HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

Μοντέλα Ανάκτησης ΙΙ
(Retrieval Models)

Γιάννης Τζίτζικας

Διάλεξη : 4
Ημερομηνία : 23-3-2007

Διάρθρωση

Μοντέλα Ανάκτησης βασισμένα σε:

- Θεωρία Ασαφών Συνόλων (Fuzzy Set-based Retrieval Models)
- Νευρωνικά Δίκτυα (Neural Network Retrieval Model)
- Λανθάνουσα Σημασιολογική Ευρετηρίαση (LSI - Latent Semantic Indexing)

Information Retrieval Models

Fuzzy Set-based Retrieval Model

Μοντέλα Βασισμένα στη Θεωρία Ασαφών Συνόλων

(Fuzzy Set-based Retrieval Models)

Κίνητρο

- Επέκταση του Boolean model με **μερικό** ταίριασμα (και άρα με δυνατότητες διαβάθμισης των στοιχείων των απαντήσεων)
  - το Εκτεταμένο Λογικό Μοντέλο (Extended Boolean Model) που είδαμε, είναι επίσης μια προσπάθεια προς την ίδια κατεύθυνση

Έχουν προταθεί αρκετά μοντέλα που βασίζονται σε fuzzy sets. Εδώ θα δούμε δύο:

- Ένα απλό μοντέλο που βασίζεται σε TF-IDF και fuzzy theory
- Το μοντέλο που προτάθηκε στο [Ogawa, Morita, and Kobayashi (1991)]

Background: Fuzzy Set Theory [Zadeh 1965]

- Framework for representing classes whose boundaries are not well defined
- Key idea is to introduce the notion of a ***degree of membership*** associated with the elements of a set
- This degree of membership varies from 0 to 1 and allows modeling the notion of *marginal* membership
- Thus, membership is now a *gradual* notion, contrary to the crispy notion enforced by classic Boolean logic

- U: universe of discourse
- A fuzzy subset A of U is characterized by a membership function

$$\mu_A(u) : U \rightarrow [0,1]$$

  which associates with each element *u* of U a number $\mu_A(u)$ in [0,1]
- Let A and B be two fuzzy subsets of U, and ¬A be the complement of A. Then,
  - $\mu_{\neg A}(u) = 1 - \mu_A(u)$
  - $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
  - $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

A Simple Retrieval Model based on Fuzzy Theory

Παράσταση εγγράφων

$$\begin{array}{c|cccc} & k_1 & k_2 & \ldots & k_t \\ d_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ d_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ d_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{array}$$

$$w_{i,j} \in [0,1]$$

- $K=\{k_1,\ldots,k_t\}$: σύνολο όλων των λέξεων ευρετηρίασης
- Κάθε <u>έγγραφο</u> $\mathbf{d_j}$ παριστάνεται με το διάνυσμα $d_j=(w_{1,j},\ldots,w_{t,j})$ όπου
  - $w_{i,j}$ το <u>βάρος</u> της λέξης k_i για το κείμενο $\mathbf{d_j}$
  - για παράδειγμα $w_{i,j}$ = $\mathbf{tf_{ij}}$ $\mathbf{idf_i}$

A Simple Retrieval Model based on Fuzzy Theory

Boolean Queries and Ranking Function

- Μια επερώτηση q είναι μια λογική έκφραση στο K, πχ:
  - q = "**k1 and (k2 or not k3))**" δηλαδή q = "**k1 $\wedge$ (k2 $\vee \neg$ k3))**"
- R(dj,q) = $\mu_q(dj)$, άρα είναι ο βαθμός συμμετοχής του dj στο σύνολο που προσδιορίζεται από τη λογική έκφραση q.
- Μπορούμε να υπολογίσουμε το R(dj,q) βάσει των κανόνων της θεωρίας των Fuzzy sets, θεωρώντας ότι R(dj,ti) = $\mu_{ti}(dj) = w_{i,j}$
- Για παράδειγμα
  - R(dj, t1 v t2) = max (R(dj, t1), R(dj, t2)) = max (w1j, w2j).
  - R(dj, t1 $\wedge$ t2) = min (R(dj, t1), R(dj, t2)) = min (w1j, w2j).

A Simple Retrieval Model based on Fuzzy Theory

Παρατηρήσεις

- Έστω q = $k_x \wedge k_y$. Σύμφωνα με το Boolean model ένα έγγραφο που περιέχει **μόνο έναν** από τους όρους k_x, k_y είναι **μη-συναφές**, και μάλιστα τόσο μη-συναφές, όσο ένα έγγραφο που δεν περιέχει **κανένα** από τους 2 όρους.
  - Ερώτηση: Τι συμβαίνει εδώ;
  - Απάντηση: Το ίδιο
- Έστω q = k_x v k_y. Σύμφωνα με το Boolean model ένα έγγραφο που περιέχει **και τους δύο όρους** (k_x, k_y) είναι **το ίδιο συναφές**, με ένα έγγραφο που περιέχει **έναν** από τους 2 όρους.
  - Ερώτηση: Τι συμβαίνει εδώ;
  - Απάντηση: ...
  - Άρα το παρόν μοντέλο διαβαθμίζει τα στοιχεία της απάντησης του q = k_x v k_y (κάτι που δεν είναι δυνατό με το Boolean Μοντέλο).
- Το παρόν είναι μια ειδική περίπτωση του Extended Boolean Model (συγκεκριμένα αντιστοιχεί στην περίπτωση που p = ∞).

[Ogawa, Morita, and Kobayashi,1991]

Fuzzy Set Retrieval Model [Ogawa, Morita, and Kobayashi,1991]

Εδώ θα δούμε το μοντέλο που προτάθηκε στο [Ogawa,Morita, Kobayashi,1991)

- Βασική Ιδέα:
  - Έγγραφα και επερωτήσεις παριστάνονται με **σύνολα** όρων ευρετηρίου (εδώ δεν έχουμε βάρη στο [0,1])
  - Κάθε **όρος** συσχετίζεται με ένα **fuzzy set**
  - Κάθε έγγραφο έχει ένα degree of membership σε αυτό το fuzzy set
- Παράδειγμα:
  - Έστω επερώτηση ***q=αυτοκίνητο***
  - Έστω έγγραφο d1 που *δεν περιέχει* τη λέξη αυτοκίνητο αλλά περιέχει τη λέξη «όχημα».
  - Αν υπάρχουν **πολλά** έγγραφα που περιέχουν και τις δυο λέξεις, τότε, υπάρχει ισχυρή συσχέτιση των δυο αυτών λέξεων, και
  - => άρα το d1 μπορεί να θεωρηθεί **συναφές** με την επερώτηση q.
- Η παραπάνω ιδέα θεμελιώνεται με Fuzzy Theory

Fuzzy Set Retrieval Model
Μορφή Ευρετηρίου: όπως και στο Boolean model.

	k_1	k_2	….	k_t
d_1	w_{11}	w_{21}	…	w_{t1}
d_2	w_{12}	w_{22}	…	w_{t2}
:	:	:		:
:	:	:		:
d_n	w_{1n}	w_{2n}	…	w_{tn}

$w_{i,j} \in \{0,1\}$

- K={k_1,…,k_t}: σύνολο όλων των λέξεων ευρετηρίασης
- Κάθε έγγραφο **d_j** παριστάνεται με το διάνυσμα d_j=($w_{1,j}$,…,$w_{t,j}$) όπου:
  - $w_{i,j}$ = **1** αν η λέξη k_i εμφανίζεται στο κείμενο **d_j** (αλλιώς $w_{i,j}$ =**0**)

Βάσει αυτού του πίνακα θα δημιουργήσουμε έναν πίνακα συσχέτισης όρων
(για να καταχωρήσουμε σχέσεις όπως «αυτοκίνητο» ≈ «όχημα»)

Fuzzy Set Retrieval Model
Πίνακας Συσχέτισης (correlation matrix) και εγγύτητα όρων

	k_1	k_2	….	k_t
k_1	c_{11}	c_{21}	…	c_{t1}
k_2	c_{12}	c_{22}	…	c_{t2}
:	:	:		:
:	:	:		:
k_t	c_{1n}	c_{2n}	…	c_{tn}

$$\mathbf{c(i,l) = \frac{n(i,l)}{ni + nl - n(i,l)}}$$

where:
n(i,l): number of docs which contain both ki and kl
ni: number of docs which contain ki
nl: number of docs which contain kl

Πχ
n(i,l)=0 => c(i,l)=0
n(i,l)=3, ni=3, nl=9 => c(i,l)=0.3
n(i,l)=3, ni=3, nl=30 => c(i,l)=0.1
n(i,l)=3, ni=3, nl=3 => c(i,l)=1

Έτσι έχουμε ορίσει ποσοτικά την εγγύτητα (proximity) μεταξύ των όρων
(συγκεκριμένα την συνεμφάνισή τους στα έγγραφα της συλλογής)

Fuzzy Set Retrieval Model

Fuzzy Information Retrieval

- Σε κάθε όρο ki αντιστοιχούμε ένα fuzzy set με χαρ/κή συνάρτηση μ_i
- Οι συντελεστές συσχέτισης μας επιτρέπουν να ορίσουμε το βαθμό συμμετοχής ενός εγγράφου dj στα fuzzy σύνολα των όρων.
- Για παράδειγμα έστω ότι το έγγραφο dj δεν περιέχει τον όρο ki
- Αν το έγγραφο dj περιέχει έναν όρο k_w που σχετίζεται ισχυρά με τον k_i τότε
  - θα έχουμε c(i,w) ~ 1
  - και άρα θα μπορούσαμε να θεωρήσουμε ότι $\mu_i(j)$ ~ 1. Με άλλα λόγια, αν και ο όρος ki δεν εμφανίζεται στο dj, εντούτοις περιγράφει το περιεχόμενο του dj

$$\mu_i(j) = \sum_{k_w \in dj} c(i,w)$$

Άθροισμα του βαθμού συσχέτισης του ki με τους όρους που εμφανίζονται στο dj

$$= 1 - \prod_{k_w \in dj} (1 - c(i,w))$$

Βασίζεται στο:

$$(\cup A_i)^c = \cap A_i^c$$

$$\cup A_i = \Omega - (\cup A_i)^c = \Omega - \cap A_i^c$$

Fuzzy Set Retrieval Model

Fuzzy Information Retrieval

Έστω q σε DNF q = cc1 v … v cck, όπου cci είναι μια συζευκτική συνιστώσα
Σύμφωνα με τη fuzzy set theory:

$$\mu_q(j) = \max(\mu_{cc1}(j), \ldots, \mu_{cck}(j))$$

Παρά ταύτα, εδώ προτείνεται η χρήση αθροίσματος αντί του του μεγίστου.

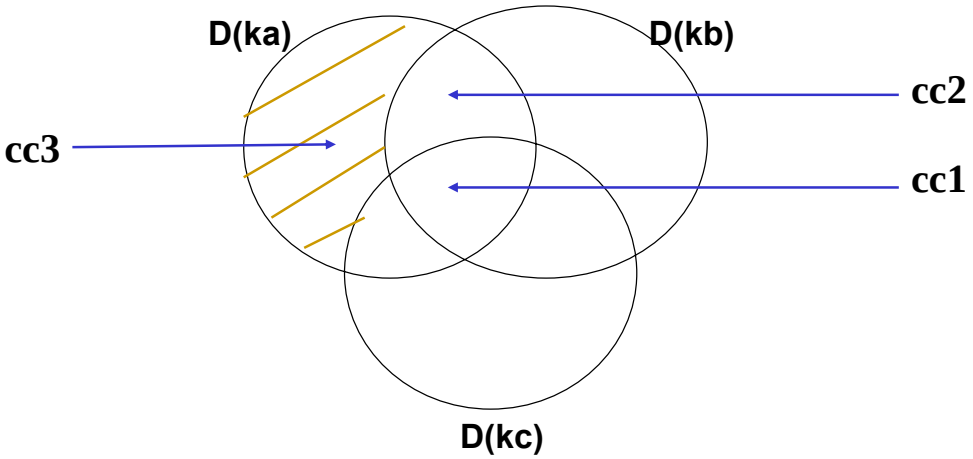
$R(dj,q) = \mu_q(dj) = \sum \mu_{cc}(dj)$ για κάθε συζευκτική συνιστώσα cc του q_{DNF}

Παράδειγμα

$q = ka \wedge (kb \vee \neg kc)$

$vec(q_{dnf}) = (1,1,1) + (1,1,0) + (1,0,0)$

$= vec(cc1) + vec(cc2) + vec(cc3)$



Παράδειγμα (II)

$q = ka \wedge (kb \vee \neg kc)$

$vec(q_{dnf}) = (1,1,1) + (1,1,0) + (1,0,0)$

$= vec(cc1) + vec(cc2) + vec(cc3)$

$$\mu_q(dj) = \mu_{cc1+cc2+cc3}(dj) = 1 - \prod_{i=1..3} (1 - \mu_{cci}(dj))$$

$= 1 - (1- [1,1,1]) * (1 - [1,1,0]) * (1 - [1,0,0])$

$\mu_a(dj)\ \mu_b(dj)\ \mu_c(dj))$ $\mu_a(dj)\ \mu_b(dj)\ (1-\mu_c(dj)))$ $\mu_a(dj)\ (1-\mu_b(dj))\ (1-\mu_c(dj)))$

Fuzzy Set Retrieval Model
Σύνοψη

- $K=\{k_1,\ldots,k_t\}$: σύνολο όλων των λέξεων ευρετηρίασης
- Κάθε έγγραφο $\mathbf{d_j}$ παριστάνεται με το διάνυσμα $d_j=(w_{1,j},\ldots,w_{t,j})$ όπου:
  - $w_{i,j} = 1$ αν η λέξη k_i εμφανίζεται στο κείμενο $\mathbf{d_j}$ (αλλιώς $w_{i,j} = 0$)
- Μια επερώτηση q είναι μια λογική έκφραση στο K, πχ:
  - q = "**k1 and (k2 or not k3))**" δηλαδή q = "**k1 ∧ (k2 ∨ ¬ k3))**"
  - q_{DNF} = "**(k1 ∧ k2 ∧ k3) ∨(k1 ∧ k2 ∧ ¬ k3) ∨(k1 ∧ ¬ k2 ∧ ¬ k3)**"
  - q_{DNF} = "**(1,1,1) ∨(1,1,0) ∨(1,0,0)**"
- $R(dj,q) = \mu_q(dj) = \Sigma\, \mu_{cc}(dj)$ για κάθε συζευκτική συνιστώσα cc του q_{DNF}
  - $\mu_{ki}(dj) = 1 - \prod_{k_w \in dj} (1 - c(ki,kw))$
  - c(ki,kj) καθορίζεται από την συνεμφάνιση των όρων ki και kj στη συλλογή

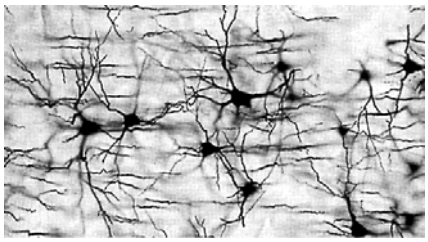
Fuzzy Set Retrieval Model
Γενικά σχόλια

- Έχουν συζητηθεί κυρίως στο χώρο της fuzzy theory
- Δεν έχουμε επαρκή αποτελέσματα πειραματικής αξιολόγησης για να τα αντιπαραβάλλουμε με τα προηγούμενα μοντέλα

Information Retrieval Models

Neural Network Model

(Μοντέλο Νευρωνικού Δικτύου)



Μοντέλο Ανάκτησης Νευρωνικού Δικτύου

- Στα “κλασσικά” μοντέλα ανάκτησης πληροφοριας:
  - τα έγγραφα και οι επερωτήσεις ευρετηριάζονται από όρους
  - η ανάκτηση βασίζεται στο “ταίριασμα” όρων
- Η ιδέα:
  - Είναι γνωστό ότι τα Νευρωνικά Δίκτυα είναι καλοί pattern matchers

Human Brain is a Neural Network

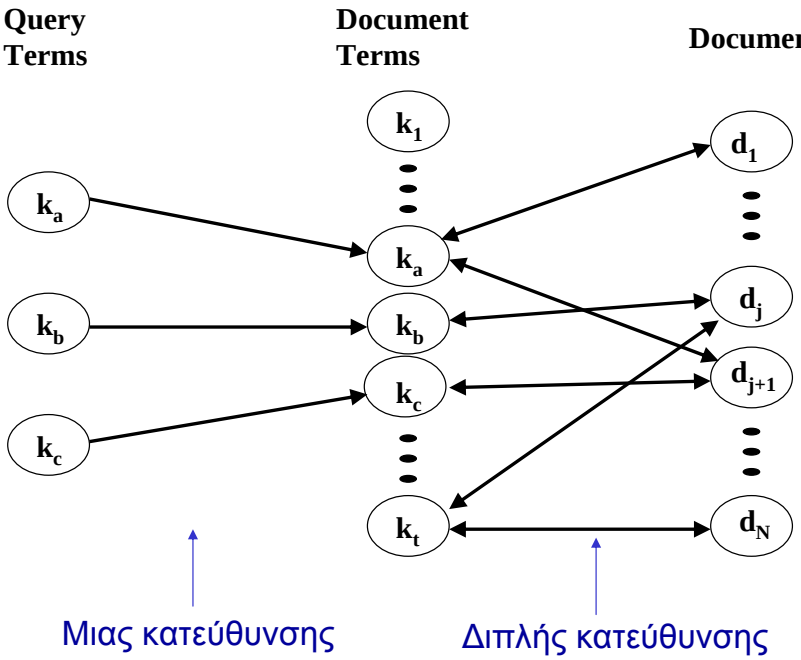
- The human brain is composed of billions of neurons
  - (1 million millions of nodes where each node has one thousands edges)
- Each neuron can be viewed as a small processing unit
- A neuron is stimulated by input signals and emits output signals in reaction
- A chain reaction of propagating signals is called a *spread activation process*
- As a result of spread activation, the brain might command the body to take physical reactions

Neural Networks

- A neural network is an oversimplified representation of the neuron interconnections in the human brain:
  - **nodes** are processing units
  - **edges** are synaptic connections
  - the **strength** of a propagating **signal** is modelled by a **weight** assigned to each edge
  - the **state** of a node is defined by its *activation level*
  - depending on its activation level, a node might issue an **output** signal

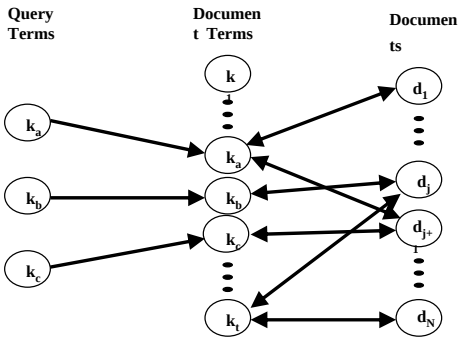
Neural Network for IR

[From the work by Wilkinson & Hingston, SIGIR'91]



Neural Network for IR

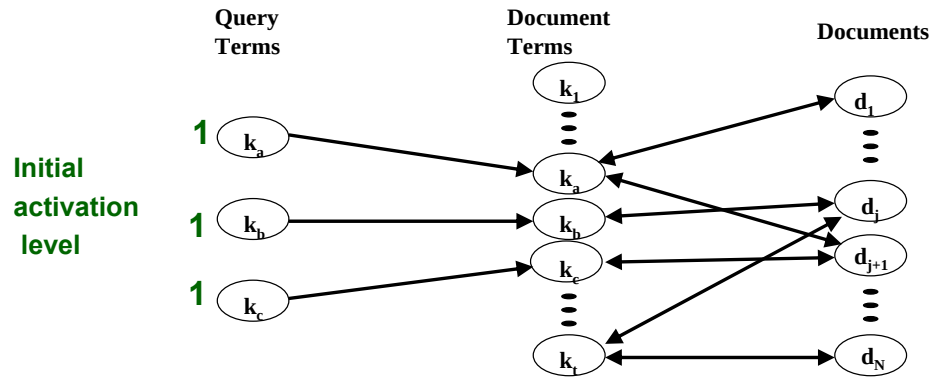
- Δίκτυο τριών επιπέδων
- Τα σήματα διαδίδονται (propagate) στο δίκτυο
- 1ο στάδιο διάδοσης:
  - Query terms issue the first signals
  - These signals propagate accross the network to reach the document nodes
- 2ο στάδιο διάδοσης:
  - Document nodes might themselves generate new signals which affect the document term nodes
  - Document term nodes might respond with new signals of their own, and so on



Μετάδοση σημάτων

- Μέγιστη τιμή σήματος =1 (άρα κάνουμε κανονικοποίηση)
- Οι όροι της επερώτησης εκπέμπουν το αρχικό σήμα ίσο με 1
- Πρέπει να καθορίσουμε τα βάρη των ακόλουθων ακμών:
  - των ακμών από τους όρους επέρώτησης στους όρους εγγράφων
    - (query terms => terms)
  - των ακμών από τους όρους εγγράφων στους κόμβους εγγράφων
    - (terms => docs)

Μετάδοση σημάτων

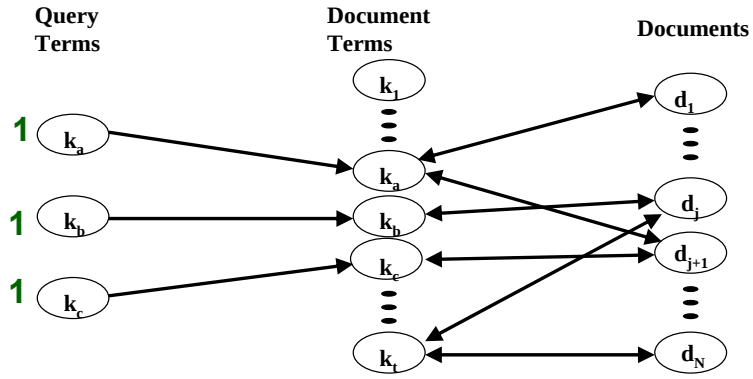


$$\vec{w}_{iq} = \frac{w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad \vec{w}_{ij} = \frac{w_{ij}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}}$$

Σημείωση: τα αρχικά wiq kai wij όπως στο διανυσματικό μοντέλο (tf-idf)

Αυτή η κανονικοποίηση μπορεί να γίνει βάζοντας αυτά τα βάρη πάνω στις ακμές

Μετάδοση σημάτων (II)



Άθροιση σημάτων. Άρα το activation level στο dj (μετά τον 1ο γύρο), είναι:

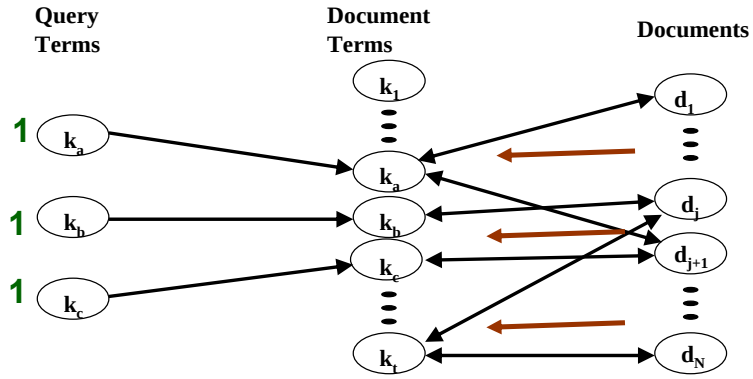
$$\sum_{i=1}^{t} \overline{w}_{iq}\,\overline{w}_{ij}$$

$$\overline{w}_{iq} = \frac{w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad \overline{w}_{ij} = \frac{w_{ij}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}}$$

The ranking of the classic Vector Space Model !

Σημείωση: τα αρχικά wiq kai wij όπως στο διανυσματικό μοντέλο (tf-idf)

Μετάδοση σημάτων (III)



Άθροιση σημάτων. Άρα το activation level στο dj (μετά τον 1ο γύρο), είναι:

$$\sum_{i=1}^{t} \overline{w}_{iq}\,\overline{w}_{ij}$$

$$\overline{w}_{iq} = \frac{w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad \overline{w}_{ij} = \frac{w_{ij}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}}$$

- Η ανάκτηση μπορεί να **βελτιωθεί** αν επιτρέψουμε στους κόμβους των εγγράφων να εκπέμψουν σήμα
  - (λειτουργία ανάλογη της ανάδρασης συνάφειας)
  - A minimum threshold should be enforced to avoid spurious signal generation

Μοντέλο Νευρωνικού Δικτύου: Επίλογος

- Model provides an interesting formulation of the IR problem
- Model has not been tested extensively
- It is not clear the improvements that the model might provide

Information Retrieval Models

Latent Semantic Indexing (LSI)

Λανθάνουσα Σημασιολογική Ευρετηρίαση

Σκεπτικό / Κίνητρο

- Classic IR might lead to poor retrieval due to:
  - relevant documents that do not contain at least one index term are not retrieved
  - A document that shares concepts with another document known to be relevant might be of interest
- The user information need is more related to **concepts and ideas** than to index terms
- We want to capture the concepts instead of the words.
- Concepts are reflected in the words. However:
  - One term may have ***multiple*** meanings (**polysemy**)
  - *Different* terms may have the *same* meaning (**synonymy**)

LSI: The approach

- LSI approach tries to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a **statistical problem**.
- The goal is to find effective models to represent the relationship between terms and documents.
- Hence a set of terms, which is by itself incomplete and unreliable, will be replaced by some set of entities which are more reliable indicants.

Γιατί λέγεται “Latent …”

- Διότι γίνεται η υπόθεση ότι υπάρχει μια «λανθάνουσα» δομή στον τρόπο χρήσης των λέξεων στα έγγραφα
- Το LSI αξιοποιεί στατιστικές τεχνικές για την εκτίμηση της

LSI: The idea

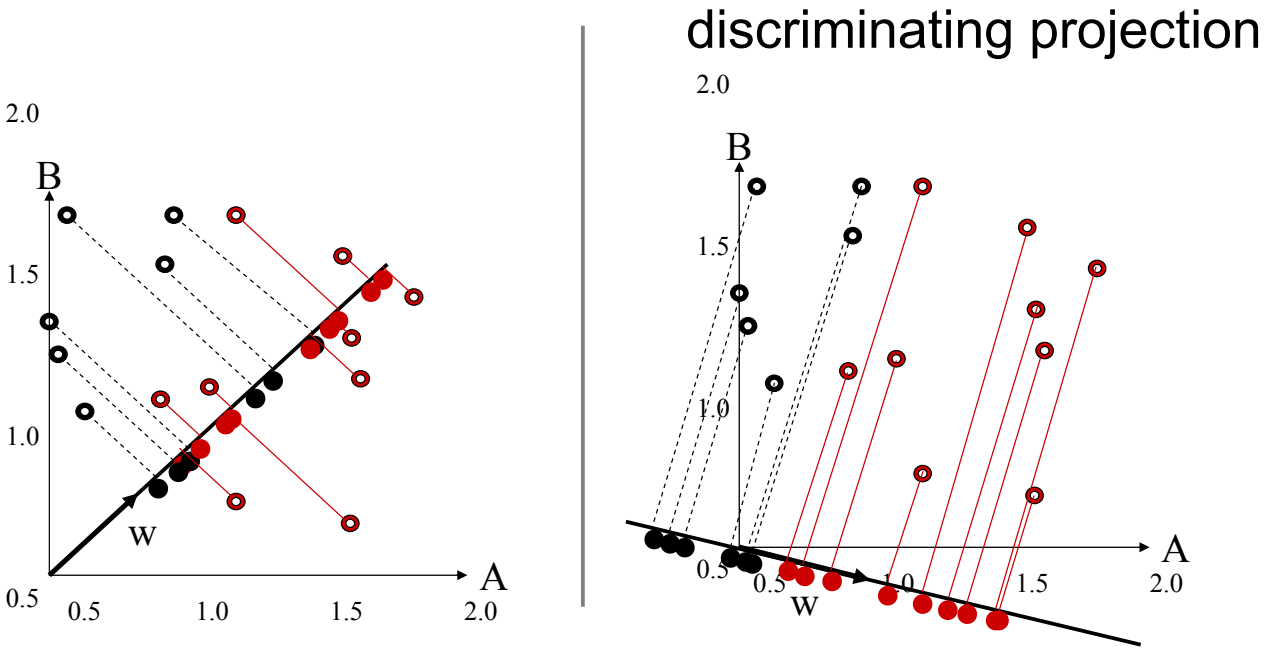
- The key idea is to map documents and queries into a **lower dimensional space**
  - (i.e., composed of higher level concepts which are fewer in number than the index terms)

- Retrieval in the reduced concept space might be superior to retrieval in the space of index terms

- But how to learn the concepts from data?

Μείωση Διαστάσεων και Διακριτική Ικανότητα
(μπορεί να έχουμε μείωση της διακριτικής ικανότητας, μπορεί όμως και όχι)

Παράδειγμα προβολής 2 διαστάσεων σε μία



SVD (Singular Value Decomposition)

- LSI is based on SVD (Singular Value Decomposition)
- So SVD is applied to derive the latent semantic structure model.

- What is SVD?
  - A dimensionality reduction technique

  - For more about matrices and SVD see:
  - The Matrix Cookbook
    http://www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf
  - http://kwon3d.com/theory/jkinem/svd.html
  - http://mathworld.wolfram.com/SingularValueDecomposition.html
  - http://www.cs.ut.ee/~toomas_l/linalg/lin2/node13.html#SECTION00013200000000000000

(TO CHECK THESE)

Definitions

- t: total number of index terms
- d: total number of documents

- (Xij): be a term-document matrix with t rows and d columns
  - To each element of this matrix is assigned a weight **wij** associated with the pair **[ki,dj]**
  - The weight wij can be freqij
    - (or based on a **tf-idf** weighting scheme)

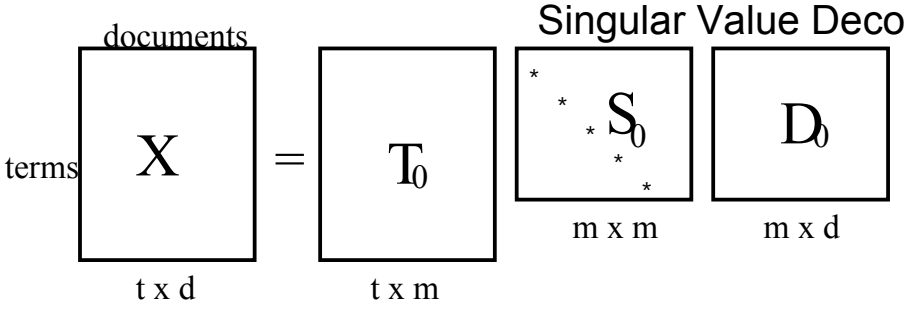
$$\mathbf{x} = \begin{array}{c} \\ k_1 \\ k_2 \\ \vdots \\ \vdots \\ k_t \end{array} \begin{pmatrix} d_1 & d_2 & \ldots. & d_d \\ w_{11} & w_{21} & \ldots & w_{d1} \\ w_{12} & w_{22} & \ldots & w_{d2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ w_{1t} & w_{2t} & \ldots & w_{dt} \end{pmatrix}$$

$w_{i,j} \in [\mathbf{0},\mathbf{1}]$

Latent Semantic Indexing: Ο τρόπος

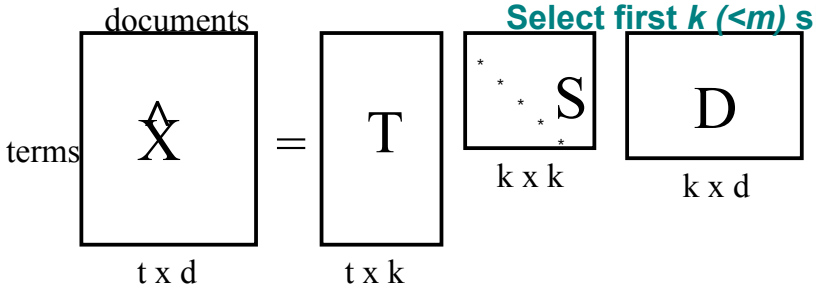
t: total number of index terms
d: total number of documents

Singular Value Decomposition

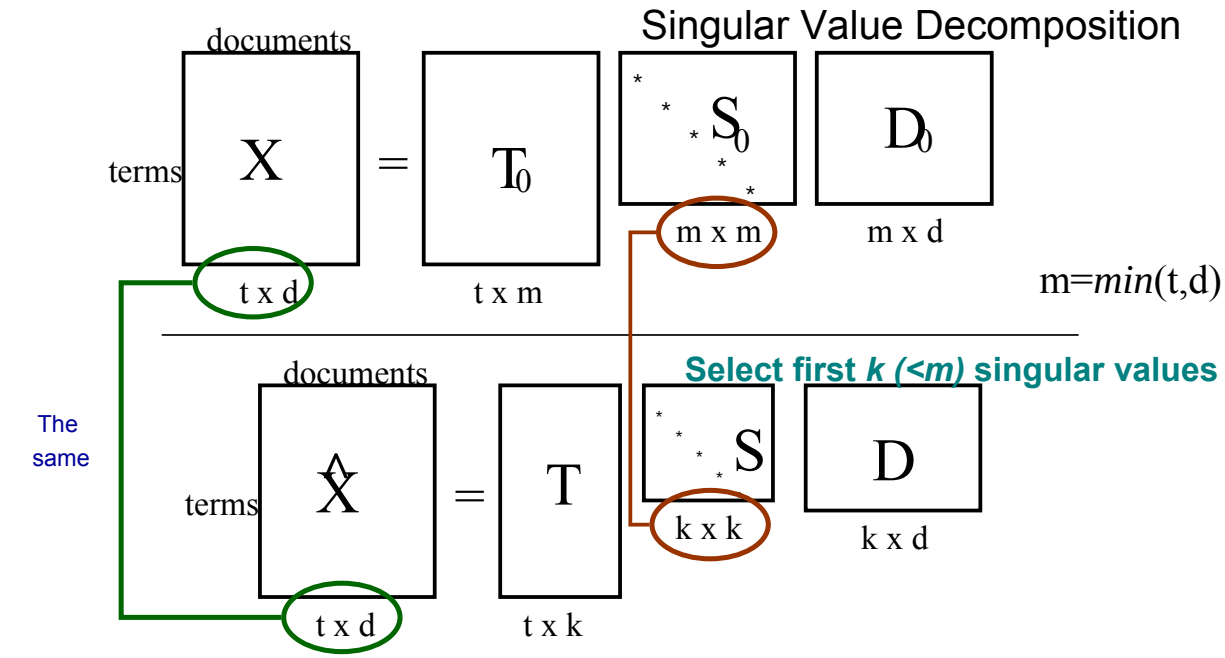


m=*min*(t,d)

Select first *k (<m)* singular values



t: total number of index terms
d: total number of documents



SVD

- SVD of the term-by-document matrix X:

$$X = T_0 S_0 D_0'$$

- If the singular values of S_0 are ordered by size, we only keep the first k largest values and get a reduced model:

$$\hat{X} = TSD'$$

  - $\hat{X}$ doesn't exactly match X and it gets closer as more and more singular values are kept
  - This is what we want. We don't want perfect fit since we think some of 0's in X should be 1 and vice versa.
  - It reflects the major associative patterns in the data, and ignores the smaller, less important influence and noise.

LSI Paper example

Index terms in italics

Titles:

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

term-document Matrix

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Weight = number of occurrences

T_0

$$\begin{bmatrix}
0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\
0.20 & -0.07 & 0.14 & -0.55 & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\
0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\
0.40 & 0.06 & -0.34 & 0.10 & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\
0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\
0.30 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\
0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\
0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\
0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & -0.68 & 0.23 \\
0.03 & 0.45 & 0.14 & -0.01 & -0.30 & 0.28 & 0.34 & 0.68 & 0.18
\end{bmatrix}$$

S_0

$$\begin{bmatrix}
3.34 & & & & & & & & \\
& 2.54 & & & & & & & \\
& & 2.35 & & & & & & \\
& & & 1.64 & & & & & \\
& & & & 1.50 & & & & \\
& & & & & 1.31 & & & \\
& & & & & & 0.85 & & \\
& & & & & & & 0.56 & \\
& & & & & & & & 0.36
\end{bmatrix}$$

D_0

$$\begin{bmatrix}
0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\
0.61 & 0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & 0.24 \\
0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & 0.02 \\
0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & -0.08 \\
0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & -0.26 \\
0.00 & 0.19 & 0.10 & 0.02 & 0.39 & -0.30 & -0.34 & 0.45 & -0.62 \\
0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & 0.02 \\
0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0.00 & 0.25 & 0.45 & 0.52 \\
0.08 & 0.53 & 0.08 & -0.03 & -0.60 & 0.36 & 0.04 & -0.07 & -0.45
\end{bmatrix}$$

SVD with minor terms dropped

$$T \quad S \quad D'$$

$$\begin{bmatrix}
0.22 & -0.11 \\
0.20 & -0.07 \\
0.24 & 0.04 \\
0.40 & 0.06 \\
0.64 & -0.17 \\
0.27 & 0.11 \\
0.27 & 0.11 \\
0.30 & -0.14 \\
0.21 & 0.27 \\
0.01 & 0.49 \\
0.04 & 0.62 \\
0.03 & 0.45
\end{bmatrix}
\begin{bmatrix}
3.34 & \\
 & 2.54
\end{bmatrix}
\begin{bmatrix}
0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\
-0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53
\end{bmatrix}$$

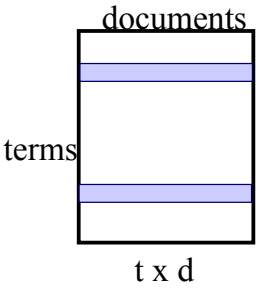
TS define coordinates for documents in latent space

Παρατηρήσεις

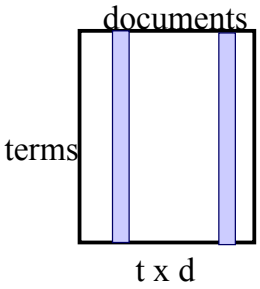
- Η παράμετρος k (<m) πρέπει να είναι:
  - large enough to allow fitting the characteristics of the data
  - small enough to filter out the non-relevant representational details

Τρόπος Σύγκρισης Όρων και Εγγράφων

- Τρόπος σύγκρισης 2 όρων:
  - the **dot product** (or cosine) between two **row vectors** reflects the extent to which two terms have a similar pattern of occurrence across the set of document.

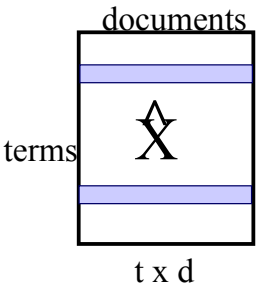


- Τρόπος σύγκρισης δύο εγγράφων:
  - **dot product** (or cosine) between two **column vectors**

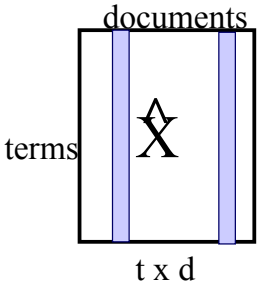


Τρόπος Σύγκρισης Όρων και Εγγράφων $\hat{X}$

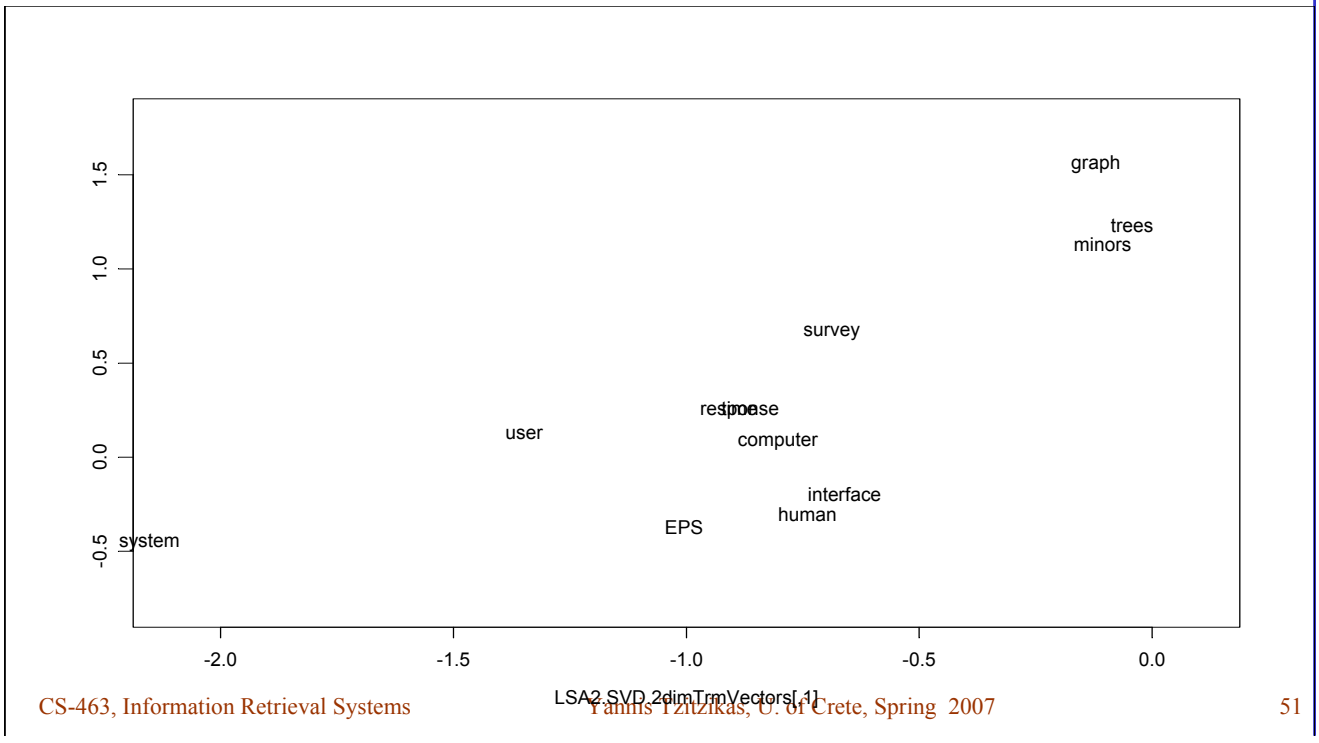
- Τρόπος σύγκρισης 2 όρων:
  - the **dot product** (or cosine) between two **row vectors** reflects the extent to which two terms have a similar pattern of occurrence across the set of document.



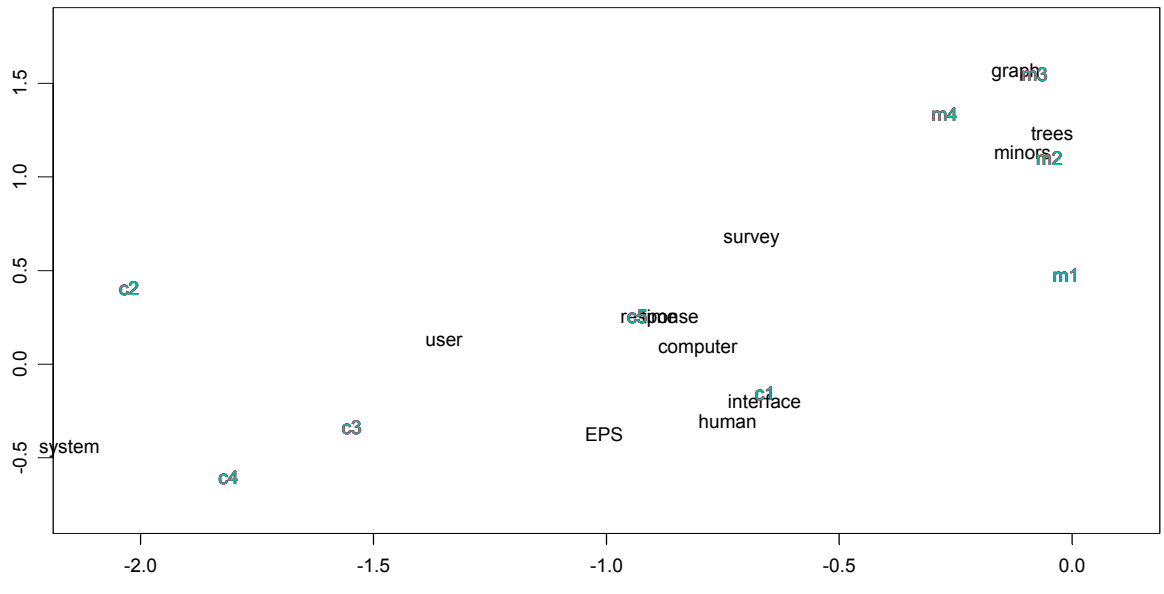
- Τρόπος σύγκρισης δύο εγγράφων:
  - **dot product** (or cosine) between two **column vectors**



Terms Graphed in Two Dimensions



Documents and Terms



Change in Text Correlation

Correlations between text in raw data

	c1	c2	c2	c4	c5	m1	m2	m3	m4
c1	1.000								
c2	-0.192	1.000							
c3	0.000	0.000	1.000						
c4	0.000	0.000	0.472	1.000					
c5	-0.333	0.577	0.000	-0.309	1.000				
m1	-0.174	-0.302	-0.213	-0.161	-0.174	1.000			
m2	-0.258	-0.447	-0.316	-0.239	-0.258	0.674	1.000		
m3	-0.333	-0.577	-0.408	-0.309	-0.333	0.522	0.775	1.000	
m4	-0.333	-0.192	-0.408	-0.309	-0.333	-0.174	0.258	0.556	1.000

Correlations in two-dimensional space

	c1	c2	c2	c4	c5	m1	m2	m3	m4
c1	1.000								
c2	0.910	1.000							
c3	1.000	0.912	1.000						
c4	0.998	0.884	0.998	1.000					
c5	0.842	0.990	0.844	0.809	1.000				
m1	-0.858	-0.568	-0.856	-0.887	-0.445	1.000			
m2	-0.853	-0.562	-0.851	-0.883	-0.438	1.000	1.000		
m3	-0.852	-0.559	-0.850	-0.881	-0.435	1.000	1.000	1.000	
m4	-0.811	-0.497	-0.809	-0.845	-0.368	0.996	0.997	0.997	1.000

Latent Semantic Indexing: Ranking

- Η επερώτηση ***q*** του χρήστη μοντελοποιείται ως ένα **ψευδο-έγγραφο** στον αρχικό πίνακα X

$$\mathbf{X} = \begin{array}{c} \\ k_1 \\ k_2 \\ \vdots \\ \vdots \\ k_t \end{array} \begin{pmatrix} d_1 & d_2 & \ldots. & d_d & \boldsymbol{q} \\ w_{11} & w_{21} & \ldots & w_{d1} & w_{q1} \\ w_{12} & w_{22} & \ldots & w_{d2} & w_{q2} \\ \vdots & \vdots & & \vdots & \\ \vdots & \vdots & & \vdots & \\ w_{1t} & w_{2t} & \ldots & w_{dt} & w_{qt} \end{pmatrix}$$

LSI: Συμπεράσματα

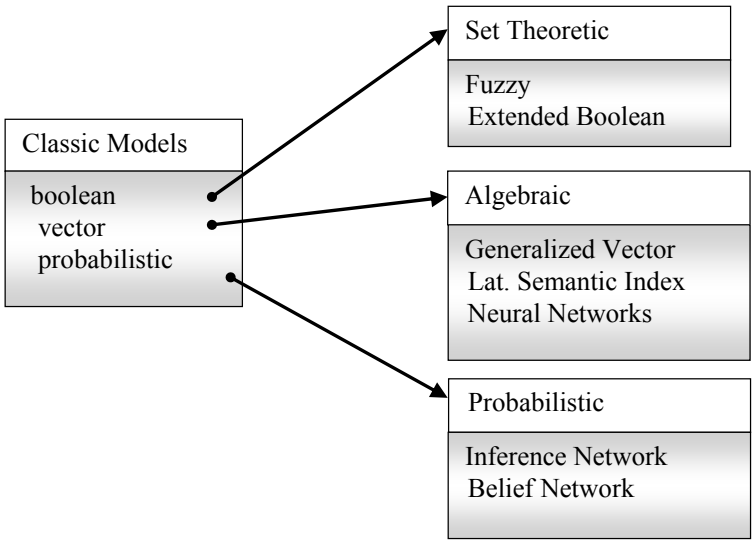
- Latent semantic indexing provides an interesting conceptualization of the IR problem
- It allows reducing the complexity of the underline representational framework which might be explored, for instance, with the purpose of interfacing with the user
- Problems
  - If new documents are added then we have to recompute X^

LSI: Παρατηρήσεις

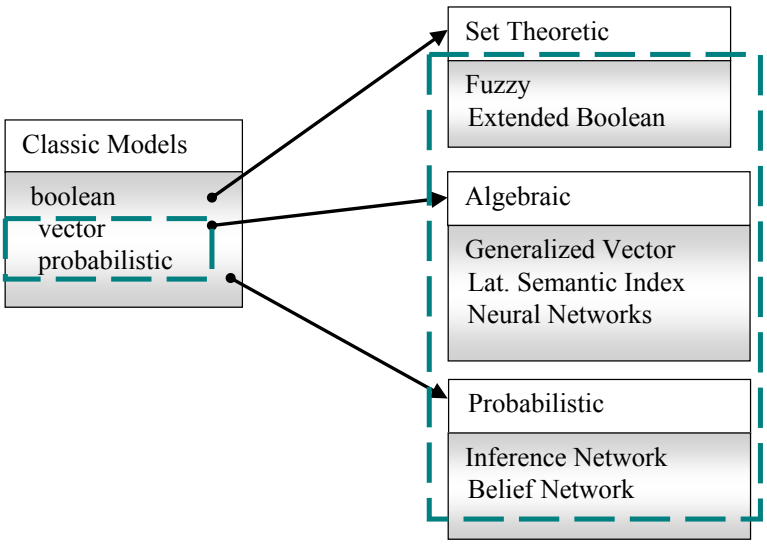
- What is the common and difference between PCA (Principle Component Analysis) and SVD?
  - Both are related to standard eigenvalue-eigenvector, to remove noise and get the most important info.
  - PCA is on covariance matrix and SVD works on original matrix.

Επισκόπηση των Μοντέλων Ανάκτησης που έχουμε εξετάσει μέχρι τώρα

Ταξινομία Μοντέλων που εξετάσαμε

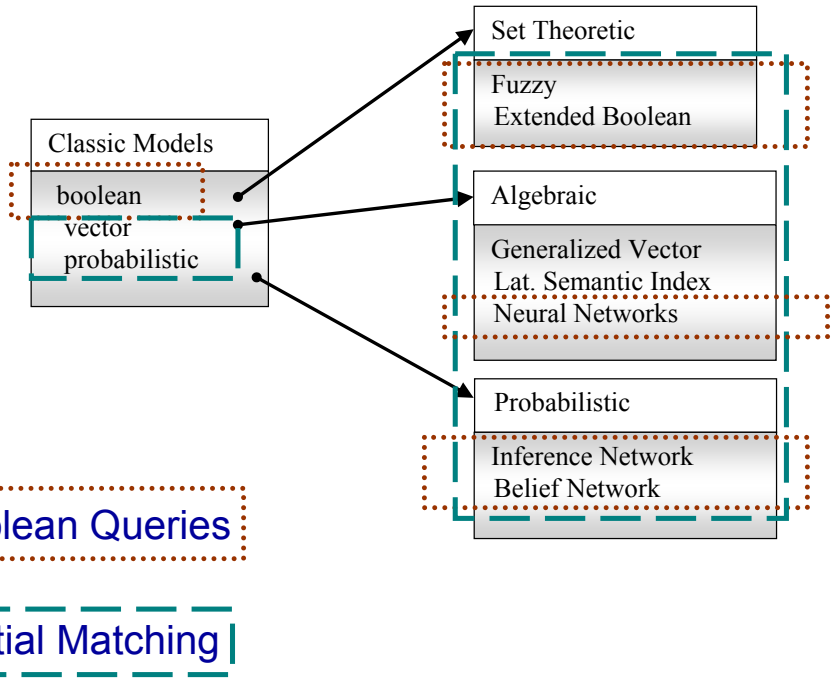


Ταξινομία Μοντέλων που εξετάσαμε

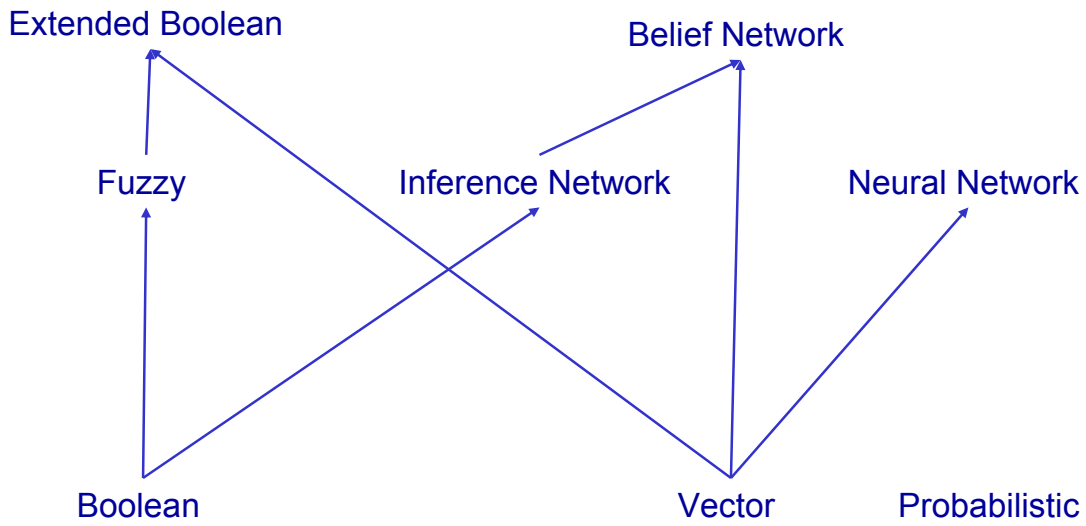


Partial Matching

Ταξινομία Μοντέλων που εξετάσαμε



Βάσει της εκφραστικής τους ικανότητας (incomplete)



Άλλοι τύποι Μοντέλων Ανάκτησης που θα δούμε αργότερα

Αργότερα

- Μοντέλα Ανάκτησης Πληροφοριών από **Ιστοσελίδες**
  - Έμφαση στους συνδέσμους
- Μοντέλα Ανάκτησης **Πολυμέσων**
- Μοντέλα Ανάκτησης βασισμένα στις **Πιθανότητες**
- Μοντέλα Ανάκτησης **Δομημένων** Εγγράφων (π.χ. XML)
- Μοντέλα Βασισμένα στη **Λογική**
  - Carlo Meghini and Umberto Straccia, A Relevance Terminological Logic for Information Retrieval,Proceedings of SIGIR'96, Zurich, Switzerland, 1996