



## HY463 - Συστήματα Ανάκτησης Πληροφοριών Information Retrieval (IR) Systems

### Εξατομίκευση: Προφίλ Χρηστών και Συνεργατική Επιλογή/Διήθηση (Personalization: User Profiles and Collaborative Selection/Filtering)



Γιάννης Τζιτζίκας  
Διάλεξη : 18  
Ημερομηνία : 5-7-2006



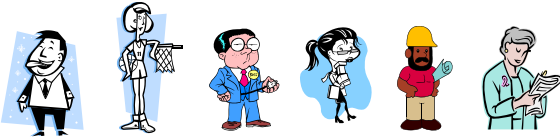
## Διάρθρωση Παρουσίασης

- Κίνητρο
- Προφίλ Χρηστών
  - μετα-διήθηση (Post-Filters)
  - προ-διήθηση (Pre-Filters)
  - Ξεχωριστά σημεία αναφοράς
- Συνεργατική Επιλογή/Διήθηση (Collaborative Selection/Filtering)



## Κίνητρο

- Διαπιστώσεις
  - Δεν έχουν όλοι οι χρήστες τα ίδια χαρακτηριστικά
  - Άρα δεν έχουν ούτε τις ίδιες πληροφοριακές ανάγκες
- Σκοπός: Προσαρμογή της λειτουργικότητας στα χαρακτηριστικά και τις ανάγκες διαφορετικών χρηστών



## Παραδείγματα Κριτηρίων Διάκρισης Χρηστών

- Εξοικείωση με την περιοχή της επερώτησης
  - Χρήστης με ΔΔ στην Πληροφορική ψάχνει για ιατρικές πληροφορίες
  - $q = \text{"theory of groups"}$ 
    - sociologist: behaviour of a set of people
    - mathematician: a particular type of algebraic structure
- Γλωσσικές Ικανότητες
  - Σελίδες στα γαλλικά (ok για δρομολόγια πλοίων, διασκέδαση, όχι όμως για φιλοσοφικά κείμενα), σελίδες με γιαπωνέζικα (τίποτα)
- Συγκεκριμένες προτιμήσεις
  - εγγραφή σε περιοδικό
  - παρακολούθηση δουλειάς συγκεκριμένων συγγραφέων (π.χ. Salton)
- Μορφωτικό επίπεδο
  - Χρήστης με Παν/κό Πτυχίο vs Χρήστης με Γνώσεις Δημοτικού



## Προφίλ Χρηστών

- Προφίλ Χρηστών:
  - μέσο διάκρισης των χρηστών βάσει των χαρακτηριστικών και προτιμήσεών τους
- Μορφή
  - Δεν υπάρχει κάποια τυποποιημένη μορφή
  - Μπορούμε να θεωρήσουμε ότι έχει τη μορφή μιας επερώτησης

### Προφίλ Χρηστών και Ηθική

- (α) Είναι «ορθό» να περιορίζουμε τα αποτελέσματα;  
(β) Ιδιωτικότητα και προστασία προσωπικών δεδομένων (Privacy)
- Αν έχουμε πολύ λεπτομερή προφίλ
    - Ποιος έχει δικαίωμα να βλέπει τα προφίλ;
    - Ποιος μπορεί να ελέγχει και να αλλάξει τα προφίλ;



## Γενικοί Τρόποι Αξιοποίησης των Προφίλ κατά την Ανάκτηση Πληροφοριών

- A) Μετα-διήθηση βάσει προφίλ (User Profile as a post-filter)
  - Εδώ το προφίλ χρησιμοποιείται **κατόπιν** της αποτίμησης της αρχικής επερώτησης
  - Η χρήση προφίλ αυξάνει το υπολογιστικό κόστος της ανάκτησης
- B) Προ-διήθηση βάσει προφίλ (User Profile as a pre-filter)
  - Εδώ το προφίλ χρησιμοποιείται για να **τροποποιήσει** την αρχική επερώτηση του χρήστη
  - Η χρήση προφίλ και η τροποποίηση επερωτήσεων δεν αυξάνει κατά ανάγκη το υπολογιστικό κόστος της ανάκτησης
- C) Επερώτηση και Προφίλ ως ξεχωριστά σημεία αναφοράς
  - (Query and Profile as Separate Reference Points)



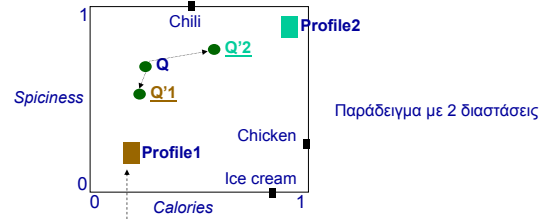
## (A) Μετα-διήθηση βάσει Προφίλ (User Profile as a Post-filter)

### • Μέθοδος:

- Η αρχική επερώτηση υπολογίζεται κανονικά
- Τα αποτελέσματα οργανώνονται βάσει του προφίλ
  - Αναδιάταξη στοιχείων απάντησης
  - Αποκλεισμός ορισμένων εγγράφων
- Υπολογιστικό κόστος
  - Η χρήση προφίλ δεν μειώνει το υπολογιστικό κόστος
  - Αντίθετα, εισαγάγει ένα παραπάνω υπολογιστικό στάδιο



## Β) Προ-διήθηση βάσει Προφίλ (User Profile as a Pre-filter) Παράδειγμα Τροποποίησης Επερωτήσεων:



Παράδειγμα με 2 διαστάσεις  
Προφίλ χρήστη που προτιμάει ελαφριά και όχι πικάντικα φαγητά



## Τεχνικές τροποποίησης επερωτήσεων

### (B.1) Simple Linear Transformation

- Μετακινεί το διάνυσμα προς την κατεύθυνση του προφίλ

### (B.2) Piecewise Linear Transformation

- Μετακινεί το διάνυσμα προς την κατεύθυνση του προφίλ βάσει περιπτώσεων



## (B.1) Simple Linear Transformation (απλός γραμμικός μετασχηματισμός)

Έστω  $q = \langle q_1, \dots, q_t \rangle$ ,  $p = \langle p_1, \dots, p_t \rangle$  ( $q_i$ ,  $p_i$  τα βάρη των διανυσμάτων)

Τροποποίηση επερώτησης  $q$  (και ορισμός της  $q'$ ):

$$q'_i = k p_i + (1-k) q_i \text{ για ένα } 0 \leq k \leq 1$$

Περιπτώσεις

- Αν  $k=0$  τότε  $q' = q$  (η επερώτηση μένει αναλλοίωτη)
- Αν  $k=1$  τότε  $q' = p$  (η νέα επερώτηση ταυτίζεται με το προφίλ)
- Οι **ενδιάμεσες** τιμές του  $k$  είναι ενδιαφέρουσες



## (B.2) Piecewise Linear Transformation

- Εδώ η τροποποίηση των βαρών προσδιορίζεται με ένα σύνολο περιπτώσεων
- Περιπτώσεις:
  - (1) όρος που εμφανίζεται **και** στην επερώτηση **και** στο προφίλ
    - εφαρμόζουμε τον απλό γραμμικό μετασχηματισμό
  - (2) όρος που εμφανίζεται μόνο στην επερώτηση
    - αφήνουμε το βάρος του όρου αμετάβλητο ή το μειώνουμε ελαφρά (πχ 5%)
  - (3) όρος που εμφανίζεται μόνο στο προφίλ
    - δεν κάνουμε τίποτα, ή εισαγάγουμε τον όρο στην επερώτηση αλλά με μικρό βάρος
  - (4) όρος που δεν εμφανίζεται ούτε στην επερώτηση ούτε στο προφίλ
    - δεν κάνουμε τίποτα
- Παράδειγμα
  - $p = \langle 5, 0, 0, 3 \rangle$
  - $q = \langle 0, 2, 0, 7 \rangle$
  - $q' = \langle 1.25, 1.5, 0, 6 \rangle$



## (C) Επερώτηση και Προφίλ ως ξεχωριστά σημεία αναφοράς (Query and Profile as Separate Reference Points)

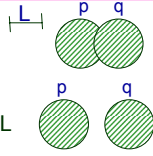
- Προσέγγιση
  - Εδώ **δεν τροποποιείται** η αρχική επερώτηση
  - Αντίθετα και η επερώτηση και το προφίλ λαμβάνονται **ξεχωριστά** υπόψη κατά τη διαδικασία της βαθμολόγησης των εγγράφων
- Ερωτήματα
  - Πώς να συνδυάσουμε αυτά τα δυο;
  - Σε ποιο να δώσουμε περισσότερο βάρος και πως;
- Υπόθεση εργασίας
  - Έστω ότι η ανάκτηση γίνεται βάσει μιας **συνάρτησης απόστασης** Dist



## Τρόποι συνδυασμού προφίλ και επερωτήσης

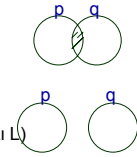
### (1) Το διαζευκτικό μοντέλο (το λιγότερο αυστηρό)

- Ένα  $d$  ανήκει στην απάντηση αν:
  - $(\text{Dist}(d,q) \leq L)$  **OR**  $(\text{Dist}(d,p) \leq L)$
  - Εναλλακτική διατύπωση:  $\min(\text{Dist}(d,q), \text{Dist}(d,p)) \leq L$
  - είναι το λιγότερο αυστηρό



### (2) Το συζευκτικό μοντέλο (το αυστηρότερο)

- $(\text{Dist}(d,q) \leq L)$  **AND**  $(\text{Dist}(d,p) \leq L)$
- $\max(\text{Dist}(d,q), \text{Dist}(d,p)) \leq L$
- είναι το πιο αυστηρό
- η απάντηση είναι η τομή των  $\text{ans}(p)$  και  $\text{ans}(q)$  (με κατώφλι  $L$ )
  - αν το  $q$  απέχει πολύ από το  $p$ , τότε η απάντηση θα είναι κενή



## Τρόποι συνδυασμού προφίλ και επερωτήσης (II)

### (3) Το ελλειψοειδές μοντέλο

- $\text{Dist}(d,q) + \text{Dist}(d,p) \leq L$
- καλό αν το  $d$  και το  $p$  δεν απέχουν πολύ
  - αν απέχουν πολύ τότε μπορεί να ανακτηθούν πολλά μη συναφή με κανένα



## Τρόποι συνδυασμού προφίλ και επερωτήσης (III)

### (4) Το οβάλ μοντέλο του Casini

- $\text{Dist}(d,q) * \text{Dist}(d,p) \leq L$
- αν το  $d$  και το  $p$  είναι κοντά, τότε μοιάζει με το ελλειψοειδές
- αν απέχουν λίγο τότε μοιάζει με φυστίκι
- αν απέχουν πολύ τότε έχει τη μορφή του 8



## Πώς μπορούμε καθορίσουμε τη σχετική βαρύτητα επερωτήσεων και προφίλ;

- **Βάρη** μπορούν να προστεθούν στα προηγούμενα μοντέλα:
  - $\min(w1 * \text{Dist}(d,q), w2 * \text{Dist}(d,p)) \leq L$  //διαζευκτικό
  - $\max(w1 * \text{Dist}(d,q), w2 * \text{Dist}(d,p)) \leq L$  //συζευκτικό
  - $w1 * \text{Dist}(d,q) + w2 * \text{Dist}(d,p) \leq L$  //ελλειψοειδές
- στο Cassini δεν είναι πολύ καλό αφού έχει τη μορφή γινομένου. Εδώ είναι προτιμότερο τα βάρη να εκφραστούν ως εκθέτες:
  - $\text{Dist}(d,q)^{w1} * \text{Dist}(d,p)^{w2} \leq L$  //Cassini



## Προφίλ Χρηστών και Αξιολόγηση Αποτελεσματικότητας Ανάκτησης

- Μόνο πειραματικά μπορούμε να αποφανθούμε για το ποια προσέγγιση είναι καλύτερη, ή για το αν αυτές οι τεχνικές βελτιώνουν την αποτελεσματικότητα της ανάκτησης
- Η πειραματική αξιολόγηση [Sung Myaeng] απέδειξε ότι οι τεχνικές αυτές βελτιώνουν την αποτελεσματικότητα



## Συστήματα Πολλαπλών Σημείων Αναφοράς (Multiple Reference Point Systems)

### Κίνητρο

- Δυνατότητα χρήσης **περισσότερων των 2 σημείων αναφοράς**
  - Στην προηγούμενη συζήτηση είχαμε δυο σημεία αναφοράς: την επερωτήση και το προφίλ.

### Ορισμός:

- **Σημείο Αναφοράς (reference point of point of interest) = Ένα ορισμένο σημείο ή έννοια ως προς την οποία μπορούμε να κρίνουμε ένα έγγραφο**

### Παραδείγματα σημείων αναφοράς:

- ένα γνωστό έγγραφο
- ένα σύνολο γνωστών εγγράφων
- ένας συγγραφέας ή ένα σύνολο συγγραφέων
- ένα γνωστό περιοδικό
- μια χρονική περίοδος

- Πώς μπορούμε να ορίσουμε ένα σημείο αναφοράς από ένα σύνολο εγγράφων  $C \subseteq D$ ;
- Απάντηση: Θεωρούμε ότι υπάρχει ένα **τεχνητό έγγραφο**, το centroid document
  - το βάρη του διανύσματος του προκύπτουν παίρνοντας τον μέσο όρο των βαριών των εγγράφων του  $C$



## Συστήματα Πολλαπλών Σημείων Αναφοράς (Multiple Reference Point Systems)

- Σημεία αναφοράς:  $R_1, \dots, R_n$
- Βάρη:  $w_1, \dots, w_n$ ,  $\sum w_i = 1$
- $\| \cdot \|$  μετρική (συνάρτηση απόστασης)
- Παρατηρήσεις
  - Τα παρακάτω είναι ανεξάρτητα της μετρικής που χρησιμοποιούμε
  - μπορούμε να χρησιμοποιήσουμε οποιαδήποτε μετρική απόστασης ή ομοιότητας επιθυμούμε
- Διαισθητικά: *Είναι σαν να κάνουμε Ανάκτηση Πληροφορίας χρησιμοποιώντας ΠΟΛΛΕΣ επερωτήσεις ταυτόχρονα*



## Multiple Reference Points: Mathematical Basis

- Θα γενικεύσουμε τα μοντέλα του δισδιάστατου χώρου που έχουμε ήδη δει:
  - $\min (w_1 * \text{Dist}(d,q), w_2 * \text{Dist}(d,p)) \leq L$  //διαζευκτικό
  - $\max (w_1 * \text{Dist}(d,q), w_2 * \text{Dist}(d,p)) \leq L$  //συζευκτικό
  - $w_1 * \text{Dist}(d,q) + w_2 * \text{Dist}(d,p) \leq L$  //ελλειψοειδές
  - $\text{Dist}(d,q)^{w_1} * \text{Dist}(d,p)^{w_2} \leq L$  //Cassini
- Δηλαδή:
  - $\min (w_1 * \text{Dist}(d,R_1), \dots, w_n * \text{Dist}(d,R_n)) \leq L$  //διαζευκτικό
  - $\max (w_1 * \text{Dist}(d, R_1), \dots, w_n * \text{Dist}(d, R_n)) \leq L$  //συζευκτικό
  - $w_1 * \text{Dist}(d, R_1) + \dots + w_n * \text{Dist}(d, R_n) \leq L$  //ελλειψοειδές
  - $\text{Dist}(d, R_1)^{w_1} * \dots * \text{Dist}(d, R_n)^{w_n} \leq L$  //Cassini
- ή συνδυασμός των παραπάνω



## Άλλες τεχνικές (που έχουμε ήδη δει) που βοηθούν την εξατομίκευση

- **Ομαδοποίηση (Clustering):** θυμηθείτε το μάθημα περί ομαδοποίησης και επιτόπιας ανάλυσης
  - Μπορεί να δώσει λύση στο παράδειγμα:
    - $q = \text{"theory of groups"}$ 
      - sociologist: behaviour of a set of people
      - mathematician: a particular type of algebraic structure
- Τεχνικές Βελτίωσης Απάντησης Επερωτήσεων (ανατροφοδότηση συνάφειας)



## Εξατομίκευση μέσω Συνεργατικής Επιλογής/Διήθησης Personalization using Collaborative Selection/Filtering



## Παράδειγμα

**Customers who bought this book also bought:**

- Reinforcement Learning: An Introduction; R. S. Sutton, A. G. Barto
- Advances in Knowledge Discovery and Data Mining; U. M. Fayyad
- Probabilistic Reasoning in Intelligent Systems; J. Pearl



## Product Rating by Users

Product rating



## Συνεργατική Επιλογή/Διήθηση

Πρόβλεψη προτιμήσεως ενός χρήστη βάσει των καταγεγραμμένων προτιμήσεων του ίδιου και άλλων χρηστών.



## Παράδειγμα: Επιλογή Εστιατορίου

- Κλασσική Προσέγγιση:
  - Χαρακτηρίζουμε τα εστιατόρια βάσει ενός πεπερασμένου συνόλου κριτηρίων (κουζίνα, κόστος, τοποθεσία). Οι προτιμήσεις ενός χρήστη εκφράζονται με μια συνάρτηση αξιολόγησης πάνω σε αυτά τα κριτήρια.
- Μειονεκτήματα
  - Στην επιλογή όμως ενός εστιατορίου εμπλέκονται και άλλοι παράγοντες (απεριόριστοι στον αριθμό) που δύσκολα θα μπορούσαν να εκφραστούν με σαφήνεια, όπως:
    - το στυλ και η ατμόσφαιρα, η διακόσμηση
    - η υπόλοιπη πελατεία, το πάρκινγκ
    - η γειτονιά, η διαδρομή προς το εστιατόριο
    - η εξυπηρέτηση, οι ώρες λειτουργίας, τα ... σερβίτσια
- Θα θέλαμε να μπορούμε να προβλέψουμε τις προτιμήσεις χωρίς να περιοριζόμαστε σε ένα σταθερό σύνολο κριτηρίων
  - χωρίς καν να χρειαστεί να αναλύσουμε τον τρόπο που σκέφτεται ο χρήστης



## Η Κλασσική Ανάκτηση Κειμένων

Ομοιότητα όρων  
βάσει των εγγράφων

$$sim(k1, k2)$$

Όροι

	$k_1$	$k_2$	...	$k_i$
$d_1$	$w_{11}$	$w_{21}$	...	$w_{i1}$
$d_2$	$w_{12}$	$w_{22}$	...	$w_{i2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d_n$	$w_{1n}$	$w_{2n}$	...	$w_{in}$
$q$	$w_{1q}$	$w_{2q}$	...	$w_{iq}$

$sim(d1, d2)$

- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων  
βάσει των λέξεων

$$w_{ij} = \{0,1\}$$

$$w_{ij} = tf_{ij} idf_i$$



## Χρήστες αντί Όρων

Ομοιότητα χρηστών  
βάσει των προτιμήσεων τους

$$sim(u1, u2)$$

Χρήστες

	$u_1$	$u_2$	...	$u_i$
$d_1$	$w_{11}$	$w_{21}$	...	$w_{i1}$
$d_2$	$w_{12}$	$w_{22}$	...	$w_{i2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d_n$	$w_{1n}$	$w_{2n}$	...	$w_{in}$
$q$	$w_{1q}$	$w_{2q}$	...	$w_{iq}$

$sim(d1, d2)$

- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων  
βάσει των (προτιμήσεων)  
των χρηστών

$$w_{ij} = \{0,1\} \implies 0: \text{Bad}, 1: \text{Good}$$

$$w_{ij} = tf_{ij} idf_i \implies w_{ij}: \text{βαθμός προτίμησης του χρήστη } i \text{ στο έγγραφο } j, \text{ } \pi \chi \{1,2,3,4,5\}$$



## Χρήστες αντί Όρων

Ομοιότητα χρηστών  
βάσει των προτιμήσεων τους

$$sim(u1, u2)$$

Χρήστες

	$u_1$	$u_2$	...	$u_i$
$d_1$	$w_{11}$	$w_{21}$	...	$w_{i1}$
$d_2$	$w_{12}$	$w_{22}$	...	$w_{i2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d_n$	$w_{1n}$	$w_{2n}$	...	$w_{in}$
$q$	$w_{1q}$	$w_{2q}$	...	$w_{iq}$

$sim(d1, d2)$

- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων  
βάσει των (προτιμήσεων)  
των χρηστών

- Αφού δεν χρησιμοποιούμε λέξεις, τα «έγγραφα» μπορεί να είναι οτιδήποτε:
  - Φωτογραφίες, Βιβλία
  - Ηλεκτρικές Συσκευές
  - Εστιατόρια, Μεζεδοπωλεία
  - Κινηματογραφικές ταινίες
  - Τηλεοπτικά Προγράμματα
  - ...

$$w_{ij} = \{0,1\} \implies 0: \text{Bad}, 1: \text{Good}$$

$$w_{ij} = tf_{ij} idf_i \implies w_{ij}: \text{βαθμός προτίμησης του χρήστη } i \text{ στο έγγραφο } j, \text{ } \pi \chi \{1,2,3,4,5\}$$



## Μαντεύοντας τις προτιμήσεις ενός χρήστη

Ομοιότητα χρηστών  
βάσει των προτιμήσεων τους

$$sim(u1, u2)$$

Χρήστες

	$u_1$	$u_2$	...	$u_i$
$d_1$	$w_{11}$	$w_{21}$	...	$w_{i1}$
$d_2$	$w_{12}$	$w_{22}$	...	$w_{i2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d_n$	$w_{1n}$	$w_{2n}$	...	$w_{in}$
$q$	$w_{1q}$	$w_{2q}$	...	$w_{iq}$

$sim(d1, d2)$

- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων  
βάσει των (προτιμήσεων)  
των χρηστών

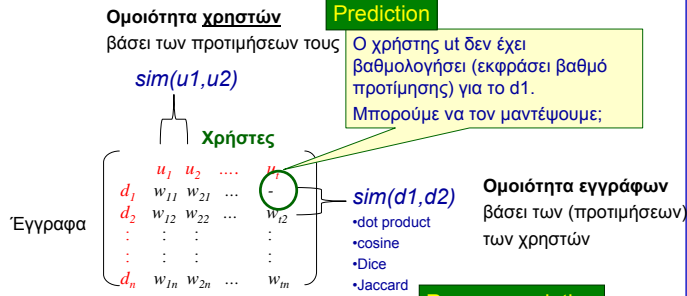
Ο χρήστης  $u_i$  δεν έχει βαθμολογήσει (εκφράσει βαθμό προτίμησης) για το  $d_1$ . Μπορούμε να τον μαντέψουμε;

$$w_{ij} = \{0,1\} \implies 0: \text{Bad}, 1: \text{Good}$$

$$w_{ij} = tf_{ij} idf_i \implies w_{ij}: \text{βαθμός προτίμησης του χρήστη } i \text{ στο έγγραφο } j, \text{ } \pi \chi \{0,1,2,3,4,5\}$$



## Υπολογισμός Προβλέψεων και Συστάσεων



Computing recommendations for a user  $u$ :

- 1/ Predict values for those cells of  $u$  that are empty, and
- 2/ Select (and give the user) the highest ranked elements

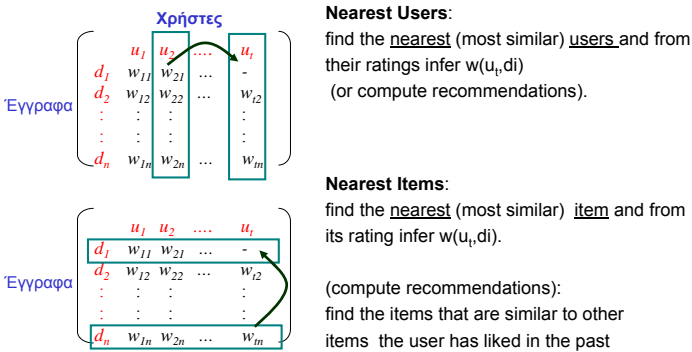


## Παράδειγμα της διαφοράς μεταξύ Πρόβλεψης και Σύστασης

- Prediction
  - e.g.: ET3 channel has tonight the movie "MATRIX", would I like it?
- Recommendation
  - e.g. recommend me what movies to rent from a Video Club



## How can we compute recommendations?



## How we can compute recommendations. Nearest Users

Objective: Compute  $w(u_t, d_i)$

- Algorithm Average
  - Let  $\text{Sim}(u_t)$  = the users that are similar to  $u_t$ 
    - E.g. k-nearest neighbours
  - $w(u_t, d_i) = \text{average}(\{w(u, d_i) \mid u \in \text{Sim}(u_t)\})$
- Algorithm Weighted Average
  - As some close neighbors are closer than others, we can assign higher weights to ratings of closer neighbors
  - $w(u_t, d_i) = \sum \text{sim}(u_t, u) * w(u, d_i)$  where  $u \in \text{Sim}(u_t)$



## Παράδειγμα πρόβλεψης βάσει των 3 κοντινότερων **χρηστών** και μέτρο απόστασης τη μετρική $L_2$

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	4
PizzaNapoli	3	3	1	1	4	3
PizzaHut	1	2	5	4	1	2
PizzaToscana	5	4	2	1	5	?

$D(\text{Tony}, \text{Yannis}) = \text{sqrt} [(4-4)^2 + (3-3)^2 + (1-2)^2] = 1$   
 $D(\text{Manos}, \text{Yannis}) = \text{sqrt} [(5-4)^2 + (3-3)^2 + (2-2)^2] = 1$   
 $D(\text{Tom}, \text{Yannis}) = \text{sqrt} [(1-4)^2 + (1-3)^2 + (5-2)^2] = 4.69$   
 $D(\text{Nick}, \text{Yannis}) = \text{sqrt} [(2-4)^2 + (1-3)^2 + (4-2)^2] = 3.46$   
 $D(\text{Titos}, \text{Yannis}) = \text{sqrt} [(5-4)^2 + (4-3)^2 + (1-2)^2] = 1.73$

Nearest 3 = Tony, Manos, Titos

$(5+4+5)/3 = 4.66$



## Παράδειγμα πρόβλεψης με βάση τις 2 κοντινότερες **πιτσαρίες** και μέτρο απόστασης τη μετρική $L_2$

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	4
PizzaNapoli	3	3	1	1	4	3
PizzaHut	1	2	5	4	1	2
PizzaToscana	5	4	2	1	5	?

$D(\text{Roma}, \text{Toscana}) = \text{sqrt} [(4-5)^2 + (5-4)^2 + (1-2)^2 + (2-1)^2 + (5-5)^2] = 2$   
 $D(\text{Napoli}, \text{Toscana}) = \text{sqrt} [(3-5)^2 + (3-4)^2 + (1-2)^2 + (1-1)^2 + (4-5)^2] = 2.65$   
 $D(\text{Hut}, \text{Toscana}) = \text{sqrt} [(1-5)^2 + (2-4)^2 + (5-2)^2 + (4-1)^2 + (1-5)^2] = 7.34$

Nearest 2 = Roma, Napoli

$(4+3)/2 = 3.5$



## Προβλήματα Εκκίνησης (I) Nearest Users

- Εισαγωγή νέου χρήστη:
  - Δεν έχει εκφράσει καμιά προτίμηση => δεν μπορούμε να του προτείνουμε τίποτα (δεν μπορούμε να εντοπίσουμε κοντινούς χρήστες)

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	-
PizzaNapoli	3	3	1	1	4	-
PizzaHut	1	2	5	4	1	-
PizzaToscana	5	4	2	1	5	?



## Προβλήματα Εκκίνησης (II) Nearest Items

- Εισαγωγή νέου αντικειμένου (new item):
  - Δεν έχουμε προτιμήσεις για αυτό => ποτέ δεν θα προταθεί σε κάποιον χρήστη

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	4
PizzaNapoli	3	3	1	1	4	3
PizzaHut	1	2	5	4	1	2
PizzaToscana	-	-	-	-	-	?



## Προβλήματα Εκκίνησης (III)

- Σε κάθε περίπτωση ποτέ δεν θα προταθεί ένα νέο στοιχείο σε ένα νέο χρήστη

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	-
PizzaNapoli	3	3	1	1	4	-
PizzaHut	1	2	5	4	1	-
PizzaToscana	-	-	-	-	-	?



## Ομοιότητα/Απόσταση Χρηστών

- Τρόποι υπολογισμού:

- εσωτερικό γινόμενο

$$sim(u1, u2) = \sum_{i=1}^t w_{1i} \cdot w_{2i}$$

Στα άδεια κελιά του πίνακα θεωρούμε ότι υπάρχει το 0

- συνημίτονο

$$\cos(\vec{u}_1, \vec{u}_2) = \frac{\vec{u}_1 \cdot \vec{u}_2}{|\vec{u}_1| \cdot |\vec{u}_2|} = \frac{\sum_{i=1}^t (w_{1i} \cdot w_{2i})}{\sqrt{\sum_{i=1}^t w_{1i}^2} \cdot \sqrt{\sum_{i=1}^t w_{2i}^2}}$$

- Mean Squared Distance

- Pearson Correlation Coefficient

- ...



## Ομοιότητα/Απόσταση Χρηστών

- Problem: Not every User rates every Item
- A solution: Determine similarity of customers  $u1$  and  $u2$  based on the similarity of ratings of those items that **both have rated**, i.e.,  $D_{u1 \cap u2}$ .

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma		5		2		
PizzaNapoli		3	1		4	3
PizzaHut	1		5			2
PizzaToscana	5		2	1	5	



## Ομοιότητα/Απόσταση Χρηστών: Mean Squared Difference

$$u1(x) \equiv w_{1x}$$

$$u2(x) \equiv w_{2x}$$

$$d_{MSD}(u1, u2) = \frac{1}{|D_{u1 \cap u2}|} \cdot \sum_{x \in D_{u1 \cap u2}} (u1(x) - u2(x))^2$$





## Ομοιότητα/Απόσταση Χρηστών: Pearson correlation

$$C_{Pearson}(u1, u2) = \frac{\sum_{x \in D_{u1 \cap u2}} (u1(x) - \bar{u1})(u2(x) - \bar{u2})}{\sqrt{\sum_{x \in D_{u1 \cap u2}} (u1(x) - \bar{u1})^2 \cdot \sum_{x \in D_{u1 \cap u2}} (u2(x) - \bar{u2})^2}}$$

$\bar{u1}$  = mean of u1  
 $\bar{u2}$  = mean of u2

$C(u1, u2) > 0$  θετική σχέση  
 $C(u1, u2) = 0$  ουδέτερη σχέση  
 $C(u1, u2) < 0$  αρνητική σχέση

The correlation coefficient measures the strength of a linear relationship between two variables. The correlation coefficient is always between -1 and +1. The closer the correlation is to +/-1, the closer to a perfect linear relationship. Here is an example of interpretation:  
-1.0 to -0.7 strong negative association.  
-0.7 to -0.3 weak negative association.  
-0.3 to +0.3 little or no association.  
+0.3 to +0.7 weak positive association.  
+0.7 to +1.0 strong positive association.



## Ομοιότητα/Απόσταση Items

- Τρόποι υπολογισμού ομοιότητας/απόστασης:
  - εσωτερικό γινόμενο
  - συνημίτονο
  - Pearson Correlation Coefficient

$$C_{Pearson}(x1, x2) = \frac{\sum_{u \in U} (u(x1) - \bar{x1})(u(x2) - \bar{x2})}{\sqrt{\sum_{u \in U} (u(x1) - \bar{x1})^2 \cdot \sum_{u \in U} (u(x2) - \bar{x2})^2}}$$

- Adjusted Pearson Correlation Coefficient

To handle the differences in rating scales of the users

$$C_{Pearson}(x1, x2) = \frac{\sum_{u \in U} (u(x1) - \bar{u1})(u(x2) - \bar{u2})}{\sqrt{\sum_{u \in U} (u(x1) - \bar{u1})^2 \cdot \sum_{u \in U} (u(x2) - \bar{u2})^2}}$$



## Ομοιότητα/Απόσταση Items

- Τρόποι υπολογισμού ομοιότητας/απόστασης:
  - εσωτερικό γινόμενο
  - συνημίτονο
  - Pearson Correlation Coefficient

$$C_{Pearson}(x1, x2) = \frac{\sum_{u \in U} (u(x1) - \bar{x1})(u(x2) - \bar{x2})}{\sqrt{\sum_{u \in U} (u(x1) - \bar{x1})^2 \cdot \sum_{u \in U} (u(x2) - \bar{x2})^2}}$$

- Adjusted Pearson Correlation Coefficient

To handle the differences in rating scales of the users

$$C_{Pearson}(x1, x2) = \frac{\sum_{u \in U} (u(x1) - \bar{u1})(u(x2) - \bar{u2})}{\sqrt{\sum_{u \in U} (u(x1) - \bar{u1})^2 \cdot \sum_{u \in U} (u(x2) - \bar{u2})^2}}$$



## Obtaining User Input

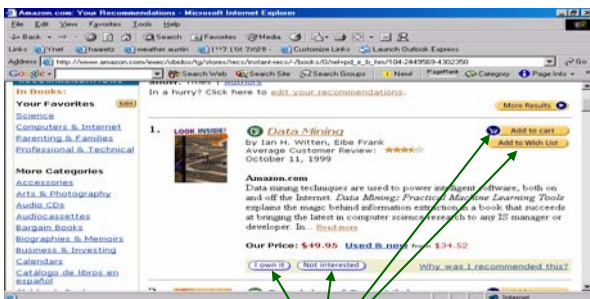
User (consumer) input is **difficult to get**

A solution:

- identify preferences that are **implicit** in *people's actions*
  - Purchase records
    - For example, people who order a book implicitly express their preference for that book (over other books)
  - Timing logs
- Works quite well (but results are not as good as with the use of rating)



## Obtaining User Input: An Example of **Implicit Rating**



**Implicit rating**



## Παρά ταύτα,

Πολύ συχνά  $|D_{u1 \cap u2}| = 0$

When thousands of items available only little overlap!  
=> Recommendations based on only a few observations

	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma		5		2		
PizzaNapoli		3	1		4	3
PizzaHut	1		5			2
PizzaToscana	5		2	1	5	

- Various solutions:

- View CF as a classification task
  - build a classifier for each user
  - employ training examples
- Reduce Dimensions
  - e.g. LSI (Latent Semantic Indexing)

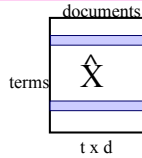




## LSI:

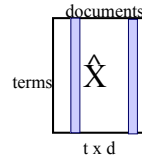
### • Τρόπος σύγκρισης 2 όρων:

- the **dot product** between two **row vectors** of  $\bar{X}$  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  $\bar{X}$



## Performance Issues

- Depends on  $|U|$  vs.  $|D|$  and their “stability”
- Typical setting
  - D stable (e.g. 5.000 movies)
  - U dynamic and  $|U| \gg |D|$  (e.g. 100.000 users)
  - A fast Item-based approach
    - **Precompute similarities** of items:
      - Requires  $O(|D|^2)$  space (very big)
      - One solution: Store only the k-rearest items of an item (this is what we need for computing recommendations)



## Evaluation Metrics

A method to evaluate a method for collaborative selection/filtering is the following:

### • Data is divided into 2 sets

- training set
- test set

### • Evaluation Metrics

- Then we compare the results of the techniques on the test set using the Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N}$$

$p_i$  : predicted rating

$q_i$  : actual rating



## Συναφή Ζητήματα που έχουμε ήδη μελετήσει

### • Ενοποίηση Διατάξεων

- Borda, Condorcet, Arrow's Impossibility Theorem
- Αν οι προτιμήσεις των χρηστών είναι ένα διατεταγμένο σύνολο επιλογών
  - Υπολογισμός συστάσεων = εύρεση ενοποιημένης διάταξης

### • Γρήγορη αποτίμηση top-k queries

- Αλγόριθμος FA (Fagin's Algorithm) και TA (Threshold Algorithm). Αν οι προτιμήσεις των χρηστών εκφράζονται με σκορ και είναι αποθηκευμένες σε άπομακρυσμένα συστήματα.



## Συνεργατική Επιλογή/Διήθηση: Σύνοψη

- **Ιδιαίτερο χαρακτηριστικό:** δεν χρειάζεται να έχουμε περιγραφή του περιεχομένου των στοιχείων
  - μπορούμε να την χρησιμοποιήσουμε για την επιλογή/διήθηση ποιημάτων, φιλοσοφικών ιδεών, μπζ, μεζεδοπωλείων, ...
- Θα μπορούσε να αξιοποιηθεί και στα πλαίσια της κλασσικής ΑΠ
  - Διάταξη στοιχείων απάντησης βάσει συνάφειας ΚΑΙ του εκτιμούμενου βαθμού τους (βάσει των αξιολογήσεων των άλλων χρηστών)
- Έχει αποδειχθεί χρήσιμη και για τους αγοραστές και για τους πωλητές (e-commerce)
- **Αδυναμίες: Sparseness & Cold Start**
  - Works well only once a "critical mass" of preference has been obtained
  - Need a very large number of consumers to express their preferences about a relatively large number of products.
  - Users' profiles don't overlap -> similarity not computable
  - Doesn't help the community forming
  - Difficult or impossible for users to control the recommendation process
- **Επεκτάσεις/Βελτιώσεις**
  - **Trust** = explicit rating of user on user



## Διάρθρωση Παρουσίασης

- Motivation
- User Profiles
  - as Post-Filters
  - as Pre-Filters (query modification)
    - Linear and Piecewise Transformations
  - as Separate Reference Points
- Collaborative Selection/Filtering

