CS-463 Information Retrieval Systems

Μοντέλα Ανάκτησης (Retrieval Models)

Part B

Yannis Tzitzikas

University of Crete

CS-463,Spring 05

Lecture : 4
Date : 3-3-2005

Διάρθρωση Διάλεξης

PART (A)

- *Ανάκτηση και Φιλτράρισμα*
- *Εισαγωγή στα Μοντέλα Άντλησης*
- *Κατηγορίες Μοντέλων*
- *Exact vs Best Match*
- *Τα κλασσικά μοντέλα ανάκτησης*
  - *Το Boolean Μοντέλο*
  - *Στατιστικά Μοντέλα - Βάρυνση Όρων*
  - *Το Διανυσματικό Μοντέλο*
  - *Το Πιθανοκρατικό Μοντέλο*

PART (B):***Εναλλακτικά μοντέλα***

- *(I) Συνολοθεωρητικά μοντέλα*
  - *Fuzzy Retrieval Model*
  - *Extended Boolean Model*
- *(II) Αλγεβρικά Μοντέλα*
  - *Latent Semantic Indexing*
  - *Neural Netwok Model*

PART (C):

- *(III) Πιθανοκρατικά Μοντέλα*
  - *Bayesian Network Model*
  - *Inference Network Model*

Συνολοθεωρητικά Μοντέλα

- Κίνητρο
  - Επέκταση του Boolean model με **μερικό** ταίριασμα
- Μοντέλα:
  - Fuzzy Set Model
  - Extended Boolean Model

Information Retrieval Models

Fuzzy Set Retrieval Model

Fuzzy Set Model

- Βασική Ιδέα:
  - Έγγραφα και επερωτήσεις παριστάνονται σε **σύνολα** όρων ευρετηρίου
  - Κάθε **όρος** σχετίζεται με ένα **fuzzy set**
  - Κάθε έγγραφο έχει ένα degree of membership σε αυτό το fuzzy set
- Υπάρχουν αρκετά μοντέλα που θεμελιώνονται έτσι, εδώ θα δούμε το μοντέλο που προτάθηκε από Ogawa, Morita, and Kobayashi (1991)

Fuzzy Set Model: Η γενική ιδέα

- Η γενική ιδέα με ένα 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.
- Θεμελίωση της ιδέας με Fuzzy Theory

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,
  - $\mu_{\neg A}(u) = 1 - \mu_A(u)$
  - $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
  - $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

Πίνακας Συσχέτισης (correlation matrix) και εγγύτητα όρων

$$\begin{array}{c|cccc} & k_1 & k_2 & \ldots & k_t \\ k_1 & c_{11} & c_{21} & \ldots & c_{t1} \\ k_2 & c_{12} & c_{22} & \ldots & c_{t2} \\ : & : & : & & : \\ : & : & : & & : \\ k_t & c_{1n} & c_{2n} & \ldots & c_{tn} \end{array}$$

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

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 την εγγύτητα (proximity) μεταξύ των όρων

Fuzzy Information Retrieval

- Οι συντελεστές συσχέτισης μας επιτρέπουν να ορίσουμε το βαθμό membership ενός εγγράφου dj.

- Σε κάθε όρο i αντιστοιχούμε ένα fuzzy set με χαρ/κή συνάρτηση μ_i
- Αν το doc dj περιέχει τον όρο k_w που σχετίζεται ισχυρά με τον k_i τότε
  - c(i,w) ~ 1
  - $\mu_i(j)$ ~ 1, με άλλα λόγια και ο όρος ki είναι καλός για το dj
- Τυπικά:

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

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

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

Fuzzy Information Retrieval

Έστω q σε DNF q = c1 v … v ck

Σύμφωνα με τη fuzzy set theory:

$$\mu_q(j) = \max(\mu_{c1}(j), \ldots, \mu_{ck}(j))$$

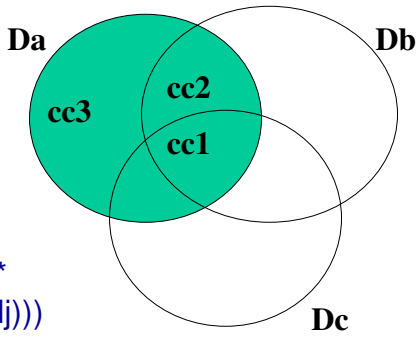
Παρά ταύτα, εδώ προτείνεται η χρήση αθροίσματος αντί του του max

$\mu_q(j)$= ….

Παράδειγμα

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

$vec(q_{dnf}) = (1,1,1) + (1,1,0) + (1,0,0)$
$= vec(cc1) + vec(cc2) + vec(cc3)$



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

Fuzzy Retrieval Model: Σύνοψη

- K={k_1,…,k_t} : σύνολο όλων των λέξεων ευρετηρίασης
- Κάθε έγγραφο **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 **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 Information Retrieval Models: Conclusion

- 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

Extended Boolean Model

Extended Boolean Model

- Κίνητρο
  - Το Boolean model είναι απλό και κομψό αλλά δεν παρέχει κατάταξη
- Προσσέγιση
  - Όπως και στο fuzzy model, μπορούμε να πάρουμε κατάταξη χαλαρώνοντας τη χαρακτηριστική συνάρτηση των συνόλων
  - Επέκταση του Boolean model με **βάρυνση όρων** και **μερικό ταίριασμα**
  - Συνδιασμός χαρακτηριστικών του Vector model και ιδιοτήτων της Boolean algebra

[Salton, Fox, and Wu, 1983]

Κίνητρο

Έστω $q = k_x \wedge ky$.

Σύμφωνα με το Boolean model ένα έγγραφο που περιέχει **μόνο ένα** από τα k_x, k_y είναι **μη-συναφές**, και μάλιστα τόσο μη-συναφές, όσο ένα έγγραφο που δεν περιέχει **κανένα** από τους 2 όρους.

Έστω δύο όροι k_x , k_y

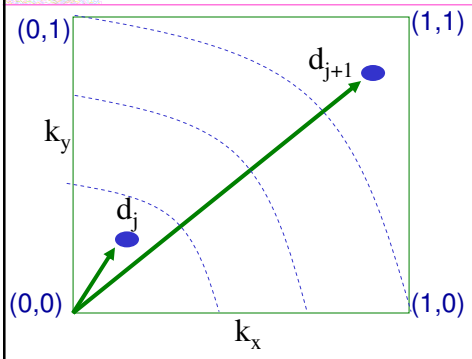
Μπορούμε να απεικονίσουμε επερωτήσεις και έγγραφα στο 2D χώρο.

Ένα έγγραφο d_j τοποθετείται βάσει των, βαρών $w_{x,j}$ και $w_{y,j.}$
Έστω ότι τα βάρη αυτά είναι κανονικοποιημένα στο [0,1], π.χ. :

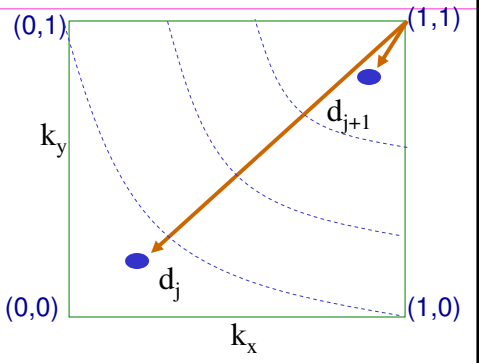
$$w_{x,j} = tf_{x,j}\, idf_x$$
$$w_{y,j} = tf_{y,j}\, idf_y$$

For brevity let $x = w_{x,j}$ and $y = w_{y,j}$
Άρα οι συντεταγμένες του dj είναι οι (x,y)

Η γενική ιδεά

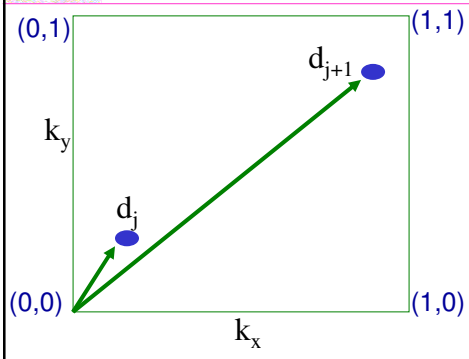


Έστω $q_{OR}=k_x \vee k_y$
Το σημείο (0,0) είναι η θέση προς αποφυγή.
Άρα μπορούμε να θεωρήσουμε την *απόσταση* του dj από αυτό το σημείο ως το **βαθμό ομοιότητας**



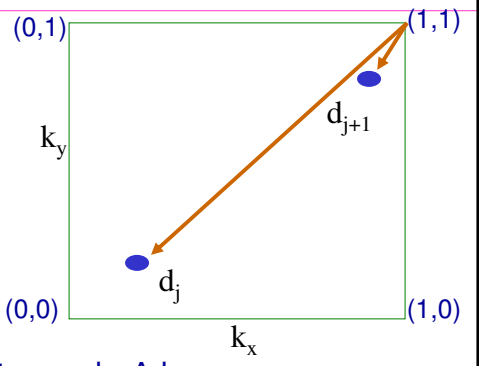
Έστω $q_{AND}=k_x \wedge k_y$
Το σημείο (1,1) είναι η πιο επιθυμητή θέση.
Άρα μπορούμε να θεωρήσουμε το *συμπλήρωμα της απόστασης* του dj από αυτό το σημείο ως **βαθμό ομοιότητας**

Η γενική ιδεά (II)



Let $q_{OR}=k_x \vee k_y$

$$sim(q_{OR}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$



Let $q_{AND}=k_x \wedge k_y$

$$sim(q_{AND}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

("2" for normalisation to [0,1])

Γενικεύοντας την ιδέα (για >2 όρους)

- Μπορούμε να γενικεύσουμε το προηγούμενο μοντέλο χρησιμοποιώντας την Ευκλείδεια απόσταση στον **t-διάστατο χώρο**
- Αυτό μπορεί να γίνει χρησιμοποιώντας ***p-norms*** που γενικεύουν την έννοια της απόστασης, όπου $1 \leq p \leq \infty$.
  - *Διαζευκτικές* επερωτήσεις
    - q_{OR} = k1 V k2 V . . V km
  - *Συζευκτικές* επερωτήσεις
    - q_{AND} = k1 Λ k2 Λ . . . Λ km

$$sim(q_{OR}, d) = \left(\frac{x_1^p + x_2^p + ... + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$sim(q_{AND}, d) = 1 - \left(\frac{(1-x_1)^p + ... + (1-x_m)^p}{m} \right)^{\frac{1}{p}}$$

Μερικές ενδιαφέρουσες ιδιότητες

- Μεταβάλλοντας το *p*, μπορούμε να κάνουμε το μοντέλο να συμπεριφέρεται όπως το Vector, το Fuzzy, ή ενδιάμεσα σε αυτά τα δυο.
- Αν p = 1 τότε (Vector like)
  - $sim(q_{OR},dj) = sim(q_{AND},dj) = \frac{x1 + \ldots + xm}{m}$
- Αν p = ∞ τότε (Fuzzy like)
  - $sim(q_{OR},dj) = \max(x_i)$
  - $sim(q_{AND},dj) = \min(x_i)$

Ερώτηση: Που πήγαν οι όροι της επερώτησης;

Πιο γενικές επερωτήσεις

- Έστω q = (k1 Λ k2) V k3
- Εφαρμόζουμε τους ορισμούς σεβόμενοι τη σειρά, εδώ:

$$sim(q\ ,d) = \left(\frac{(1-(\frac{(1-x_1)^p + (1-x_2)^p}{2})^{1/p})^p + x_3^{\ p}}{2} \right)^{\frac{1}{p}}$$

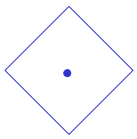
- Έστω q = (k1 V^2 k2) Λ^∞ k3
  - K1 and k2 should be used as in a vector system but the presence of k3 is required

Μερικές Παρατηρήσεις

- Είναι αρκετά ισχυρό μοντέλο με ενδιαφέρουσες ιδιότητες
- Η επιμεριστική ιδιότητα δεν ισχύει:
  - q1 = (k1 ∨ k2) ∧ k3
  - q2 = (k1 ∧ k3) ∨ (k2 ∧ k3)
  - sim(q1,dj) ≠ sim(q2,dj)

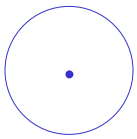
Ισομετρικές καμπύλες $\sqrt[p]{(x^p + y^p)}$

L_1



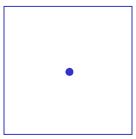
$x + y = 1$

L_2

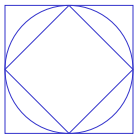


$\sqrt[2]{(x^2 + y^2)} = 1$

L_∞



$\max(x, y) = 1$



Διάρθρωση Διάλεξης

PART (A)

- *Ανάκτηση και Φιλτράρισμα*
- *Εισαγωγή στα Μοντέλα Άντλησης*
- *Κατηγορίες Μοντέλων*
- *Exact vs Best Match*
- *Τα κλασσικά μοντέλα ανάκτησης*
  - *Το Boolean Μοντέλο*
  - *Στατιστικά Μοντέλα - Βάρυνση Όρων*
  - *Το Διανυσματικό Μοντέλο*
  - *Το Πιθανοκρατικό Μοντέλο*

PART (B): ***Εναλλακτικά μοντέλα***

- *(I) Συνολοθεωρητικά μοντέλα*
  - *Fuzzy Retrieval Model*
  - *Extended Boolean Model*
- ***(II) Αλγεβρικά Μοντέλα***
  - ***Latent Semantic Indexing***
  - *Neural Netwok Model*

PART (C):

- *(III) Πιθανοκρατικά Μοντέλα*
  - *Bayesian Network Model*
  - *Inference Network Model*

Information Retrieval Models

Latent Semantic Indexing

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**)

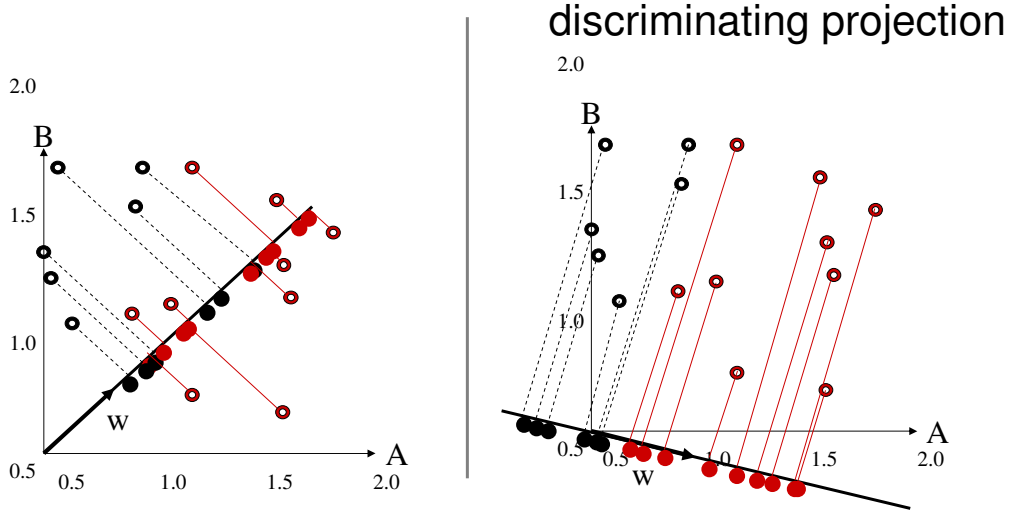
LSI: Η γενική προσέγγιση

- 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.

Η ιδέα

- The key idea is to map documents and queries into a **lower dimensional space**
  - (i.e., composed of higher level concepts which are in fewer number than the index terms)
- Retrieval in this reduced concept space might be superior to retrieval in the space of index terms
- But how to learn the concepts from data?

Παράδειγμα προβολής 2 διαστάσεων σε μία



SVD (Singular Value Decomposition)

- SVD is applied to derive the latent semantic structure model.
- What is SVD?
  - A dimensionality reduction technique
  - For more about matrices 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

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 based on a **tf-idf** weighting scheme

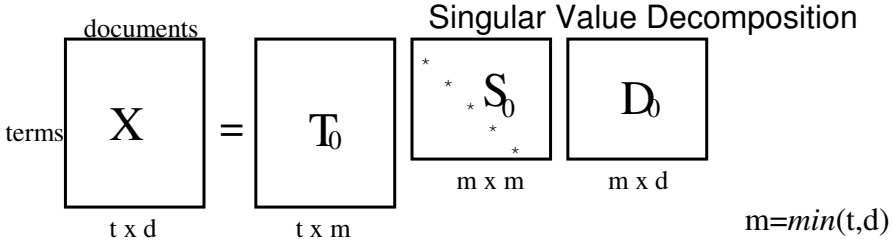
$$\mathbf{X} = \begin{pmatrix} & d_1 & d_2 & \ldots & d_d \\ k_1 & w_{11} & w_{21} & \ldots & w_{d1} \\ k_2 & w_{12} & w_{22} & \ldots & w_{d2} \\ : & : & : & & : \\ : & : & : & & : \\ k_t & w_{1t} & w_{2t} & \ldots & w_{dt} \end{pmatrix}$$

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

Latent Semantic Indexing: Ο τρόπος

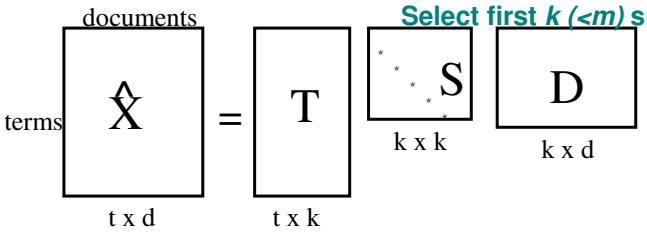
t: total number of index terms
d: total number of documents

Singular Value Decomposition



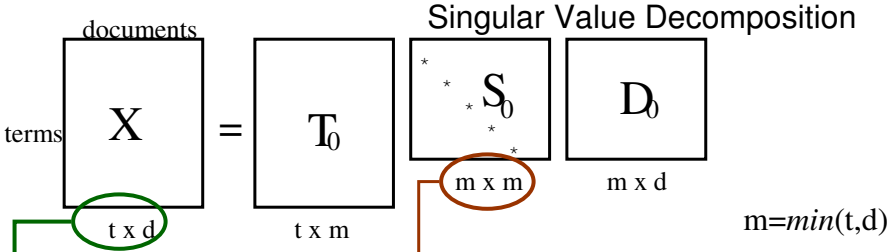
m=*min*(t,d)

Select first *k (<m)* singular values



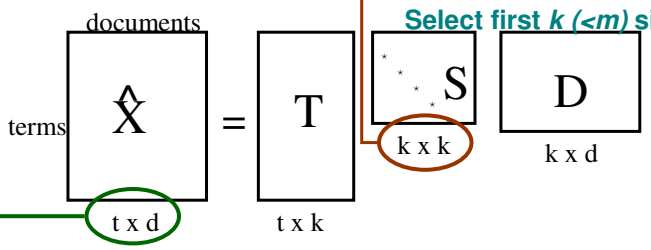
t: total number of index terms
d: total number of documents

Singular Value Decomposition



m=*min*(t,d)

Select first *k (<m)* singular values



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

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

$$\begin{bmatrix} 0.22 & -0.11 \\ 0.20 & -0.07 \\ 0.24 & 0.04 \\ 0.40 & 0.06 \\ 0.64 & -0.17 \\ 0.27 & 0.11 \\ 0.27 & 0.11 \\ 0.30 & -0.14 \\ 0.21 & 0.27 \\ 0.01 & 0.49 \\ 0.04 & 0.62 \\ 0.03 & 0.45 \end{bmatrix}$$

S

$$\begin{bmatrix} 3.34 & \\ & 2.54 \end{bmatrix}$$

D'

$$\begin{bmatrix} 0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{bmatrix}$$

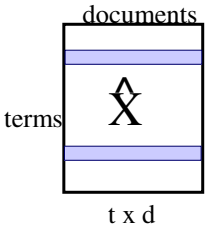
TS define coordinates for documents in latent space

Παρατηρήσεις

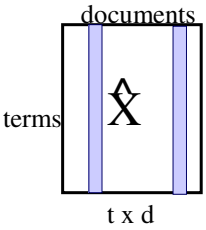
- Η παράμετρος k (<m) πρέπει να είναι:
  - large enough to allow fitting the characteristics of the data
  - small enough to filter out the non-relevant representational details

Τρόπος Σύγκρισης Όρων και Εγγράφων

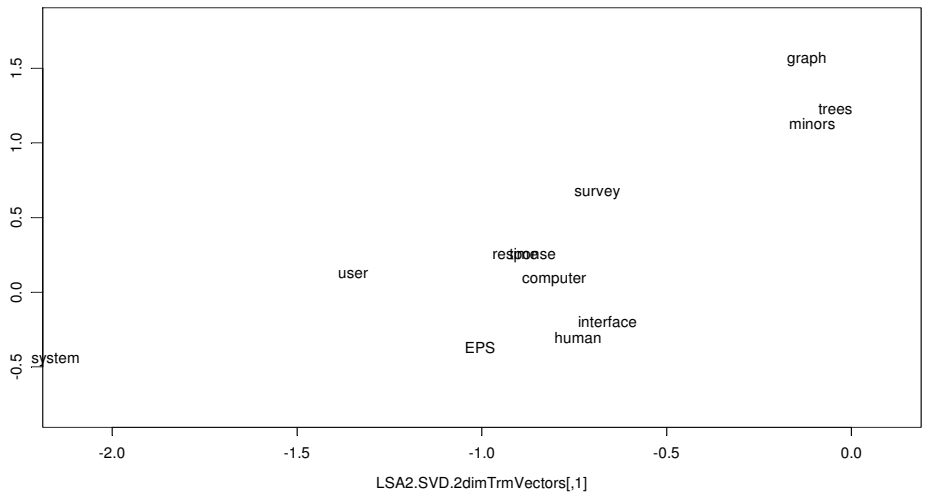
- Τρόπος σύγκρισης 2 όρων:
  - the **dot product** between two **row vectors of** X reflects the extent to which two terms have a similar pattern of occurrence across the set of document.



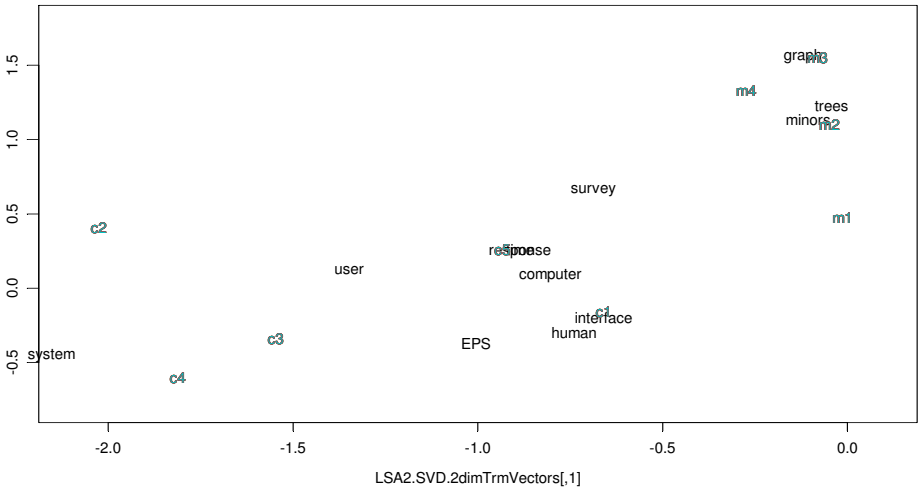
- Τρόπος σύγκρισης δύο εγγράφων:
  - **dot product** between two **column vectors** of X



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

$$\mathbf{X}$$

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

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

Παρατηρήσεις

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

Διάρθρωση Διάλεξης

PART (A)

- *Ανάκτηση και Φιλτράρισμα*
- *Εισαγωγή στα Μοντέλα Άντλησης*
- *Κατηγορίες Μοντέλων*
- *Exact vs Best Match*
- *Τα κλασσικά μοντέλα ανάκτησης*
  - *Το Boolean Μοντέλο*
  - *Στατιστικά Μοντέλα - Βάρυνση Όρων*
  - *Το Διανυσματικό Μοντέλο*
  - *Το Πιθανοκρατικό Μοντέλο*

PART (B):***Εναλλακτικά μοντέλα***

- *(I) Συνολοθεωρητικά μοντέλα*
  - *Fuzzy Retrieval Model*
  - *Extended Boolean Model*
- *(II) Αλγεβρικά Μοντέλα*
  - *Latent Semantic Indexing*
  - *Neural Netwok Model*

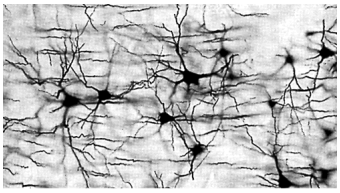
PART (C):

- *(III) Πιθανοκρατικά Μοντέλα*
  - *Bayesian Network Model*
  - *Inference Network Model*

Information Retrieval Models

Neural Network Model

(Μοντέλο Νευρωνικού Δικτύου)



Neural Network Model

- Κλασσικά μοντέλα ΑΠ:
  - Έγγραφα και επερωτήσεις ευρετηριάζονται από όρους
  - Η ανάκτηση βασίζεται στο “ταίριασμα” όρων
- Κίνητρο:
  - Είναι γνωστό ότι τα Νευρωνικά Δίκτυα είναι καλοί pattern matchers

Human Brain is a Neural Network

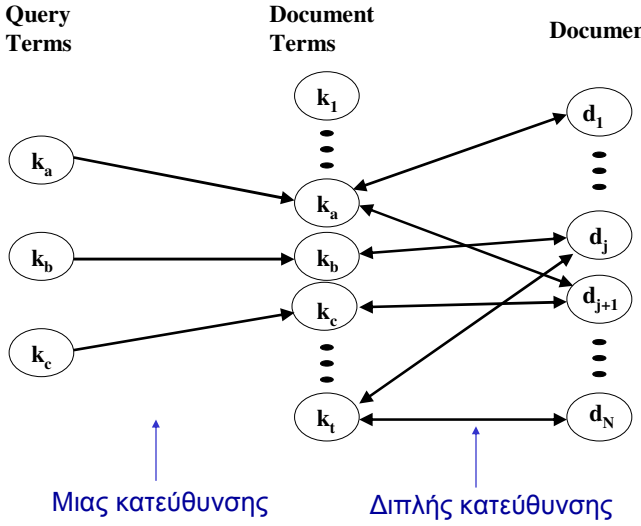
- The human brain is composed of billions of neurons
  - (1 million millions of nodes where each node has one thousands edges)
- Each neuron can be viewed as a small processing unit
- A neuron is stimulated by input signals and emits output signals in reaction
- A chain reaction of propagating signals 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

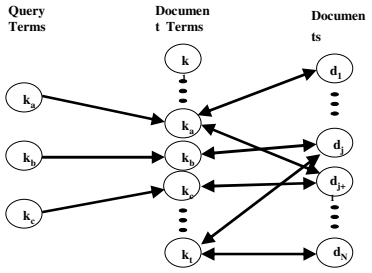
Neural Network for IR

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



Neural Network for IR

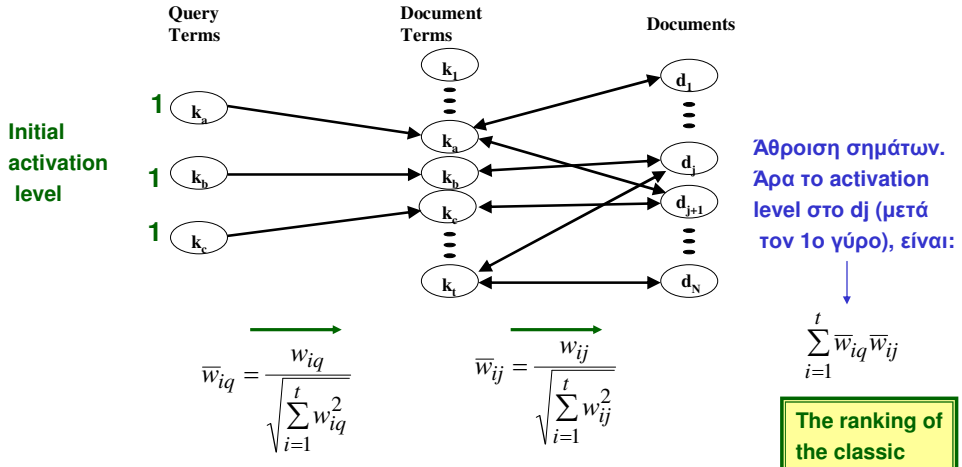
- Δίκτυο τριών επιπέδων
- Τα σήματα διαδίδονται (propagate) στο δίκτυο
- 1ο στάδιο διάδοσης:
  - Query terms issue the first signals
  - These signals propagate accross the network to reach the document nodes
- 2ο στάδιο διάδοσης:
  - Document nodes might themselves generate new signals which affect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



Μετάδοση σημάτων

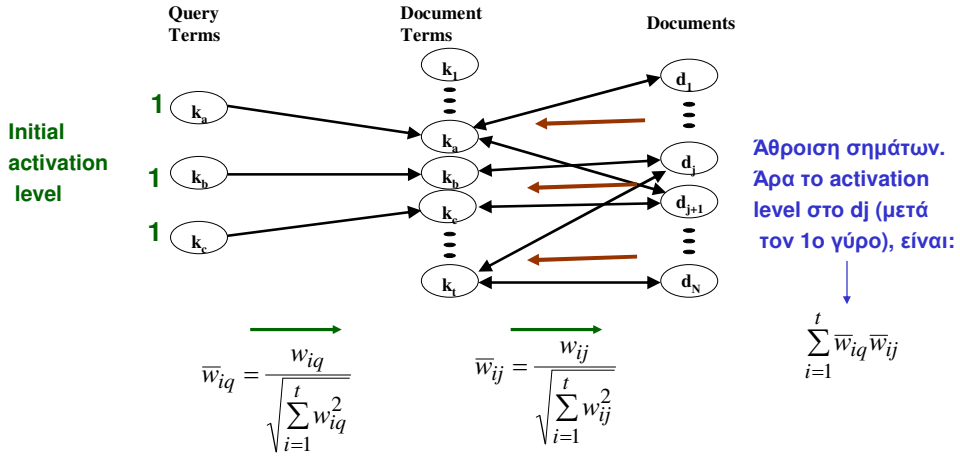
- Μέγιστη τιμή σήματος =1 (άρα κάνουμε κανονικοποίηση)
- Οι όροι της επερώτησης εκπέμπουν το αρχικό σήμα ίσο με 1
- Καθορισμός των βαρών στις ακμές:
  - query terms => terms
  - terms => docs

Μετάδοση σημάτων (II)



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

Μετάδοση σημάτων (III)



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

- 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