

Recommender Systems

Blaž Zupan, University of Ljubljana

amazon.de

Recommended items other customers often buy again

More recommendations for you [See more](#)

Best backpacks

Men's backpacks

New products for kitchen & household

Handmade jewellery

[See more](#)

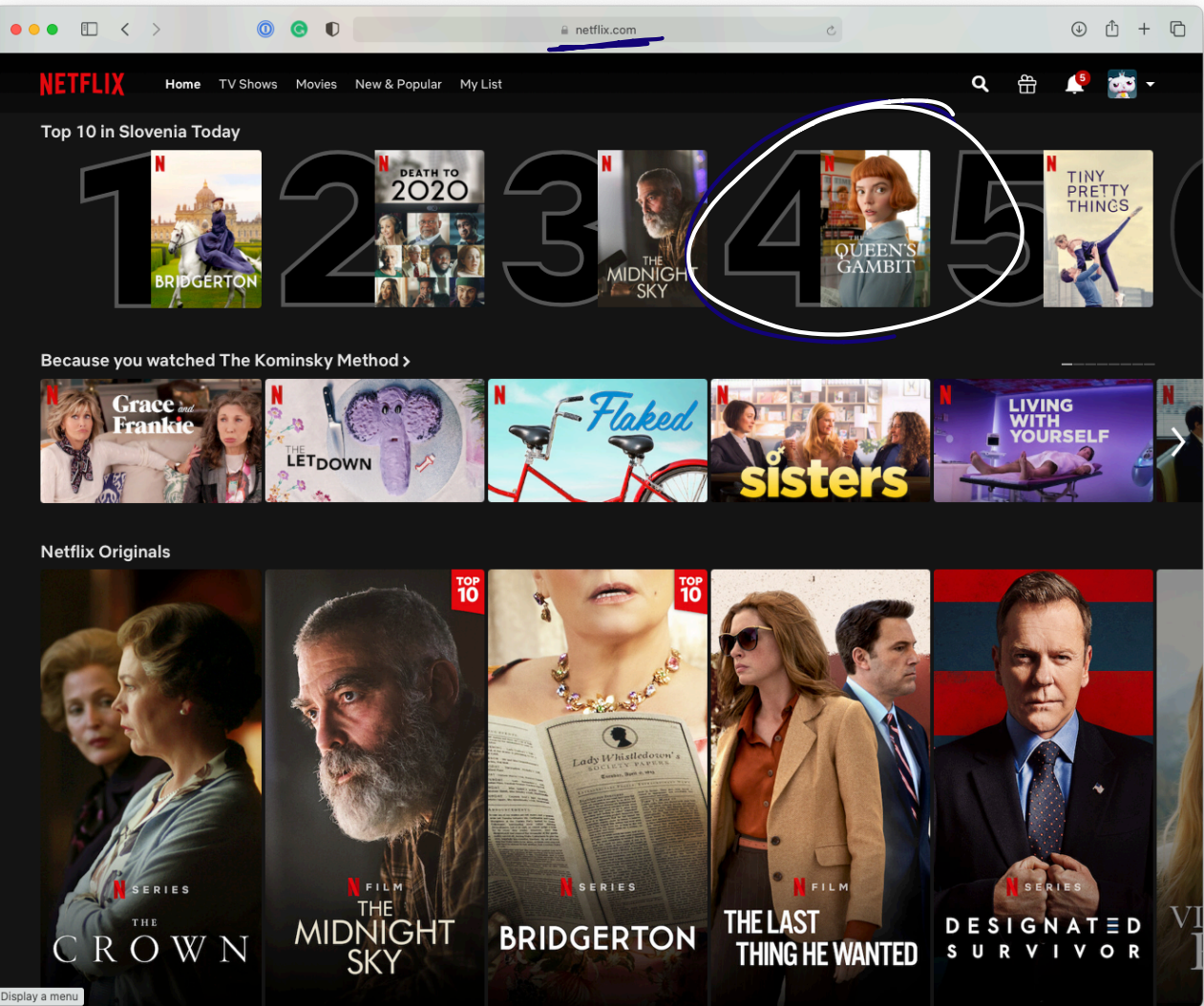
[See more](#)

[See more](#)

[To the store](#)

The screenshot displays the Amazon Germany homepage with several recommendation sections. Handwritten blue annotations are present throughout the page:

- A circle around the text "other customers often buy again" in the first recommendation section.
- An arrow pointing to a black weight plate in the first section.
- Arrows pointing to a black beanie, a pair of green socks, and a black long-sleeved shirt in the first section.
- An arrow pointing to a blue water bottle in the second section.
- Arrows pointing to a grey and a yellow backpack in the second section.
- An arrow pointing to a black long-sleeved shirt in the second section.
- An arrow pointing to a yellow backpack in the second section.
- An arrow pointing to the "New products for kitchen & household" section header.
- An arrow pointing to the "Handmade jewellery" section header.




Spletišča Mercator




Prezem v HM SMARTINSKA KlirinDvig


Zasedenost 88%

Prijava

**Mercator**
Spletna trgovina

Vpišite iskano besedo ali kliknite ikono desno za napredno iskanje

 Prijava


 Moja košarica **1**


VSI IZDELKI


RADI IMAMO DOMAČE


AKCIJE IN UGODNOSTI

ZNAMKE MERCATOR

 MMM...RECEPTI



 Priporočamo varno in priročno plačilo že ob naročilu. Za dostavo v nedeljo od 8. do 16. ure oddajte naročilo najkasneje v soboto do 18. ure.

 FILTRI

SHRANJENI

> ISKANJE

pršut

Eko/Bio izdelki

> KATEGORIJE IZDELKOV

> BLAGOVNE ZNAMKE

> AKCIJE


> CENA



> ALERGENI


> LASTNOSTI

HM SMARTINSKA KlirinDvig / VSI IZDELKI / ŠT. NAJDENIH (48)


Razvrsti po: **Prijubljenosti** ceni Najnižji ceni Najvišji ceni Najnižji teži Najvišji teži Novosti






Narezek kuhan pršut, Mercator, pakiran, 150 g
1,79 €
1 kos   V košarico






POPOLNA ČISTOČA!
Kupi zdaj




IZJAZNO -31%
Narezek pečen pršut, Kras, 100 g, pakirano
2,89 € **1,99 €**
1 kos   V košarico






IZJAZNO -31%
Narezek kuhan pršut, Kras, 100 g, pakirano
2,89 € **1,99 €**
1 kos   V košarico





Prelistajte TEDENSKI KATALOG




Narezek kuhan pršut pakiran, Citterio, 100 g
2,49 €
1 kos   V košarico




Pečen pršut z zelišči Citterio, cca 6 kg, cena za kg
22,49 €
6,00 kg   V košarico


Več


 Info

Display a menu for "https://trgovina.mercator.si/market/brskaj"

Dodaj obrnjene

 v košarico (1)



 Na blagajno (1)



data mining tools



All

Images

Videos

News

Maps

More

Settings

Tools

About 543,000,000 results (0,65 seconds)



View all

This article lists out 10 comprehensive data mining tools widely used in the big data industry.

- Rapid Miner. ...
- Oracle Data Mining. ...
- IBM SPSS Modeler. ...
- KNIME. ...
- Python. ...
- Orange. ...
- Kaggle. ...
- Rattle.

More items... • Sep 17, 2018

www.analyticsinsight.net › the-top-10-data-mining-tools...

The Top 10 Data Mining Tools of 2018 | Analytics Insight

About featured snippets • Feedback

People also ask

What are the five major types of data mining tools?



Which is the best data mining tool?



What are the most popular free data mining tools?



Is Excel a data mining tool?



Feedback

Display a menu

www.softwaretestinghelp.com › data-mining-tools

2006

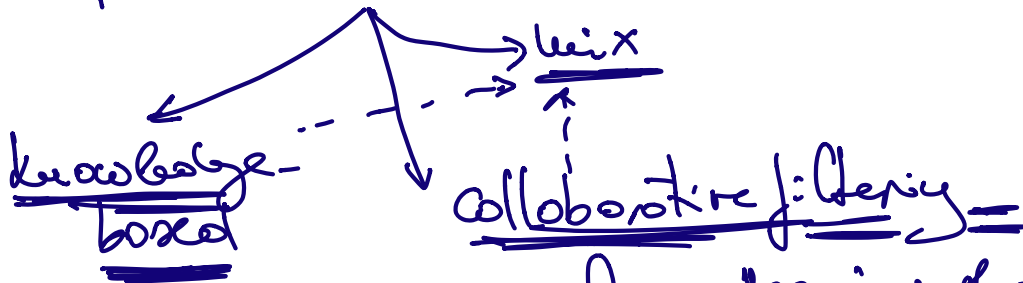


Types of Recommendations

- editorial
list of "best of" items

- simple aggregators
most popular items

- tailored to individual users



from previous shopping carts
data from previous purchases

Example of the Data / Formalism

person	salami	proscutto	eggs	jogurt	jam	cornflakes	cheese	cerials	banana	strawberries
blaž	<u>1</u>	<u>1</u>	<u>5</u>	<u>4</u>		<u>5</u>	<u>3</u>			
ivan		<u>4</u>	5							
milana	5	5	5		5	2	3	5		4
nina	2			<u>5</u>			4			
andraž	1	1	<u>4</u>	<u>4</u>	<u>3:7</u>	<u>2:3</u>	<u>1:7</u>	5	5	2
klemen	1	1	3	1	1	4	1	4	2	3
monsarat		1		4				5	5	5
mehmet	3	5		<u>5</u>	<u>5</u>	<u>5</u>	3	2	2	<u>1</u>
naja	1	1		<u>5</u>	<u>1</u>			5		
chiara	1	1			5			5		

set of users $u \in \underline{U}$
 set of items $i \in \underline{I}$
 utility scores $r \in \underline{\mathbb{R}}$

$r \in [1..5]$
 $r \in \{0, 1\}$
 $r \in \{1\}$

The Goal

data: realization of (U, I, R) \mathcal{D}
that is, we have a set of triplets
 (u, i, r)

→ estimate R with \hat{R}

$$\mathcal{L}MSE = \sqrt{\mathbb{E}\{(\hat{R} - R)^2\}}$$

that is, for every r_{ui} we would like to
predict \hat{r}_{ui} so that the error of estimator
is minimal

Some More Notation

distribution of (U, I, R) is not known
all we have is the data J' , a sample

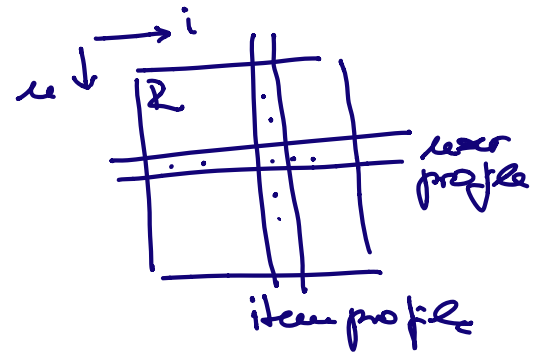
$$\underline{J} = \{ (u, i), \exists r : (u, i, r) \in J' \}$$

↑ training data

$$\hat{RMSE} = \sqrt{\frac{1}{|J|} \sum_{(u, i) \in J} (\hat{r}_{ui} - r_{ui})^2}$$

: minimize this

given u, i
estimate \hat{r}_{ui}



Key Problems

- gather the data,
the preferential data
 ↙ → implicitly
 explicitly
- extrapolate known relations (u, i)
to the unknown ones
 ↳ model, obstructions
 of the data
- evaluation,
score the quality of estimates

Similarity-Based Techniques

Baselines

= mean of the ratings

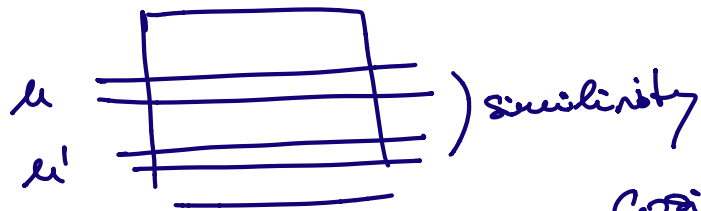
$$\hat{r}_{ui} = \frac{\sum_{(u,i') \in J} r_{ui'}}{|J|}$$

= mean of the user

$$\hat{\bar{r}}_{ui} = \frac{\sum_{i', i' \in J, i' = u} r_{ui'}}{|\{i', i' \in J, i' = u\}|} = \bar{r}_u$$

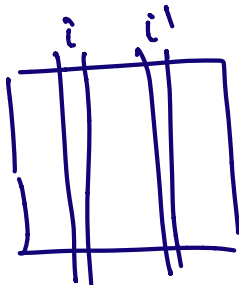
= mean of the item, \bar{r}_i

= ~~weighted~~ mean $\hat{r}_{ui} = \frac{\bar{r}_u + \bar{r}_i}{2}$



$$\rho(\mu, \mu') = \frac{\vec{r}_\mu \cdot \vec{r}_{\mu'}}{|\vec{r}_\mu| \cdot |\vec{r}_{\mu'}|} = \cos \varphi$$

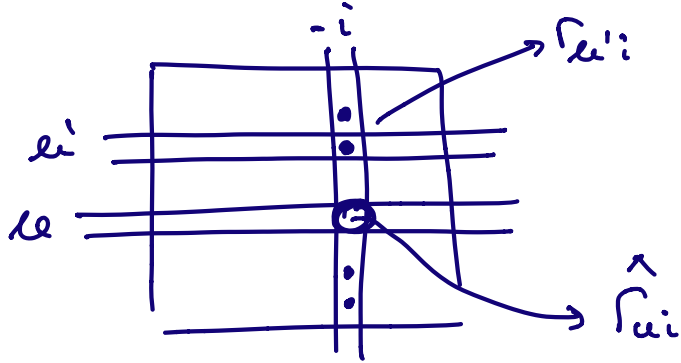
Cosine
Similarity



$$\rho(i, i') = \frac{\vec{r}_i \cdot \vec{r}_{i'}}{|\vec{r}_i| |\vec{r}_{i'}|}$$



estimate of the similarity



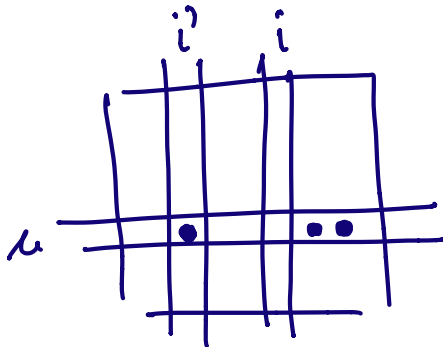
user-based estimate

$$\hat{r}_{ui} = \frac{\sum_{(u',i) \in J; u' \neq u} s(u',u) \cdot r_{u'i}}{\sum_{(u',i) \in J; u' \neq u} s(u',u)}$$

user
similarity-based
estimate

$$\sum_{(u',i) \in J; u' \neq u} s(u',u)$$

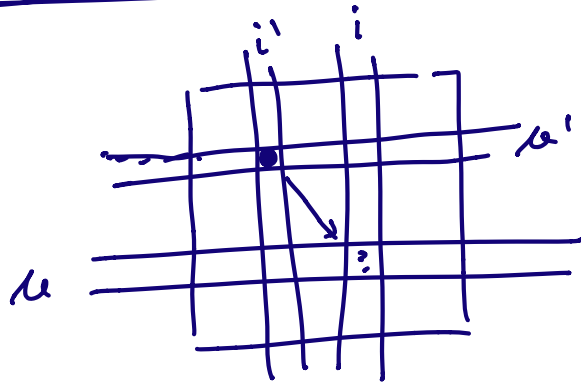
item-based estimate



$$\hat{r}_{ui} = \frac{\sum_{(u,i') \in J, i' \neq i} s(i',i) r_{ui'}}{\sum_{(u,i') \in J, i' \neq i} s(i',i)}$$

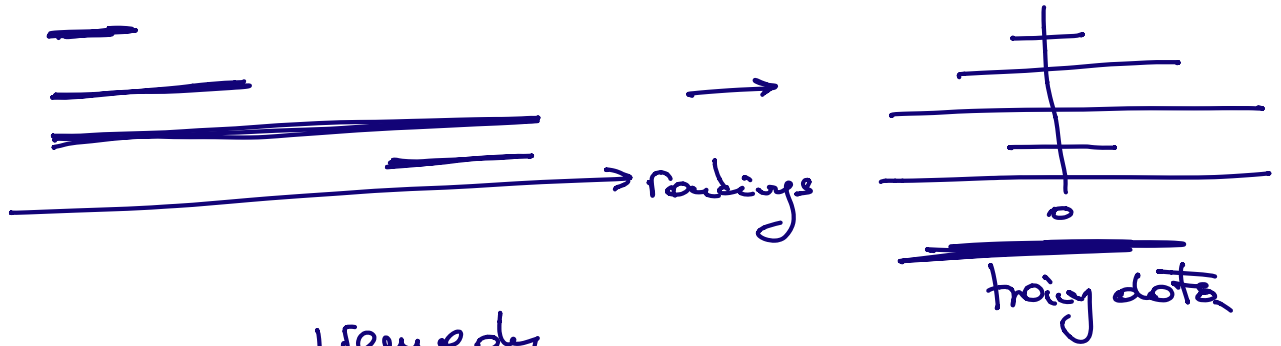
$$(\mu, i') \in \mathcal{I}; i' \neq i$$

user-item bond \leftarrow test set



$$\frac{\sum \Lambda(\mu, \mu') \cdot \Lambda(i, i') \cdot r_{\mu' i'}}{\sum \epsilon(\mu, \mu') \Lambda(i, i')}$$

Bias (Users)



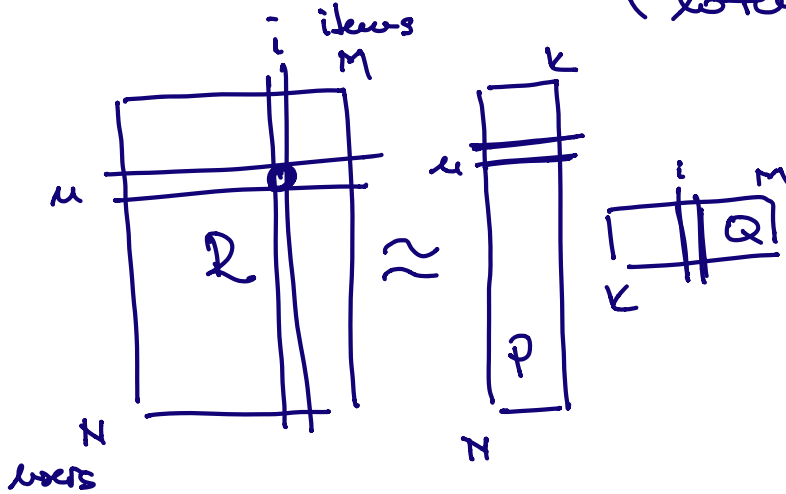
remedy

subtract the mean score
of each user from her/his scores

When we estimate the score,
we have to add the mean score of the user

Factorization-Based Approaches

(latent factors)



$$k \ll M \quad (k = 10 \dots 200)$$

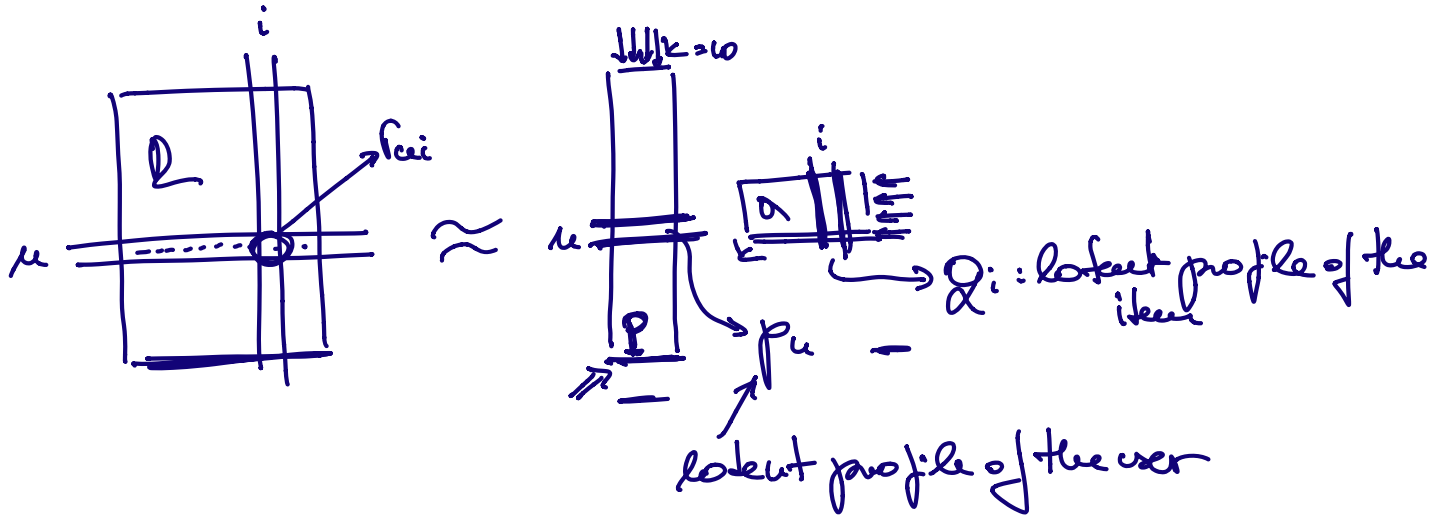
$$k \ll N \quad (\underline{k \approx 50})$$

$$R \approx PQ$$

$$\hat{R} = PQ$$

netflix prize
 480.000 users
 17.000 items

 100.000.000 ratings



$$\hat{r}_{ui} = \vec{p}_u \cdot \vec{Q}_i = \sum_{k=1}^K p_{uk} Q_{ki}$$

$$Q \xrightarrow{?} \underline{P}, \underline{Q}$$

goal: minimize the error

$$\frac{1}{2} \underline{e}_{ui}^2 = \frac{1}{2} (r_{ui} - \hat{r}_{ui})^2$$

$$\rightarrow \frac{1}{2} (r_{ui} - \sum_{k=1}^K p_{uk} Q_{ki})^2$$


define $\underline{P}, \underline{Q}$ so that we would like to minimize the error

Factorization-Based Approaches: The Gradient

$$\frac{\partial \mathcal{L}_{ui}^2}{\partial p_{uk}} = -(r_{ui} - \hat{r}_{ui}) q_{ki}$$
$$= - \underline{\underline{e_{ui} q_{ki}}}$$

$$\frac{\partial \mathcal{L}_{ui}^2}{\partial q_{ki}} = - \underline{\underline{e_{ui} p_{uk}}}$$

gradient descent
optimization
P, Q



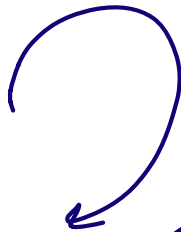
Factorization-Based Approaches: Update Rules

initialize P, Q (small random values)

$$p_{uk} \leftarrow p_{uk} - \alpha \frac{\partial e_{ui}^2}{\partial p_{uk}}$$

learning rate
 $\alpha = 0.001$

$$= p_{uk} + \alpha \underline{e}_{ui} \underline{q}_{ki} \quad \left| \text{for } k=1 \dots K \right.$$


$$q_{ki} \leftarrow q_{ki} + \alpha \underline{e}_{ui} p_{uk}$$

Scalable Collaborative Filtering Approaches for Large Recommender Systems



Gábor Takács*

GTAKACS@SZE.HU

*Dept. of Mathematics and Computer Science
Széchenyi István University
Egyetem tér 1.
Győr, Hungary*

István Pilászy*

PILA@MIT.BME.HU

*Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary*

Bottyán Németh*

BOTTYAN@TMIT.BME.HU

Domonkos Tikk*

TIKK@TMIT.BME.HU

*Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary*

stochastic
gradient descent
↓

ensemble approach
↓ 100x, PQ

Input: T' : training set, η : learning rate (λ : regularization factor)
Output: $\mathbf{P}^*, \mathbf{Q}^*$: the user and item feature matrices

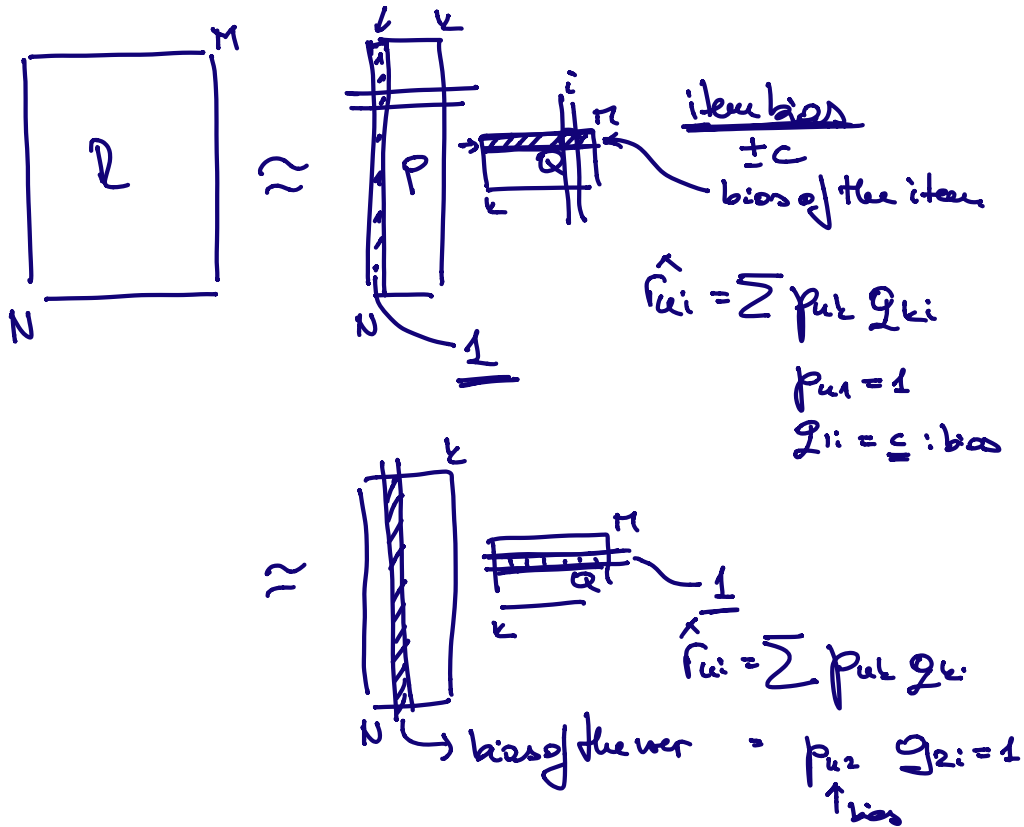
- 1 Partition T' into two sets: T'_I T'_II (validation set);
- 2 Initialize \mathbf{P} and \mathbf{Q} with small random numbers. ←
- 3 **loop** until the terminal condition is met. One epoch:
- 4 iterate over each (u, i, r_{ui}) element of T'_I :
- 5 compute $e'_{ui} = r_{ui} - \hat{r}_{ui}$
- 6 compute the gradient of e'_{ui} , according to Eq. (6);
- 7 for each k
- 8 update \mathbf{p}_u , the u -th row of \mathbf{P} ,
- 9 and \mathbf{q}_i , the i -th column of \mathbf{Q} according to Eq. (7);
- 10 calculate the RMSE on T'_II ;
- 11 if the RMSE on T'_II was better than in any previous epoch:
- 12 Let $\mathbf{P}^* = \mathbf{P}$ and $\mathbf{Q}^* = \mathbf{Q}$.
- 13 terminal condition: RMSE on T'_II does not decrease during two epochs.
- 14 **end**

Algorithm 1: Training algorithm for RISMF

RMSE $\sqrt{\frac{1}{|T'_II|} \sum_{(u,i) \in T'_II} (r_{ui} - \hat{r}_{ui})^2}$

Robert
model
↓
 \mathbf{P}, \mathbf{Q}

Factorization-Based Approaches: Bias



oct 2006 , goal : 10% improvement → june 2008



march 2010, second prize
cancelled

tricky data set 100.000.000 ratings
480.000 users
17.000 items

↓ leaves 2.700 leaves $RMSE = 0.8563$

around \$1.000.000, for 10% improvement

$RMSE$

1.1296 + global mean

1.0651 + user mean

0.9514 + netflix*

0.94 + similarity-based techniques (collaborative filtering)

0.90 + latent model

0.89 + latent model + bias

0.876 + latent model + bias + time (u, i, r, t)

0.8563 + ensembles