Fundamental concepts of similarity

Euclidean Distance

Cosine Similarity:

Cosine Similarity is the cosine of the angle between the 2 vectors A and B

Closer the vectors, smaller will be the angle and hence larger their cosine value.

Pearson Similarity:

Pearson similarity is also another measure of finding similarity between two vectors.

Pearson correlation and cosine similarity are invariant to scaling.

Cosine similarity is NOT invariant to shifts. Pearson correlation is invariant to shifts.

$$Inner(x,y) = \sum_i x_i y_i = \langle x, y \rangle$$

$$CosSim(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x, y \rangle}{||x||\ ||y||}$$

$$\begin{aligned} Corr(x,y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{||x - \bar{x}||\ ||y - \bar{y}||} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

Correlation is the cosine similarity between centered versions of x and y & between -1 to 1

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Fundamentals

Jaccard Similarity:

Similarity is defined as the number of users which have rated item A and B divided by the number of users who have rated either A or B

Intersection over Union

It is used for comparing both similarity and diversity of two sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

If x,y are two vectors with all real xi, yi then jaccardian similarity coefficient:

$$J(\mathbf{x},\mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)},$$

The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Collaborative Filtering

Idea: If a person X likes items A, B, C and Y like B,C,D then they have similar interests and **X should like item D and Y should like item A.**

This algorithm is entirely based on the user’s past behaviour and not on the context.

This makes it one of the **most commonly used algorithm** and is not dependent on any additional information.

Basic **assumptions:**

- Customers who had similar tastes in the past, will have similar tastes in the future
- Users give ratings to catalog items (implicitly or explicitly)

Examples:

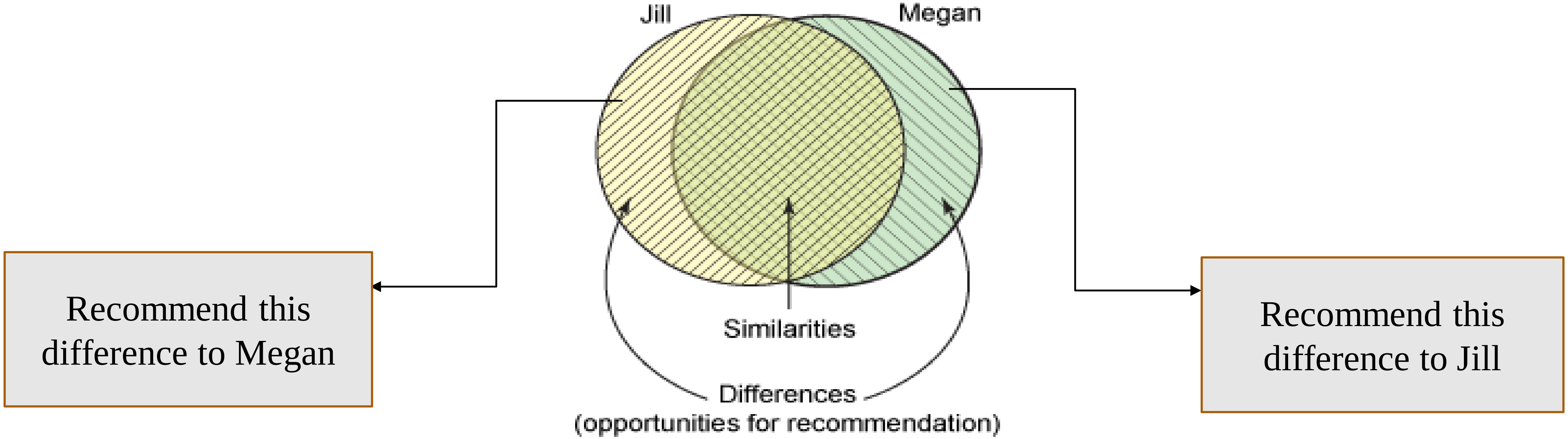
- Product recommendations by e-commerce player like Amazon and merchant recommendations by banks like American Express.

User-User Collaborative filtering

Item-Item Collaborative filtering

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Collaborative Filtering



Recommend this difference to Megan

Recommend this difference to Jill

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

greatlearning
Learning for Life

User based nearest neighbour Collaborative filtering

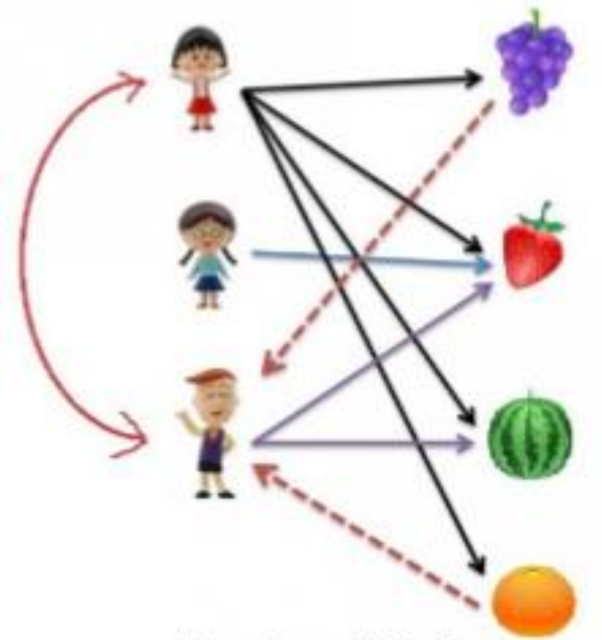
Find out the users who have a similar taste of products as the current user chosen.

Similarity is based upon similarity in their purchasing behaviour.

"User A is similar to user B because both purchased items x, y and z."

Memory-based: the rating matrix is directly used to find neighbours / make predictions

Does not scale well for most real-world scenarios

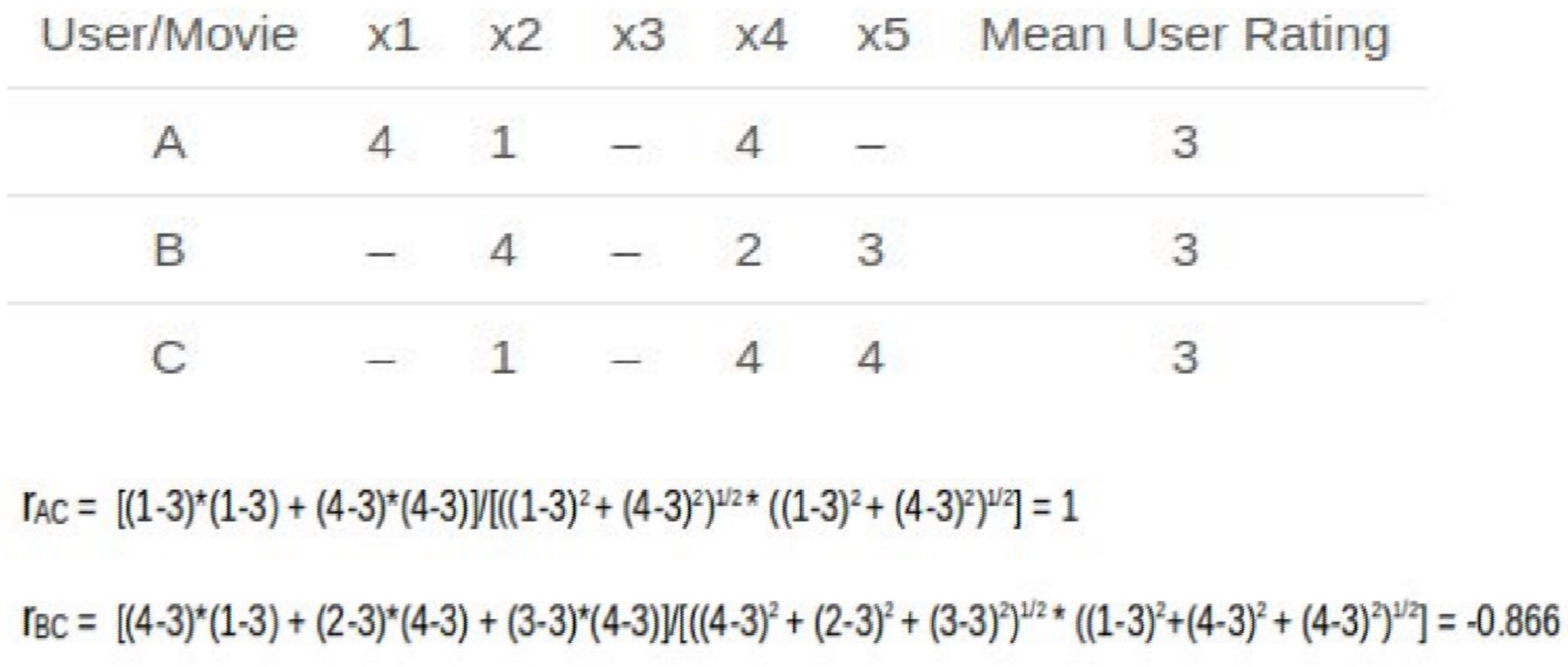


User-based filtering

Source: http://www.salemmarafi.com/code/collaborative-filtering-with-python/

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

User-User collaborative filtering



User/Movie	x1	x2	x3	x4	x5	Mean User Rating
A	4	1	–	4	–	3
B	–	4	–	2	3	3
C	–	1	–	4	4	3

$$r_{AC} = [(1-3)^*(1-3) + (4-3)^*(4-3)]/[((1-3)^2 + (4-3)^2)^{1/2} * ((1-3)^2 + (4-3)^2)^{1/2}] = 1$$

$$r_{BC} = [(4-3)^*(1-3) + (2-3)^*(4-3) + (3-3)^*(4-3)]/[((4-3)^2 + (2-3)^2 + (3-3)^2)^{1/2} * ((1-3)^2+(4-3)^2 + (4-3)^2)^{1/2}] = -0.866$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

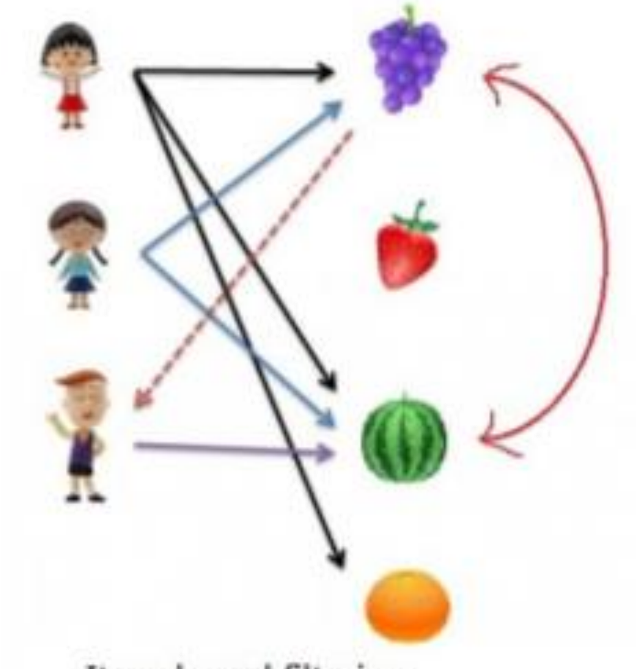
Item based nearest-neighbour collaborative filtering

Reverse of user based collaborative filtering

Recommend items to the user that are similar to the items the user has bought.

Similarity is based upon co-occurence of purchases:

- Use the similarity between items (and not users) to make predictions



Item-based filtering

“Items **X** and y were purchased by both users A and B, so they are similar.”

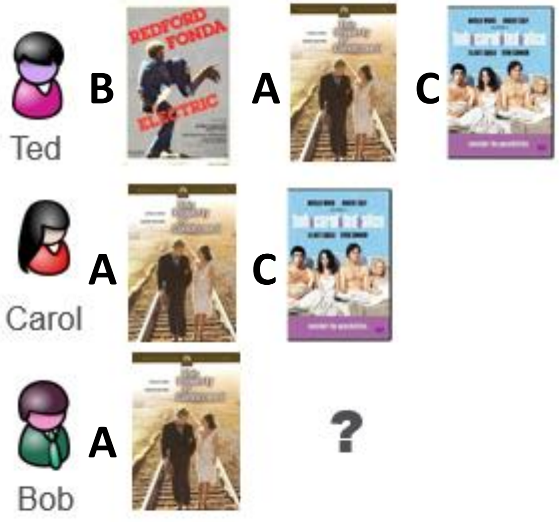
- *Item*-based CF is an example for model-based approaches

Source: http://www.salemmarafi.com/code/collaborative-filtering-with-python/

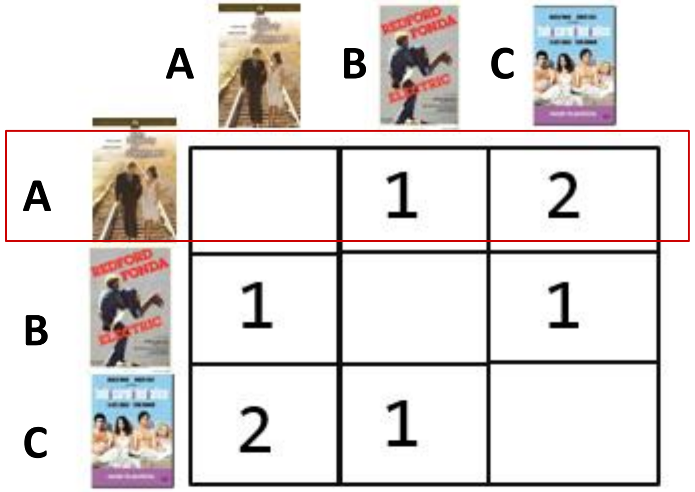
Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Item based collaborative filtering: Example

History Matrix



Co-occurrence Matrix: Items by Items



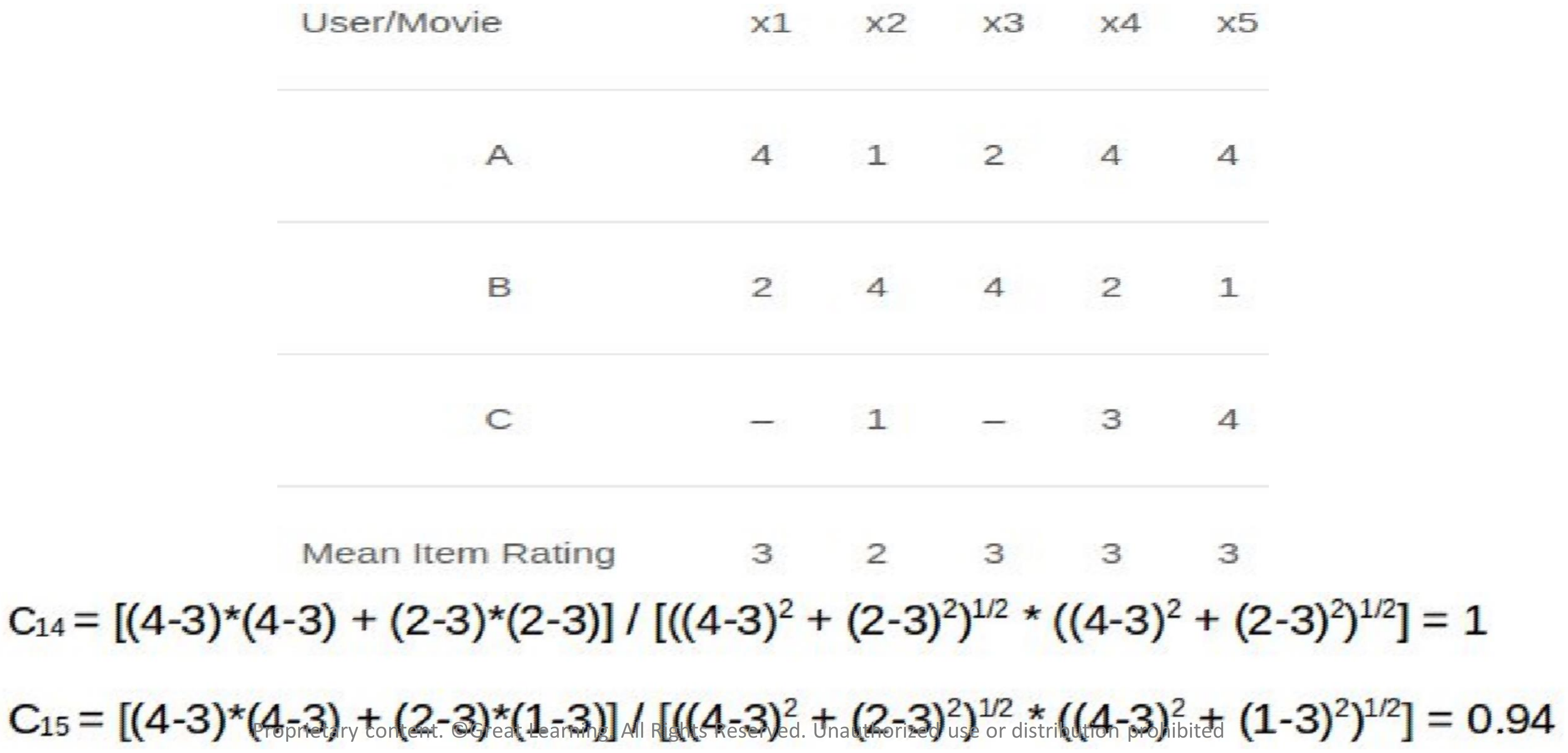
	A	B	C
A		1	2
B	1		1
C	2	1	

Bob’s Recommendations= [C, B]

Source: https://www.mapr.com/blog/inside-look-at-components-of-recommendation-engine

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Item-Item collaborative filtering



User/Movie	x1	x2	x3	x4	x5
A	4	1	2	4	4
B	2	4	4	2	1
C	–	1	–	3	4
Mean Item Rating	3	2	3	3	3

$$C_{14} = [(4-3)*(4-3) + (2-3)*(2-3)] / [((4-3)^2 + (2-3)^2)^{1/2} * ((4-3)^2 + (2-3)^2)^{1/2}] = 1$$

$$C_{15} = [(4-3)*(4-3) + (2-3)*(1-3)] / [((4-3)^2 + (2-3)^2)^{1/2} * ((4-3)^2 + (1-3)^2)^{1/2}] = 0.94$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Market basket analysis

Discovers co-occurrence relationships among activities performed.

Market basket analysis can be used to **divide customers into groups**

Market basket analysis may provide the retailer with information **to understand the purchase behavior of a buyer**.

“customers who bought book A also bought book B”

When one super market chain discovered in its analysis that male customers that bought diapers often bought beer as well, have put the diapers close to beer coolers, and their sales increased dramatically.

Might tell a retailer that customers often purchase shampoo and conditioner together, so putting both items on promotion at the same time would not create a significant increase in revenue, while **a promotion involving just one of the items** would likely drive sales of the other.

Association rule mining

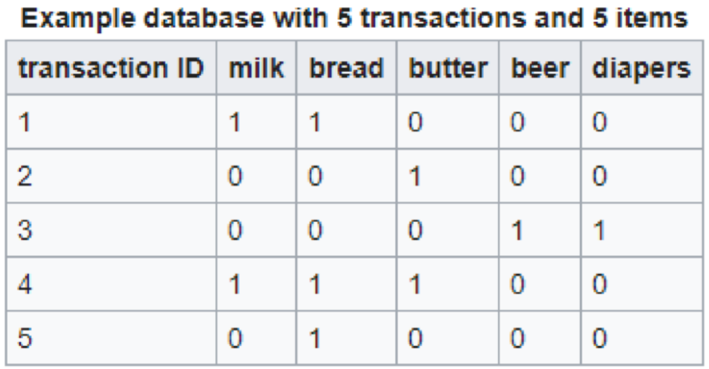
Rules of the form x -> Y

from a set of sale transactions.

It's common use in shopping behaviour analysis

Can be used as the basis for d**ecisions about marketing activities** such as, e.g., promotional pricing or product placements.

It is NOT sequence mining: Association rule learning typically does not consider the order of items either within a transaction or across transactions.



Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

$$I = \{\mathrm{milk}, \mathrm{bread}, \mathrm{butter}, \mathrm{beer}, \mathrm{diapers}\}$$

$$\{\mathrm{butter}, \mathrm{bread}\} \Rightarrow \{\mathrm{milk}\}$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Performance metrics

- RMSE
- MAE –Mean Absolute Error
- Accuracy
- ROC curve
- Precision
- Recall
- Precision Recall Curve: Evaluation of top n recommendations

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Evaluation measures

Out of all the recommended items, how many user actually liked the recommendations?

What ratio of items that a user likes were actually recommended.

- **Precision**: Out of all items retrieved, how many are relevant
  - It is based on true positives and false positives.
  - It shows how many selected items are actually also relevant.
  - True positives (TP) are the positive guesses made that were actually correct while the false positives (FP) are the positive guesses made that were incorrect.

$$Precision = \frac{TP}{(TP + FP)}$$

- **Recall**: How many relevant items retrieved from all relevant items
  - is based on true positives and false negatives.
  - It shows how many of the relevant information is selected.
  - The true positives (TP) are again the positive guesses made that were actually correct while the false negatives (FN) are negative guesses while they should have been positive.

$$Recall = \frac{TP}{(TP + FN)}$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Evaluation measures

- **Accuracy**: Percentage guessed correct from total guesses
  - method that is based on exact matches.
  - This is the most simple evaluation method, but may perform poorly based on the data.
  - If the labels of your data mostly consist of one specific cluster, the classifiers may classify all the data as that cluster. Since the data contains mostly that cluster, its accuracy will be high due to one classification.
  - This may cause promising results, but they are not correct. So for this measure, it is necessary to have information about the labels of your data. The number itself is not enough for performance conclusions.

$$Accuracy = \frac{Good\ guesses}{Total\ Guesses} \cdot 100$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

RoC and Precision Recall curve

Receiver operating characteristics curve.

A plot of true positive fraction (= sensitivity) vs. false positive fraction (= 1 – specificity) for all potential cut-offs for a test.

A ROC curve plots recall (true positive rate) against fallout (false positive rate) for increasing recommendation set size.

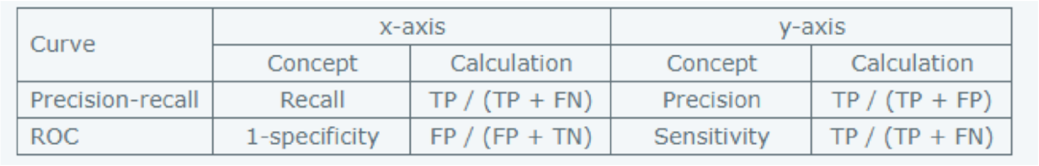
In Top-20 recommender example,

The 20 items you recommend for a user are the Positive items, and the unrecommended items are Negative.

Precision Recall Curve:

A precision-recall curve shows the relationship between precision (= positive predictive value) and recall (= sensitivity) for every possible cut-off.

The main difference between ROC curves and precision-recall curves is that the number of true-negative results is not used for making a PRC.



Curve	x-axis		y-axis	
	Concept	Calculation	Concept	Calculation
Precision-recall	Recall	TP / (TP + FN)	Precision	TP / (TP + FP)
ROC	1-specificity	FP / (FP + TN)	Sensitivity	TP / (TP + FN)

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Hybrid recommender systems

Multiple recommender systems are combined to improve recommendations

- Although any type of recommender systems can be combined a common approach in industry is to combine **content based approaches and collaborative filtering approaches**
- Content based models can be used to solve **the Cold Start and Gray Sheep problems** in Collaborative Filtering
- Some of the typical methods of Hybridization include
  - **Weighted** –Recommendations from each system is weighted to calculate final recommendation
  - **Switching** –System switches between different recommendation model
  - **Mixed** - Recommendations from different recommenders are presented together

A common approach is to use **Latent Factor models for high level recommendation** and then **improving them using content based systems** by using information on users or items

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Summary

- Basics of recommendation system
- Popularity based recommendations
- Classification model based
- Content based recommendations
- Nearest neighbour collaborative filtering:
  - User based
  - Item based
- Hybrid approches
- Python examples
- Association rule mining

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Industry example of e-commerce

Events:-

- Tracks and stores on all consumer activity and behavior
- Each click on product, adding to wishlist, adding to cart, save for later and purchase

Ratings:-

- Assign implicit values on user actions
- Ratings of products from the users
- And Feedback

Filtering:-

Hybrid approach of Collaborative filtering and user based filtering

Considerations

- Sometimes not recommending or simple recommendation is the best option
- Privacy concerns in Recommendation systems
  - The case of Target Corporation
- Computational challenges in recommendation systems, can be costly to implement
  - •The final model built by winning Netflix team could not be implemented due to engineering challenges
  - •Good Data Collection, well thought out metrics are a must

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

References

1. http://dataconomy.com/an-introduction-to-recommendation-engines/

2. https://goo.gl/ehBnhf

3. https://github.com/dvysardana/RecommenderSystems_PyData_2016

4. https://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/

5. https://wiki.epfl.ch/edicpublic/documents/Candidacy%20exam/Evaluation.pdf

6. http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/

7. Gunawardana, Asela, and Guy Shani. "A survey of accuracy evaluation metrics of recommendation tasks." *Journal of Machine Learning Research*10.Dec (2009): 2935-2962.

8. Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited

Further reading

Book: Recommender Systems An Introduction by Dietmar Jannach

Book: Mining Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman (www.mmds.org)

Coursera course on Recommender Systems, by University of Washington

Coursera course on Recommender Systems, by University of Minnesota

https://dl.acm.org/citation.cfm?doid=2959100.2959166

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited