technique leads to creation of a hierarchy of terms, where
  - General terms appear as top-level categories
  - More specific terms appear as lower-level categories
- Pros
  - Simple and effective
- Cons
  - Requires n^2 computations of conditional probabilities, where n is the number of terms in the collection
  - Requires the terms to have a unique meaning
    - However If we use this technique only on query results and by using only terms that appear more frequently in the query results than in the whole collection. then this lessens the problem of ambiguity and reduces the number of terms that form the subsumption hierarchy.

Αυτόματη Κατασκευή Πολυεδρικών Ιεραρχιών [Automatic Construction of Multifaceted Browsing Interfaces,W. Dakka, P. Ipeirotis, K. Wood, CICK'05]

- It describes an approach for constructing multifaceted hierarchies.
- Includes methods for selecting the best parts of the generated hierarchies when it is not possible to fit all the categories on screen
- Experiments with real-life data sets indicate that automatic construction of multifaceted interfaces is feasible, and generates high-quality hierarchies

A Data Mining approach (to organize the set of terms hierarchically)

- Let $I=\{i_1,...i_m\}$ be a set of items
- Let D be a set of transactions where each transaction is a subset of I
- An association rule is an implication of the form $X \rightarrow Y$ where X, Y are subsets of I and $X \cap Y=\varnothing$
- A rule $X \rightarrow Y$ holds in the transaction set D with
  - confidence c if c% of the transactions in D that contain X also contain Y
  - support s if s% of the transactions in D contain $X \cup Y$

Consider the case of an IR system. In that case

- The set I could be the set of all terms (the vocabulary)
- The set D could be the set of binary vectors of the documents
- A rule $X \rightarrow Y$ would be an implication between set of terms
  - If $|X|=|Y|=1$ then the implications are between single terms
  - If $|X|=|Y|=2$ then the implications are between pairs of terms
- So we could exploit data mining algorithms to get a taxonomy from an IR system

Περίληψη

- The complete landscape
  - Global methods
    - Query expansion
      - Thesauri
      - Automatic thesaurus generation
  - Local methods
    - Relevance feedback
    - Pseudo relevance feedback

Relevance Feedback -- Summary

- Relevance feedback has been shown to be very effective at improving relevance of results.
  - Requires enough judged documents, otherwise it's unstable (≥ 5 recommended)
  - Requires queries for which the set of relevant documents is medium to large
- Full relevance feedback is painful for the user.
- Full relevance feedback is not very efficient in most IR systems.
- Other types of interactive retrieval may improve relevance by as much with less work.