Introduction to

Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 11: Εξατομίκευση. Συστήματα Συστάσεων.

Τι θα δούμε σήμερα

- Εξατομίκευση
- Συστήματα Συστάσεων

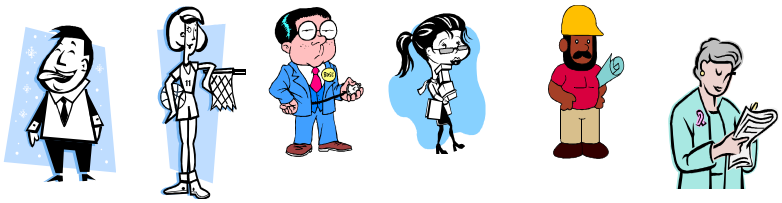
Διαφάνειες βασισμένες στις διαφάνειες του

Γιάννη Τζίτζικα για το μάθημα

HY463 - Συστήματα Ανάκτησης Πληροφοριών, Πανεπιστήμιο Κρήτης

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 του ερωτήματος
  - Παράδειγμα 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
- Τοποθεσία χρήστη

User Profile

- *Μέσο διάκρισης των χρηστών βάσει των χαρακτηριστικών και προτιμήσεών τους*

Μορφή

- Δεν υπάρχει κάποια τυποποιημένη μορφή
- Μπορούμε να θεωρήσουμε ότι έχει τη μορφή ενός ερωτήματος

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;

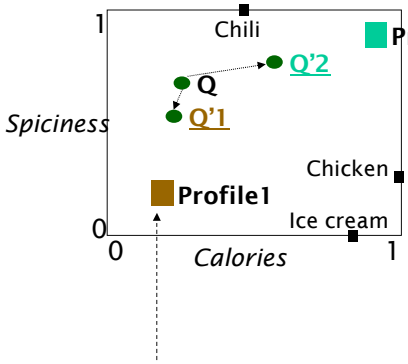
Γενικοί Τρόποι Αξιοποίησης κατά την Ανάκτηση Πληροφοριών

1. **Μετα-διήθηση** (**post-filter**)
   - Το προφίλ χρησιμοποιείται *κατόπιν* της αποτίμησης της αρχικής επερώτησης
   - Η χρήση προφίλ αυξάνει το υπολογιστικό κόστος της ανάκτησης
2. **Προ-διήθηση** (**pre-filter**)
   - 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
3. Επερώτηση και Προφίλ ως **ξεχωριστά σημεία αναφοράς** (Query and Profile as Separate Reference Points)

Μετα-διήθηση (post-filtering)

- 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, προσθέτει ένα παραπάνω υπολογιστικό στάδιο

Προ-διήθηση (Pre-filter)



Παράδειγμα με 2 διαστάσεις

Προφίλ χρήστη που προτιμάει ελαφριά και όχι πικάντικα φαγητά

Τεχνικές τροποποίησης επερωτήσεων

(B.1) Simple Linear Transformation (απλός γραμμικός μετασχηματισμός)
- Μετακινεί το διάνυσμα προς την κατεύθυνση του προφίλ

(B.2) Piecewise Linear Transformation
- Μετακινεί το διάνυσμα προς την κατεύθυνση του προφίλ βάσει περιπτώσεων

Απλός γραμμικός μετασχηματισμός

Έστω ερώτημα q = $<q_1,\ldots,q_t>$, προφίλ p = $<p_1,\ldots,p_t>$
(q_i, p_i τα βάρη των διανυσμάτων)

Τροποποίηση ερωτήματος q (και ορισμός της q') :

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

Περιπτώσεις
- Αν k=0 τότε q' = q (το ερώτημα μένει αναλλοίωτο)
- Αν k=1 τότε q' = **p** (το νέο ερώτημα ταυτίζεται με το προφίλ)
- Οι **ενδιάμεσες** τιμές του k είναι ενδιαφέρουσες

Piecewise Linear Transformation

Εδώ η τροποποίηση των βαρών προσδιορίζεται με ένα σύνολο περιπτώσεων

Διαφορετική συμπεριφορά με βάση αν ο όρος εμφανίζεται ή όχι στο ερώτημα και στο προφίλ

Περιπτώσεις:

1. όρος που εμφανίζεται και στο ερώτημα και στο προφίλ

   εφαρμόζουμε τον απλό γραμμικό μετασχηματισμό

2. όρος που εμφανίζεται μόνο στο ερώτημα

   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 στο ερώτημα ούτε στο προφίλ

   δεν κάνουμε τίποτα

Παράδειγμα

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

Ερώτημα και Προφίλ ως ξεχωριστά σημεία αναφοράς

Προσέγγιση

- Εδώ **δεν τροποποιείται** το αρχικό ερώτημα
- Αντίθετα και το ερώτημα και το προφίλ λαμβάνονται ξεχωριστά υπόψη κατά τη διαδικασία της βαθμολόγησης των εγγράφων

Θέματα

- 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

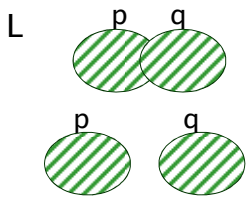
Τρόποι συνδυασμού προφίλ και ερωτήματος

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

- Ένα έγγραφο 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

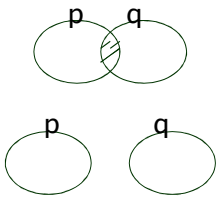


(2) Το *συζευκτικό μοντέλο* (το αυστηρότερο)

- Ένα έγγραφο d ανήκει στην απάντηση αν:

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



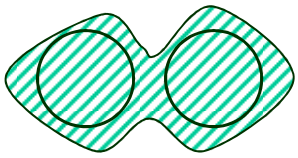
Τρόποι συνδυασμού προφίλ και ερωτήματος

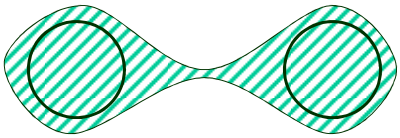
(3) Το ελλειψοειδές μοντέλο

- $\text{Dist}(d, q) + \text{Dist}(d, p) \leq L$
- 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 ερωτήματος

(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





Τρόποι συνδυασμού προφίλ και ερωτήματος

- Για να καθορίσουμε τη σχετική βαρύτητα του ερωτήματος και του προφίλ μπορούμε να προσθέσουμ βάρη 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: το ερώτημα 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

Σημεία αναφοράς: R_1,… R_n

Βάρη: w_1,…w_n, $\Sigma w_i=1$

|| || 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 ΠΟΛΛΑ ερωτήματα ταυτόχρονα*

19

Συστήματα Πολλαπλών Σημείων Αναφοράς

Θα γενικεύσουμε τα μοντέλα του δισδιάστατου χώρου που έχουμε ήδη δει:

- min ($w1$*Dist(d, q), $w2$* Dist(d, p)) $\leq L$ //διαζευκτικό
- max ($w1$* Dist(d, q), $w2$* Dist(d, p)) $\leq L$ //συζευκτικό
- $w1$*Dist(d, q) + $w2$* Dist(d, p) $\leq L$ //ελλειψοειδές
- Dist(d, q)^{w1} * Dist(d, p)^{w2} $\leq L$ //Cassini

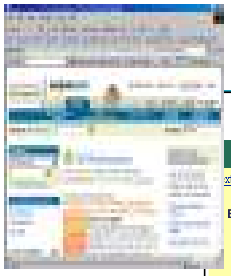
Συγκεκριμένα:

- min ($w1$*Dist(d, R_1), …, wn* Dist(d, R_n)) $\leq L$ //διαζευκτικό
- max ($w1$* Dist(d, R_1),…, wn* Dist(d, R_n)) $\leq L$ //συζευκτικό
- $w1$*Dist(d, R_1) + … + wn* Dist(d, R_n) $\leq L$ //ελλειψοειδές
- Dist(d, R_1)^{w1} *…* Dist(d, R_n)^{wn} $\leq L$ //Cassini

ή συνδυασμός των παραπάνω

20

ΣΥΣΤΑΣΕΙΣ (RECOMMENDATIONS)



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

xt Only

Book 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.

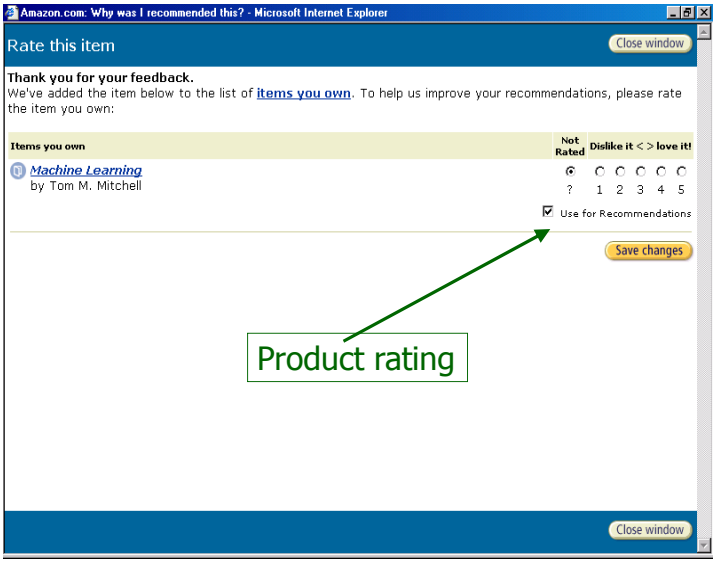
Add to Shopping Cart
(you can always remove it later)

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



Βασικές κατηγορίες συστημάτων συστάσεων

1. Content-based recommendations (με βάση την ομοιότητα των αντικειμένων): The user will be recommended items similar to the ones the user preferred in the past;

2. Collaborative recommendations (συνεργατικές συστάσεις): The user will be recommended items that people with similar tastes and preferences liked in the past

Συστάσεις βάσει περιεχομένου

Content-based recommendations (συστάσεις με βάση την ομοιότητα μεταξύ των αντικειμένων)

Πως ορίζεται η ομοιότητα;
Κλασική Προσέγγιση: κάθε αντικείμενο μπορεί να περιγραφεί με βάση κάποια χαρακτηριστικά του

Συστάσεις

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 να χρειαστεί να αναλύσουμε τον τρόπο που σκέφτεται ο χρήστης

Κλασική ανάκτηση κειμένου

Ομοιότητα όρων *βάσει των εγγράφων*

$sim(k_1, k_2)$

Όροι

Έγγραφα

	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}
q	w_{1q}	w_{2q}	...	w_{tq}

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

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

Πχ, αν ξέρουμε τα έγγραφα (ή άλλα αντικείμενα) (διανύσματα) που επέλεξε ο χρήστης, προτείνουμε όμοια

Ομοιότητα εγγράφων *βάσει των όρων*

$sim(d_1, d_2)$

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

Ομοιότητα ερωτήματος-εγγράφου

Πίνακας εγγράφων-χρηστών

Ομοιότητα χρηστών *βάσει των προτιμήσεων τους*

$sim(u_1, u_2)$

Χρήστες

Χρήστες αντί Όρων

Ομοιότητα εγγράφων *βάσει των (προτιμήσεων) των χρηστών*

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}

$sim(d_1, d_2)$

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

$w_{i,j} = \{0, 1\}$ ===> 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}**

Πίνακας εγγράφων-χρηστών

Ομοιότητα χρηστών *βάσει των προτιμήσεων τους*

$sim(u_1, u_2)$

Χρήστες

Χρήστες αντί Όρων

Ομοιότητα εγγράφων *βάσει των (προτιμήσεων) των χρηστών*

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}

$sim(d_1,d_2)$

- dot product
- cosine
- Dice
- Jaccard
- …

Αφού δεν χρησιμοποιούμε λέξεις, τα «έγγραφα» μπορεί να είναι οτιδήποτε:

- Φωτογραφίες, Βιβλία
- Ηλεκτρικές Συσκευές
- Εστιατόρια, Μεζεδοπωλεία
- Κινηματογραφικές ταινίες
- Τηλεοπτικά Προγράμματα
- ..

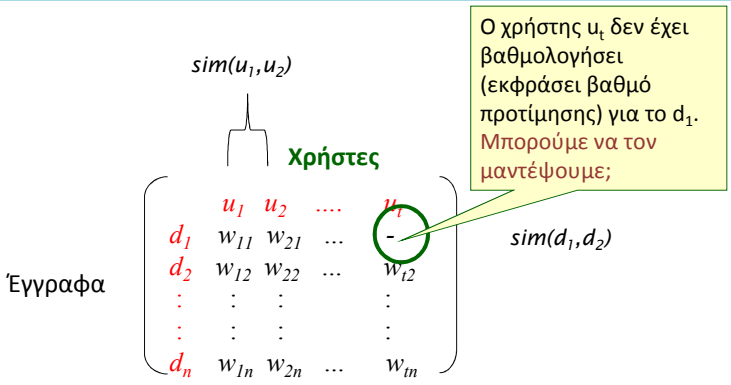
Πίνακας αντικειμένων-χρηστών

Χρήστες

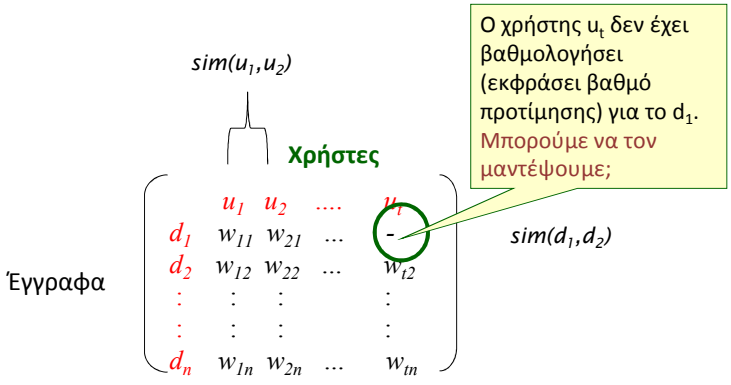
Αντικείμενα (items)

	u_1	u_2	....	u_t
i_1	w_{11}	w_{21}	…	w_{t1}
i_2	w_{12}	w_{22}	…	w_{t2}
:	:	:		:
:	:	:		:
i_n	w_{1n}	w_{2n}	…	w_{tn}

Πρόβλεψη (prediction)



Σύσταση (recommendation)

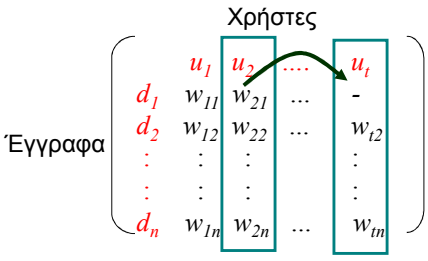


Computing recommendations for a user u:
1/ Predict values for those cells of u that are empty, and
2/ Select (and give the user) the highest ranked elements

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

Υπολογισμός Συστάσεων



Με ποιον χρήστη έχω παρόμοιες προτιμήσεις;

Τι λέει αυτός ο χρήστης για το d_1;

Nearest Users (κοντινότεροι (ποιο όμοιοι χρήστες):

find the nearest (most similar) users and from their ratings infer $w(u_t,d_i)$

Υπολογισμός βάσει ομοιότητας χρηστών

Objective: Compute $w(u_t, di)$

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}

- 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 ∈ $Sim(u_t)$

Υπολογισμός βάσει ομοιότητας χρηστών

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

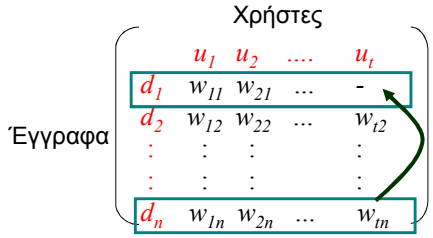
	Tony	Manos	Tom	Nick	Titos	Yannis
PizzaRoma	4	5	1	2	5	4
PizzaNapoli	3	3	1	1	4	3
PizzaHut	1	2	5	4	1	2
PizzaToscana	5	4	2	1	5	?

D(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

Υπολογισμός Συστάσεων



Nearest Items (κοντινότερα (ποιο όμοια) αντικείμενα:

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

Υπολογισμός βάσει ομοιότητας αντικειμένων

Παράδειγμα πρόβλεψης βάσει των 2 κοντινότερων αντικειμένων 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

Πρόβλεψη και Σύσταση

Μέθοδος συστάσεων

1. Προβλέπουμε το βαθμό όλων των αντικειμένων που δεν έχει βαθμολογήσει ο χρήστης
   - Χρήση ομοιότητας χρηστών, ή
   - Χρήση ομοιότητας αντικειμένων
2. Συστήνουμε τα *k* αντικείμενα με το μεγαλύτερο βαθμό

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

Παράδειγμα: mean squared distance

$$u1(x) \equiv w_{1x}$$
$$u2(x) \equiv w_{2x}$$

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

	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

Τρόποι υπολογισμού:

- εσωτερικό γινόμενο $sim(u_1,u_2) = \sum_{i=1}^{t} w_{1i} \cdot w_{2i}$
- συνημίτονο

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

- Mean Squared Distance

Στα άδεια κελιά του πίνακα θεωρούμε ότι υπάρχει το 0

Ομοιότητα/Απόσταση χρηστών: Pearson coefficient

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

$\overline{u1}$ = 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

measures the strength of a linear relationship between two variables

Pearson coefficient

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

Between -1 and +1.
The closer the correlation is to +/-1, the closer to a perfect linear relationship.
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 Αντικειμένων

Τρόποι υπολογισμού

- 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

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

45

Προβλήματα Εκκίνησης

Κοντινότερος γείτονας

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	?

46

Προβλήματα Εκκίνησης

Κοντινότερο αντικείμενο

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	-	-	-	-	-	?

47

Προβλήματα Εκκίνησης

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	-	-	-	-	-	?

48

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

Αραιός Πίνακας

Πολύ συχνά $|D_{u1 \cap u2}|=0$

When thousands of items available only little overlap!

=> Recommendations based on only a few observations

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)

Performance Issues

- Depends on size of set of users|U| vs. size of set of items|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 : predictedrating

q_i : actualrating

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)

Συνεργατική Επιλογή/Διήθηση: Σύνοψη

- 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

ΤΕΛΟΣ 11ου Μαθήματος

Ερωτήσεις?