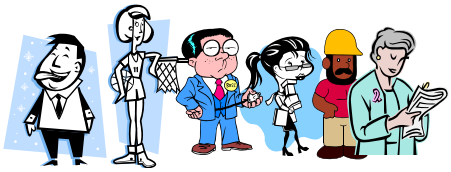
HY463 - Συστήματα Ανάκτησης Πληροφοριών
Information Retrieval (IR) Systems

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



Γιάννης Τζίτζικας

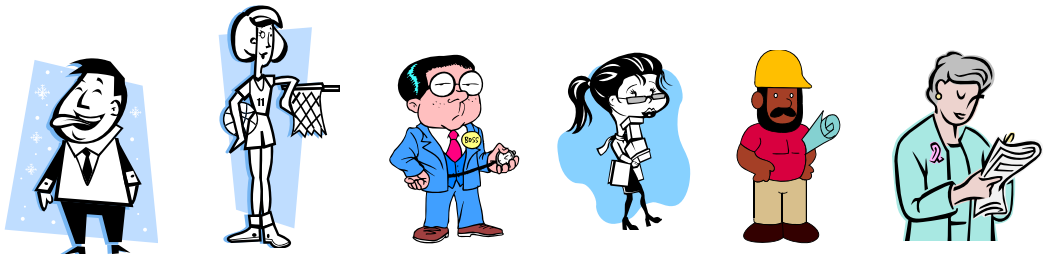
Ενότητα : 16
Ημερομηνία :

Διάρθρωση Παρουσίασης

- Κίνητρο
- Προφίλ Χρηστών
  - μετα-διήθηση (Post-Filters)
  - προ-διήθηση (Pre-Filters)
  - Πολλαπλά 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="theory of groups"
    - sociologist: behaviour of a set of people
    - mathematician: a particular type of algebraic structure
- Γλωσσικές Ικανότητες
  - Ιστοσελίδες στη γαλλική γλώσσα (οκ για εύρεση δρομολογίων πλοίων, όχι όμως για φιλοσοφικά κείμενα), σελίδες στην ιαπωνική (τίποτα)
- Συγκεκριμένες προτιμήσεις
  - εγγραφή σε περιοδικό
  - παρακολούθηση δουλειάς συγκεκριμένων συγγραφέων (π.χ. Salton)
- Μορφωτικό επίπεδο
  - Χρήστης με Παν/κό Πτυχίο έναντι Χρήστη 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 p

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 αλλάζει τα προφίλ;

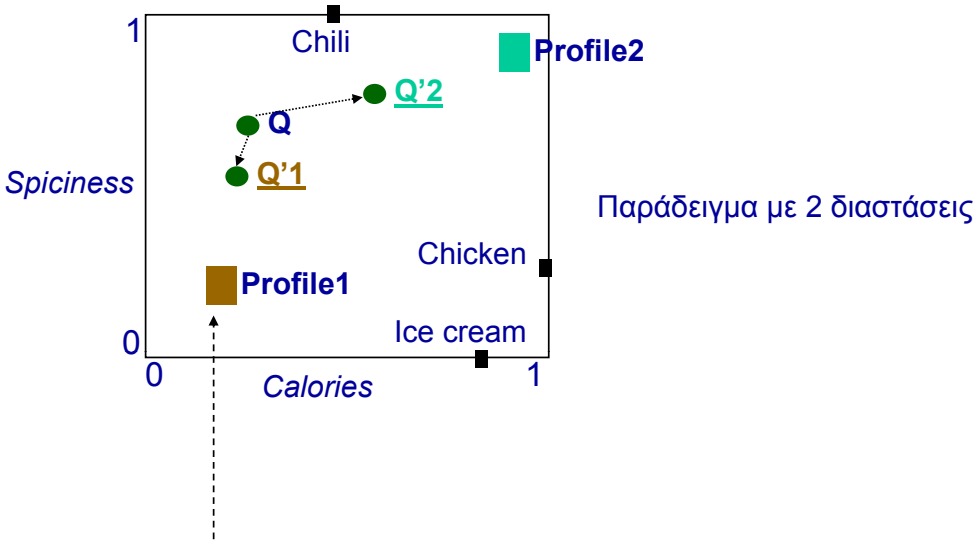
Γενικοί Τρόποι Αξιοποίησης των Προφίλ κατά την Ανάκτηση Πληροφοριών

- Α) **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
- Β) **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 προφιλ
    - Αναδιάταξη στοιχείων απάντησης
    - Αποκλεισμός ορισμένων εγγράφων
- Υπολογιστικό κόστος
  - Η χρήση προφίλ δεν μειώνει το υπολογιστικό κόστος
  - Αντίθετα, προσθέτει 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 ότι το προφίλ είναι ένα διάνυσμα p και λαμβάνουμε μία επερώτηση q όπου q=<q1,…,qt>, p=<p1,…,pt> (qi, pi τα βάρη των διανυσμάτων)

Τροποποίηση επερώτησης q (και ορισμός της q') :

$\mathbf{q_i' = k\ p_i + (1-k)\ q_i}$ όπου κ μία σταθερά $0 \leq \kappa \leq 1$

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 = <5, 0, 0, 3>
  - q = <0, 2, 0, 7>
  - q' = <1.25, 1.5, 0, 6>

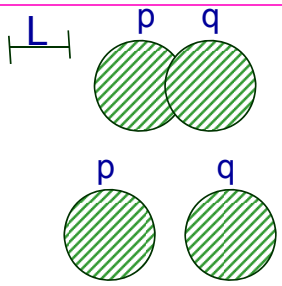
(C) Επερώτηση και Προφίλ ως ξεχωριστά σημεία αναφοράς (Query and Profile as Separate Reference Points)

- 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 **συνάρτηση απόστασης** Dist

Τρόποι συνδυασμού προφίλ και επερώτησης

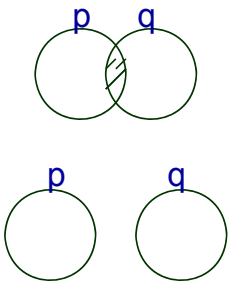
(1) Το διαζευκτικό μοντέλο (το λιγότερο αυστηρό)

- Ένα d ανήκει στην απάντηση αν:
- (Dist(d,q) ≤ L) OR (Dist(d,p) ≤ L)
- Εναλλακτική διατύπωση: min (Dist(d,q), Dist(d,p)) ≤ L
- είναι το λιγότερο αυστηρό



(2) Το συζευκτικό μοντέλο (το αυστηρότερο)

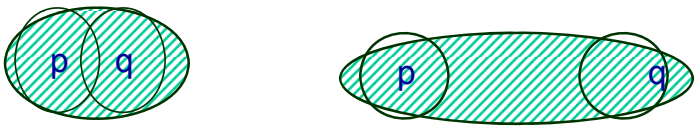
- (Dist(d,q) ≤ L) AND (Dist(d,p) ≤ L)
- max (Dist(d,q), Dist(d,p)) ≤ L
- είναι το πιο αυστηρό
- η απάντηση είναι η τομή των ans(p) και ans(q) (με κατώφλι L)
  - αν το q απέχει πολύ από το p, τότε η απάντηση θα είναι κενή



Τρόποι συνδυασμού προφίλ και επερώτησης (II)

(3) Το ελλειψοειδές μοντέλο

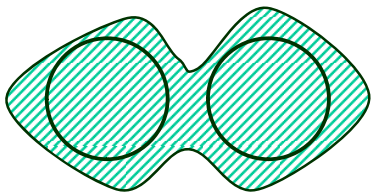
- Dist(d,q) + Dist(d,p) ≤ L
- καλό αν το d και το p δεν απέχουν πολύ
  - αν απέχουν πολύ τότε μπορεί να ανακτηθούν πολλά μη συναφή με κανένα

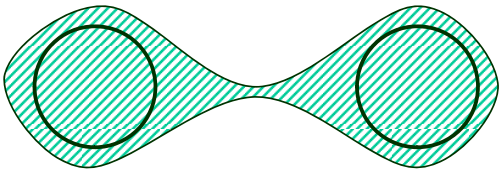


Τρόποι συνδυασμού προφίλ και επερώτησης (III)

(4) Το οβάλ μοντέλο του Casini

- Dist(d,q) * Dist(d,p) ≤ L
- αν το d και το p είναι κοντά, τότε ομοιάζει 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****Dist(d,q), ***w2**** Dist(d,p)) ≤ L //διαζευκτικό
  - max (***w1**** Dist(d,q), ***w2**** Dist(d,p)) ≤ L //συζευκτικό
  - ***w1****Dist(d,q) + ***w2**** Dist(d,p) ≤ L //ελλειψοειδές
- Στο μοντέλο Cassini τα βάρη είναι καλύτερα να εκφρασθούν ως εκθέτες:
  - $Dist(d,q)^{w1} * Dist(d,p)^{w2} \leq L$ //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 επερώτηση 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

- Πως 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 βαρών των εγγράφων του C

Συστήματα Πολλαπλών Σημείων Αναφοράς (Multiple Reference Point Systems)

- Σημεία αναφοράς: R1,… Rn
- Βάρη: w1,…wn, Σ wi=1
- || || 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 (***w1****Dist(d,q), ***w2**** Dist(d,p)) ≤ L //διαζευκτικό
  - max (***w1**** Dist(d,q), ***w2**** Dist(d,p)) ≤ L //συζευκτικό
  - ***w1****Dist(d,q) **+** ***w2**** Dist(d,p) ≤ L //ελλειψοειδές
  - Dist(d,q)^{w1} * Dist(d,p)^{w2} ≤ L //Cassini

- Συγκεκριμένα:
  - min (***w1****Dist(d,R1), …, ***wn**** Dist(d,Rn)) ≤ L //διαζευκτικό
  - max (***w1**** Dist(d, R1),…, ***wn**** Dist(d, Rn)) ≤ L //συζευκτικό
  - ***w1****Dist(d, R1) **+** … **+** ***wn**** Dist(d, Rn) ≤ L //ελλειψοειδές
  - Dist(d, R1)^{w1} *…* Dist(d, Rn)^{wn} ≤ L //Cassini
- ή συνδυασμός των παραπάνω

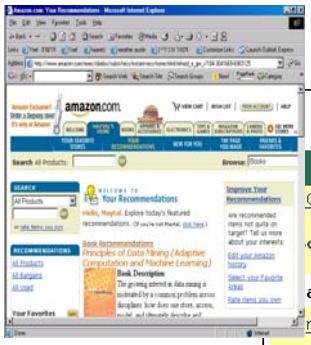
Άλλες τεχνικές (που έχουμε ήδη δει) που βοηθούν την εξατομίκευση

- **Ομαδοποίηση (Clustering):** θυμηθείτε το μάθημα περί ομαδοποίησης και επιτόπιας ανάλυσης
  - Μπορεί να δώσει λύση στο παράδειγμα:
  - q="**theory of groups**"
    - sociologist: behaviour of a set of people
    - mathematician: a particular type of algebraic structure
  - υπο την έννοια οτι η διαδικασία της ομαδοποίησης θα μας δώσει διαφορετικές ομάδες και ο εκάστοτε χρήστης θα μπορεί επιλέξει την κατάλληλη
- Τεχνικές Βελτίωσης Απάντησης Επερωτήσεων (ανατροφοδότηση συνάφειας)

Εξατομίκευση μέσω Συνεργατικής Επιλογής/Διήθησης

Personalization using Collaborative Selection/Filtering

Παράδειγμα



amazon.com

BOOKS | MUSIC | VIDEO | GIFTS | e-CARDS | AUCTIONS | HELP | YOUR ACCOUNT

BOOK SEARCH | BROWSE SUBJECTS | BESTSELLERS | FEATURED IN THE MEDIA | AWARD WINNERS | COMPUTERS & INTERNET | KIDS | BUSINESS & INVESTING

Only

ook Information

at a glance

reviews

customer comments

if you like this book...

Machine Learning (McGraw-Hill Series in Computer Science)

by Tom M. Mitchell, Thomas M. Mitchell

Our Price: $85.15

Availability: Usually ships within 24 hours.

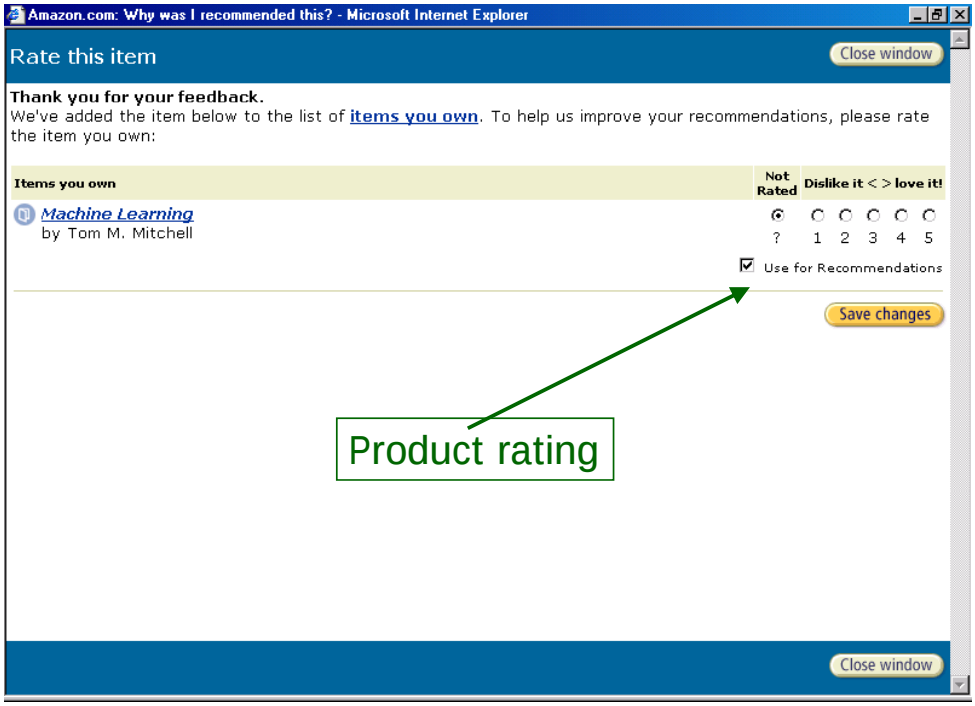
Add to Shopping Cart
(you can always remove it later)

Shopping with us is 100% safe.
Guaranteed.

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



Συνεργατική Επιλογή/Διήθηση

Πρόβλεψη προτιμήσεως ενός χρήστη

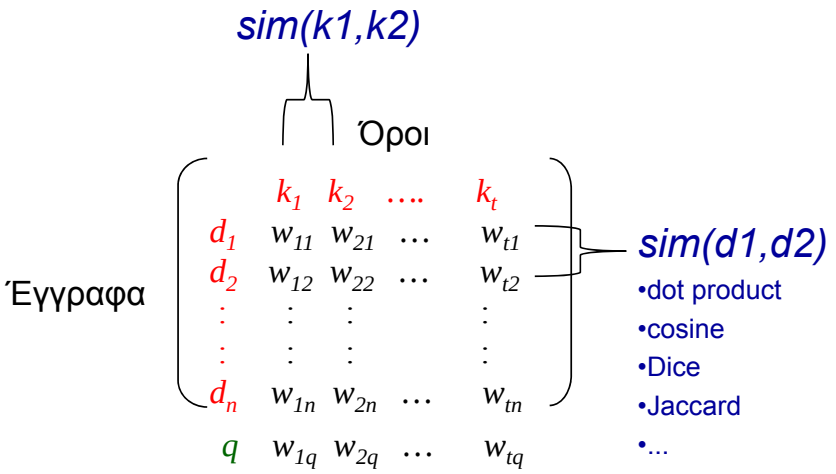
βάσει των καταγεγραμμένων προτιμήσεων του ιδίου και **άλλων** χρηστών.

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 πάρκινγ
    - 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

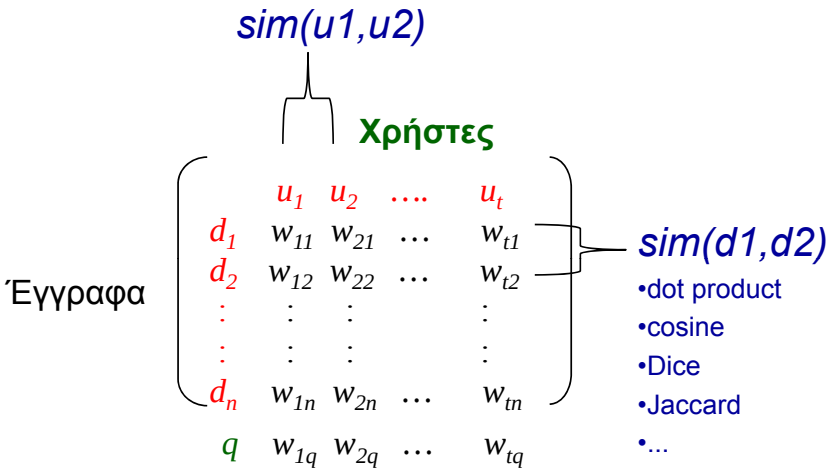


Ομοιότητα εγγράφων
βάσει των λέξεων

$w_{i,j} = \{0,1\}$
$w_{i,j} = tf_{i,j}\, idf_i$

Χρήστες αντί Όρων

Ομοιότητα χρηστών
βάσει των προτιμήσεων τους



Ομοιότητα εγγράφων
βάσει των (προτιμήσεων) των χρηστών

$w_{i,j} = \{0,1\}$ ===> 0: Bad, 1: Good
$w_{i,j} = tf_{i,j}\, idf_i$ ===> **$w_{i,j}$:βαθμός προτίμησης του χρήστη i στο έγγραφο j, πχ {1,2,3,4,5}**

Χρήστες αντί Όρων

Ομοιότητα χρηστών
βάσει των προτιμήσεων τους

sim(u1,u2)

Χρήστες

Έγγραφα

	u_1	u_2	….	u_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}
q	w_{1q}	w_{2q}	…	w_{tq}

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 (γενικά πολυμέσα), Βιβλία
  - Ηλεκτρικές Συσκευές
  - Εστιατόρια, Μεζεδοπωλεία
  - Κινηματογραφικές ταινίες
  - Τηλεοπτικά Προγράμματα
  - Λογισμικό
  - ...

$w_{i,j} = \{0,1\}$ ===> 0: Bad, 1: Good

$w_{i,j} = tf_{i,j}\, idf_i$ ===> $w_{i,j}$:βαθμός προτίμησης του χρήστη i στο έγγραφο j, πχ {1,2,3,4,5}

Μαντεύοντας τις προτιμήσεις ενός χρήστη

Ομοιότητα χρηστών
βάσει των προτιμήσεων τους

sim(u1,u2)

Χρήστες

Έγγραφα

	u_1	u_2	….	u_t
d_1	w_{11}	w_{21}	…	-
d_2	w_{12}	w_{22}	…	w_{t2}
:	:	:		:
:	:	:		:
d_n	w_{1n}	w_{2n}	…	w_{tn}

Ο χρήστης ut δεν έχει βαθμολογήσει (εκφράσει βαθμό προτίμησης) για το d1.
Μπορούμε να τον μαντέψουμε;

sim(d1,d2)
- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων
βάσει των (προτιμήσεων) των χρηστών

$w_{i,j} = \{0,1\}$ ===> 0: Bad, 1: Good

$w_{i,j} = tf_{i,j}\, idf_i$ ===> $w_{i,j}$:βαθμός προτίμησης του χρήστη i στο έγγραφο j, πχ {0,1,2,3,4,5}

Υπολογισμός Προβλέψεων και Συστάσεων

Ομοιότητα χρηστών
βάσει των προτιμήσεων τους

sim(u1,u2)

Χρήστες

Prediction

Ο χρήστης ut δεν έχει βαθμολογήσει (εκφράσει βαθμό προτίμησης) για το d1.
Μπορούμε να τον μαντέψουμε;

Έγγραφα

	u_1	u_2	….	u_t
d_1	w_{11}	w_{21}	…	-
d_2	w_{12}	w_{22}	…	w_{t2}
:	:	:		:
:	:	:		:
d_n	w_{1n}	w_{2n}	…	w_{tn}

sim(d1,d2)
- dot product
- cosine
- Dice
- Jaccard
- ...

Ομοιότητα εγγράφων
βάσει των (προτιμήσεων) των χρηστών

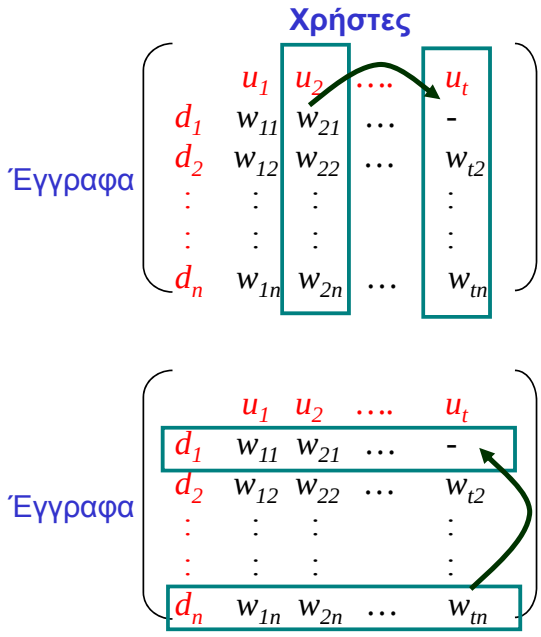
Recommendation

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?



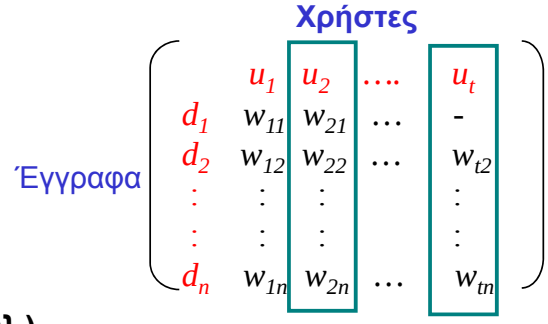
Nearest Users:
find the nearest (most similar) users and from their ratings infer $w(u_t,di)$
(or compute recommendations).

Nearest Items:
find the nearest (most similar) item and from its rating infer $w(u_t,di)$.

(compute recommendations):
find the items that are similar to other items the user has liked in the past

How we can compute recommendations. Nearest Users

Objective: Compute $w(u_t,di)$



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

Παράδειγμα πρόβλεψης βάσει των 3 κοντινότερων **χρήστων** 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(Tony, Yannis) = sqrt $[(4-4)^2+(3-3)^2+(1-2)^2]$=**1**
D(Manos, Yannis) = sqrt $[(5-4)^2+(3-3)^2+(2-2)^2]$=**1**
D(Tom, Yannis) = sqrt $[(1-4)^2+(1-3)^2+(5-2)^2]$=**4.69**
D(Nick, Yannis) = sqrt $[(2-4)^2+(1-3)^2+(4-2)^2]$=**3.46**
D(Titos, Yannis) = 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(Roma,Toscana) = sqrt $[(4-5)^2+(5-4)^2+(1-2)^2+(2-1)^2+(5-5)^2]$= **2**
D(Napoli, Toscana) = sqrt $[(3-5)^2+(3-4)^2+(1-2)^2+(1-1)^2+(4-5)^2]$= **2.65**
D(Hut, Toscana) = 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
  - συνημίτονο
  - Mean Squared Distance
  - Pearson Correlation Coefficient
  - ...

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

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

Στα άδεια κελιά του πίνακα θεωρούμε ότι υπάρχει το 0

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**

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

$\overline{u1}$ = mean of u1

$\overline{u2}$ = 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

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

  - Adjusted Pearson Correlation Coefficient

To handle the differences in rating scales of the users

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

Ομοιότητα/Απόσταση **Items**

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

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

  - Adjusted Pearson Correlation Coefficient

To handle the differences in rating scales of the users

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

Obtaining User Input

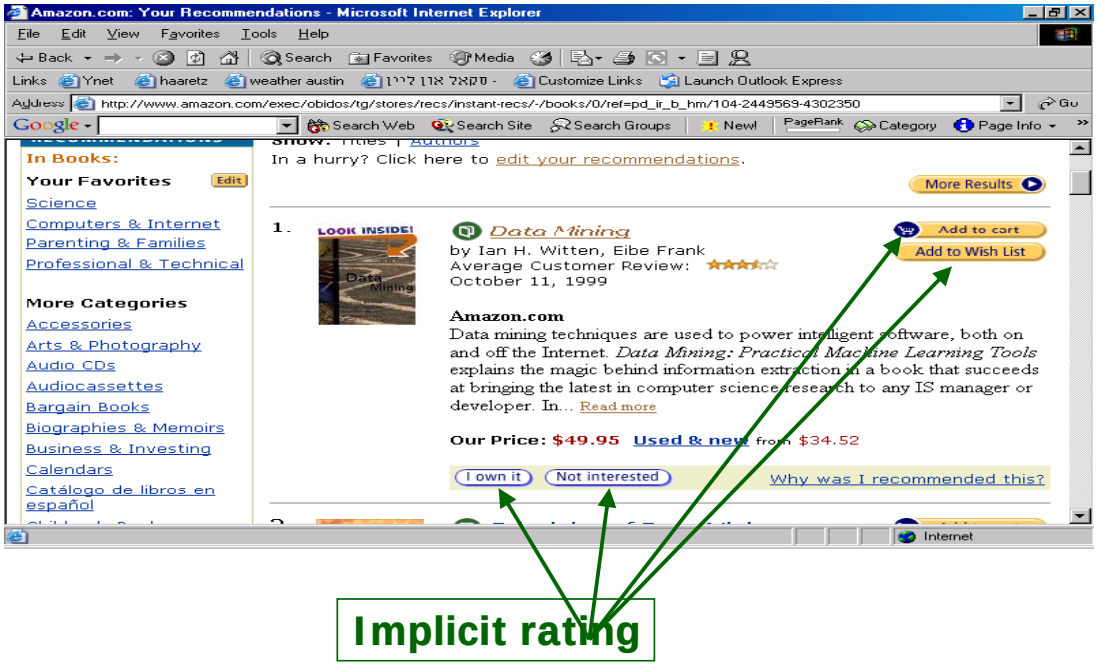
User (consumer) input is **difficult to get**

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**



Παρά ταύτα,

Πολύ συχνά $|D_{u1 \cap u2}|=0$

When thousands of items available only little overlap!

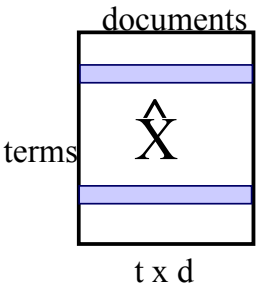
=> Recommendations based on only a few observations

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

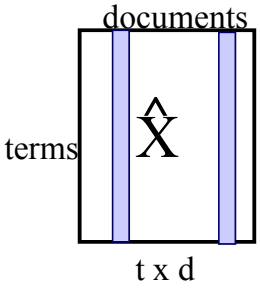
- Various solutions:
  - View CF as a classification task
    - build a classifier for each user
    - employ training examples
  - Reduce Dimensions
    - e.g. LSI (Latent Semantic Indexing)

LSI:

- Τρόπος σύγκρισης 2 όρων:
  - the **dot product** between two **row vectors of** 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



Performance Issues

- Depends on |U| vs. |D| and their "stability"
- Typical setting
  - D stable (e.g. 5.000 movies)
  - U dynamic and |U| >> |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 recommendatins)

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 techiques 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 απομακρυσμένα 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, mp3, 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: Sparceness & 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

Bibliography

- User's Profiles (Reference Points)
  - [4] Korfhage's Book Chapter 6 and 7
- Collaborative Filtering
  - Shardanand, U., and Mayes, P. (1995) Social Information Filtering: Algorithms for Automating 'Word of Mouth', in Proceedings of CHI95, 210-217. http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/us_bdy.htm
  - Billsus, D., & Pazzani, M.J. (1998) Learning Collaborative Information Filters. In: The 15th International Conference on Machine Learning, ICML-98. http://www.ics.uci.edu/~dbillsus/papers/icml98.pdf
  - Smyth B., Cotter P., *'Surfing the Digital Wave, Generating Personalised TV Listings using Collaborative, Case-Based Recommendation'*, In: Proceedings of the Third International Conference on Case-Based Reasoning ICCBR99', Springer.
  - Berkeley School of Information Systems, Link Collection on Collaborative Filtering. http://www.sims.berkeley.edu/resources/collab/