

# Apunts del Taller de Nous Usos de la Informàtica

Jordi Vitrià

Universitat de Barcelona

10 de setembre de 2019



UNIVERSITAT  
DE  
BARCELONA

# Lliçó: Sistemes de Recomanació



# El problema de la recomanació

## Objectius i Usos

L'objectiu d'un sistema de **recomanació** és posar en correspondència un *usuari* amb *ítems* en funció de les seves preferències i interessos. Poden servir per filtrar informació, assistir en una compra, etc.

**amazon.com**

Hello, Jordi Vitrià. We have [recommendations](#) for you. ([Not Jordi?](#))  
[Jordi's Amazon.com](#) | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

**FREE 2-Day Shipping: See details**

Your Digital Items | Your Account | Help

[Shop All Departments](#) [Search](#) [All Departments](#) [GO](#) [Cart](#) [Wish List](#)

Your Amazon.com | Your Browsing History | Recommended For You | Amazon Betterizer | Improve Your Recommendations | Your Profile | Learn More

Jordi, Welcome to Your Amazon.com ([If you're not Jordi Vitrià, click here.](#))

### Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).

Page 1 of 44

[LOOK INSIDE!](#)



[The Beginning of Infinity: Expl...](#)

(Hardcover) by David Deutsch  
★★★★★ (13) \$17.24

[Fix this recommendation](#)

[LOOK INSIDE!](#)



[The Filter Bubble: What the Inter...](#)

(Hardcover) by Eli Pariser  
★★★★★ (26) \$14.91

[Fix this recommendation](#)

[LOOK INSIDE!](#)



[Networks: An Introduction](#)

(Hardcover) by Mark Newman  
★★★★★ (4) \$65.38

[Fix this recommendation](#)

[LOOK INSIDE!](#)



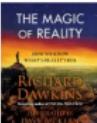
[The Shallows: What the Internet...](#)

(Paperback) by Nicholas Carr  
★★★★★ (112) \$9.16

[Fix this recommendation](#)

### Coming Soon for You

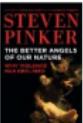
Page 1 of 2



[The Magic of Reality: How We...](#)

(Hardcover) by Richard Dawkins  
\$18.48

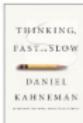
[Fix this recommendation](#)



[The Better Angels of Our Nature...](#)

(Hardcover) by Steven Pinker  
\$24.39

[Fix this recommendation](#)



[Thinking, Fast...Slow](#)

(Hardcover) by Daniel Kahneman  
\$17.64

[Fix this recommendation](#)

### Tap into Your Friends

BETA



Connect to Facebook to get Amazon recommendations for you and discover your friends' Favorites and Likes

[Sign in and Connect](#)

(You can disconnect at any time)

[See more recommended future releases](#)

Page 1 of 6

### New For You<sup>®</sup>



### Improve Your Recommendations

[The Information: A History, a Theory, a Flood](#)

[Rate this item](#)



**movieLens - Mozilla Firefox**

File Edit View Go Bookmarks Tools Help

http://movielens.umn.edu/main

Mozilla Firebird Help User Support Forum Plug-in FAQ Kayak LAKAWA- Lakes Area... Yahoo! Calendar - jtr... Slashdot: News for n...

**movieLens**  
helping you find the *right* movies

Welcome riedl@cs.umn.edu  
You've rated 205 movies.  
You're the 31st visitor in the past hour.

★★★★★ = Must See  
★★★★☆ = Will Enjoy  
★★★★☆ = It's OK  
★★☆☆☆ = Fairly Bad  
★☆☆☆☆ = Awful

Home | Manage Buddies | Your Preferences | Help | Publish | Logout

**Shortcuts** **Search**

- Your Ratings
- Your Wishlist
- Newest Additions
- Rate Random Movies
- Most Often Rated
- Suggest Title
- New Drama
- New DVDs
- New Movies

How to create your own shortcuts

Welcome to MovieLens!

**Advanced Search** now allows you to search for movies by director and/or actors in addition to its other features: Multiple genres, exclude genres, date ranges, language, hide predictions, and more! Check it out.

Also, don't forget about the new **publish** feature that lets you publish predictions in HTML/RSS 2.0 format. This and other info about recently added features is available in our **archived announcements**.

Did you know...? 4909 people joined MovieLens the same day you did.

New movies	New DVDs
★★★★★ Spider-Man 2 (a.k.a. Spiderman 2) (2004)	★★★★★ Great Escape, The (1963)
★★★★★ Shrek 2 (2004)	★★★★★ Fog of War: Eleven Lessons from the Life of Robert S. McNamara, The (2003)
★★★★★ Kill Bill: Vol. 2 (2004)	★★★★★ Miracle (2004)
★★★★★ Anchorman (2004)	★★★★★ Last Samurai, The (2003)
★★★★★ Super Size Me (2004)	★★★★★ Boat, The (Das Boot) (1981)
★★★★★ Fahrenheit 9/11 (2004)	★★★★★ Field of Dreams (1989)
★★★★★ Zatoichi (Zatōichi) (2003)	★★★★★ Suddenly (1954)
★★★★★ Man on Fire (2004)	★★★★★ Lord of the Rings: The Return of the King, The (2003)
★★★★★ Mean Girls (2004)	

Done

**movieLens - Mozilla Firefox**

File Edit View Go Bookmarks Tools Help

http://movielens.umn.edu/main

Mozilla Firebird Help User Support Forum Plug-in FAQ Kayak LAKAWA- Lakes Area... Yahoo! Calendar - jtr... Slashdot: News for n...

**movieLens**  
helping you find the *right* movies

Welcome riedl@cs.umn.edu  
You've rated 205 movies.  
You're the 31st visitor in the past hour.

★★★★★ = Must See  
★★★★☆ = Will Enjoy  
★★★★☆ = It's OK  
★★☆☆☆ = Fairly Bad  
★★☆☆☆ = Awful

Home | Manage Buddies | Your Preferences | Help | Publish | Logout

**Shortcuts** **Search**

Search Titles  **Go!**  
 Use selected buddies!

Search Genres    
Domain: All movies  Use selected buddies!  
**Search Genres!**

Advanced Search

Welcome to MovieLens!

**Advanced Search** now allows you to search for movies by director and/or actors in addition to its other features: Multiple genres, exclude genres, date ranges, language, hide predictions, and more! Check it out.

Also, don't forget about the new **publish** feature that lets you publish predictions in HTML/RSS 2.0 format. This and other info about recently added features is available in our [archived announcements](#).

**Did you know...?** 4909 people joined MovieLens the same day you did.

New movies	New DVDs
★★★★★ Spider-Man 2 (a.k.a. Spiderman 2) (2004)	★★★★★ Great Escape, The (1963)
★★★★★ Shrek 2 (2004)	★★★★★ Fog of War: Eleven Lessons from the Life of Robert S. McNamara, The (2003)
★★★★★ Kill Bill: Vol. 2 (2004)	★★★★★ Miracle (2004)
★★★★★ Anchorman (2004)	★★★★★ Last Samurai, The (2003)
★★★★★ Super Size Me (2004)	★★★★★ Boat, The (Das Boot) (1981)
★★★★★ Fahrenheit 9/11 (2004)	★★★★★ Field of Dreams (1989)
★★★★★ Zatoichi (Zatōichi) (2003)	★★★★★ Suddenly (1954)
★★★★★ Man on Fire (2004)	★★★★★ Lord of the Rings: The Return of the King, The (2003)
★★★★★ Mean Girls (2004)	

javascript:showTab('Search')

- Els recomanadors es poden veure com un pas més enllà dels cercadors en la direcció del descobriment: la *cerca* és el que fas quan busques alguna cosa, el *descobriment* és quan alguna cosa que tu no sabies que existia (o no sabies com cercar) et troba.
- Els recomanadors són efectius: el 60% dels films que lloga Netflix i el 35% de les vendes d'Amazon són recomanacions.

# El problema de la recomanació

- Sigui  $C$  el conjunt de tots els usuaris i sigui  $S$  el conjunt de tots els ítems que es poden recomanar (llibres, restaurants o compres).
- Els espais  $S$  i  $C$  poden tenir una cardinalitat molt gran!

## Definició

Sigui  $u : C \times S \rightarrow \mathbb{R}$  una funció que mesura el grau d'utilitat que pot tenir un determinat ítem per un determinat usuari. El **problema de la recomanació** és escollir, per a cada usuari  $c \in C$ , l'ítem  $s'_c \in S$  que maximitza la funció d'utilitat:

$$\forall c \in C, s'_c = \operatorname{argmax}_{s \in S} u(c, s) \quad (1)$$

$$\begin{matrix} U & s_1 & s_2 & \dots & s_n \\ c_1 & u_{11} & u_{12} & \dots & u_{1n} \\ c_2 & u_{21} & u_{22} & \dots & u_{2n} \\ \dots & \vdots & \vdots & \ddots & \vdots \\ c_m & u_{m1} & u_{m2} & \dots & u_{mn} \end{matrix}$$

- Els elements de  $C$  s'acostumen a definir amb un perfil, que inclou un cert nombre de característiques definitòries de l'usuari (edat, sexe, etc.).
- Els elements de  $S$  també es defineixen amb un conjunt de característiques (p.e. per una pel·lícula, tota la informació associada).
- El principal problema pels sistemes de recomanació és que  $u$  no està definida per tot l'espai, sinó que només en tenim una mostra i per tant hem d'*extrapolar* els seus valors.
- **Les extrapolacions es poden fer de diverses maneres, però la més important és estimant la funció d'utilitat de manera que optimitzi algun criteri relacionat amb l'error sobre la part de  $u$  que coneixem (error empíric).**

$$\begin{array}{c}
 U & s_1 & s_2 & \dots & s_n \\
 \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{matrix} &
 \left( \begin{array}{ccccc}
 ? & u_{12} & \dots & ? \\
 ? & ? & \dots & u_{2n} \\
 \vdots & \vdots & \ddots & \vdots \\
 u_{m1} & ? & \dots & u_{mn}
 \end{array} \right)
 \end{array}$$

Volem que la diferència (error) entre els valors  $u_{ij}$  que coneixem i el valor surt de la funció que els prediu,  $u(c, s)$ , sigui mínima, amb l'esperança de que també ho serà pels que no coneixem.

## Sistemes de recomanació

- ① **Recomanacions col·laboratives.** Recomanarem a l'usuari ítems que tenen un alt valor d'utilitat, sent la utilitat un concepte definit exclusivament a partir dels elements  $u_{ij}$  coneguts.
- ② **Recomanacions (no col·laboratives) basades en el contingut.** Recomanarem a l'usuari ítems semblants als que ha triat en el passat o ben valorats per usuaris semblants a ell. En aquest cas la semblaça es defineix a partir del *contingut* dels ítems o de la descripció dels usuaris.
- ③ **Aproximacions híbrides.**

# Mètodes Col·laboratius basats en la semblança d'usuaris

## Mètodes Col·laboratius basats en la semblança d'usuaris

La utilitat  $u(c, s)$  de l'ítem  $s$  per l'usuari  $c$  s'estima a partir de les utilitats  $u(c_i, s)$  assignades a l'ítem  $s$  pels altres usuaris  $c_i$ , ponderades segons la semblança entre els  $c_i$  i  $c$ . En aquest cas el problema és definir què entenem per semblança entre usuaris!

# Mètodes Col·laboratius basats en la semblança d'usuaris

Si volem saber la utilitat  $u_{c_qs_p}$  de l'ítem  $s_p$  per l'usuari  $c_q$ , analitzem la columna  $s_p$  de la matriu:

$$\begin{matrix} U & s_1 & \dots & s_p & \dots & s_n \\ c_1 & ? & \dots & u_{1p} & \dots & ? \\ \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_q & ? & \dots & ? & \dots & u_{2n} \\ \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_m & u_{m1} & \dots & u_{mp} & \dots & u_{mn} \end{matrix}$$

$$u_{c_qs_p} = \sum_{j=1}^m \alpha_{c_q c_j} u_{c_j s_p}$$

on  $\alpha_{c_q c_j}$  és 0 si no coneixem  $u_{c_j s_p}$  i un pes que depèn de la semblança entre els usuaris  $c_q$  i  $c_j$  en el cas contrari.

# Representació dels usuaris

*Have you ever wondered what you look like to Amazon? Here is the cold, hard truth: You are a very long row of numbers in a very, very large table. This row describes everything you've looked at, everything you've clicked on, and everything you've purchased on the site; the rest of the table represents the millions of other Amazon shoppers. Your row changes every time you enter the site, and it changes again with every action you take while you're there. That information in turn affects what you see on each page you visit and what e-mail and special offers you receive from the company.*

**Deconstructing Recommender Systems.** Joseph Konstan, John Riedl.  
IEEE Spectrum, 24 September 2012.

# Representació dels usuaris

Donada aquesta representació, la semblança entre dos usuaris  $c_q, c_j$  s'ha de definir en funció de les files que els representen a la matriu  $U$ .

- Intuïtivament, direm que dos usuaris són semblants si tendeixen a valorar (directa o indirectament) els items de la mateixa manera, i són diferents si tendeixen a valorar els items de manera diferent.
- Per això, la mesura de semblança entre dos usuaris està definida sobre els items que ambdós han valorat (que pot tenir diferent per cada parella d'usuaris).

Les limitacions principals del model col·laboratiu són:

- ① El problema dels nous usuaris o *cold start*: per a ser útil per l'usuari, hem de tenir una bona quantitat d'ítems valorats per ell mateix. En cas contrari la funció de semblança entre usuaris no serà precisa.
- ② El problema del nou ítem: fins que no tenim prous valoracions de l'ítem, no serà recomanat!
- ③ El problema de la *long tail function*: hi ha molts ítems que estaran, per definició, recomanats per poca gent (usuaris amb gustos no massius!).

# Long tail function

*In statistics, a long tail of some distributions of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution.* (Font: Wikipedia)

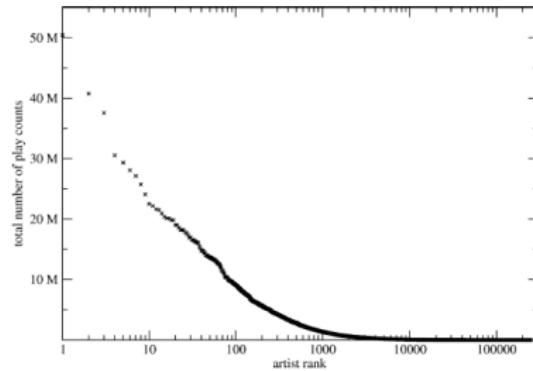


Fig. 4.2 The music Long Tail effect. A log-linear plot depicting the total number of plays per artist. Data gathered during July, 2007, for a list of 260,525 artists.

El negoci dels darrers 200.000 usuaris pot ser més gran que els dels primers 1000!

# Implementació

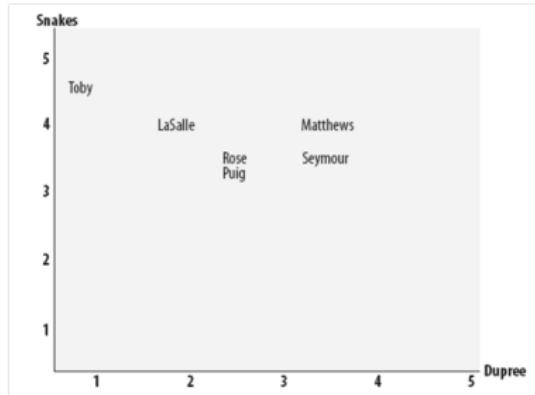
- Suposem que tenim un fitxer anomenat `recommendations.py` amb dades sobre les preferències (en una escala de 1 a 5) d'un conjunt d'usuaris sobre les pel·lícules que han vist.

```
# A dictionary of movie critics and their ratings of a small
# set of movies
critics={'Lisa Rose': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.5,
    'Just My Luck': 3.0, 'Superman Returns': 3.5, 'You, Me and Dupree': 2.5,
    'The Night Listener': 3.0},
    'Gene Seymour': {'Lady in the Water': 3.0, 'Snakes on a Plane': 3.5,
    'Just My Luck': 1.5, 'Superman Returns': 5.0, 'The Night Listener': 3.0,
    'You, Me and Dupree': 3.5},
    'Michael Phillips': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.0,
    'Superman Returns': 3.5, 'The Night Listener': 4.0},
    'Claudia Puig': {'Snakes on a Plane': 3.5, 'Just My Luck': 3.0,
    'The Night Listener': 4.5, 'Superman Returns': 4.0,
    'You, Me and Dupree': 2.5},
    'Mick LaSalle': {'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0,
    'Just My Luck': 2.0, 'Superman Returns': 3.0, 'The Night Listener': 3.0,
    'You, Me and Dupree': 2.0},
    'Jack Matthews': {'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0,
    'The Night Listener': 3.0, 'Superman Returns': 5.0, 'You, Me and Dupree': 3.5},
    'Toby': {'Snakes on a Plane': 4.5, 'You, Me and Dupree': 1.0, 'Superman Returns': 4.0}}
```

# Com podem calcular la similitud entre dos usuaris?

- Representarem l'usuari  $i$  amb un vector numèric  $(u_{i1}, u_{i2}, \dots, u_{im})$ , on  $m$  és el nombre d'ítems a la nostra base de dades.
- Les dues mesures més importants per avaluar la similitud entre dos vectors són la **distància euclidiana** i el **coeficient de correlació de Pearson**, tot i que en casos concrets hi ha moltes més mesures útils, com la **Earth Mover Distance**.

- Cada usuari és un punt en un espai de dimensió  $m$  que anomenem *l'espai de preferències*.



## Distància Euclidiana

- Per calcular la distància entre dos usuaris  $X = (x_1, \dots, x_m)$ ,  $Y = (y_1, \dots, y_m)$ , la **distància euclidiana** calcula l'arrel quadrada de la suma dels quadrats de les diferències entre cada una de les components dels usuaris:

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_m - y_m)^2} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (2)$$

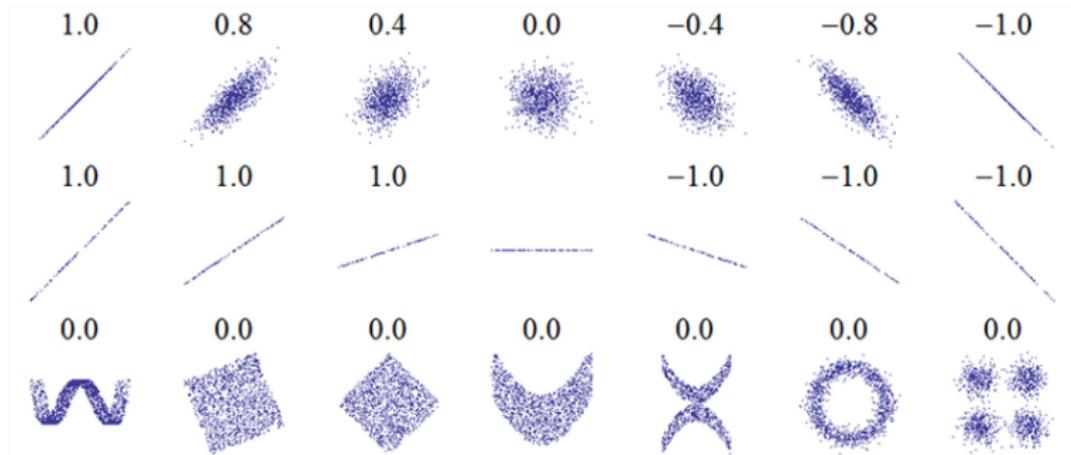
- Ho podem invertir per tenir un nombre que sigui més gran com més semblants, acotat entre 0 i 1.

- Alternativament, també podem analitzar la semblança entre dos usuaris  $X, Y$  calculant el **Coeficient de correlació de Pearson** del conjunt  $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ .

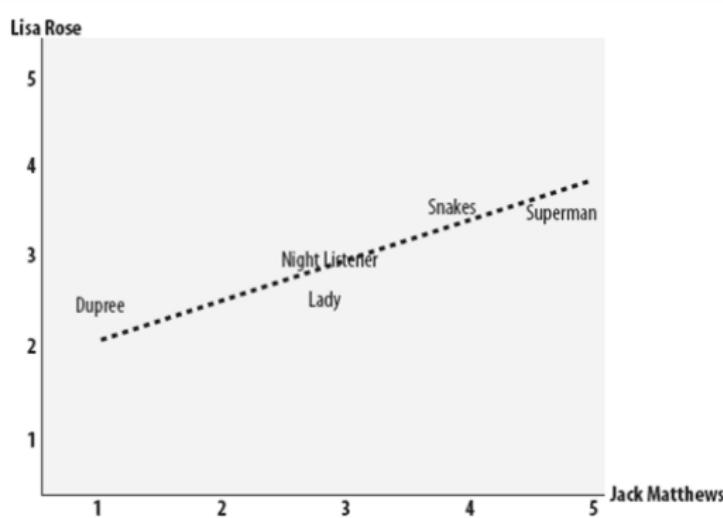
### Coeficient de correlació de Pearson

El **coeficient de correlació de Pearson** calcula una mesura de l'ajust d'un conjunt de punts a una recta. Funciona millor que la distància euclidiana si les components no estan ben normalitzades. La seva fòrmula és:

$$r(X, Y) = \frac{\sum_{i=1}^m (x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sqrt{\sum_{i=1}^m (x_i - \hat{x}_i)^2} \sqrt{\sum_{i=1}^m (y_i - \hat{y}_i)^2}} \quad (3)$$

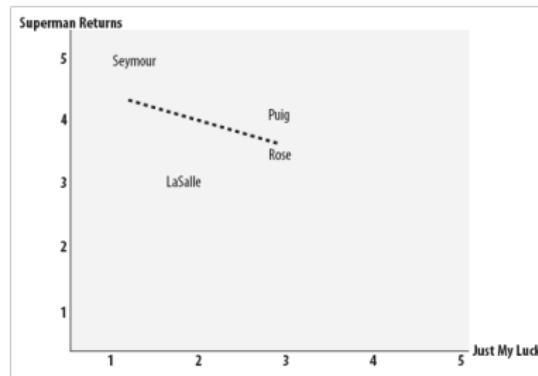


- En els gràfics es pot veure com aquesta mesura és insensible al fet que hi pot haver usuaris que puntuïn més alt (o més baix) de forma sistemàtica (**inflació de puntuació**): en Jack Mathews puntuïa més alt que la Lisa Rose, però correlacionen perfectament els seus gustos.



# Observació

- Les correlacions negatives indiquen que aquells a qui agrada un ítem (Superman) tendeixen a no sentir-se atrats per un altre (Just My Luck).



# Recomanacions basades en la semblança entre usuaris

- Ara ja estem en posició de fer una primera proposta (simplista) de recomanació a un usuari  $A$ : *Podem buscar un usuari semblant a  $A$ ,  $A'$ , i recomanar alguna pel·lícula que hagi agradat a  $A'$  i que  $A$  no hagi vist.*
- **Observació:** Aquesta estratègia no és perfecte, atès que podríem recomanar pel·lícules que tothom a puntuat malament excepte  $A'$ !
- Una possible solució a aquest problema és fer una ponderació de les puntuacions amb tots els usuaris.

- La taula mostra un conjunt d'usuaris, la seva similitud respecte a mi i les pel·lícules que jo no he vist (Night, Lady i Luck).
- La columna S.xNight mostra la similitud multiplicada per la puntuació, de manera que una persona que s'assembla a mi produeix puntuacions més altes que una que no s'assembla.
- Finalment, la suma de les puntuacions que ha rebut cada pel·lícula (Total) es normalitza per eliminar el biaix introduït a les pel·lícules que han revisat molts crítics dividint-ho per la suma de les seves similituds.

Critic	Similarity	Night	S.xNight	Lady	S.xLady	Luck	S.xLuck
Rose	0.99	3.0	2.97	2.5	2.48	3.0	2.97
Seymour	0.38	3.0	1.14	3.0	1.14	1.5	0.57
Puig	0.89	4.5	4.02			3.0	2.68
LaSalle	0.92	3.0	2.77	3.0	2.77	2.0	1.85
Matthews	0.66	3.0	1.99	3.0	1.99		
Total			12.89		8.38		8.07
Sim. Sum			3.84		2.95		3.18
Total/Sim. Sum			3.35		2.83		2.53

- Ara ja tenim el sistema de recomanació complet basat en buscar usuaris semblants:

```
>>> reload(recommendations)
>>> recommendations.getRecommendations(recommendations.critics, 'Toby')
[(3.3477895267131013, 'The Night Listener'), (2.8325499182641614, 'Lady in the
Water'), (2.5309807037655645, 'Just My Luck')]
>>> recommendations.getRecommendations(recommendations.critics, 'Toby',
...     similarity=recommendations.sim_distance)
[(3.5002478401415877, 'The Night Listener'), (2.7561242939959363, 'Lady in the
Water'), (2.4619884860743739, 'Just My Luck')]
```

# Recomanacions col·laboratives basades en ítems

- Anem ara a veure una alternativa al mètode anterior que ens permet fer recomanacions col·laboratives basades en la **semblança** entre ítems:

---

## Customers who bought this item also bought

[Learning Python, Second Edition](#) by Mark Lutz

[Python Cookbook](#) by Alex Martelli

[Python in a Nutshell](#) by Alex Martelli

[Python Essential Reference \(2nd Edition\)](#) by David Beazley

[Foundations of Python Network Programming \(Foundations\)](#) by John Goerzen

► [Explore similar items : Books](#) (42)

---

- Això és especialment útil quan no tenim molta informació de l'usuari (és més normal tenir informació sobre un ítem que sobre un usuari).

# Recomanacions col·laboratives basades en ítems

## Mètodes col·laboratius basats en ítems

La utilitat  $u(c, s)$  de l'ítem  $s$  per l'usuari  $c$  s'estima a partir de les utilitats  $u(c, s_i)$  assignades per l'usuari  $c$  als items  $s_i$ , ponderades segons la **semblança entre els ítems**. En aquest cas el problema és definir què entenem per semblança entre items!

- La idea intuitiva és que dos ítems seran semblants si han estat valorats de la mateixa forma per un nombre important d'usuaris.

# Recomanacions col·laboratives basades basades en ítems

Si volem saber la utilitat  $u_{c_qs_p}$  de l'ítem  $s_p$  per l'usuari  $c_q$ , analitzem la fila  $c_p$  de la matriu:

$$\begin{matrix} U & s_1 & \dots & s_p & \dots & s_n \\ c_1 & ? & \dots & u_{1p} & \dots & ? \\ \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_q & ? & \dots & ? & \dots & u_{qn} \\ \dots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_m & u_{m1} & \dots & u_{mp} & \dots & u_{mn} \end{matrix}$$

$$u_{c_qs_p} = \sum_{j=1}^n \alpha_{s_ps_j} u_{c_qs_j}$$

on  $\alpha_{s_ps_j}$  és 0 si no coneixem  $u_{c_qs_j}$  i un pes que depèn de la semblança entre els items  $s_p$  i  $s_j$  en el cas contrari.

# Implementació

- Si calculem la matriu  $U^t$ , o el que és el mateix, podem invertir el fitxer de preferències que teníem:

```
{'Lady in the Water': {'Lisa Rose': 2.5, 'Gene Seymour': 3.0},  
 'Snakes on a Plane': {'Lisa Rose': 3.5, 'Gene Seymour': 3.5}} etc..
```

podem iterar sobre tots els ítems i guardar la llista d'ítems més semblants per cada ítem.

- Llavors, aplicant els mateixos mètodes que hem vist pels usuaris podem fer una recomanació basada en ítems i no en usuaris.

# Implementació

- La gran diferència entre els dos mètodes que hem vist és que, tot i que els dos casos hem de processar tota la taula (i això és costós!), les comparacions entre ítems canvien en el temps més lentament que les comparacions entre usuaris i per tant es pot fer *off-line*.
- Per tant, podem crear de forma *off-line* una llista ponderada amb els ítems més a cada ítem, i quan fem una recomanació a un usuari només cal mirar aquesta llista.

# Resum sobre recomanació col·laborativa

Recomanació col·laborativa basada en usuaris:

- ① Identificar  $I$ , el conjunt d'items que ha valorat l'usuari objectiu.
- ② Identificar  $N$ , el conjunt d'usuaris que han valorat 1 o més items del conjunt  $I$ .
- ③ Calcular la semblança de cada usuari d' $N$  a l'usuari objectiu.
- ④ Predir les valoracions de l'usuari objectiu per  $I^c$ , el conjunt d'items que no ha valorat.
- ⑤ Recomanar els  $n$  productes de  $I^c$  amb valoració més alta.

Recomanació col·laborativa basada en ítems:

- ① Identificar  $U$ , el conjunt d'usuaris que han valorat l'ítem objectiu.
- ② Identificar  $N$ , el conjunt d'ítems que han estat valorat pels usuaris de  $U$ .
- ③ Calcular la semblança entre cada element de  $N$  i l'ítem objectiu.
- ④ Predir la valoració de l'ítem objectiu.

# El problema de la recol·lecció de dades

- Quan construïm la base de dades podem usar dos estratègies: l'explícita i la implícita.
- La **recol·lecció explícita** de dades pot ser: demanar a un usuari avaluar un ítem segons una escala numèrica, demanar a un usuari una ordenació d'un conjunt d'ítems segons les seves preferències, presentar a l'usuari dos ítems i preguntar-li quin prefereix, demanar a l'usuari que faci una llista dels ítems que li agraden, etc.
- La **recol·lecció implícita** de dades pot ser: anotar els ítems que consulta a la base de dades, analitzar els períodes de temps d'aquestes consultes, analitzar la xarxa social de l'usuari i descobrir gustos semblants, etc.

# Recomanació (no col·laborativa) basada en el contingut

- Els sistemes col·laboratius purs usen exclusivament la matriu de puntuacions dels usuaris, però és evident que podem millorar-ho si tenim informació de l'usuari (p.e. dades demogràfiques) i dels ítems (director de la pel·lícula, gènere, etc.).
- Els **mètodes no col·laboratius basats en el contingut** recomanen ítems a partir de comparar descripcions del contingut de cada ítem a representacions dels continguts dels ítems que sabem interessen a l'usuari.

# Recomanació (no col·laborativa) basada en el contingut

- Per a certs ítems pot ser difícil, però per **ítems textuais** (llibres, notícies, pàgines web, blogs, etc.) és un camp bastant explotat.



# Models Híbrids

- Els models col·laboratius i els no col·laboratius es poden combinar de moltes maneres.
- La més obvia és que cada model produeixi el seu ranking i llavors produir un ranking agregat.
- També podem calcular un ranking agregat ponderat, en el que el pes del component col·laboratiu s'incrementa a mesura que creix el nombre d'usuaris que accedeixen a l'item.

# Mètriques d'avaluació

- La qualitat d'un sistema de recomanació es pot avaluar comparant les recomanacions que fa el mètode desenvolupat amb les avaluacions d'un conjunt de test (amb avaluacions conegeudes d'usuaris) que no s'han fet servir per la construcció del sistema (**mètrica de precisió de la predicció**).
- La mètrica més usada és l'**error absolut mig**, que es defineix com la diferència absoluta mitja entre les avaluacions predites i les reals:

$$MAE = \frac{\sum_{\{u,i\}} |p_{u,i} - r_{u,i}|}{N} \quad (4)$$

on  $p_{u,i}$  és l'avaluació predicta de l'usuari  $u$  per l'item  $i$ ,  $r_{u,i}$  és l'avaluació real, i  $N$  és el nombre d'avaluacions del conjunt de test.

# Avaluació

*So how well do recommenders ultimately work? They certainly are increasing online sales; analyst Jack Aaronson of the Aaronson Group estimates that investments in recommenders bring in returns of 10 to 30 percent, thanks to the increased sales they drive. And they still have a long way to go.*

**Deconstructing Recommender Systems.** Joseph Konstan, John Riedl.  
IEEE Spectrum, 24 September 2012.

# Conclusions

- La recomanació col·laborativa basada en ítems és molt més **eficient** que la recomanació col·laborativa basada en usuaris, però s'ha de mantenir una taula adicional de semblances, que pot ser gran.
- El mètode col·laboratiu basat en ítems funciona millor que el basat en usuaris en bases de dades *sparse* i el basat en usuaris funciona millor en bases de dades denses.
- La recomanació basada en el contingut pot ajudar, sobretot en determinats escenaris.
- El principal problema és com tractar els nous usuaris i els nous ítems (*cold start problem*).
- Els sistemes de recomanació són un objectiu clar per la manipulació fraudulenta i generen problemes ètics importants!

# Problemes ètics

- Com fer que els usuaris revelin les seves preferències?
- Com aconseguir puntuacions per tots els productes (no només aquells que els usuaris odien o estimen!)
- Quines dades personals és ètic demanar?
- Etc.

**Sense Networks**

Macrosense Citysense Technology Principles Media Center About Us

Indexing the real world using location data  
for predictive analytics.

**News** **Events** **Solutions**

**Forbes**  
08.12.10 - Cofounder Sandy Pentland featured in Forbes

**Sense Networks**  
07.27.10 - Sense Networks Inc. Announces Hiring of New CEO David Petersen

**thewherebusiness.com**  
07.01.10 - Using location analytics to mine mobile location data for user segmentation

**CNN**

**macrōsense™**  
Platform for analyzing large amounts of mobile location data in real-time to drive relevant recommendation, personalization and discovery.

**CabSense™**  
Consumer application for iPhone and Android analyzes tens of millions of data points to help you find the best corner to catch a cab in New York City.

**Citysense™**  
Consumer application for real-time nightlife

**CabSense™**

New York



TheSmartestWaytoFindaCab

as featured in **am NEW YORK**

- CabSense analyzes tens of millions of GPS data points from NYC taxis to help you find the best corner to catch a cab
- Use [Map View](#) or [Radar View](#) to find the best corner
- Plan ahead with the [Time Slider](#) and see the best locations at a future time

[>> More Features](#)

iPad available now!

free download!

Available on the **App Store**

Get it at the **ANDROID Marketplace**

Follow CabSense on **twitter**

Sense Networks

Macrosense Citysense Technology Principles Media Center About Us

Quick Links

- Citysense Site

## Citysense™

Citysense is an innovative mobile application for local nightlife discovery and social navigation, answering the question, "Where is everybody?"

Citysense shows the overall activity level of the city, top activity hotspots, and places with unexpectedly high activity, all in real-time. Then it links to Yelp and Google to show what venues are operating at those locations. Citysense is a free demonstration of the Macrosense platform that everyone can enjoy.

Currently, local discovery depends on proactive searching for relevant locations. Users are challenged to input specific location data into mobile interfaces with small screens.

Currently, local discovery depends on proactive searching for relevant locations. Users are challenged to input specific location data into mobile interfaces with small screens.

#### **Citysense eliminates the need to search**

Instead, it enables serendipitous discovery. Citysense passively "senses" the most popular places based on actual real-time activity and displays a live heat map. The application intelligently leverages the inherent wisdom of crowds without any change in existing user behavior, in order to navigate people to the hottest spots in a city. And it's not dependent on having a critical mass of users on the system.

#### **Citysense is an application that learns**

The application learns about where each user likes to spend time – and it processes the movements of other users with similar patterns. In its next release, Citysense will not only answer "where is everyone right now" but "where is everyone like me right now." Four friends at dinner discussing where to go next will see four different live maps of hotspots and unexpected activity. Even if they're having dinner in a city they've never visited before.

#### **powerful back-end infrastructure**

Sense Networks has built a unique back-end infrastructure that processes years of data encompassing billions of points of positioning data. Created on the Macrosense platform, Citysense leverages this historical data analysis to normalize live location data originating from tens of thousands of devices and users moving throughout a given city.

#### **Citysense never shares your location or asks for personal information**

No logins, passwords, or phone numbers. Users actually own any historical data that they create while the system will never be used to personalize the service. There are buttons in Citysense to "delete any data acquired in the last 24 hours" and to "delete all historical data." After these deletions are made, personalized services will no longer operate, but users should always have this choice. Sense Networks has developed a revolutionary new approach to data ownership and privacy. Read our Corporate Principles for more thoughts on what it means to own your data.



#### **Citysense is available now**

Citysense is available in the city of San Francisco for alpha testing, and will eventually rollout to major metropolitan areas in the US and abroad.

For more information on Citysense and to download the free application, visit [www.citysense.com](http://www.citysense.com). It is currently available on BlackBerry devices and will be released for the Apple iPhone soon.

**Macrosense**  
Macrosense Login

**Citysense**  
Citysense Site

**Technology**  
Machine Learning  
MVE Algorithm

**Principles**

**Media Center**  
Press Coverage

**About Us**  
Executive Team  
Advisors  
Careers  
Contact

### Company Principles

"Sense Networks has deeply rooted principles that drive every business and technology decision. The company has built its systems from scratch to introduce the following new paradigm of data ownership and privacy."

**Professor Alex (Sandy) Pentland, Sense Networks Co-Founder, Chief Privacy Advocate and Director of Human Dynamics Research at MIT**



#### People should own their own data

People should have full control over the use of any data that they generate. All data collection should be "opt-in," and users should be able to easily remove themselves and their data from the system without questions or hassle. The system doesn't "remember" a user for later, but completely deletes data at the user's discretion.

#### People have a right to privacy

Sense Networks respects the privacy and anonymity of its users and requires no personally identifiable information to access its consumer applications. We never share specific user data with anyone. And we use best practices to ensure the safekeeping of the data we receive.

#### People should receive a meaningful benefit in exchange for sharing data

Meaningful benefits include compelling applications to help manage life better, or personalized services based on anonymous learning from "users like me." People should be able to enjoy the benefits of these services simply in exchange for their data.

#### Aggregate anonymous location data should be used for common good

Sense Networks is working with thought leaders at institutions such as MIT and Columbia University to explore ways of leveraging aggregate, anonymous data for the common good. For example, we're forging innovative partnerships with recycling companies that can use the data to more efficiently direct recycling resources to high activity locations in a given city.

We're looking for additional common good uses of aggregate, anonymous location data. If you would like to submit a project for consideration, please contact us at: [commongood@sensenetworks.com](mailto:commongood@sensenetworks.com).