ΜΟΝΤΕΛΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Διαφάνειες του καθ. **Γιάννη Τζίτζικα** (Παν. Κρήτης)

http://www.ics.forth.gr/~tzitzik/

Για το πιθανοκρατικό του καθ. **Απ. Παπαδόπουλου** (Αριστοτέλειο Παν.)

Κεφάλαιο 2 του βιβλίου

2ο ΜΕΡΟΣ

Information Retrieval Models

Probabilistic Model

Κλασικά Μοντέλα Ανάκτησης

Τρία είναι τα, λεγόμενα, κλασικά μοντέλα ανάκτησης:

Λογικό (Boolean) που βασίζεται στη Θεωρία Συνόλων

Διανυσματικό (Vector) που βασίζεται στη Γραμμική Άλγεβρα

Πιθανοκρατικό (Probabilistic) που βασίζεται στη Θεωρία Πιθανοτήτων

Το Διανυσματικό και το Πιθανοκρατικό έχουν σημαντική επικάλυψη αν και στηρίζονται σε εντελώς διαφορετικές θεωρίες.

Πιθανοκρατικό Μοντέλο

Στόχος: να ορίσουμε το IR πρόβλημα σε πιθανοτικό πλαίσιο

- Για κάθε ερώτηση ***q*** (επερώτημα) υπάρχει ένα **ιδανικό σύνολο κειμένων** (R) που το ικανοποιεί.
- Επεξεργαζόμαστε την ερώτηση με βάση *τις ιδιότητες αυτού του συνόλου.*
- Ποιες είναι όμως αυτές οι ιδιότητες;
- Αρχικά γίνεται μία πρόβλεψη και στη συνέχεια η πρόβλεψη βελτιώνεται.

Πιθανοκρατικό Μοντέλο

- Αρχικά επιστρέφεται ένα σύνολο εγγράφων.
- Ο χρήστης εξετάζει τα κείμενα αναζητώντας σχετικά κείμενα.
- Το σύστημα IR χρησιμοποιεί το feedback του χρήστη ώστε να προσδιοριστεί καλύτερα το ιδανικό σύνολο κειμένων.
- Η διαδικασία επαναλαμβάνεται.
- Η περιγραφή του ιδανικού συνόλου κειμένων πραγματοποιείται πιθανοτικά.

Ανεξάρτητες Μεταβλητές και Πιθανότητα υπό Συνθήκη

Έστω *a,* και *b* δύο γεγονότα με πιθανότητες να συμβούν $P(a)$ και $P(b)$ αντίστοιχα.

Ανεξάρτητα Γεγονότα

Τα γεγονότα *a* και *b* είναι ανεξάρτητα αν και μόνο αν:

$$P(a \cap b) = P(b)\, P(a)$$

Υπό Συνθήκη Πιθανότητα

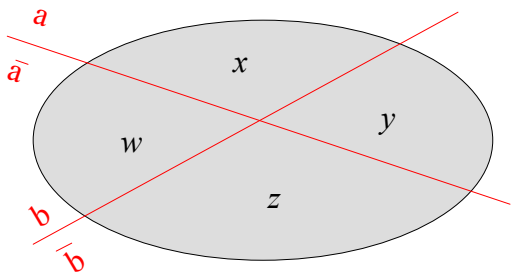
$P(a \mid b)$ είναι η πιθανότητα του *a* δεδομένου του *b*.

Τα γεγονότα $a_1, ..., a_n$ καλούνται υπό συνθήκη ανεξάρτητα αν και μόνο αν:

$$P(a_i \mid a_j) = P(a_i) \quad \text{για όλα τα } i \text{ και } j$$

Παράδειγμα I

$\bar{a}$ είναι η άρνηση του γεγονότος a



$P(a) = x + y$

$P(b) = w + x$

$P(a \mid b) = x / (w + x)$

$\boldsymbol{P(a \mid b)\, P(b) = P(a \cap b) = P(b \mid a)\, P(a)}$

Παράδειγμα II

Ανεξάρτητα γεγονότα

Έστω a και b οι τιμές που φέρνουν δύο ίδια ζάρια. Ισχύει:

$P(a{=}5 \mid b{=}3) = P(a{=}5) = 1/6$

Μη ανεξάρτητα

Έστω a και b οι τιμές που φέρνουν δύο ίδια ζάρια και t το άθροισμά τους. Τότε ισχύει:

$t = a + b$

$P(t{=}8 \mid a{=}2) = 1/6$
$P(t{=}8 \mid a{=}1) = 0$

Θεώρημα του Bayes

Έστω *a* και *b* δύο γεγονότα.

$P(a \mid b)$ είναι η πιθανότητα να συμβεί το γεγονός *a* δεδομένου ότι έχει συμβεί το γεγονός *b*.

Θεώρημα Bayes

$$P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

Ισχύει επίσης ότι:

$$P(a \mid b)\, P(b) = P(a \cap b) = P(b \mid a)\, P(a)$$

Θεώρημα Bayes: παράδειγμα

Example

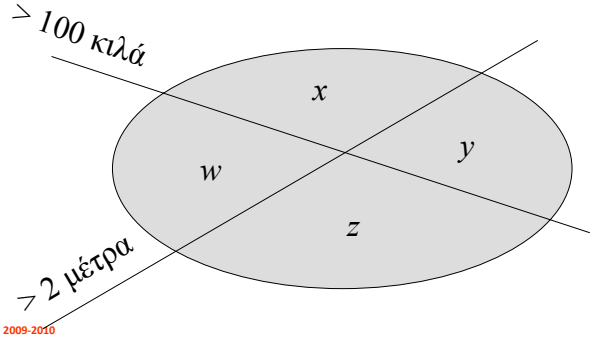
a βάρος πάνω από 100 κιλά

b ύψος πάνω από 2 μέτρα.

$P(a \mid b) = x / (w+x) = x / P(b)$

$P(b \mid a) = x / (x+y) = x / P(a)$

$x = P(a \cap b)$



Αρχή Πιθανοκρατικής Κατάταξης
Probabilistic Ranking Principle (PRP)

"If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing **probability of usefulness** to the user who submitted the request, where the probabilities are estimated as accurately a possible on the basis of whatever data is made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data."

Εάν η απάντηση ενός συστήματος ανάκτησης σε κάθε ερώτημα είναι μία λίστα εγγράφων ταξινομημένη με φθίνουσα διάταξη ως προς την **πιθανότητα σχετικότητας** του κάθε εγγράφου ως προς το χρήστη, όπου οι πιθανότητες υπολογίζονται όσο γίνεται ακριβέστερα με βάση τα δεδομένα που είναι διαθέσιμα, η συνολική αποτελεσματικότητα του συστήματος θα είναι η καλύτερη δυνατή.

W.S. Cooper

Πιθανοκρατική Βαθμολόγηση

"Για ένα δεδομένο ερώτημα, εάν γνωρίζουμε κάποια από τα σχετικά έγγραφα, οι **όροι** που εμφανίζονται σε αυτά θα πρέπει να έχουν μεγαλύτερη **βαρύτητα** κατά την αναζήτηση άλλων σχετικών εγγράφων.

Κάνοντας διάφορες παραδοχές σχετικά με την κατανομή των όρων και χρησιμοποιώντας το θεώρημα του Bayes είναι δυνατόν να υπολογίσουμε τα βάρη αυτά."

Van Rijsbergen

Βασικές Έννοιες

- Η πιθανότητα ένα έγγραφο να είναι σχετικό ως προς το ερώτημα θεωρείται ότι εξαρτάται μόνο από τους όρους που περιέχονται στο έγγραφο και από τους όρους που περιέχονται στο ερώτημα.

- Η σχετικότητα ενός εγγράφου d ως προς το ερώτημα q δεν εξαρτάται από τη σχετικότητα άλλων εγγράφων της συλλογής.

- Για κάποιο ερώτημα q το σύνολο των σχετικών εγγράφων ***R*** είναι το **ιδανικό σύνολο** που μπορούμε να έχουμε ως απάντηση.

Βασικές Έννοιες

Για ένα ερώτημα q και ένα έγγραφο d το πιθανοκρατικό μοντέλο χρειάζεται μία εκτίμηση για την πιθανότητα $P(R \mid d)$ που δηλώνει την πιθανότητα το έγγραφο d να είναι σχετικό ως προς το ερώτημα.

P(R|d) πιθανότητα το έγγραφο να είναι σχετικό με το ερώτημα

$P(\overline{R}|d)$ πιθανότητα το έγγραφο να μην είναι σχετικό με το ερώτημα

Μέτρο Ομοιότητας (odds of being relevant to q):

$S(q, d)$, ομοιότητα του εγγράφου d ως προς το ερώτημα q:

$$\frac{\text{πιθανότητα } d \text{ σχετικό}}{\text{πιθανότητα } d \text{ μη σχετικό}} = \frac{P(R \mid d)}{P(\overline{R} \mid d)}$$

Οι τιμές της $S(\ )$ μπορεί να είναι από πολύ μικρές έως πολύ μεγάλες και για αυτό χρησιμοποιείται συνήθως ο λογάριθμος για την άμβλυνση των διαφορών.

Βασικές Έννοιες

$$S(q, d) = \frac{P(R \mid d)}{P(\bar{R} \mid d)}$$

$$= \frac{P(d \mid R)\, P(R)}{P(d \mid \bar{R})\, P(\bar{R})} \qquad \text{θεώρημα Bayes}$$

$P(d \mid R)$ είναι η πιθανότητα να διαλέξουμε τυχαία το έγγραφο d από τη συλλογή των σχετικών με την ερώτηση εγγράφων R.

$$\frac{P(d \mid R)\, \cancel{P(R)}}{P(d \mid \bar{R})\, \cancel{P(\bar{R})}}$$

Ίδια (σταθερά) για όλα τα έγγραφα της συλλογής (έστω μια σταθερά k)

Άρα πρέπει να εκτιμήσουμε/υπολογίσουμε αυτές τις πιθανότητες

Πως; Κοιτάμε τους όρους (terms) που εμφανίζονται στο d

Βασικές Έννοιες

$$\frac{P(d \mid R)\, \cancel{P(R)}}{P(d \mid \bar{R})\, \cancel{P(\bar{R})}}$$

P(d | R): Πιθανότητα να επιλέξουμε το έγγραφο d από τα σχετικά με την ερώτηση

Θα χρησιμοποιήσουμε τους όρους k_i που έχει το έγγραφο d για να την υπολογίσουμε

Βασικές Έννοιες

Ανάκτηση Δυαδικής Ανεξαρτησίας

Binary Independence Retrieval (BIR)

Τα βάρη των όρων είναι δυαδικά και οι όροι είναι ανεξάρτητοι μεταξύ τους (η παρουσία ή μη κάποιου όρου δεν επηρεάζει τους υπόλοιπους).

Το βάρος ενός όρου σε ένα έγγραφο είναι είτε 1 (αν ο όρος περιέχεται στο έγγραφο) είτε 0 (σε διαφορετική περίπτωση).

Όπως και στο Λογικό αλλά και στο Διανυσματικό μοντέλο, η σχετικότητα ενός εγγράφου καθορίζεται από τους όρους που περιέχονται σε αυτό.

Naïve Bayes

Έστω **x** = (x_1, x_2, ... x_n) το διάνυσμα του εγγράφου *d* όπου
x_i = 1 αν ο i-οστός όρος περιέχεται στο έγγραφο, x_i = 0 διαφορετικά.

Η εκτίμηση της πιθανότητας $P(d \mid R)$ γίνεται χρησιμοποιώντας την πιθανότητα $P(\mathbf{x} \mid R)$

Εάν οι όροι είναι ανεξάρτητοι τότε:

$$\boldsymbol{P(x \mid R)} = P(x_1 \cap R)\, P(x_2 \cap R) \ldots P(x_n \cap R)$$

$$= P(x_1 \mid R)\, P(x_2 \mid R) \ldots P(x_n \mid R)$$

$$= \prod \boldsymbol{P(x_i \mid R)}$$

$\boldsymbol{P(x_i \mid R)}$ είναι η πιθανότητα ο όρος x_i να βρίσκεται σε ένα έγγραφο που επιλέγεται τυχαία από το ιδανικό σύνολο *R*.

Αντίστοιχα $P(x_i \mid \bar{R})$

Το μοντέλο αυτό είναι γνωστό και ως **Naive Bayes**

Συνάρτηση Ομοιότητας

$$S(q, d) = k \frac{\prod P(x_i \mid R)}{\prod P(x_i \mid \bar{R})}$$

Αφού το κάθε x_i είναι 0 ή 1 έχουμε:

$$S = k \prod_{x_i = 1} \frac{P(x_{i=1} \mid R)}{P(x_{i=1} \mid \bar{R})} \prod_{x_i = 0} \frac{P(x_{i=0} \mid R)}{P(x_{i=0} \mid \bar{R})}$$

Το σπάμε: όροι που το x_i είναι 1 και όροι που το x_i είναι 0

Συνάρτηση Ομοιότητας

Για τους όρους που εμφανίζονται στο ερώτημα θέτουμε:

$p_i = P(x_i = 1 \mid R)$

$r_i = P(x_i = 1 \mid \bar{R})$

p_i πιθανότητα ότι ένα έγγραφο που επιλέγεται από το ιδανικό σύνολο έχει τον όρο x_i -- ένας όρος εμφανίζεται σε ένα ιδανικό έγγραφο

r_i το ίδιο για το μη ιδανικό

Για τους όρους που δεν εμφανίζονται στο ερώτημα έστω:

$p_i = r_i$

όροι με $q_i = 0$ είναι ίσοι με $p_i/r_i = 1$

$$S = k \prod_{x_i = q_i = 1} \frac{p_i}{r_i} \prod_{x_i = 0, q_i = 1} \frac{1 - p_i}{1 - r_i}$$

Πολλαπλασιάζουμε το δεξί γινόμενο με τους όρους που υπάρχουν στο έγγραφο και διαιρούμε το αριστερό γινόμενο με τον ίδιο όρο

$$= k \prod_{x_i = q_i = 1} \frac{p_i (1 - r_i)}{r_i (1 - p_i)} \prod_{q_i = 1} \frac{1 - p_i}{1 - r_i}$$

σταθερή ποσότητα για δεδομένο ερώτημα (ανεξάρτητη του εγγράφου)

Συνάρτηση Ομοιότητας

Με λογαρίθμηση της σχέσης και αγνοώντας σταθερούς παράγοντες η συνάρτηση ομοιότητας $S_{prob}(q,d)$ παίρνει τη μορφή:

$$S_{prob}(q,d) = \log\left(S(q,d)\right)$$

$$\boxed{S_{prob}(q,d) = \sum_i \log \frac{p_i \cdot (1-r_i)}{r_i \cdot (1-p_i)}}$$

Όπου η άθροιση αφορά στους όρους που βρίσκονται και στο ερώτημα και στο έγγραφο.

Σχέση με το Διανυσματικό Μοντέλο

Στο Διανυσματικό μοντέλο ανάκτησης θεωρήστε ότι η *i*-οστή συνιστώσα του διανύσματος ενός εγγράφου (βάρος) ισούται με την ποσότητα

$$\log \frac{p_i \cdot (1-r_i)}{r_i \cdot (1-p_i)}$$

ενώ το διάνυσμα του ερωτήματος *q* ισούται με άσσους για τους όρους που ανήκουν στο ερώτημα και μηδενικά διαφορετικά.

Τότε, η συνάρτηση ομοιότητας $S_{prob}(q,d)$ ισούται με το εσωτερικό γινόμενο των δύο διανυσμάτων.

Αλλάζουμε μόνο τον τρόπο που υπολογίζονται τα βάρη

Αρχική Εκτίμηση των $P(x_i \mid R)$

Αρχικά θέτουμε τιμές στις πιθανότητες :

$$p_i = P(x_i \mid R) = c$$

$$r_i = P(x_i \mid \overline{R}) = n_i / N$$

p_i πιθανότητα ότι ένα έγγραφο που επιλέγεται από το ιδανικό σύνολο έχει τον όρο x_i

r_i το ίδιο για το μη ιδανικό

όπου:

c είναι μία τυχαία σταθερά (π.χ., 0.5) ίδια για όλους τους όρους *(δεν επηρεάζουν)*

η κατανομή των όρων ανάμεσα στα μη σχετικά ακολουθεί την κατανομή που ακολουθεί σε όλη τη συλλογή – δεν επηρεάζει την επιλογή

n_i είναι το πλήθος των εγγράφων που περιέχουν τον i-οστό όρο

N πλήθος εγγράφων συλλογής *(document frequency)*

The document ranking is determined simply by which query terms appear in the document scaled by their idf weighting

Προσαρμογή Τιμών των $P(x_i \mid R)$

Είναι προφανές ότι η αυθαίρετη ανάθεση τιμών δεν μπορεί να οδηγεί πάντα σε ικανοποιητικά αποτελέσματα. Για τη βελτίωση της ποιότητας των αποτελεσμάτων οι πρώτες εφαρμογές του Πιθανοκρατικού μοντέλου χρειάζονταν την παρέμβαση του χρήστη για την αναπροσαρμογή των τιμών.

Εναλλακτικά μπορεί να χρησιμοποιηθεί και αυτοματοποιημένος τρόπος. Αρχικά εκτελείται το ερώτημα με τις αρχικές εκτιμήσεις. Επιλέγονται τα k καλύτερα έγγραφα. Έστω k_i ο αριθμός των εγγράφων που περιέχουν τον i-οστό όρο. Θέτουμε:

$$p_i = P(x_i \mid R) = k_i / k$$

$$r_i = P(x_i \mid \overline{R}) = (n_i - k_i) / (N - k)$$

Υποθέσεις

Υποθέσεις

1. Δυαδική αναπαράσταση ερωτημάτων και κειμένων (0,1 (υπάρχει/δεν υπάρχει ο όρος)
2. Ανεξαρτησία όρων
3. Όροι που δεν εμφανίζονται στην ερώτηση δεν επηρεάζουν το αποτέλεσμα
4. Οι τιμές σχετικότητας των εγγράφων είναι ανεξάρτητες μεταξύ τους (επιλογή όμοιων (ή σχεδόν όμοιων) εγγράφων

Πλεονεκτήματα-Μειονεκτήματα

Πλεονεκτήματα:

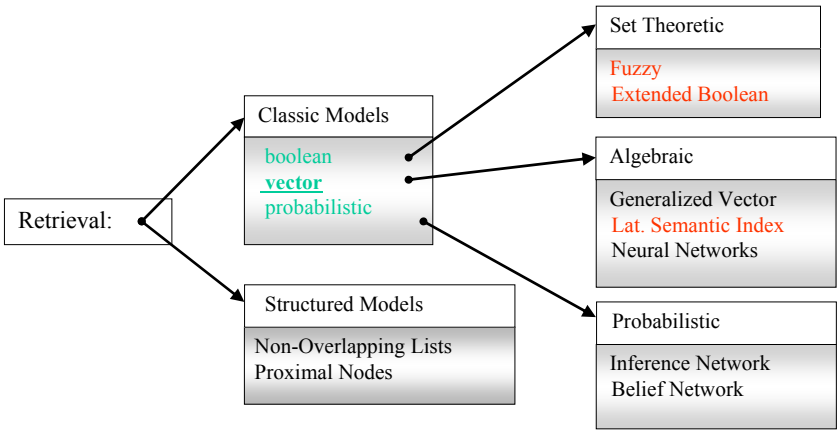
1. Απλό μοντέλο
2. Τα κείμενα ταξινομούνται σε φθίνουσα διάταξη ως προς την πιθανότητα να είναι σχετικά

Θεωρητικό τρόπο ορισμού της σχετικότητας

Μειονεκτήματα:

1. Χρειάζεται να μαντέψουμε
2. Δε λαμβάνεται υπ' όψιν η συχνότητα εμφάνισης
3. Θεωρεί ότι οι όροι είναι ανεξάρτητοι

Μια Ταξινομία των Μοντέλων Ανάκτησης



Information Retrieval Models

Extended Boolean Model

Extended Boolean Model

- Κίνητρο
  - Το Boolean model είναι απλό και κομψό αλλά δεν παρέχει κατάταξη (διαβάθμιση των συναφών εγγράφων)
- Προσέγγιση
  - Επέκταση του Boolean model με **βάρυνση όρων** και **μερικό ταίριασμα**
  - Συνδυασμός χαρακτηριστικών του Vector model και ιδιοτήτων της Boolean algebra

[Salton, Fox, and Wu, 1983]

Σκεπτικό / 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.

Extended Boolean Model

Έστω ότι έχουμε μόνο δύο όρους k_x , k_y

Μπορούμε να θεωρήσουμε κάθε όρο ως μια διάσταση
Άρα έγγραφα και επερωτήσεις απεικονίζονται στο 2D χώρο.

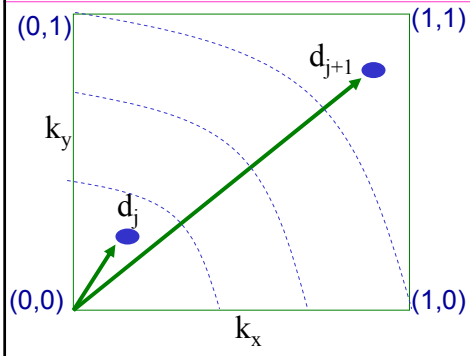
Ένα έγγραφο d_j τοποθετείται βάσει των, βαρών $w_{x,j}$ και $w_{y,j.}$
Έστω ότι τα βάρη αυτά είναι κανονικοποιημένα στο [0,1], π.χ. :

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

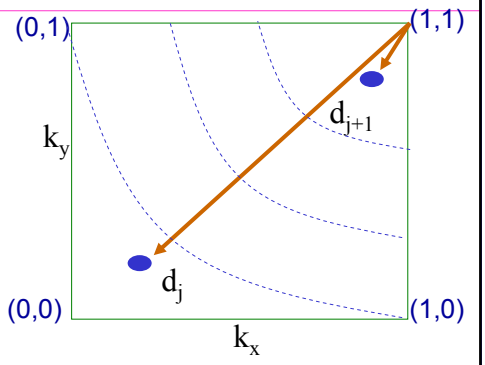
Για συντομία έστω: $x = w_{x,j}$ και $y = w_{y,j}$

Άρα οι συντεταγμένες του d_j είναι οι (x, y)

Η γενική ιδέα

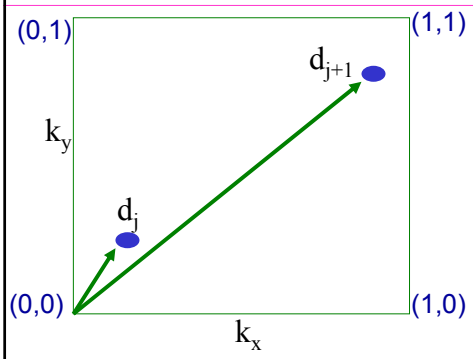


Έστω $q_{OR}=k_x \vee k_y$
Το σημείο (0,0) είναι η θέση προς αποφυγή.
Άρα μπορούμε να θεωρήσουμε την *απόσταση* του d_j από αυτό το σημείο
ως το **βαθμό ομοιότητας** *(όσο πιο μακριά, τόσο πιο όμοιο)*



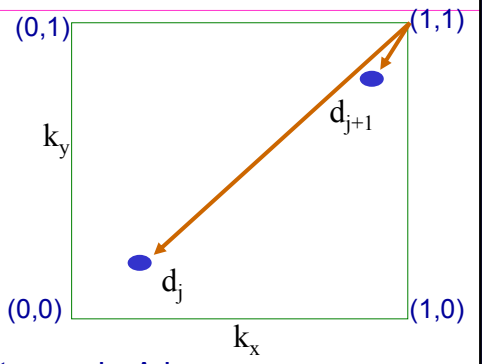
Έστω $q_{AND}=k_x \wedge k_y$
Το σημείο (1,1) είναι η πιο επιθυμητή θέση.
Άρα μπορούμε να θεωρήσουμε το *συμπλήρωμα της απόστασης* του d_j από αυτό το σημείο
ως το **βαθμό ομοιότητας** *(όσο πιο κοντά, τόσο πιο όμοιο)*

Η γενική ιδεά (ΙΙ)



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])

Όταν δυαδικά βάρη (0, 1);

Γενικεύοντας την ιδέα (για >2 όρους)

Μπορούμε να γενικεύσουμε το προηγούμενο μοντέλο χρησιμοποιώντας την *Ευκλείδεια απόσταση* στον t-διάστατο χώρο

Μπορεί να γενικευτεί επίσης χρησιμοποιώντας ***p-norms*** που γενικεύουν την έννοια της απόστασης, όπου $1 \le p \le \infty$.(Ευκλείδεια, p = 2)

- *Διαζευκτικές* επερωτήσεις
  - q_{OR} = k1 V k2 V . . V km

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

- *Συζευκτικές* επερωτήσεις
  - q_{AND} = k1 Λ k2 Λ . . . Λ km

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

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)
  - $sim(q_{OR},d_j) = sim(q_{AND},d_j) = \frac{x_1 + \ldots + x_m}{m}$

Σε αυτήν την περίπτωση απλώς αθροίζουμε τα βάρη

- Αν p = ∞ τότε (Fuzzy like)
  - $sim(q_{OR},d_j) = \max(x_i)$
  - $sim(q_{AND},d_j) = \min(x_i)$

Ερώτηση: Που πήγαν οι όροι της επερώτησης;

Σύνθετες επερωτήσεις

Έστω $q = (k_1 \wedge k_2) \vee k_3$

- 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 $q = (k_1 \vee^2 k_2) \wedge^\infty k_3$
  - k_1 and k_2 should be used as in a vector system but the presence of k_3 is required

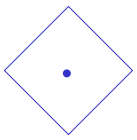
Μερικές Παρατηρήσεις

Είναι αρκετά ισχυρό μοντέλο με ενδιαφέρουσες ιδιότητες

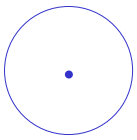
- Η επιμεριστική ιδιότητα δεν ισχύει:
  - $q_1 = (k_1 \vee k_2) \wedge k_3$
  - $q_2 = (k_1 \wedge k_3) \vee (k_2 \wedge k_3)$
  - $sim(q_1,d_j) \neq sim(q_2,d_j)$

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

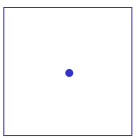
L_1 L_2 L_∞



$x + y = 1$

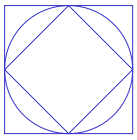


$\sqrt[2]{(x^2 + y^2)} = 1$



$\max(x, y) = 1$

Το σύνολο των διανυσμάτων που έχουν νόρμα 1



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)

Τι είναι ένα ασαφές σύνολο;

«Κλασσικά» σύνολα (crispy or Boolean sets): ένα στοιχείο ανήκει ή δεν ανήκει

Ασαφή σύνολα: ένα στοιχείο του συνόλου ανήκει με ένα βαθμό συμμετοχής (<=1)

Ιδέα:

Κάθε όρος της ερώτησης ένα ασαφές σύνολο

Ένα έγγραφο ανήκει σε αυτό το ασαφές σύνολο του όρου με ένα βαθμό

Μοντέλα Βασισμένα στη Θεωρία Ασαφών Συνόλων
(Fuzzy Set-based Retrieval Models)

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*** (βαθμός συμμετοχής) associated with the elements of a set
- This degree of membership varies from ***0 to 1***(τιμές στο διάστημα [0, 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

Background: Fuzzy Set Theory [Zadeh 1965]

- 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]

Βασικές πράξεις σε ασαφή σύνολα (συμπλήρωμα, τομή και ένωση)

- 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,…,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
  - $w_{i,j}$ το βάρος 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 ∧ (k2 ∨ ¬ k3))**"

$R(d_j, q) = \mu_q(d_j)$ άρα είναι ο βαθμός συμμετοχής του d_j στο σύνολο που προσδιορίζεται από τη λογική έκφραση q.

- Μπορούμε να υπολογίσουμε το $R(d_j,q)$ βάσει των κανόνων της θεωρίας των Fuzzy sets, θεωρώντας ότι $R(d_j, t_i) = \mu_{ti}(d_j) = w_{i,j}$
- Για παράδειγμα
  - $R(d_j, t1 \vee t2) = \max (R(d_j, t1), R(d_j, t2)) = \max (w_{1j}, w_{2j})$.
  - $R(d_j, t1 \wedge t2) = \min (R(d_j, t1), R(d_j, t2)) = \min (w_{1j}, w_{2j})$.

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 $q = k_x \vee k_y$. 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 $q = k_x \vee ky$ (κάτι που δεν είναι δυνατό με το Boolean Μοντέλο).

Το παρόν είναι μια ειδική περίπτωση του Extended Boolean Model (συγκεκριμένα αντιστοιχεί στην περίπτωση που $p = \infty$).

A Simple Retrieval Model based on Fuzzy Theory

Παρατηρήσεις

Πως θα υπολογίζουμε τη συνάρτηση συμμετοχής

[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.

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

Πίνακα συσχέτισης μεταξύ των όρων

term-term correlation matrix ή keyword connection matrix

	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}

Ορίζουμε ποσοτικά την εγγύτητα (proximity) μεταξύ δυο όρων k_i και k_l->
ως την συν-εμφάνισή τους στα έγγραφα της συλλογής

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

where:
$n_{i,l}$: number of docs which contain both k_i and k_l
n_i: number of docs which contain k_i
n_l: number of docs which contain k_l

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}

$$\mathbf{c(i,l) = c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}}$$

where:
$n_{i,l}$: number of docs which contain both k_i and k_l
n_i: number of docs which contain k_i
n_l: number of docs which contain k_l

Τέτοιες πίνακες είναι αρκετά συνηθισμένοι (θα τους ξαναδούμε σε αλγόριθμους clustering)

Πχ
$n_{il}=0$ => $c_{il}=0$
$n_{il}=3, n_i=3, n_l=9$ => $c_{il}=0.3$
$n_{il}=3, n_i=3, n_l=30$ => $c_{il}=0.1$
$n_{il}=3, n_i=3, n_l=3$ => $c_{il}=1$

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 $d_j=(w_{1,j},\ldots,w_{t,j})$ 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 [Ogawa, Morita, and Kobayashi,1991]

Έστω όρος k_i και έγγραφο d_j θέλουμε το βαθμό συμμετοχής του εγγράφου στο ασαφές σύνολο που ορίζει το k_i (συνάρτηση συμμετοχής μ_i)

$$\mu_i(j) = \sum_{k_w \in d_j} c_{i,w}$$

Άθροισμα του βαθμού συσχέτισης του k_i με τους όρους που εμφανίζονται στο d_j (θεωρούμε άθροισμα αντί για max, πιο ήπια διαβάθμιση)

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

Βασίζεται στο:

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

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

Για παράδειγμα έστω ότι το έγγραφο d_j δεν περιέχει τον όρο k_i

Αν το έγγραφο d_j περιέχει έναν όρο k_w που σχετίζεται ισχυρά με τον k_i τότε

- θα έχουμε $c_{iw} \sim 1$
- και άρα θα μπορούσαμε να θεωρήσουμε ότι $\mu_i(j) \sim 1$. Με άλλα λόγια, αν και ο όρος k_i δεν εμφανίζεται στο d_j, εντούτοις περιγράφει το περιεχόμενο του d_j

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 μεγίστου.

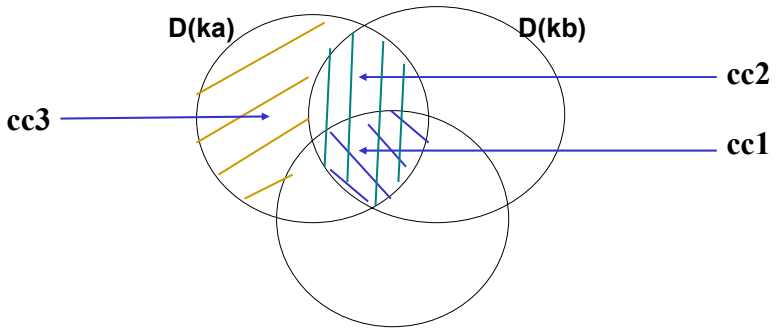
$R(d_j,q) = \mu_q(d_j) = \Sigma\, \mu_{cc}(d_j)$ για κάθε συζευκτική συνιστώσα cc του q_{DNF}

Fuzzy Set Retrieval Model
Παράδειγμα

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

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

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



D(kc) // documents containing the term kc

Fuzzy Set Retrieval Model
Παράδειγμα (II)

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

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

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

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

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

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

Fuzzy Set Retrieval Model
Σύνοψη

- K={k_1,…,k_t} : σύνολο όλων των λέξεων ευρετηρίασης
- Κάθε έγγραφο **d_j** παριστάνεται με το διάνυσμα d_j=($w_{1,j}$,…,$w_{t,j}$) 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)**"
- $R(d_j,q) = \mu_q(d_j) = \Sigma\ \mu_{cc}(d_j)$ για κάθε συζευκτική συνιστώσα cc του q_{DNF}
  - $\mu_{ki}(d_j) = 1 - \prod_{k_w \in d_j} (1 - c(k_i,k_w))$
  - $c(k_i,k_j)$ καθορίζεται από την συνεμφάνιση των όρων k_i και k_j 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

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 εκτίμησή της

LSI: The idea

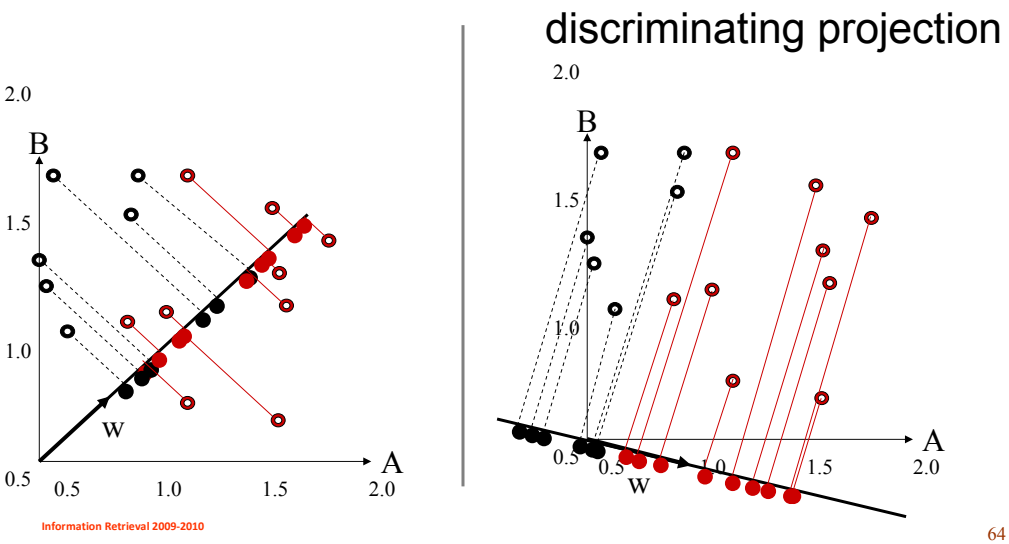
- The key idea is to map documents and queries into a **lower dimensional space**
  - (i.e., composed of higher level concepts which are fewer in number than the index terms)

- 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.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-1-understanding.html

SVD (HIDE)

- SVD: Διάσπαση σε ιδιάζουσες τιμές
- Ένας μεγάλος πίνακας όρων-εγγράφων αναλύεται σε ένα σύνολο από κ (100..200) ορθοκανονικούς παράγοντες από τους οποίους ο αρχικός πίνακας μπορεί να προσεγγιστεί με γραμμικό συνδυασμό.
- Πλέον έγγραφα και επερωτήσεις παριστάνονται βάσει αυτών των κ διαστάσεων

- Αφού οι διαστάσεις μειώθηκαν, οι λέξεις δεν μπορεί πλέον να είναι ανεξάρτητες

Definitions

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

Αρχικός Πίνακας (t x d)

X

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

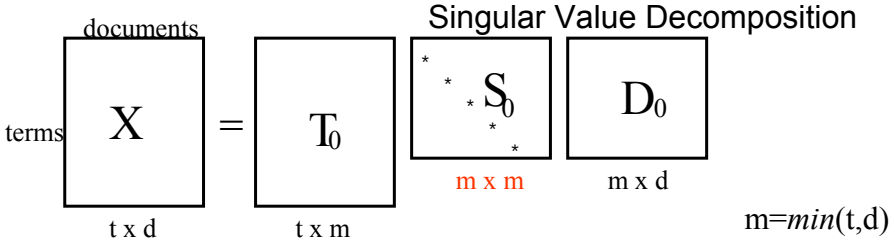
- (X_{ij}): be a term-document matrix with t rows and d columns
  - To each element of this matrix a weight $\mathbf{w_{ij}}$ associated is assigned with the pair $\mathbf{[k_i,d_j]}$
  - The weight w_{ij} can be $freq_{ij}$
    - (or based on a **tf-idf** weighting scheme)

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

Latent Semantic Indexing: Ο τρόπος

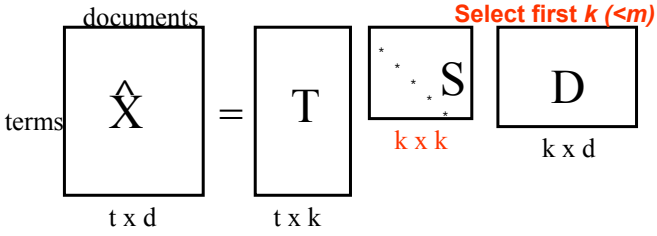
t: total number of index terms
d: total number of documents

Singular Value Decomposition

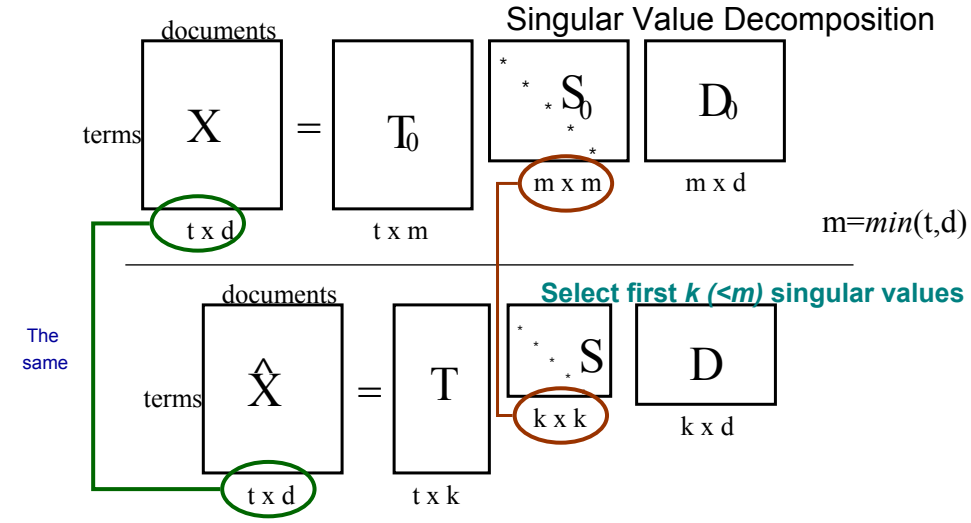


m=*min*(t,d)

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

$$\underset{T}{\begin{bmatrix}
0.22 & -0.11 \\
0.20 & -0.07 \\
0.24 & 0.04 \\
0.40 & 0.06 \\
0.64 & -0.17 \\
0.27 & 0.11 \\
0.27 & 0.11 \\
0.30 & -0.14 \\
0.21 & 0.27 \\
0.01 & 0.49 \\
0.04 & 0.62 \\
0.03 & 0.45
\end{bmatrix}}
\underset{S}{\begin{bmatrix}
3.34 & \\
 & 2.54
\end{bmatrix}}
\underset{D'}{\begin{bmatrix}
0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\
-0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53
\end{bmatrix}}$$

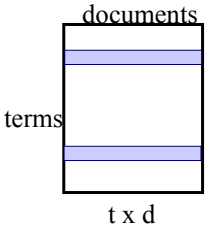
TS define coordinates for documents in latent space

Παρατηρήσεις

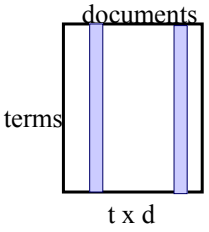
- Η παράμετρος k (<m) πρέπει να είναι:
  - large enough to allow fitting the characteristics of the data
  - small enough to filter out the non-relevant representational details

Τρόπος Σύγκρισης Όρων και Εγγράφων

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

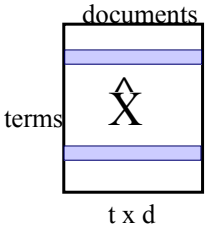


- Τρόπος σύγκρισης δύο εγγράφων:
  - **dot product** (or cosine) between two **column vectors**

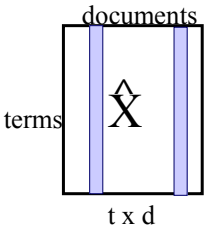


Τρόπος Σύγκρισης Όρων και Εγγράφων $\hat{X}$

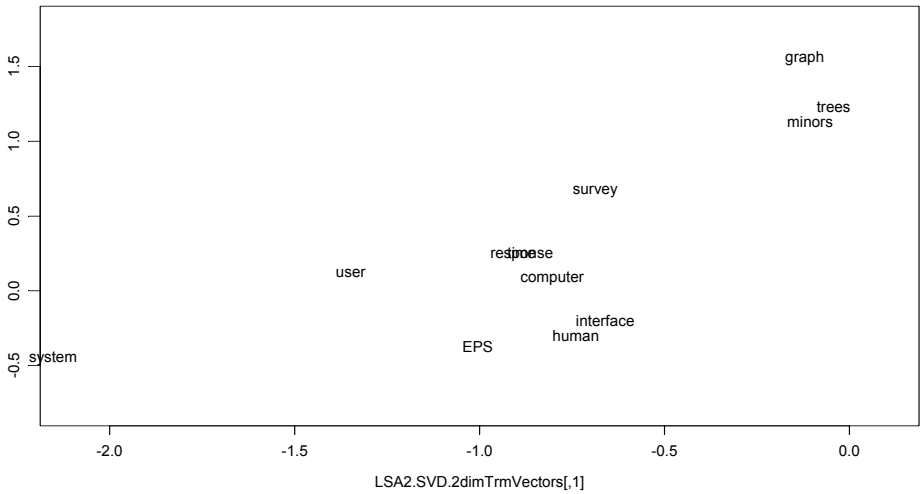
- Τρόπος σύγκρισης 2 όρων:
  - the **dot product** (or cosine) between two **row vectors** reflects the extent to which two terms have a similar pattern of occurrence across the set of document.



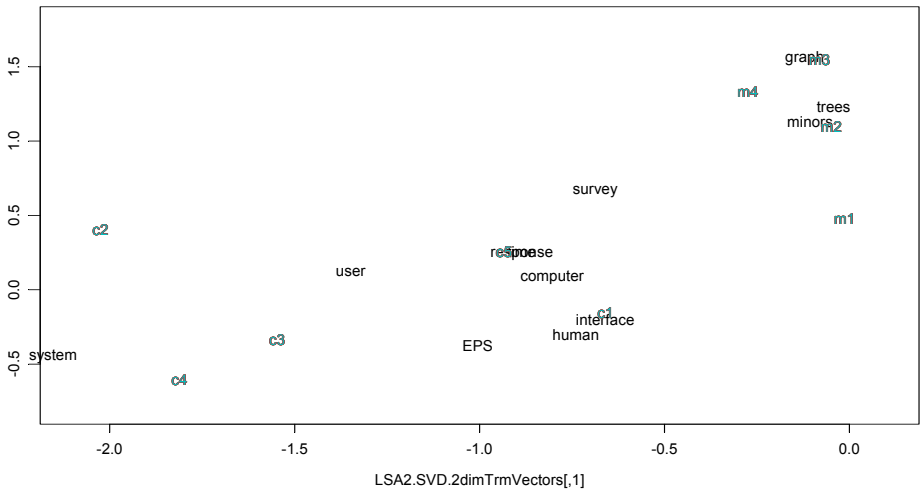
- Τρόπος σύγκρισης δύο εγγράφων:
  - **dot product** (or cosine) between two **column vectors**



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

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

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^

Το υπολογιστικό κόστος για το SVD πολύ μεγάλο

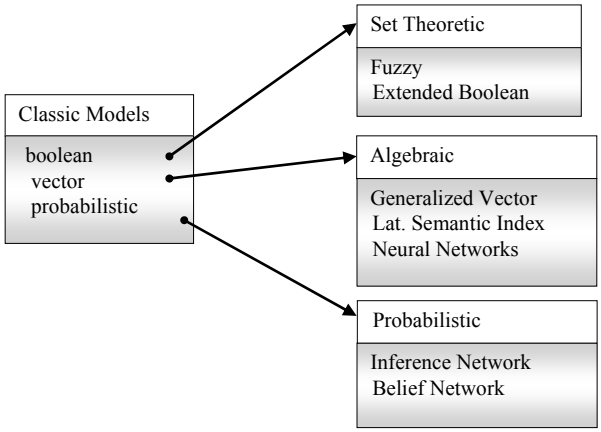
Δουλεύει καλύτερα σε εφαρμογές που υπάρχει μικρή επικάλυψη μεταξύ των ερωτημάτων και των εγγράφων

Μικρές τιμές του k (εκατοντάδες)

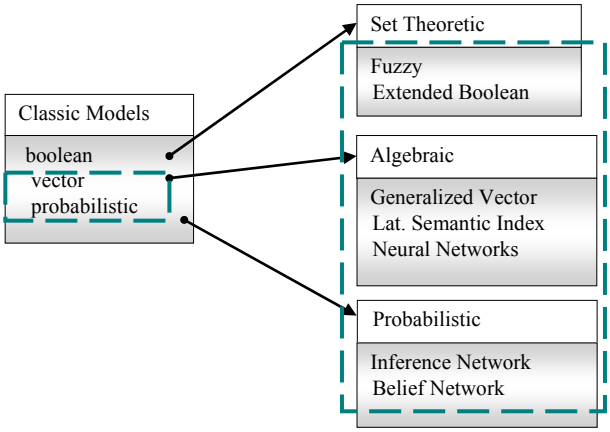
Δεν υπάρχει τρόπος να εκφραστεί απουσία όρου και exact match

Επισκόπηση των Μοντέλων Ανάκτησης που έχουμε εξετάσει μέχρι τώρα

Ταξινομία Μοντέλων που εξετάσαμε

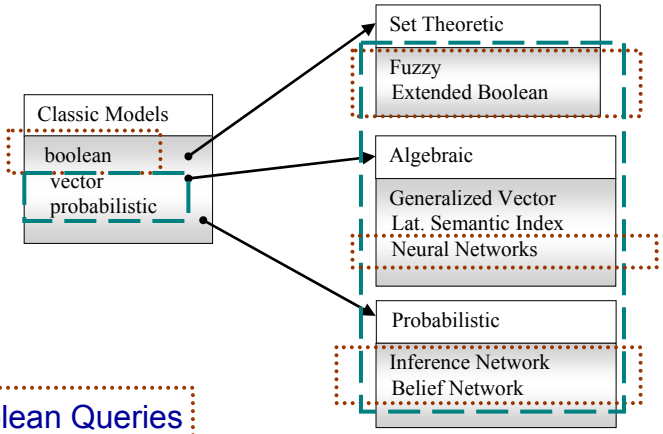


Ταξινομία Μοντέλων που εξετάσαμε



Partial Matching

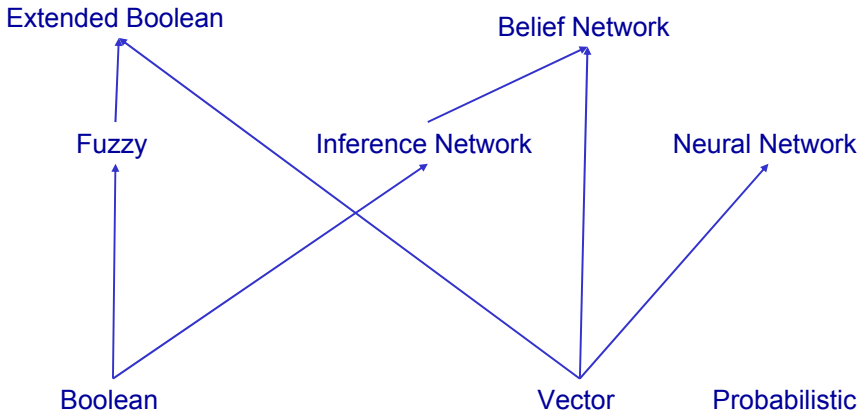
Ταξινομία Μοντέλων που εξετάσαμε



Boolean Queries

Partial Matching

Βάσει της εκφραστικής τους ικανότητας (incomplete)



Άλλοι τύποι Μοντέλων Ανάκτησης που ενδεχομένως να προλάβουμε να δούμε αργότερα

- Μοντέλα Ανάκτησης Πληροφοριών από **Ιστοσελίδες**
  - Έμφαση στους συνδέσμους
- Μοντέλα Ανάκτησης **Πολυμέσων**
- Μοντέλα Ανάκτησης **Δομημένων** Εγγράφων (π.χ. XML)
- Μοντέλα Βασισμένα στη **Λογική**

Θα δούμε τα «κόκκινα» αργότερα στο μάθημα