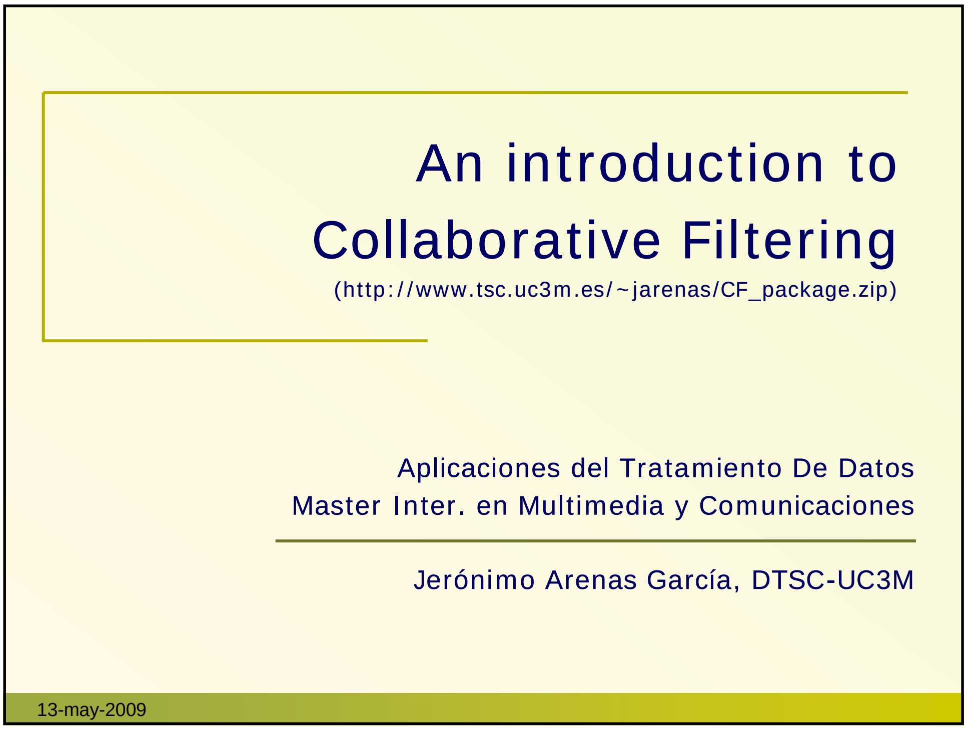
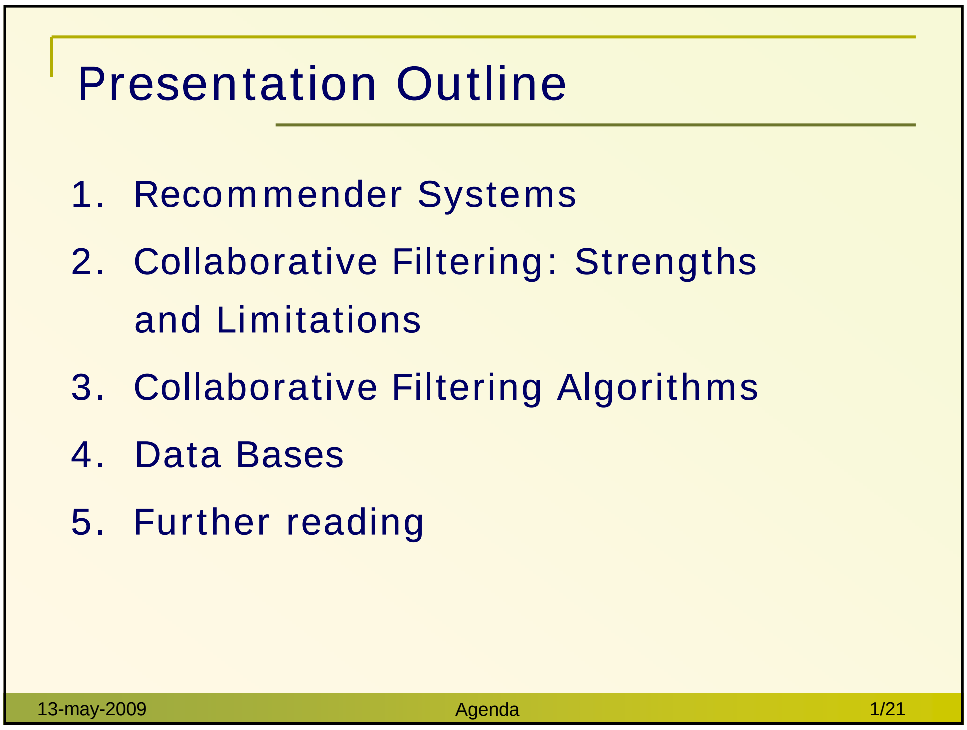


An introduction to Collaborative Filtering

(http://www.tsc.uc3m.es/~jarenas/CF_package.zip)

Aplicaciones del Tratamiento De Datos
Master Inter. en Multimedia y Comunicaciones

Jerónimo Arenas García, DTSC-UC3M

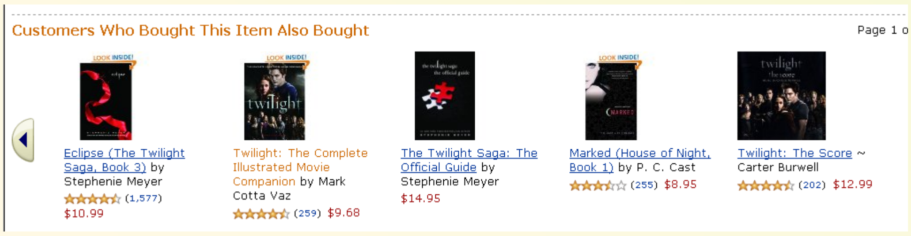


Presentation Outline

1. Recommender Systems
2. Collaborative Filtering: Strengths and Limitations
3. Collaborative Filtering Algorithms
4. Data Bases
5. Further reading

Recommender Systems: Examples

Definition: Systems that apply knowledge discovery techniques for making personalized recommendations for information, products or services, etc, usually during a live interaction



Customers Who Bought This Item Also Bought

Page 1 o

Eclipse (The Twilight Saga, Book 3) by Stephenie Meyer ★★★★☆ (1,577) $10.99

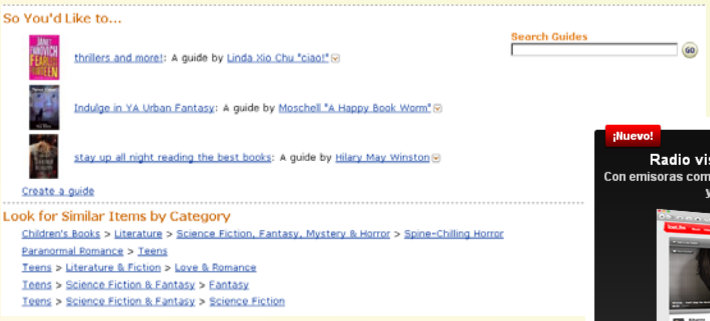
Twilight: The Complete Illustrated Movie Companion by Mark Cotta Vaz ★★★★☆ (259) $9.68

The Twilight Saga: The Official Guide by Stephenie Meyer $14.95

Marked (House of Night, Book 1) by P. C. Cast ★★★☆☆ (255) $8.95

Twilight: The Score ~ Carter Burwell ★★★★☆ (202) $12.99

Recommender Systems: Examples



So You'd Like to...

thrillers and more!: A guide by Linda Xio Chu "ciao!"

Indulge in YA Urban Fantasy: A guide by Moschell "A Happy Book Worm"

stay up all night reading the best books: A guide by Hilary May Winston

Create a guide

Search Guides

Look for Similar Items by Category

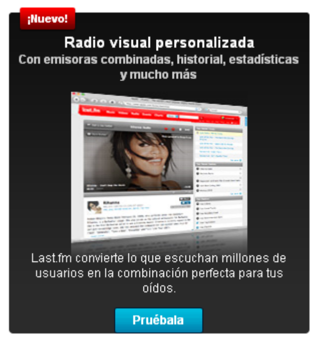
Children's Books > Literature > Science Fiction, Fantasy, Mystery & Horror > Spine-Chilling Horror

Paranormal Romance > Teens

Teens > Literature & Fiction > Love & Romance

Teens > Science Fiction & Fantasy > Fantasy

Teens > Science Fiction & Fantasy > Science Fiction



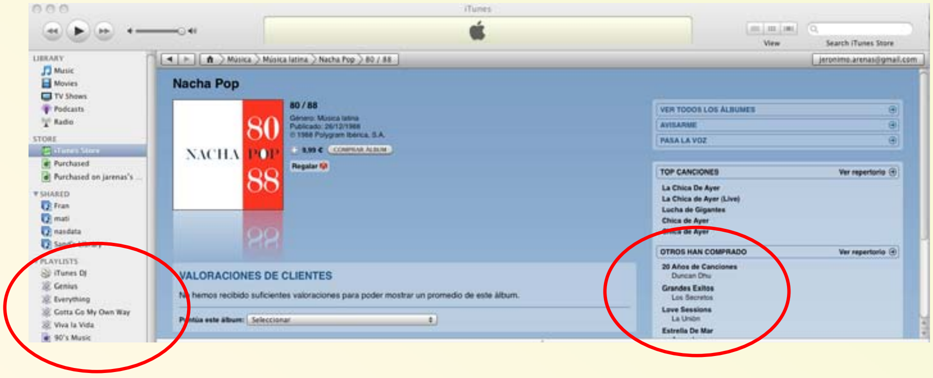
¡Nuevo!

Radio visual personalizada

Con emisoras combinadas, historial, estadísticas y mucho más

Last.fm convierte lo que escuchan millones de usuarios en la combinación perfecta para tus oídos.

Recommender Systems: Examples



Nacha Pop

80 / 88
Género: Música latina
Publicado: 26/12/1988
© 1988 Polygram Ibérica, S.A.
9,99 €

VALORACIONES DE CLIENTES

No hemos recibido suficientes valoraciones para poder mostrar un promedio de este álbum.

Puntúa este álbum: Seleccionar

TOP CANCIONES

La Chica De Ayer
La Chica de Ayer (Live)
Lucha de Gigantes
Chica de Ayer
Chica de Ayer

OTROS HAN COMPRADO

20 Años de Canciones
Duncan Dhu
Grandes Exitos
Los Secretos
Love Sessions
La Unión
Estrella De Mar

Recommender Systems: Available Info

- Contents Metadata
- Users' evaluation: either explicit or implicit
- Socio-demographic user information

Normally, there is a lot of missing information. The goal of a recommender system is, precisely, to predict users' interest on unrated items. In other words: to fill in the blank positions in the user-item evaluation table.

Recommender Systems: Taxonomy

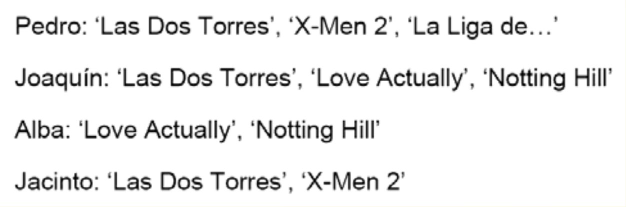
Depending on which information is used to predict users' interest, we can classify them as:

- Content filtering systems: a user is recommended items similar to those he/she liked in the past
- Collaborative filtering systems: a user is recommended items positively ranked by users that usually agree with the active user
- Socio-demographic filtering systems: a user is recommended items positively ranked by users with a similar socio-demographic profile

  (not a very generally applied approach)
- Hybrid systems: exploit strengths and limitations of the different approaches

Collaborative Filtering

Definition: trying to predict the opinion the user will have on the different items, and be able to recommend the "best" items to each user based on the user's previous likings and the opinions of other like-minded users



Pedro: 'Las Dos Torres', 'X-Men 2', 'La Liga de…'

Joaquín: 'Las Dos Torres', 'Love Actually', 'Notting Hill'

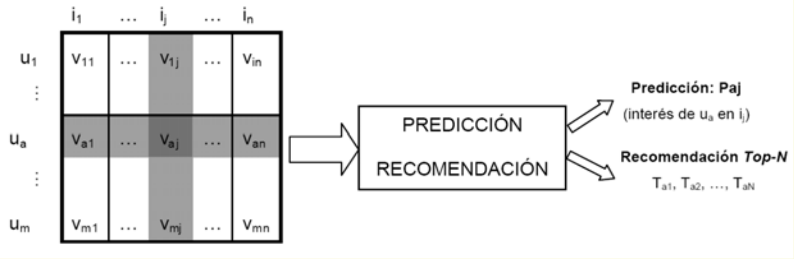
Alba: 'Love Actually', 'Notting Hill'

Jacinto: 'Las Dos Torres', 'X-Men 2'

Underlying assumption: users' who have agreed in the past will tend to agree again in the future …

Collaborative Filtering: Available data

Only user-item ratings are necessary:

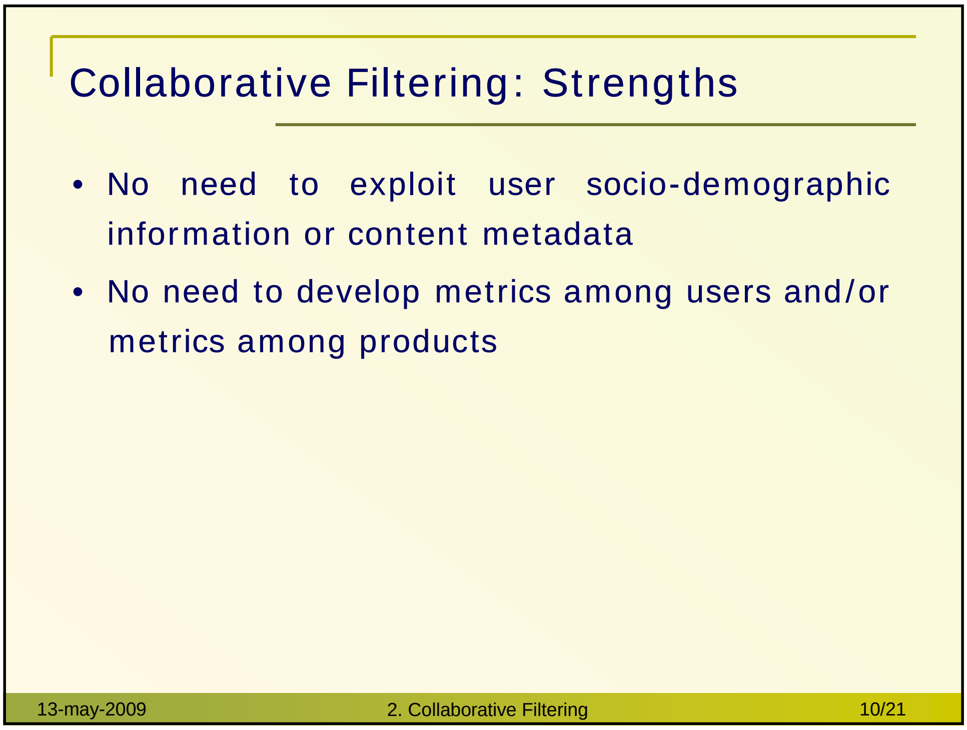


	i_1	…	i_j	…	i_n
u_1	v_{11}	…	v_{1j}	…	v_{in}
⋮					
u_a	v_{a1}	…	v_{aj}	…	v_{an}
⋮					
u_m	v_{m1}	…	v_{mj}	…	v_{mn}

- Each user has a list of items he/she expressed their opinion about (can be a null set)
- Explicit opinions: a rating score (numerical value)
- Implicit opinions (e.g., purchase records)

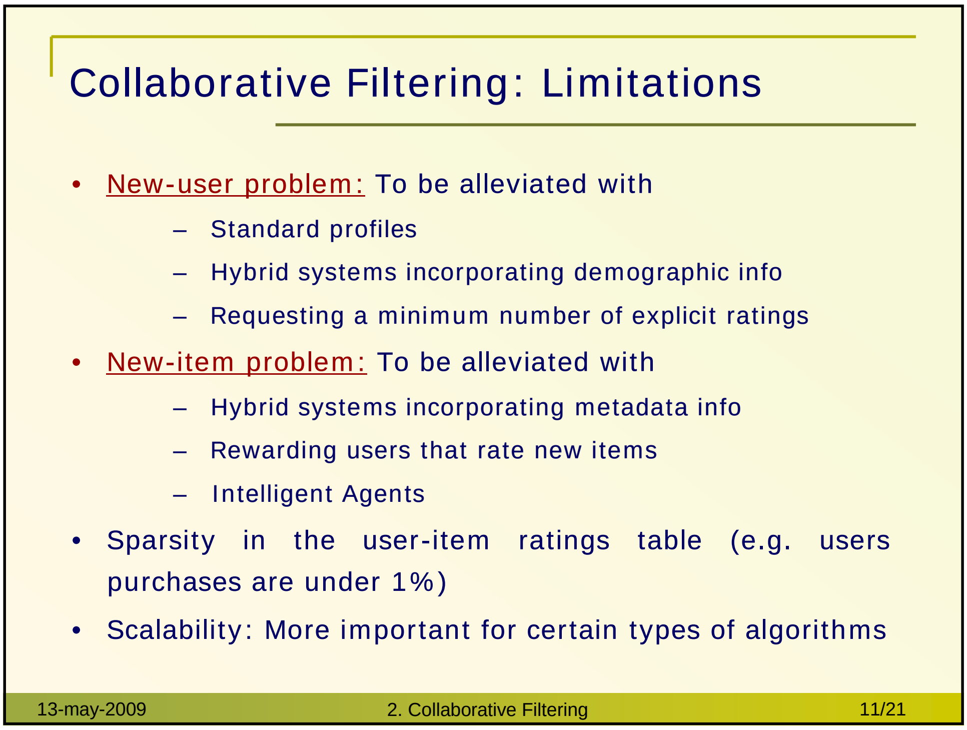
Collaborative Filtering: Applications

- Recommender Systems: inferred ratings can be used in information-pull and info-push environments. Tracking of "typical" user tastes paths
- Social Network: establishing interactions among users with similar tastes
- "Breaking reputation" systems: reducing the importance of direct marketing and well-known trademarks
- Production of new products/services
- Personalized information representation: highlighting the information which is more relevant for the user



Collaborative Filtering: Strengths

- No need to exploit user socio-demographic information or content metadata
- No need to develop metrics among users and/or metrics among products



Collaborative Filtering: Limitations

- New-user problem: To be alleviated with
  - Standard profiles
  - Hybrid systems incorporating demographic info
  - Requesting a minimum number of explicit ratings
- New-item problem: To be alleviated with
  - Hybrid systems incorporating metadata info
  - Rewarding users that rate new items
  - Intelligent Agents
- Sparsity in the user-item ratings table (e.g. users purchases are under 1%)
- Scalability: More important for certain types of algorithms

Collaborative Filtering Algorithms

CF algorithms can be classified according to 2 different criteria

CRITERION 1

- User-to-user: users similarity is first inferred based on their ratings. A user is then recommended items liked by similar users
- Item-to-item: contents similarity is first inferred based on user ratings only. A user is then recommended items similar to the ones he/she likes

CRITERION 2

- Memory-based: during the recommendation phase it is necessary to access the ratings of all users
- Model-based: the algorithm learns a model for carrying out predictions during the operational phase

User-to-user CF

We study first a “standard” memory-based algorithm:

1) Computing user similarity (off-line)

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}}$$

(Correlation)

$$w(a,i) = \frac{\sum_{j \in I_{ai}} v_{a,j} v_{i,j}}{\sqrt{\sum_{j \in I_{ai}} v_{a,j}^2} \sqrt{\sum_{j \in I_{ai}} v_{i,j}^2}}$$

(Cosine Distance)

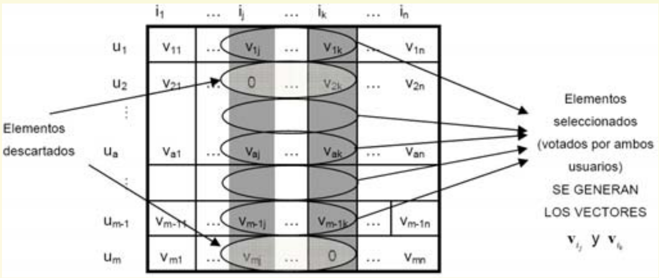
2) Computing predictions (on-line)

$$p_{a,j} = K \sum_{i=1}^{m} w(a,i) v_{i,j}$$

$$p_{a,j} = \bar{v}_a + K \sum_{i=1}^{m} w(a,i)(v_{i,j} - \bar{v}_i)$$

- Default voting: for non-rated items
- Inverse frequency: Users who like most items or items which are liked by most users are not very discriminative
- Case amplification: Extreme ratings are more informative

Item-to-item CF


i_1 … i_j … i_k … i_n
u_1
u_2
u_a
u_{m-1}
u_m
v_{11} … v_{1j} … v_{1k} … v_{1n}
v_{21} … 0 … v_{2k} … v_{2n}
v_{a1} … v_{aj} … v_{ak} … v_{an}
v_{m-11} … v_{m-1j} … v_{m-1k} … v_{m-1n}
v_{m1} … v_{mj} … 0 … v_{mn}
Elementos descartados
Elementos seleccionados (votados por ambos usuarios)
SE GENERAN LOS VECTORES
$\mathbf{v}_{i_j}$ y $\mathbf{v}_{i_k}$

1) Computing item similarity (off-line)

$$s(i_j,i_k)=\frac{\mathbf{v}_{i_j}^T\mathbf{v}_{i_k}}{\left\|\mathbf{v}_{i_j}\right\|^2\left\|\mathbf{v}_{i_k}\right\|^2}$$

$$s(i_j,i_k)=\frac{\sum_{i\in I_{jk}}\left(v_{i,j}-\bar{\mathbf{v}}_{i_j}\right)\left(v_{i,k}-\bar{\mathbf{v}}_{i_k}\right)}{\sqrt{\sum_{i\in I_{jk}}\left(v_{i,j}-\bar{\mathbf{v}}_{i_j}\right)^2\sum_{i\in I}\left(v_{i,k}-\bar{\mathbf{v}}_{i_k}\right)^2}}$$

$$s(i_j,i_k)=\frac{\sum_{i\in I_{jk}}\left(v_{i,j}-\bar{\mathbf{v}}_{u_i}\right)\left(v_{i,k}-\bar{\mathbf{v}}_{u_i}\right)}{\sqrt{\sum_{i\in I_{jk}}\left(v_{i,j}-\bar{\mathbf{v}}_{u_i}\right)^2\sum_{i\in I}\left(v_{i,k}-\bar{\mathbf{v}}_{u_i}\right)^2}}$$

2) Computing predictions (on-line)

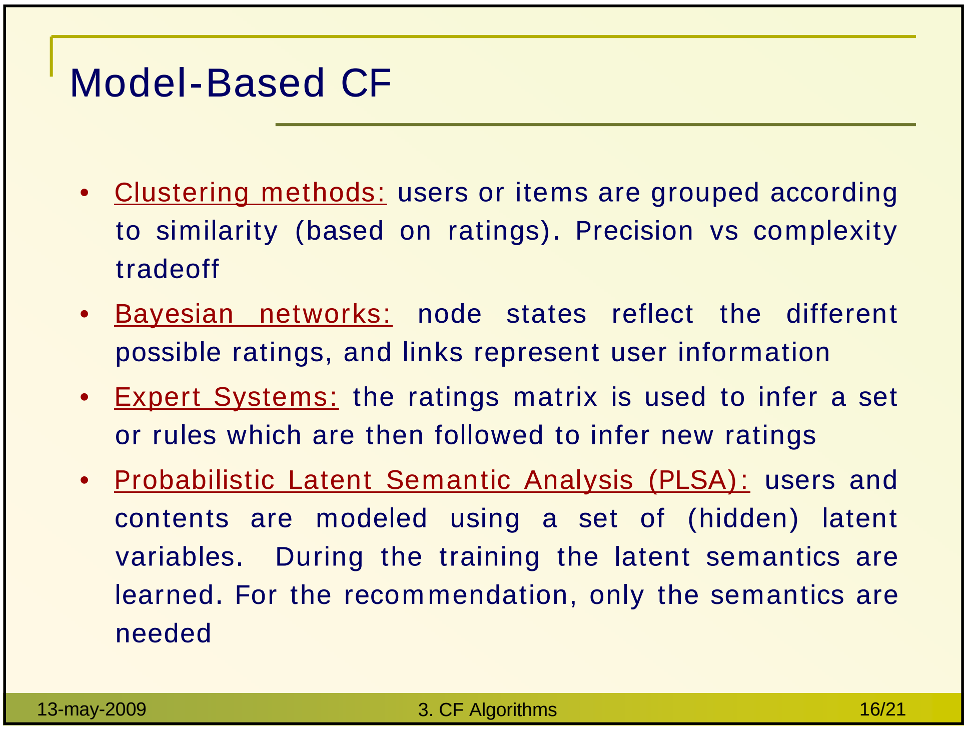
$$p_{aj}=\frac{\sum_{k\in Ia}s(i_j,i_k)v_{ak}}{\sum_{k\in Ia}\left|v_{ak}\right|}$$

Neighborhood methods

Use statistical tools to find a set of neighbors with a similar profile to the active user, simplifying the sum for computing predictions (close in philosophy to model-based CF algorithms). Improved scalability and performance.

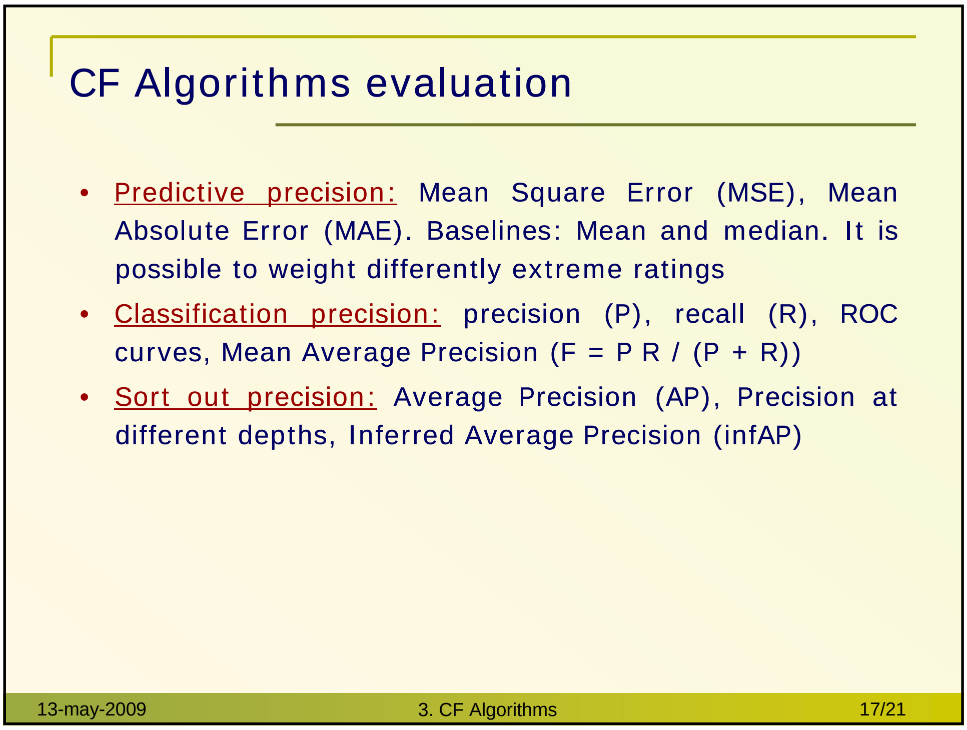
Neighbors can be found:

- Thresholding similarity weights (non-uniform neighborhood size)
- K-NN (selection of K is critical)
- Clustering methods



Model-Based CF

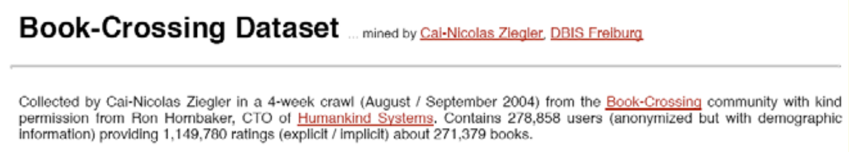
- Clustering methods: users or items are grouped according to similarity (based on ratings). Precision vs complexity tradeoff
- Bayesian networks: node states reflect the different possible ratings, and links represent user information
- Expert Systems: the ratings matrix is used to infer a set or rules which are then followed to infer new ratings
- Probabilistic Latent Semantic Analysis (PLSA): users and contents are modeled using a set of (hidden) latent variables. During the training the latent semantics are learned. For the recommendation, only the semantics are needed



CF Algorithms evaluation

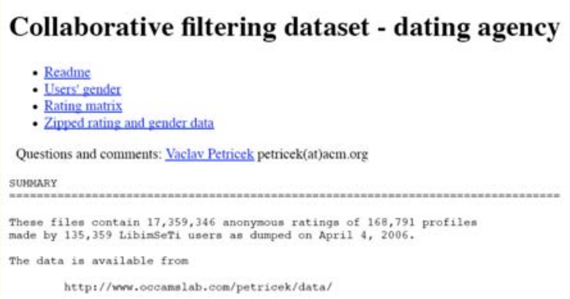
- Predictive precision: Mean Square Error (MSE), Mean Absolute Error (MAE). Baselines: Mean and median. It is possible to weight differently extreme ratings
- Classification precision: precision (P), recall (R), ROC curves, Mean Average Precision (F = P R / (P + R))
- Sort out precision: Average Precision (AP), Precision at different depths, Inferred Average Precision (infAP)

Data Bases for Collaborative Filtering



Book-Crossing Dataset ... mined by Cai-Nicolas Ziegler, DBIS Freiburg

Collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.



Collaborative filtering dataset - dating agency

- Readme
- Users' gender
- Rating matrix
- Zipped rating and gender data

Questions and comments: Vaclav Petricek petricek(at)acm.org

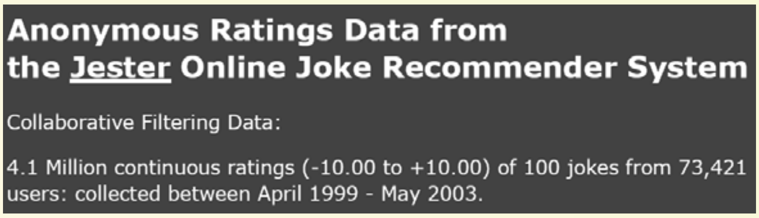
```
SUMMARY
================================================================================

These files contain 17,359,346 anonymous ratings of 168,791 profiles
made by 135,359 LibimSeTi users as dumped on April 4, 2006.

The data is available from

        http://www.occamslab.com/petricek/data/
```

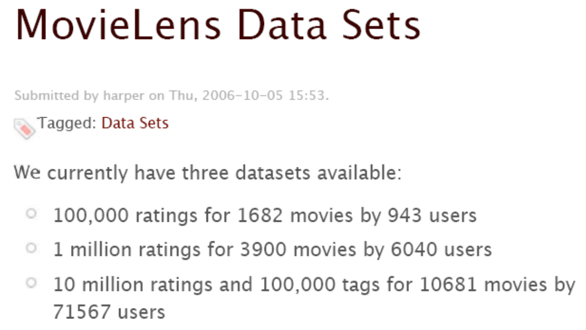
Data Bases for Collaborative Filtering (II)



Anonymous Ratings Data from the Jester Online Joke Recommender System

Collaborative Filtering Data:

4.1 Million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users: collected between April 1999 - May 2003.



MovieLens Data Sets

Submitted by harper on Thu, 2006-10-05 15:53.

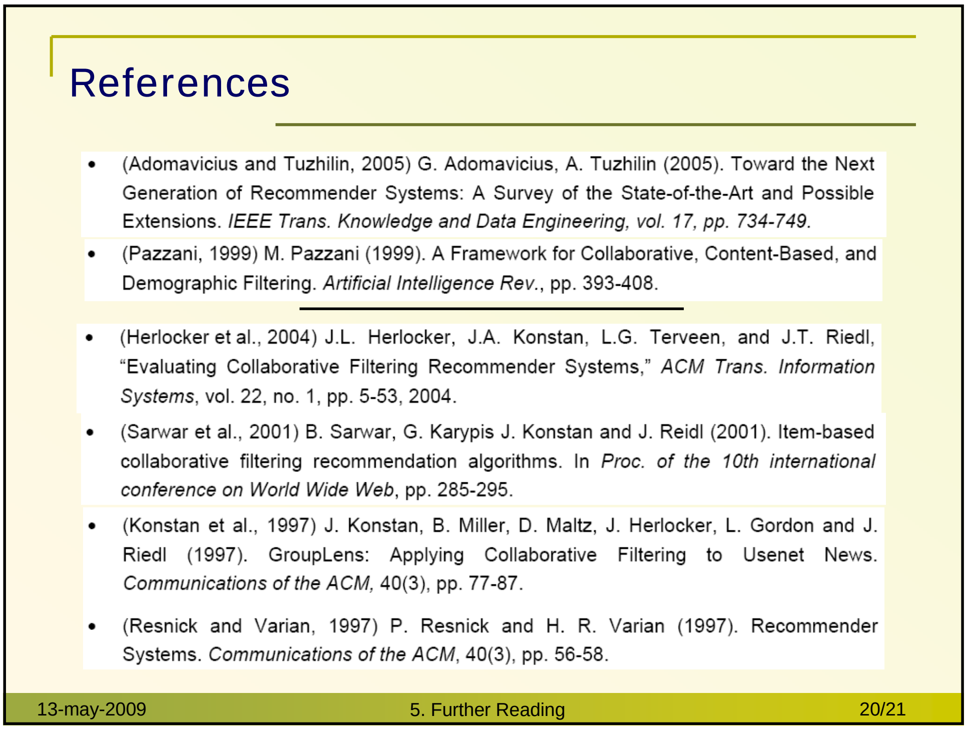
Tagged: Data Sets

We currently have three datasets available:

○ 100,000 ratings for 1682 movies by 943 users

○ 1 million ratings for 3900 movies by 6040 users

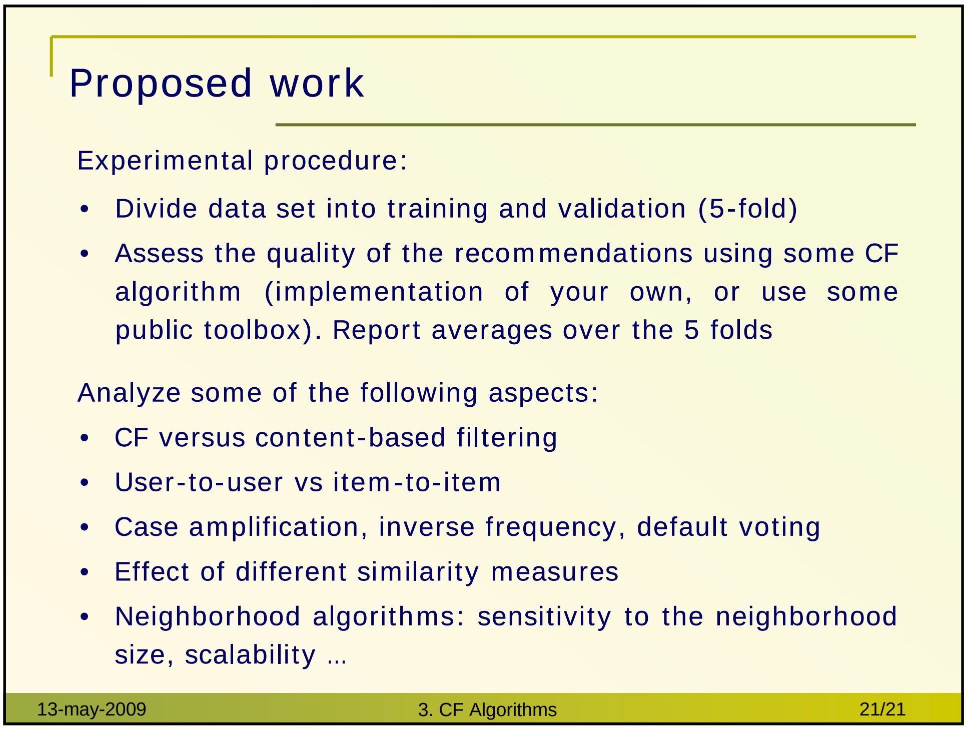
○ 10 million ratings and 100,000 tags for 10681 movies by 71567 users



References

- (Adomavicius and Tuzhilin, 2005) G. Adomavicius, A. Tuzhilin (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. Knowledge and Data Engineering, vol. 17, pp. 734-749.*
- (Pazzani, 1999) M. Pazzani (1999). A Framework for Collaborative, Content-Based, and Demographic Filtering. *Artificial Intelligence Rev.*, pp. 393-408.
- (Herlocker et al., 2004) J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Information Systems*, vol. 22, no. 1, pp. 5-53, 2004.
- (Sarwar et al., 2001) B. Sarwar, G. Karypis J. Konstan and J. Reidl (2001). Item-based collaborative filtering recommendation algorithms. In *Proc. of the 10th international conference on World Wide Web*, pp. 285-295.
- (Konstan et al., 1997) J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM,* 40(3), pp. 77-87.
- (Resnick and Varian, 1997) P. Resnick and H. R. Varian (1997). Recommender Systems. *Communications of the ACM*, 40(3), pp. 56-58.

13-may-2009 5. Further Reading 20/21



Proposed work

Experimental procedure:

- Divide data set into training and validation (5-fold)
- Assess the quality of the recommendations using some CF algorithm (implementation of your own, or use some public toolbox). Report averages over the 5 folds

Analyze some of the following aspects:

- CF versus content-based filtering
- User-to-user vs item-to-item
- Case amplification, inverse frequency, default voting
- Effect of different similarity measures
- Neighborhood algorithms: sensitivity to the neighborhood size, scalability …

13-may-2009 3. CF Algorithms 21/21