

Рекомендательные системы и Большие Данные

Авторы доклада:

к.т.н., Сергей Папулин
papulin_bmstu@mail.ru

Ольга Недильченко
nediola@yandex.ru



План доклада

- Виды рекомендательных систем
- Распределенная контентная фильтрация
- Распределенная коллаборативная фильтрация
- Распределенная факторизация матрицы рейтингов
- Решения на базе Spark

Виды рекомендательных систем



Виды рекомендательных систем

[Last Place on Earth](#) | [Earth](#) | [Antarctica](#)

The last place on Earth without human noise

Is there anywhere left utterly free of man-made sound? In the first of a series for BBC Future called Last Place on Earth, Rachel Nuwer sets out to find havens where silence still rules – but discovers that avoiding civilisation's clatter is harder than it seems. In fact, there's one human noise you will never escape.



By Rachel Nuwer
17 January 2014

A special kind of noisiness accosts passengers waiting for New York City subways. Down there, sound levels regularly exceed 100 decibels – enough to **damage a person's hearing over time**. It was on one such platform that George Foy, a journalist and New York University creative writing professor, suddenly found himself losing it one day, when four trains pulled in at once. "I kind of went momentarily crazy," he says. He hunched over and stuck his fingers in his ears, desperately trying to block out the cacophony. "I started wondering why the hell I was putting up with this," he says.

It was then that his obsession to find the quietest place on

Related Stories



Give peace (and quiet) a chance



How to make kids learn faster



Remote that reduces street noises


Content-based
(основанные на
контенте)
рекомендации на
основе предыдущих
оценок пользователя
и схожести объектов



Виды рекомендательных систем

Customers Who Bought This Item Also Bought



Recommender Systems:
The Textbook
› Charu C. Aggarwal
Hardcover
\$63.20 






Statistical Methods for
Recommender Systems
Deepak K. Agarwal
★★★★★ 1
Hardcover
\$56.99 

Collaborative filtering
(основанные на
коллаборативной
фильтрации)
рекомендации на
основе оценок других
пользователей



Виды рекомендательных систем

Sponsored Links ([What's this?](#))

1. [Russian Art Auctions](#) 
2. [Booming Art Market](#) 
3. [Bernard Buffet Paintings](#)
4. [Photographic Art](#) 

Автошкола при
МГТУ

drivemaster.ru



**Social/Demographic
based**

**(основанные на
социальных данных)**
используют данные о
поле, возрасте, стране
пользователя и др.



Виды рекомендательных систем

Есть и оценки ресурса, и
оценки пользователя?

Да

Нет

Коллаборативная
фильтрация

Чего не хватает?

Оценок ресурса

Оценок
пользователя

Использовать
контент ресурса

Использовать
демографические
данные

Hybrid
(гибридные)
используют
смешанные
подходы



Формализация задачи рекомендательных систем

Дано:

- множество пользователей $u \in U$
- множество объектов $i \in I$
- множество оценок $r_{ui} \in K$

Задача:

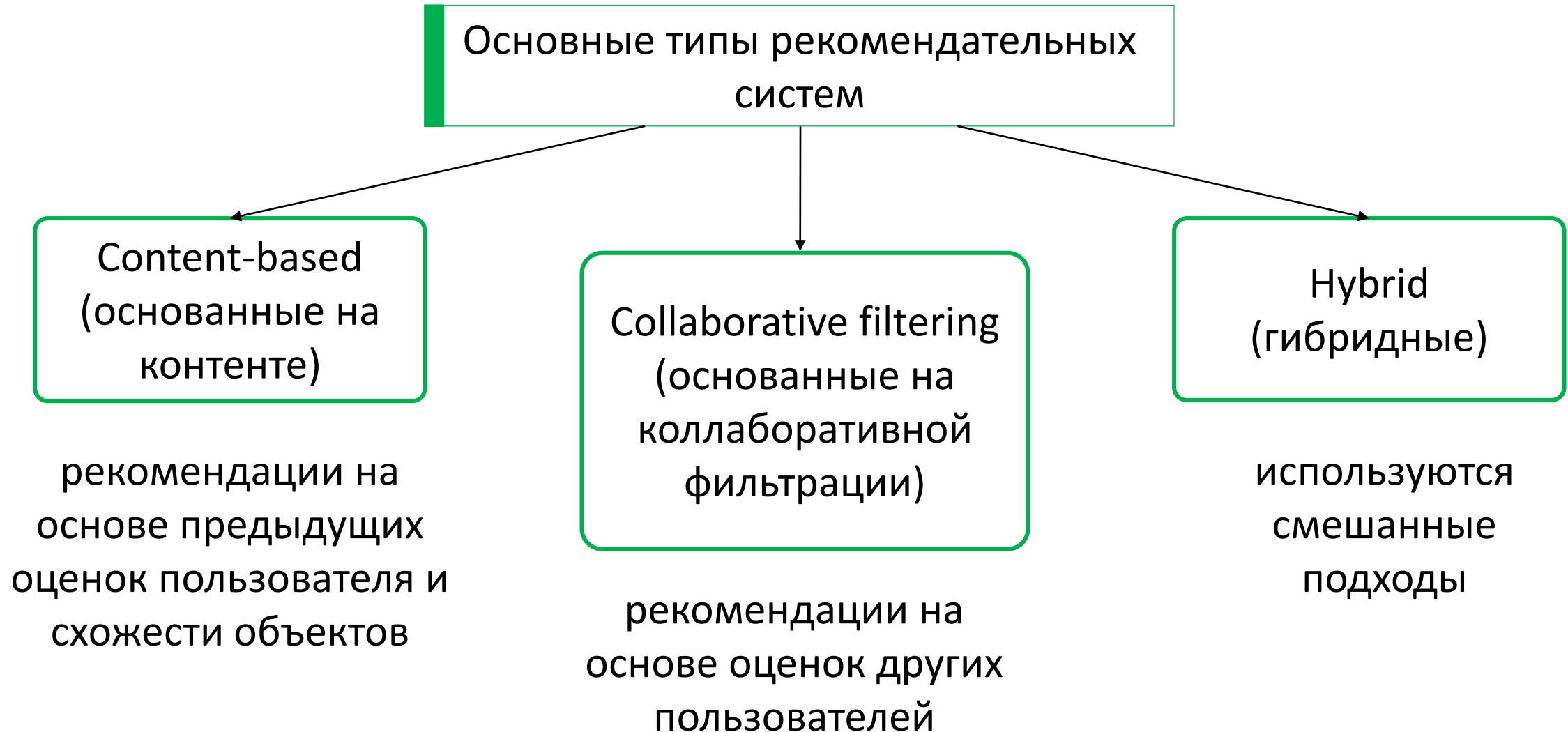
для пользователя u и объекта i
предсказать оценку \hat{r}_{ui}

Матрица оценок

Пользователи	r_{11}	?	r_{13}
	?	r_{22}	r_{23}
	?	?	r_{33}
Объекты			

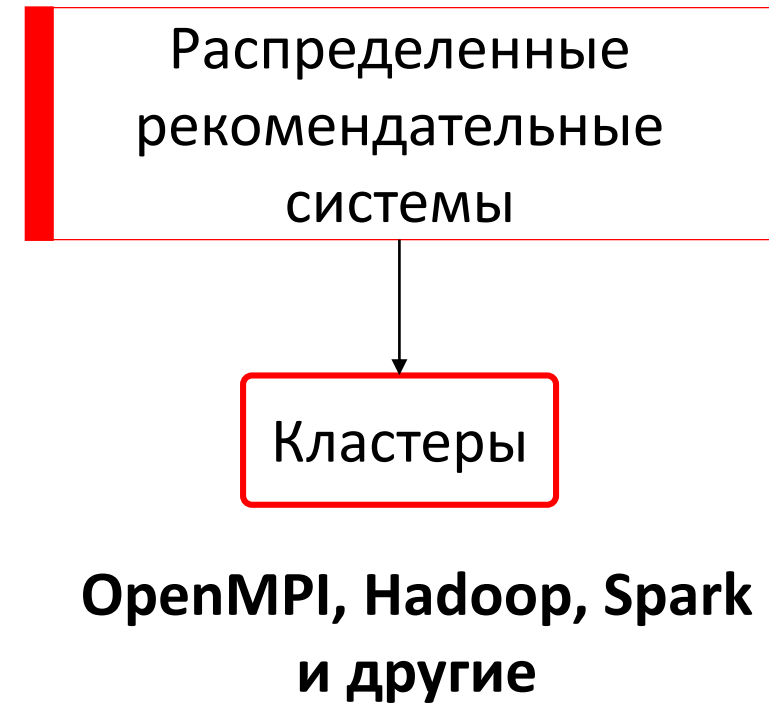
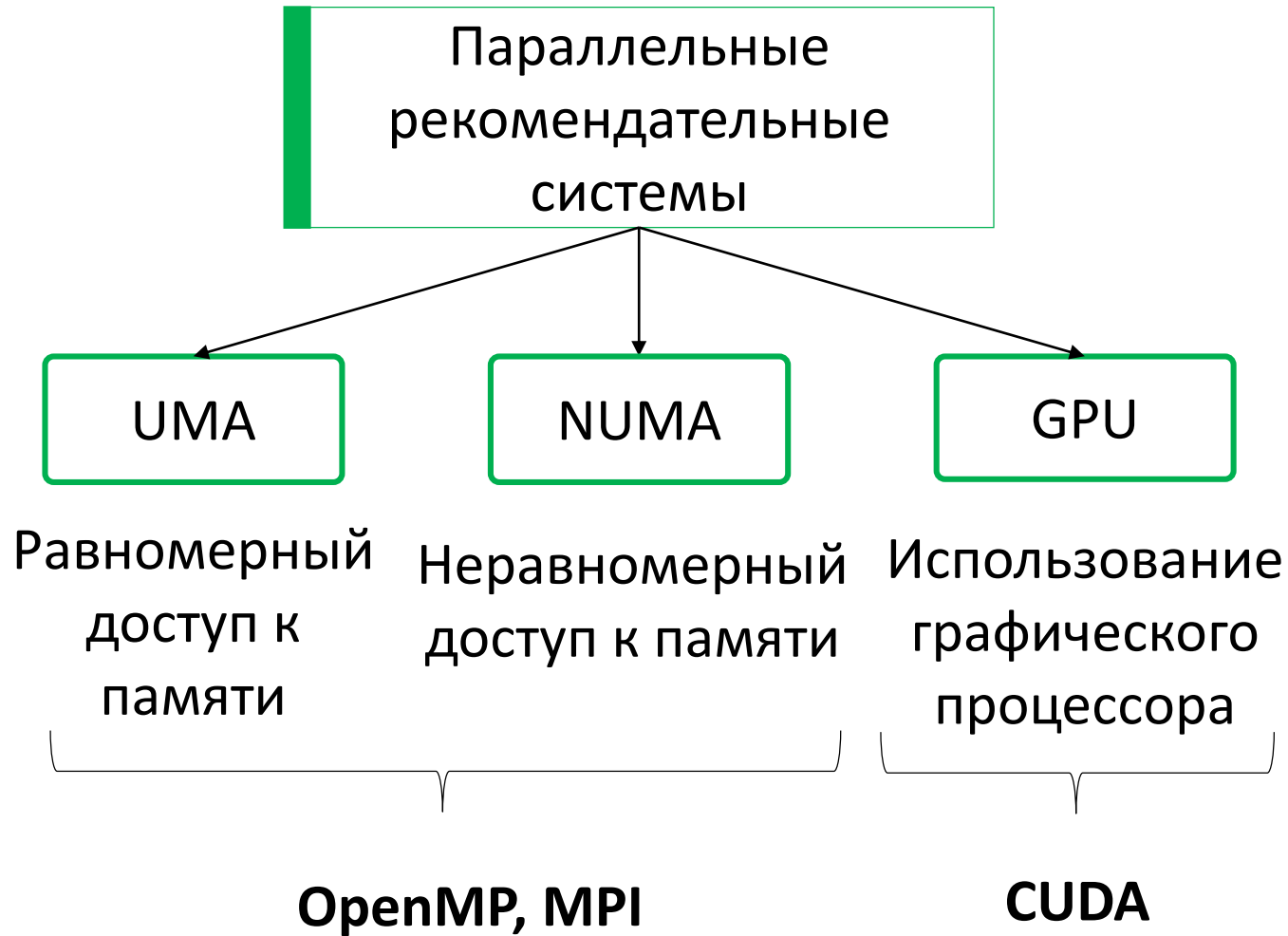


Виды рекомендательных систем





Параллелизм рекомендательных систем



Распределенная контентная фильтрация



Контентная фильтрация

Объекты

Пользователи

	i_1	i_2	i_3
u_1	r_{11}	?	r_{13}
u_2	?	?	r_{23}
u_3	r_{31}	?	r_{33}
u_4	?	r_{42}	?

Матрица оценок

Признаки объектов

	f_1	f_2	f_3	f_4
i_1		f_{1*}		
i_2		f_{2*}		
i_3		f_{3*}		

Матрица признаков объектов

Общая формула

$$\hat{r}_{ui} = \frac{\sum_{j \in I_U} r_{uj} \cdot \text{sim}(f_{i*}, f_{j*})}{\sum_{j \in I_U} \text{sim}(f_{i*}, f_{j*})}$$

$$\hat{r}_{32} = \frac{r_{31} \cdot \text{sim}(f_{1*}, f_{2*}) + r_{33} \cdot \text{sim}(f_{3*}, f_{2*})}{\text{sim}(f_{1*}, f_{2*}) + \text{sim}(f_{3*}, f_{2*})}$$



Контентная фильтрация

Преимущества

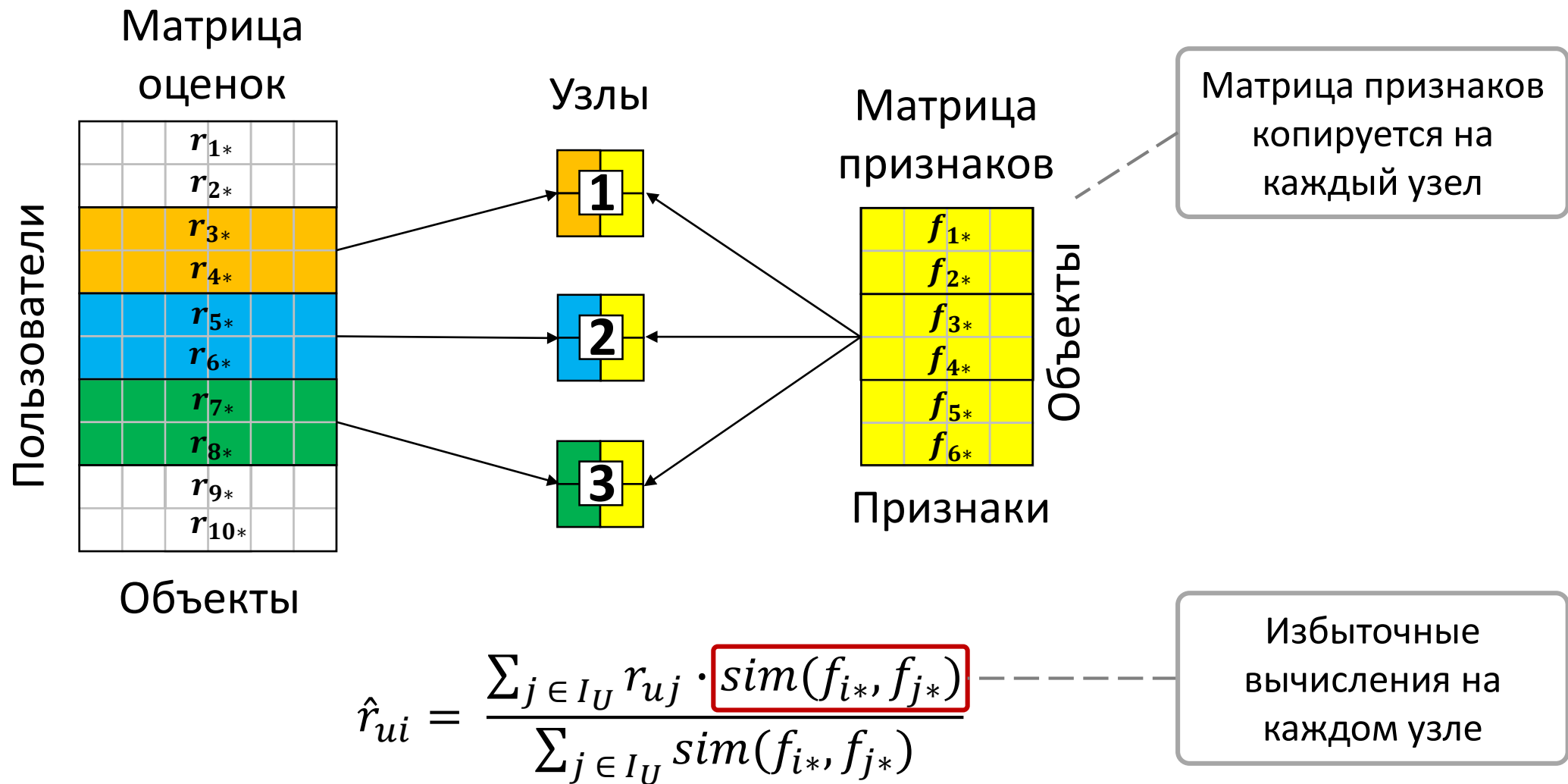
- простота вычислений
- гибкость (можно выбирать, какие признаки объектов использовать)
- прозрачность (рекомендации просто объяснить)

Недостатки

- не учитываются сложные случаи
- нельзя рекомендовать объекты вне текущих предпочтений пользователя
- проблема "холодного старта" (новые объекты никому не рекомендуются)

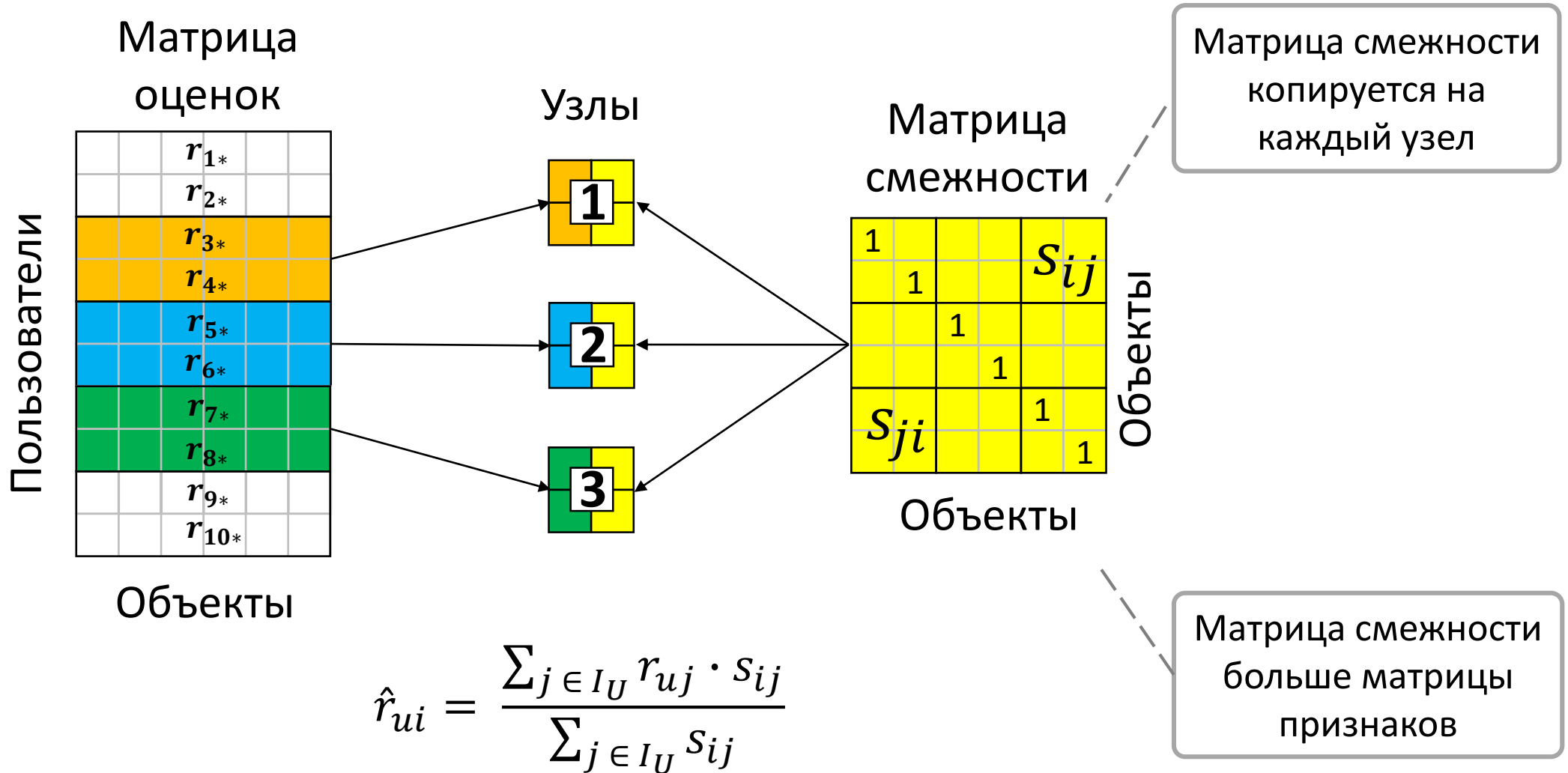


Распределенная контентная фильтрация. Вариант 1





Распределенная контентная фильтрация. Вариант 2



Распределенная коллаборативная фильтрация



Коллаборативная фильтрация

Подходы к коллаборативной фильтрации

Memory-based

Методы, основанные на
нахождении ближайших соседей
(kNN = k Nearest Neighbours)

Model-based

Методы, основанные на
факторизация матриц, байесовых
сетях, методах кластеризации



Коллаборативная фильтрация (Memory Based)

Матрица оценок

	i_1	i_2	...	i_k	...	i_m
u_1	r_{11}	r_{12}	...	r_{1k}	...	r_{1m}
u_2	r_{21}	r_{22}	...	r_{2k}	...	r_{2m}
...
u_k	r_{k1}	r_{k2}	...	r_{kk}	...	r_{km}
...
u_n	r_{n1}	r_{n2}	...	r_{nk}	...	r_{nm}

$v \in U_i$ — пользователи, оценившие объект i

$j \in I_U$ — объекты, которые оценивал пользователь u

Основанная на пользователях
(user-based)

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in U_i} \text{sim}(u, v)(r_{vi} - \bar{r}_v)}{\sum_{v \in U_i} \text{sim}(u, v)}$$

Основанная на объектах
(item-based)

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{j \in I_U} \text{sim}(i, j)(r_{uj} - \bar{r}_j)}{\sum_{j \in I_U} \text{sim}(i, j)}$$



Коллаборативная фильтрация (Memory Based)

Преимущества

- позволяет выявлять сильные зависимости между пользователями/объектами
- прозрачность (рекомендации просто объяснить)

Недостатки

- сложность вычислений (на деле берут лишь k ближайших соседей)
- нет поддержки пользователей с уникальными предпочтениями
- проблема "холодного старта" (новые объекты никому не рекомендуются)
- рекомендации часто тривиальны

Item-based

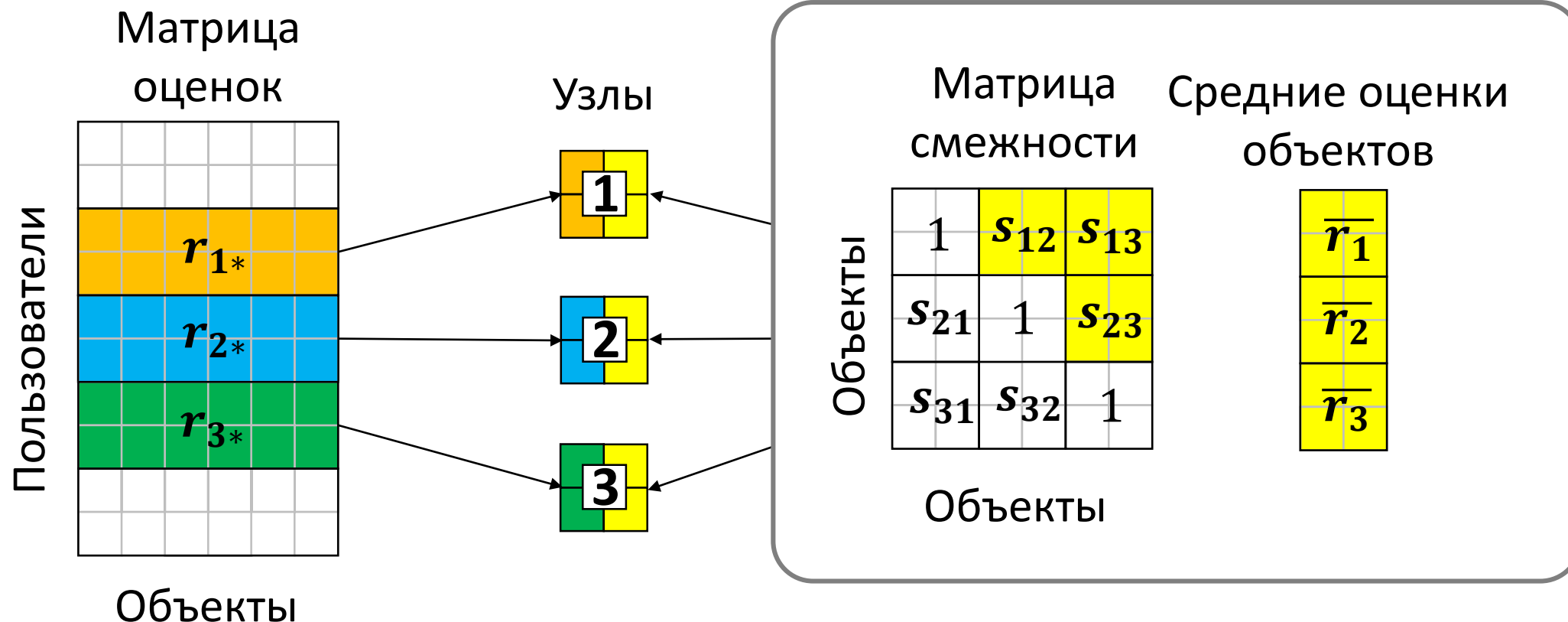
$$\text{sim}(i, j) = s_{ij}$$

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{j \in I_U} \text{sim}(i, j)(r_{uj} - \bar{r}_j)}{\sum_{j \in I_U} \text{sim}(i, j)}$$

Матрица смежности

$$A \cdot \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ \dots & \dots & \dots & \dots \\ s_{m1} & s_{m2} & \dots & s_{mm} \end{bmatrix} \cdot \begin{bmatrix} r_{u1} - \bar{r}_1 \\ r_{u2} - \bar{r}_2 \\ \dots \\ r_{um} - \bar{r}_m \end{bmatrix} + \begin{bmatrix} \bar{r}_1 \\ \bar{r}_2 \\ \dots \\ \bar{r}_m \end{bmatrix} = \begin{bmatrix} \hat{r}_{u1} \\ \hat{r}_{u2} \\ \dots \\ \hat{r}_{um} \end{bmatrix}$$

$$A = \text{diag}\left(\frac{1}{\sum_{j \in I_U} s_{1j}}, \frac{1}{\sum_{j \in I_U} s_{2j}}, \dots, \frac{1}{\sum_{j \in I_U} s_{mj}}\right)$$



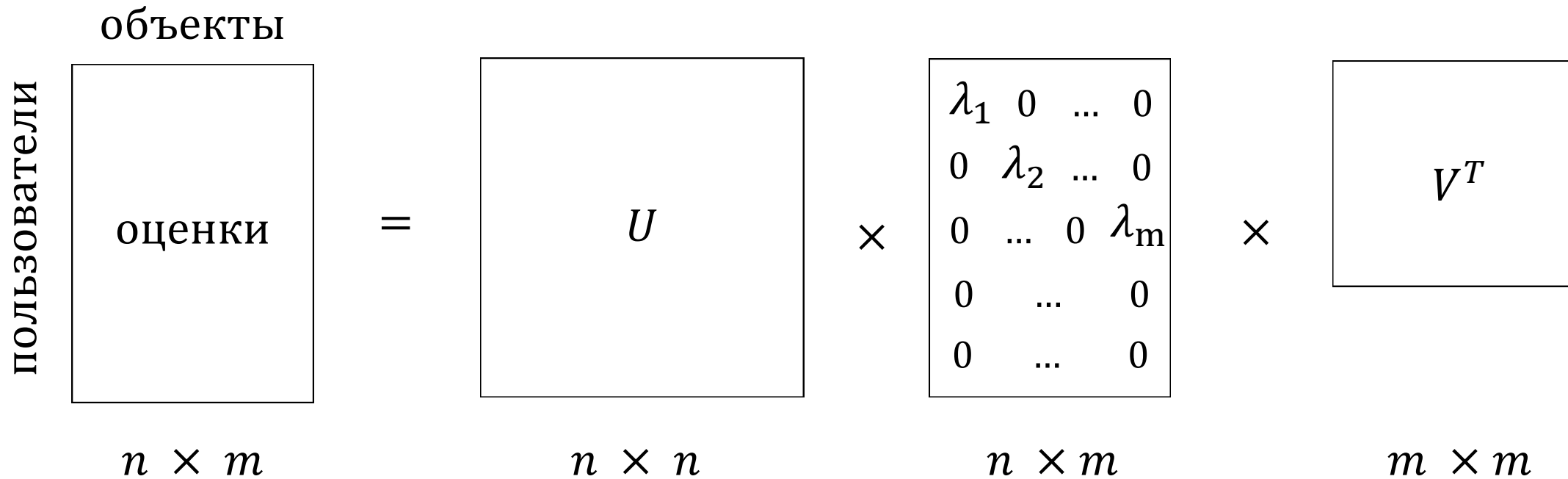
$$s_{ij} = \text{sim}(i, j) = \text{sim}(r_{*i}, r_{*j})$$



Коллаборативная фильтрация (Model Based)

$$\underset{n \times m}{A} = \underset{n \times n}{U} \times \underset{n \times m}{\Sigma} \times \underset{m \times m}{V^T}$$

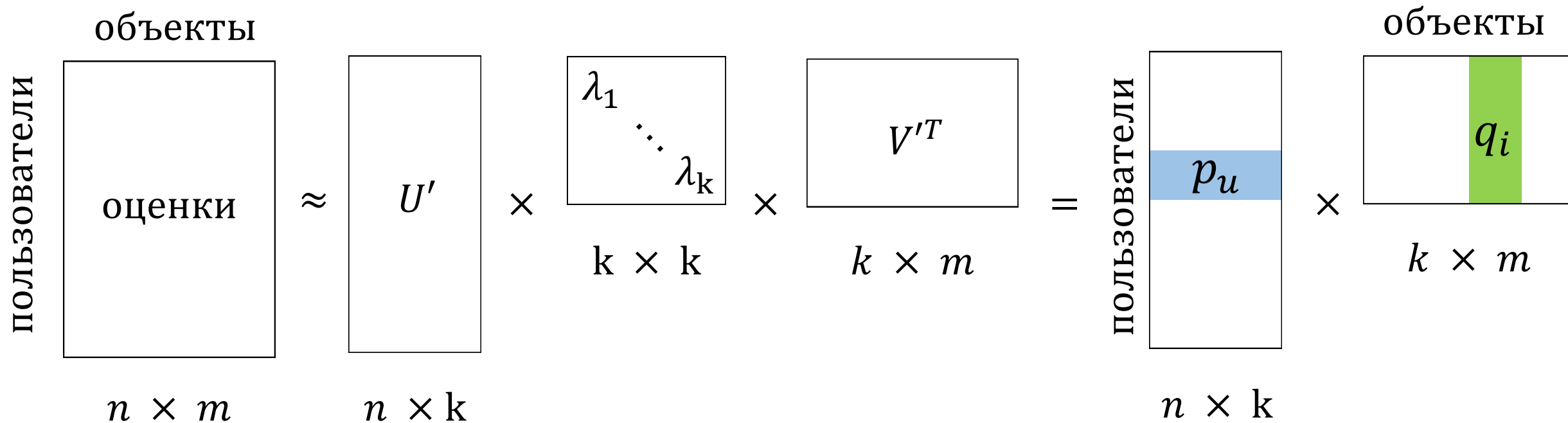
$$\lambda_1 \geq \dots \geq \lambda_{\min(n,m)} \geq 0$$





Коллаборативная фильтрация (Model Based)

$$A'_{n \times m} = U'_{n \times k} \times \Sigma'_{k \times k} \times V'^T_{k \times m}$$



$$\hat{r}_{ui} = \langle p_u, q_i \rangle$$



Модели SVD и SVD++

SVD: $\hat{r}_{ui} = \mu + b_u + b_i + p_u^T q_i$

SVD + +: $\hat{r}_{ui} = \mu + b_u + b_i + (p_u + \left[\frac{1}{\sqrt{|N(u)|}} \sum_{y \in N(u)} y_j \right]^T q_i$

μ – средняя оценка по всем объектам

b_u, b_i – отклонения средней оценки пользователя u и объекта i

b_i – отклонение средней оценки товара i

$N(u)$ – объекты, просмотренные пользователем u

y_j – дополнительный набор факторов, характеризующий пользователя на основе того, что он просматривал

$$\sum_{(u,i) \in K} (r_{ui} - \hat{r}_{ui}(\Theta))^2 + \lambda \sum_{\theta \in \Theta} \theta^2 \rightarrow \min_{\Theta} \quad \Theta = \{p_u, q_i, y_j, b_u, b_i\}$$



Преимущества

- позволяет выявлять общие зависимости, учитывать все имеющиеся данные
- высокая точность

Недостатки

- высокие временные затраты на обучение модели
- не прозрачный (рекомендации сложно объяснить)

Распределенная факторизация матрицы рейтингов



Факторизация матрицы. Задача оптимизации

Задача оптимизации

$$\widehat{W}, \widehat{H} = \arg \min_{W, H} L(W, H)$$

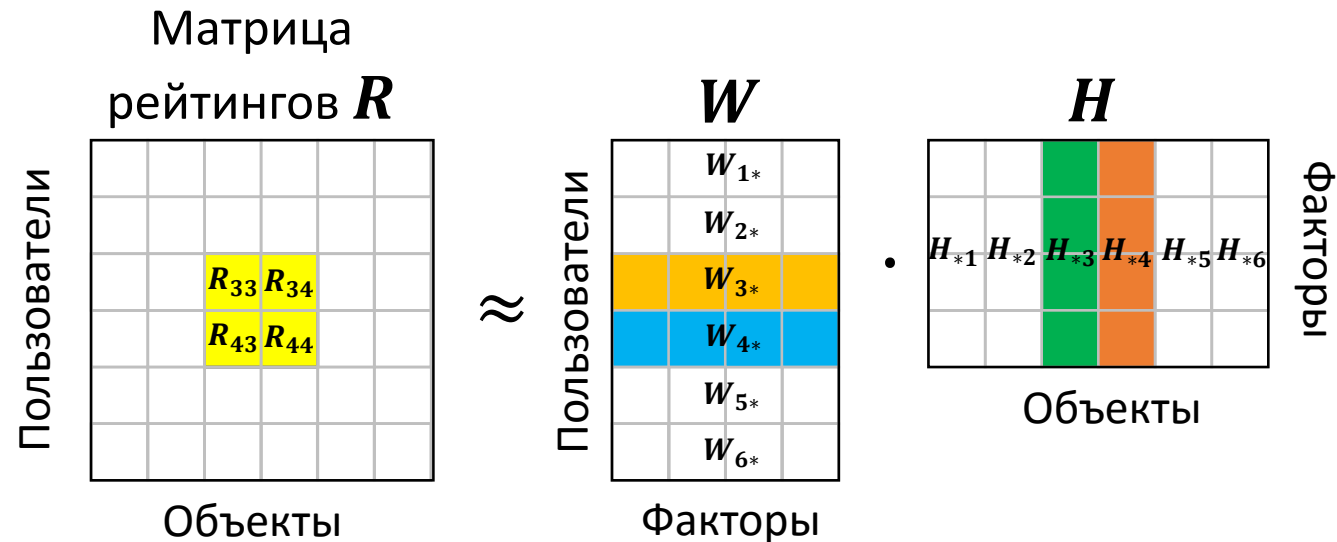
Функция потерь

$$L(W, H) = \sum_{(i,j) \in \Omega} L_{ij}(W, H)$$

$$L_{ij}(W, H) = l(R_{ij}, W_{i*}, H_{*j})$$

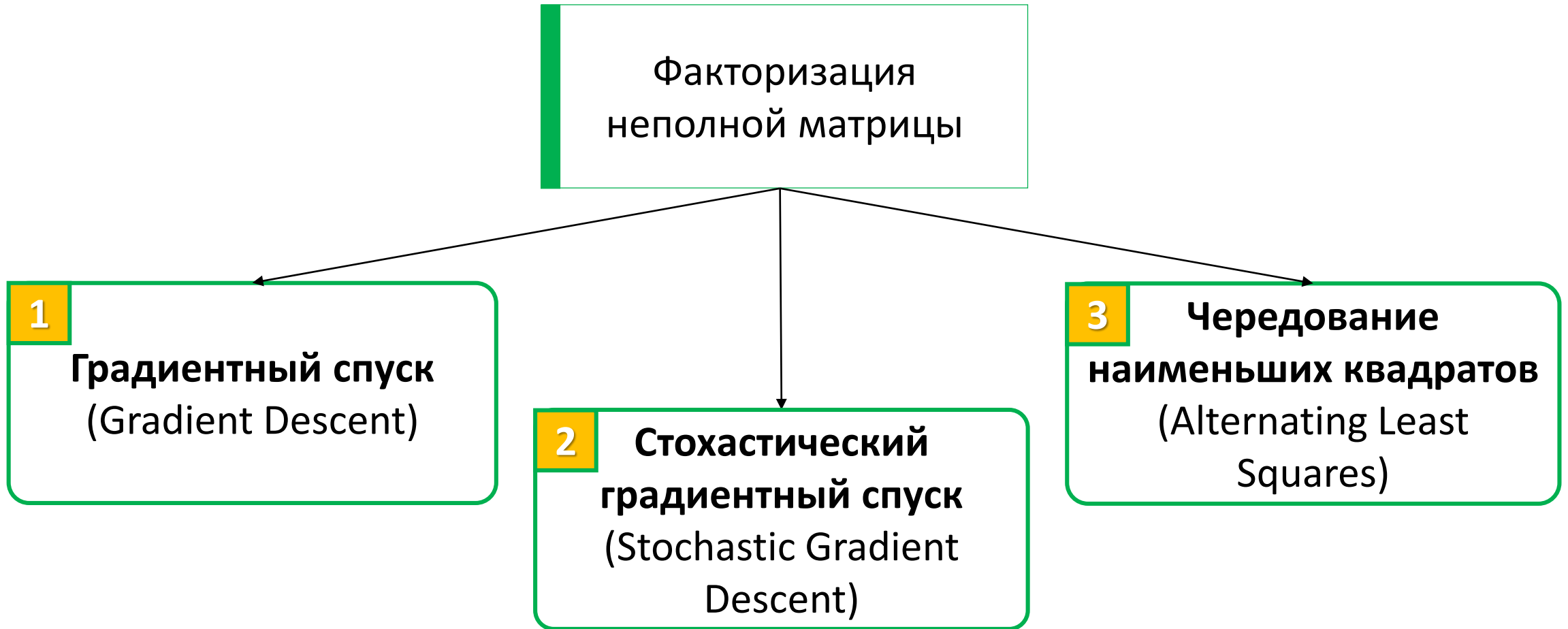
Пример функции потерь (без регуляризации)

$$L(W, H) = \sum_{(i,j) \in \Omega} (R_{ij} - W_{i*} H_{*j})^2$$





Факторизация матрицы. Распределенные алгоритмы

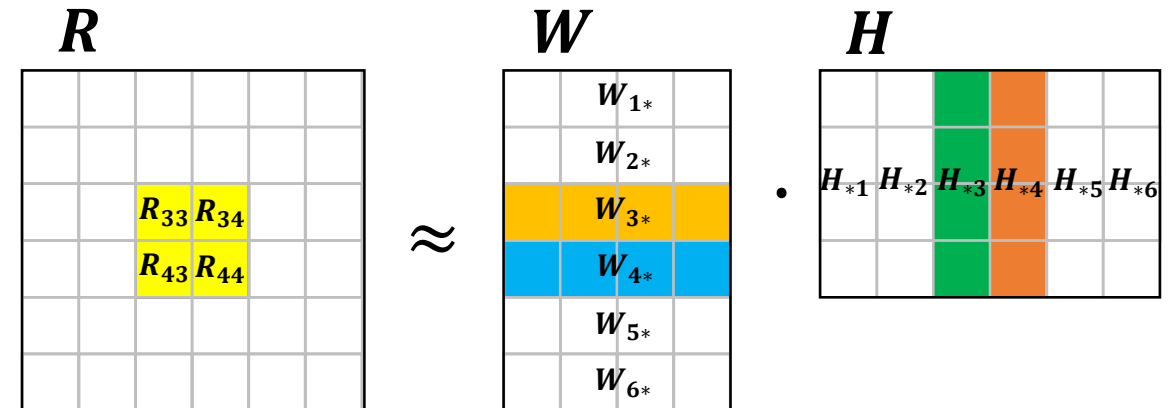




Градиентный спуск

Задача оптимизации

$$\widehat{W}, \widehat{H} = \arg \min_{W, H} L(W, H)$$



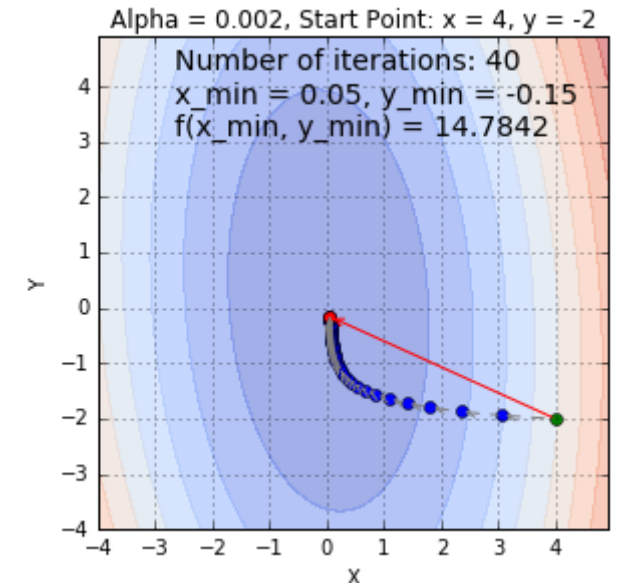
Градиентный спуск для минимизации функции потерь

$$W_{n+1} = W_n - \alpha_n W_n^\nabla$$

n – итерация

$$H_{n+1} = H_n - \alpha_n H_n^\nabla$$

α_n – шаг обучения





Градиентный спуск

Градиентный спуск

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \alpha_n \mathbf{W}_n^\nabla$$

$$\mathbf{H}_{n+1} = \mathbf{H}_n - \alpha_n \mathbf{H}_n^\nabla$$

Частные производные

$$\frac{\partial}{\partial \mathbf{W}_{i*}} L(\mathbf{W}_n, \mathbf{H}_n) = \sum_{j \in \Omega_j} \frac{\partial}{\partial \mathbf{W}_{i*}} L_{ij}(\mathbf{W}_n, \mathbf{H}_n)$$

$$\Omega_j = \{j | (i, j) \in \Omega\}$$

Матрица

рейтингов \mathbf{R}

Пользователи

		R_{33}	R_{34}		
		R_{43}	R_{44}		

Объекты

\approx

Пользователи

	W_{1*}				
	W_{2*}				
	W_{3*}				
	W_{4*}				
	W_{5*}				
	W_{6*}				

Факторы

\cdot

\mathbf{H}

Объекты

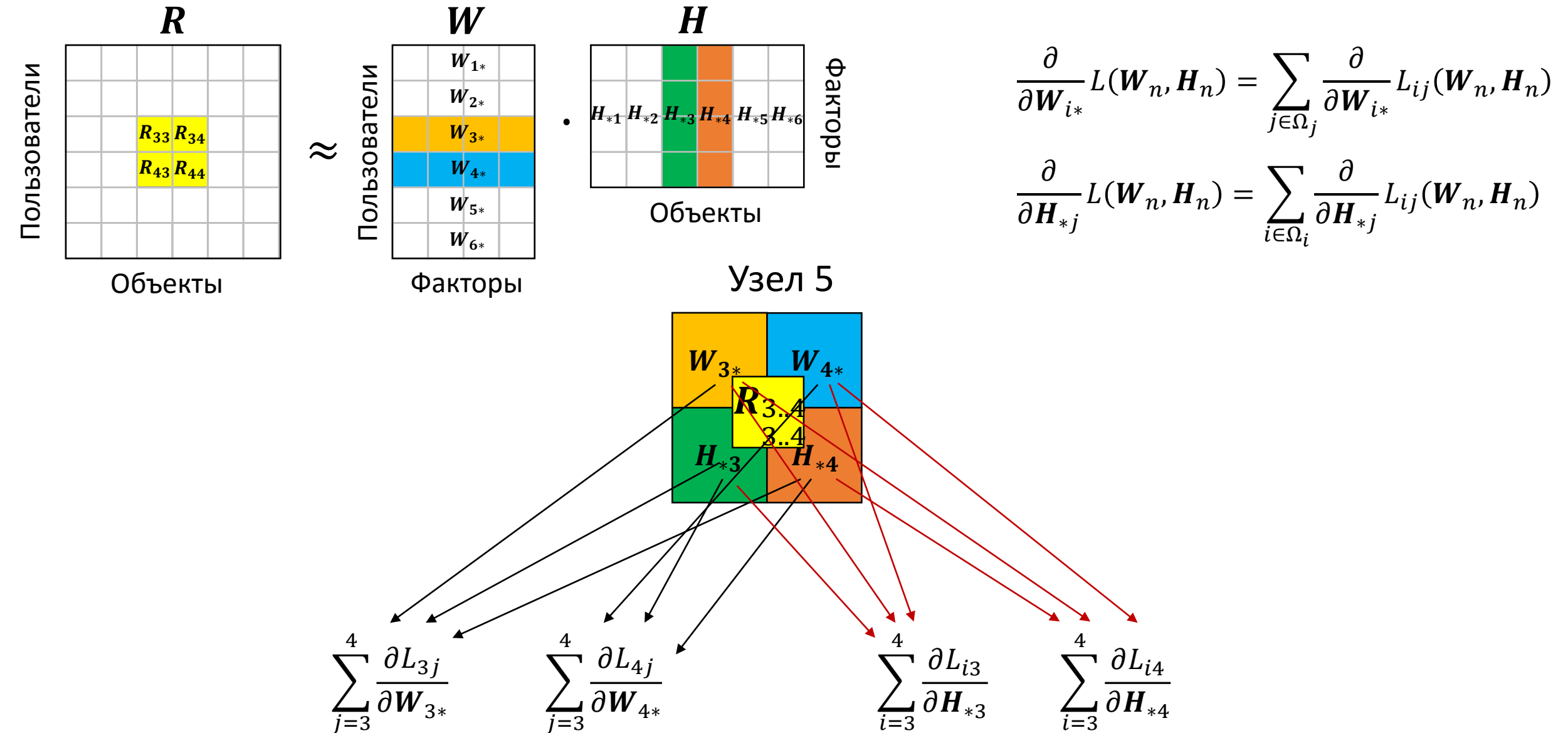
Факторы

Пример частной производной для функции потерь (без регуляризации)

$$\frac{\partial}{\partial \mathbf{W}_{ik}} L_{ij}(\mathbf{W}_n, \mathbf{H}_n) = -2(\mathbf{R}_{ij} - \mathbf{W}_{i*} \mathbf{H}_{*j}) \mathbf{H}_{kj}$$



Распределенный градиентный спуск



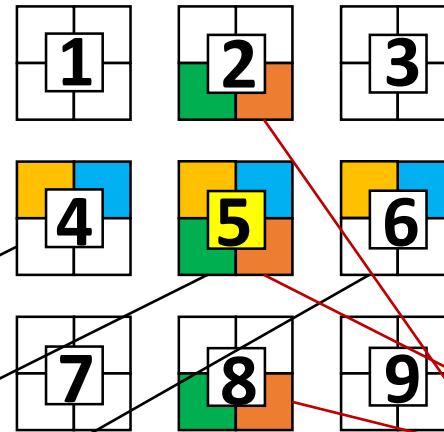


Распределенный градиентный спуск

W		
	W_{1*}	
	W_{2*}	
	W_{3*}	
	W_{4*}	
	W_{5*}	
	W_{6*}	

H						

Узлы



$$\frac{\partial}{\partial W_{i*}} L(W_n, H_n) = \sum_{j \in \Omega_j} \frac{\partial}{\partial W_{i*}} L_{ij}(W_n, H_n)$$

$$\frac{\partial}{\partial H_{*j}} L(W_n, H_n) = \sum_{i \in \Omega_i} \frac{\partial}{\partial H_{*j}} L_{ij}(W_n, H_n)$$

$$\frac{\partial L}{\partial W_{3*}} = \sum_{j=1}^2 \frac{\partial L_{3j}}{\partial W_{3*}} + \sum_{j=3}^4 \frac{\partial L_{3j}}{\partial W_{3*}} + \sum_{j=5}^6 \frac{\partial L_{3j}}{\partial W_{3*}} = \sum_{j=1}^6 \frac{\partial L_{3j}}{\partial W_{3*}}$$

$$\frac{\partial L}{\partial H_{*3}} = \sum_{i=1}^2 \frac{\partial L_{i3}}{\partial H_{*3}} + \sum_{i=3}^4 \frac{\partial L_{i3}}{\partial H_{*3}} + \sum_{i=4}^6 \frac{\partial L_{i3}}{\partial H_{*3}} = \sum_{i=1}^6 \frac{\partial L_{i3}}{\partial H_{*3}}$$



Распределенный градиентный спуск

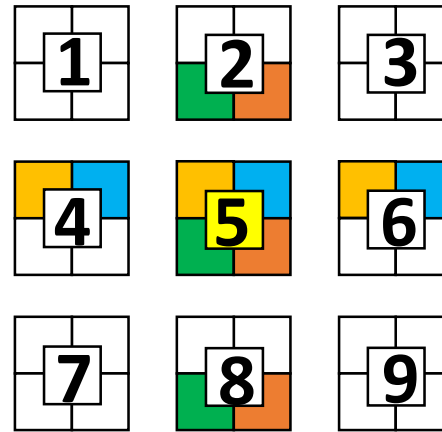
W

	W_{1*}	
	W_{2*}	
	W_{3*}	
	W_{4*}	
	W_{5*}	
	W_{6*}	

H

H_{*1}	H_{*2}	H_{*3}	H_{*4}	H_{*5}	H_{*6}

Узлы



$\frac{\partial L}{\partial W_{1*}}$	$\frac{\partial L}{\partial W_{2*}}$	$\frac{\partial L}{\partial W_{3*}}$	$\frac{\partial L}{\partial W_{4*}}$	$\frac{\partial L}{\partial W_{5*}}$	$\frac{\partial L}{\partial W_{6*}}$
--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------

$\frac{\partial L}{\partial H_{*1}}$	$\frac{\partial L}{\partial H_{*2}}$	$\frac{\partial L}{\partial H_{*3}}$	$\frac{\partial L}{\partial H_{*4}}$	$\frac{\partial L}{\partial H_{*5}}$	$\frac{\partial L}{\partial H_{*6}}$
--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------	--------------------------------------



Распределенный градиентный спуск

- 1 $W \leftarrow$ случайные значения
- 2 $H \leftarrow$ случайные значения
- 3 Цикл: критерий остановки
- 4 $W^\nabla, H^\nabla \leftarrow$ Распределенное вычисление градиента на d узлах
- 5 $W \leftarrow W - \alpha W^\nabla$
- 6 $H \leftarrow H - \alpha H^\nabla$
- 7 Обновление W и H на d распределенных узлах



Стохастический градиентный спуск

Градиентный спуск

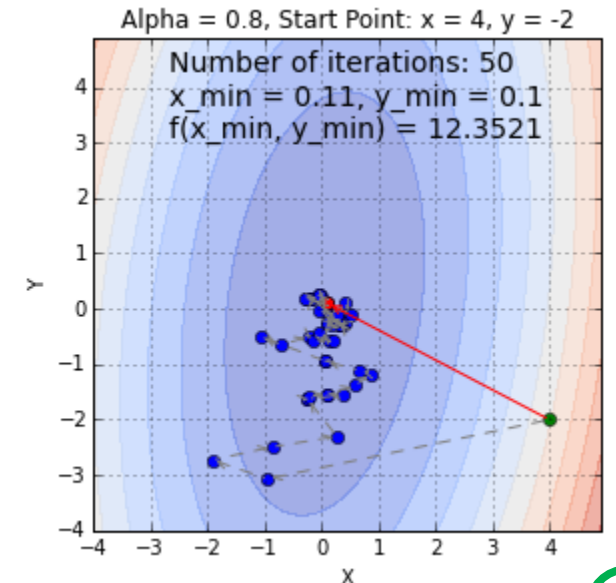
$$\mathbf{W}_{n+1} = \mathbf{W}_n - \alpha_n \mathbf{W}_n^\nabla \quad \rightarrow \quad \mathbf{W}_{i*}^{(n+1)} = \mathbf{W}_{i*}^{(n)} - \alpha^{(n)} \sum_{j \in \Omega_j} \frac{\partial}{\partial \mathbf{W}_{i*}} L_{ij}(\mathbf{W}^{(n)}, \mathbf{H}^{(n)})$$

$$\mathbf{H}_{n+1} = \mathbf{H}_n - \alpha_n \mathbf{H}_n^\nabla \quad \rightarrow \quad \mathbf{H}_{*j}^{(n+1)} = \mathbf{H}_{*j}^{(n)} - \alpha^{(n)} \sum_{i \in \Omega_i} \frac{\partial}{\partial \mathbf{H}_{*j}} L_{ij}(\mathbf{W}^{(n)}, \mathbf{H}^{(n)})$$

Стохастический градиентный спуск

$$\mathbf{W}_{i*}^{(n+1)} = \mathbf{W}_{i*}^{(n)} - \alpha^{(n)} N \frac{\partial}{\partial \mathbf{W}_{i*}} L_{ij}(\mathbf{W}^{(n)}, \mathbf{H}^{(n)})$$

$$\mathbf{H}_{*j}^{(n+1)} = \mathbf{H}_{*j}^{(n)} - \alpha^{(n)} N \frac{\partial}{\partial \mathbf{H}_{*j}} L_{ij}(\mathbf{W}^{(n)}, \mathbf{H}^{(n)})$$



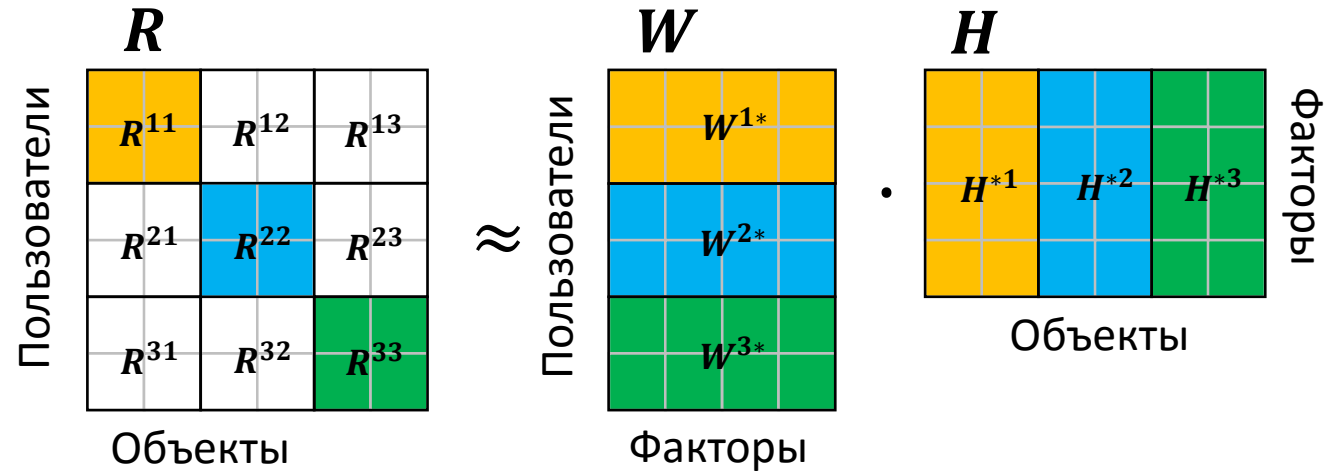


Стохастический градиентный спуск

- 1 $\mathbf{W} \leftarrow$ случайные значения
- 2 $\mathbf{H} \leftarrow$ случайные значения
- 3 Цикл: критерий остановки
- 4 | Выбор α
- 5 | Цикл: выбор значения (i, j) из Ω
- 6 | $\mathbf{W}'_{i*} \leftarrow \mathbf{W}_{i*} - \alpha N \frac{\partial}{\partial \mathbf{W}_{i*}} L_{ij}(\mathbf{W}, \mathbf{H})$
- 7 | $\mathbf{H}_{*j} \leftarrow \mathbf{H}_{*j} - \alpha N \frac{\partial}{\partial \mathbf{H}_{*j}} L_{ij}(\mathbf{W}, \mathbf{H})$
- 8 | $\mathbf{W}_{i*} \leftarrow \mathbf{W}'_{i*}$



Распределенный стохастический градиентный спуск



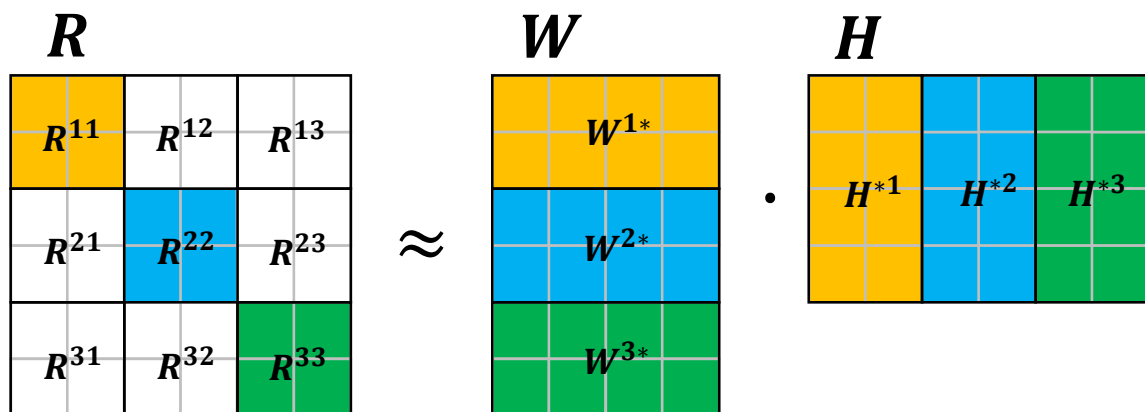
$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = R^{11}$$

$$\begin{bmatrix} W_{3*} \\ W_{4*} \end{bmatrix} = W^{2*}$$

$$\begin{bmatrix} H_{*5} & H_{*6} \end{bmatrix} = H^{*3}$$



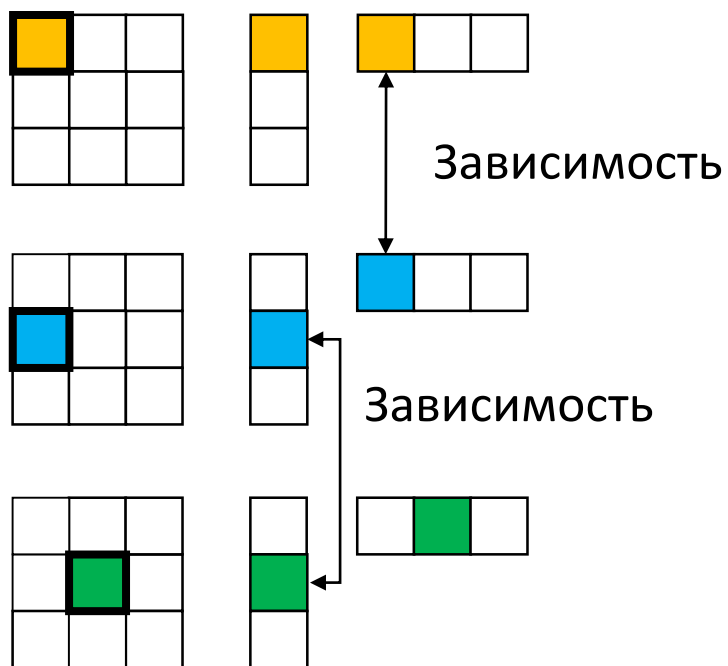
Распределенный стохастический градиентный спуск



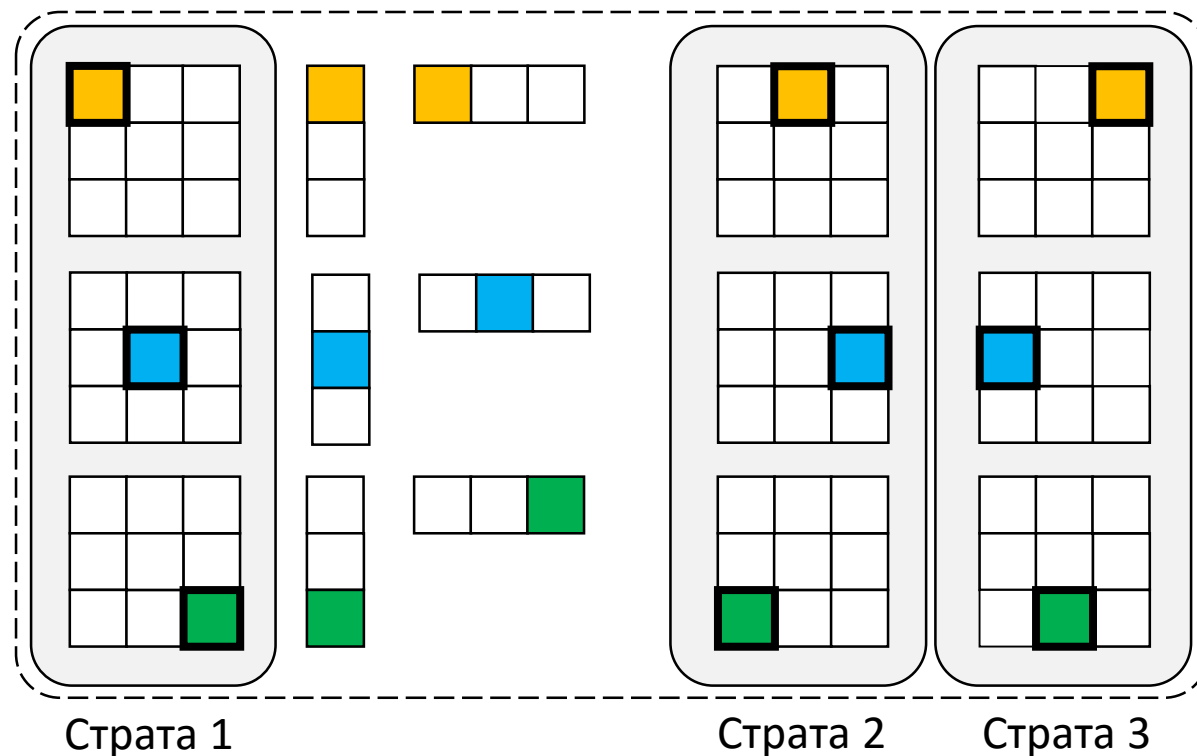
$$W_{i*}^{(n+1)} = W_{i*}^{(n)} - \alpha^{(n)} N \frac{\partial}{\partial W_{i*}} L_{ij}(W^{(n)}, H^{(n)})$$

$$H_{*j}^{(n+1)} = H_{*j}^{(n)} - \alpha^{(n)} N \frac{\partial}{\partial H_{*j}} L_{ij}(W^{(n)}, H^{(n)})$$

Проблема

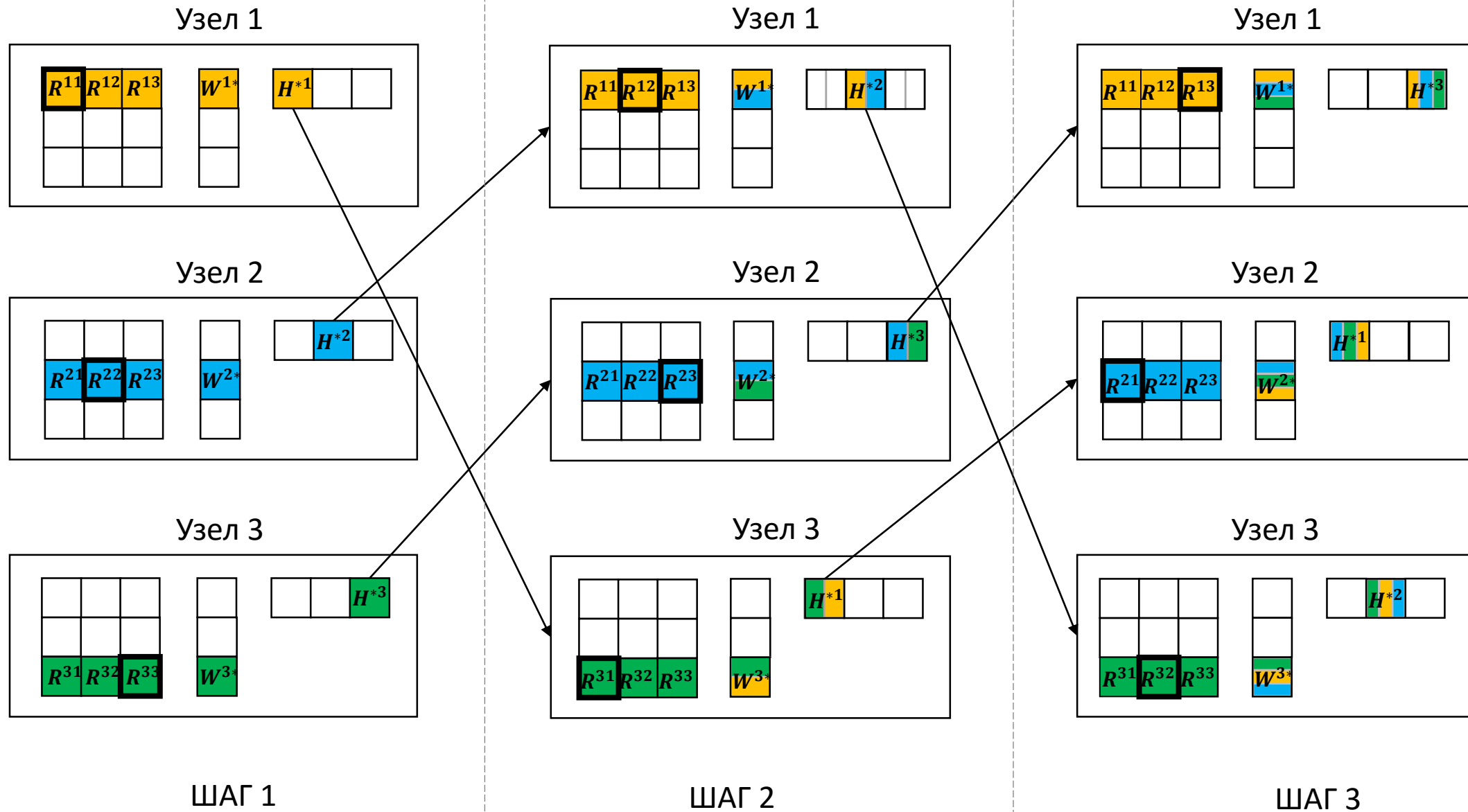


Решение





Распределенный стохастический градиентный спуск





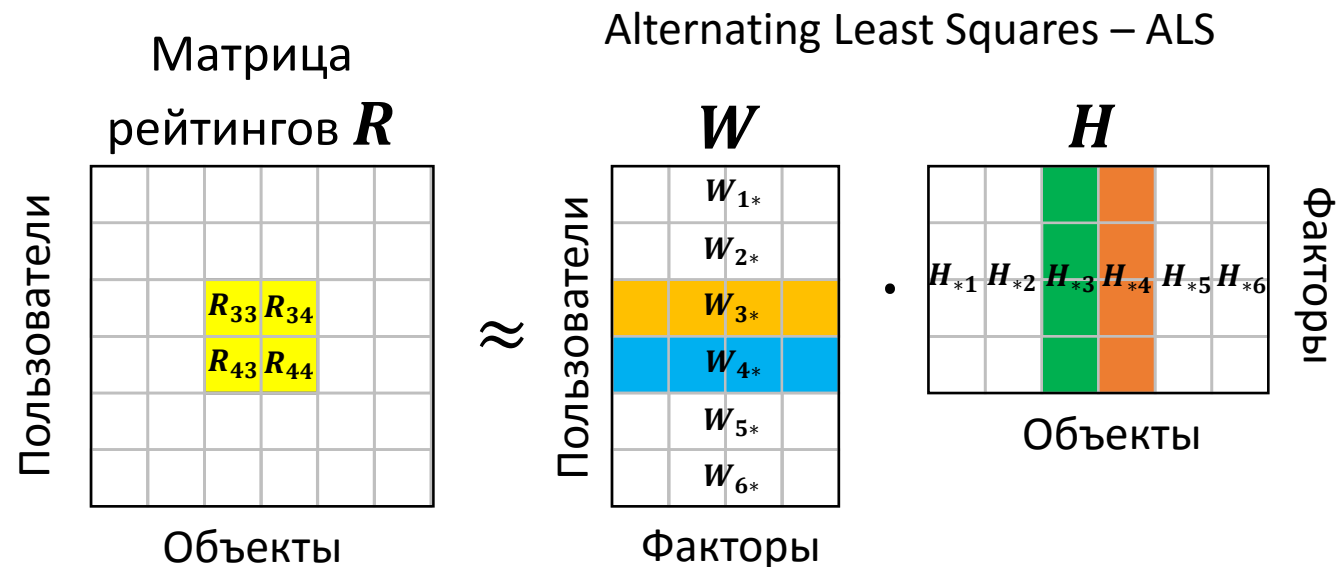
Распределенный стохастический градиентный спуск

- 1 $\mathbf{W} \leftarrow$ случайные значения
- 2 $\mathbf{H} \leftarrow$ случайные значения
- 3 Разделение на блоки
- 4 Цикл: критерий остановки
 - 5 Выбор α
 - 6 Цикл: 1 ... d
 - 7 Выбор d блоков из \mathbf{R} для формирования страты
 - 8 Стохастический град. спуск для блоков распределено на d узлах
+ обмен блоками \mathbf{W} и \mathbf{H} (при необходимости)

Метод наименьших квадратов с чередованием (ALS)

Задача оптимизации

$$\widehat{W}, \widehat{H} = \arg \min_{W, H} L(W, H)$$



Чередующиеся наименьшие квадраты (ALS)

$$\frac{\partial L(W, H)}{\partial W_{i*}} = 0$$

$$R_{i*} - \mathbf{W}_{i*} H^{(n)} = \mathbf{0}$$

$$\frac{\partial L(W, H)}{\partial H_{*j}} = 0$$

$$R_{*j} - W^{(n+1)} \mathbf{H}_{*j} = \mathbf{0}$$

Метод наименьших квадратов с чередованием (ALS)

Вычисление W при фиксированном H

$$W_{i*}^{(n+1)T} = \left(H_{\Omega_{i*}}^{(n)} H_{\Omega_{i*}}^{(n)T} \right)^{-1} H_{\Omega_{i*}}^{(n)} R_{\Omega_{i*}}^T$$

	R					
	1	2	3	4	5	6
1		2	2	3		
2			5		5	
3	4		3	5		
4		3	3			5
5	2				4	
6	5	2		3		2

$$R_{\Omega_{3*}} = \begin{matrix} & 1 & 3 & 4 \\ 3 & \begin{bmatrix} 4 & 3 & 5 \end{bmatrix} \end{matrix}$$

С регуляризацией

$$W_{i*}^{(n+1)T} = \left(H_{\Omega_{i*}}^{(n)} H_{\Omega_{i*}}^{(n)T} + \lambda n_{hi} E \right)^{-1} H_{\Omega_{i*}}^{(n)} R_{\Omega_{i*}}^T$$

	H					
	1	2	3	4	5	6
1						
2	H_{*1}	H_{*2}	H_{*3}	H_{*4}	H_{*5}	H_{*6}
3						
4						

$$H_{\Omega_{3*}} = \begin{matrix} & 1 & 3 & 4 \\ 3 & \begin{bmatrix} H_{*1} & H_{*3} & H_{*4} \end{bmatrix} \end{matrix}$$

Метод наименьших квадратов с чередованием (ALS)

Вычисление H при фиксированном W

$$H_{*j}^{(n+1)} = \left(W_{\Omega_{*j}}^{(n+1)T} W_{\Omega_{*j}}^{(n+1)} \right)^{-1} W_{\Omega_{*j}}^{(n+1)T} R_{\Omega_{*j}}$$

	R					
	1	2	3	4	5	6
1		2	2	3		
2			5		5	
3	4		3	5		
4		3	3			5
5	2				4	
6	5	2		3		2

$$R_{\Omega_{*4}} = \begin{matrix} & 4 \\ 1 & 3 \\ 3 & 5 \\ 6 & 3 \end{matrix}$$

С регуляризацией

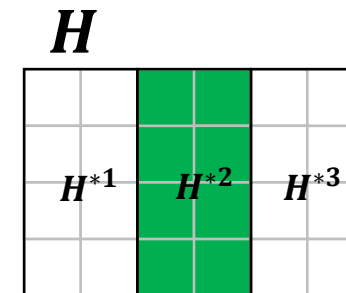
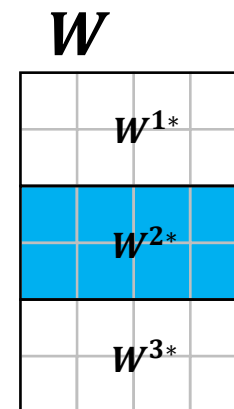
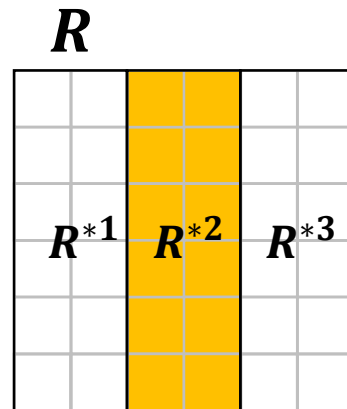
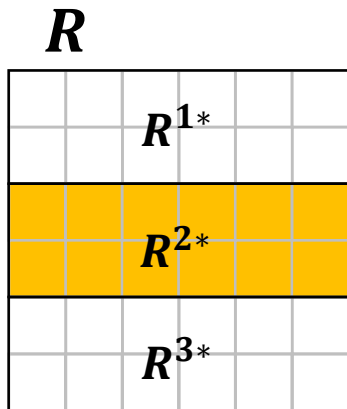
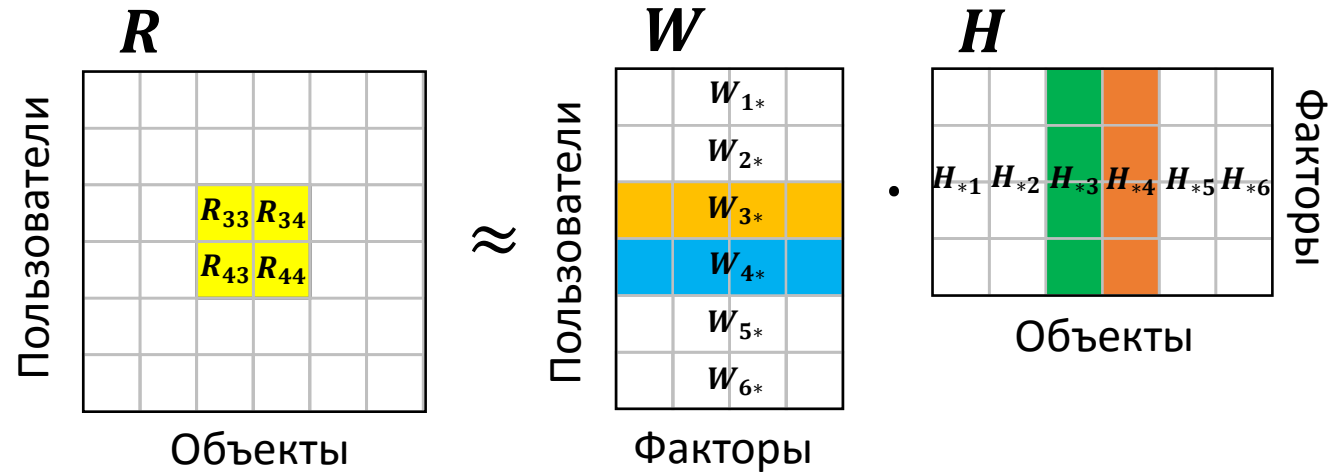
$$H_{*j}^{(n+1)} = \left(W_{\Omega_{*j}}^{(n+1)T} W_{\Omega_{*j}}^{(n+1)} + \lambda n_{wi} E \right)^{-1} W_{\Omega_{*j}}^{(n+1)T} R_{\Omega_{*j}}$$

	W			
	1	2	3	4
1		w_{1*}		
2		w_{2*}		
3		w_{3*}		
4		w_{4*}		
5		w_{5*}		
6		w_{6*}		

$$W_{\Omega_{*4}} = \begin{matrix} & 1 & 2 & 3 & 4 \\ 1 & & w_{1*} & & \\ 3 & & w_{3*} & & \\ 6 & & w_{6*} & & \end{matrix}$$



Распределенный ALS

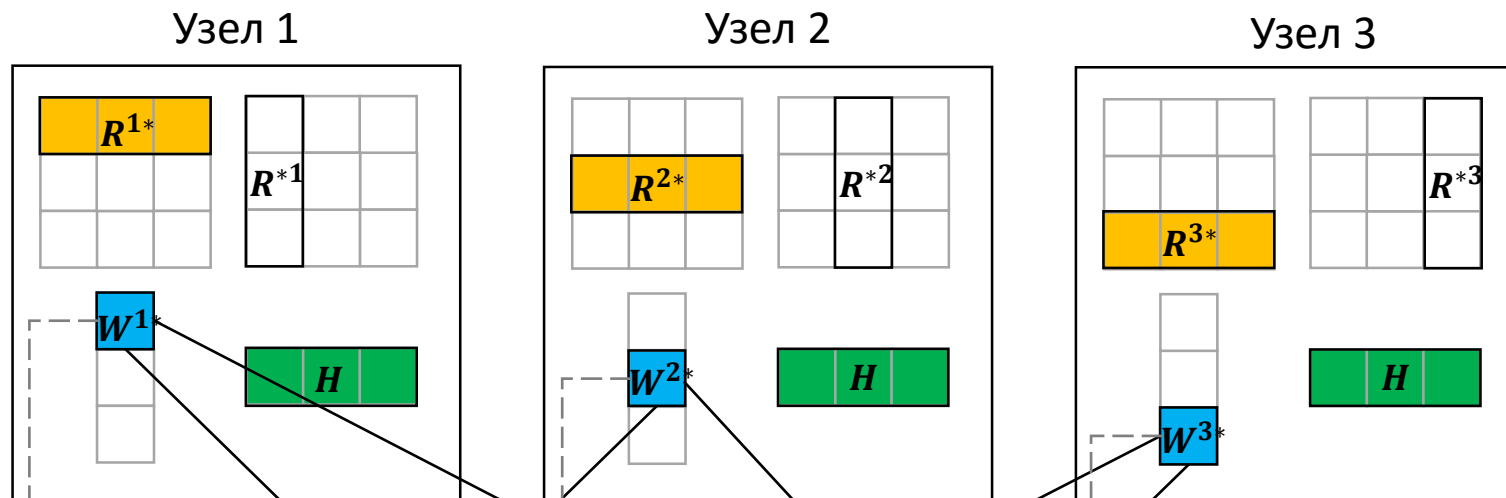




Распределенный ALS

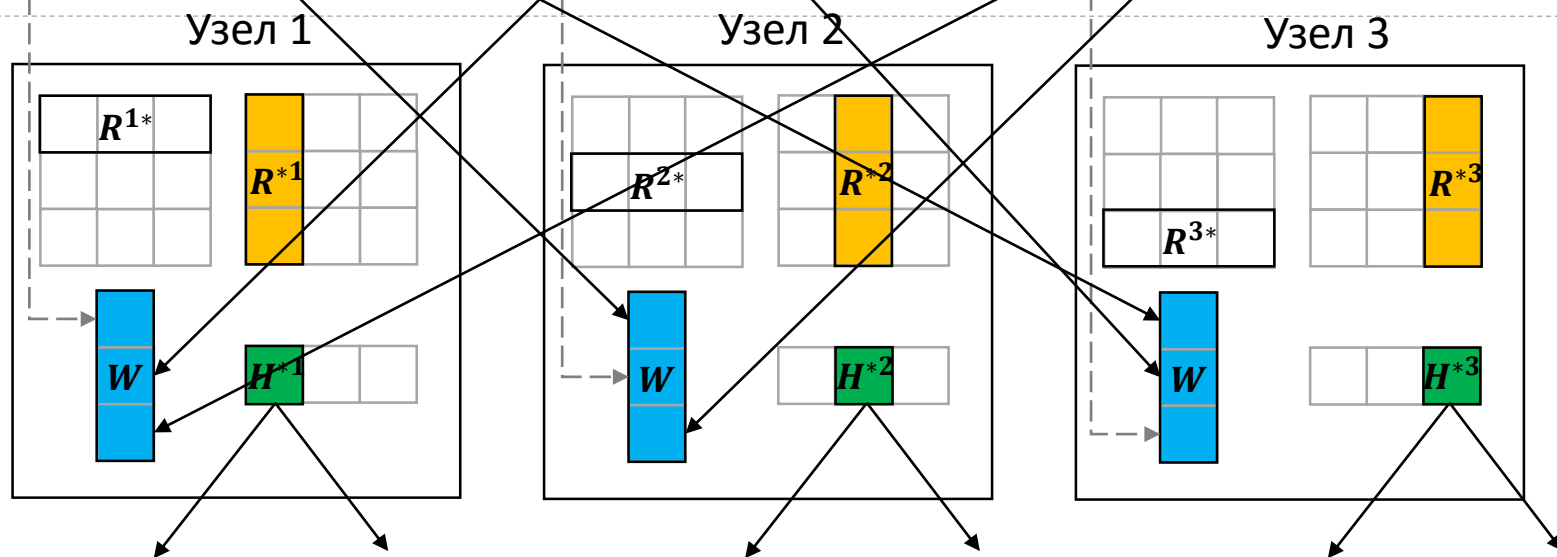
1

Вычисление W при фиксированном H



2

Вычисление H при фиксированном W



Повтор



Распределенный ALS

- 1 $\mathbf{H} \leftarrow$ средний рейтинг для первой строки и малые случайные значения для остальных
- 2 Цикл: критерий остановки
- 3 $\mathbf{W} \leftarrow$ вычисление при фиксированном \mathbf{H} распределено на d узлах
- 4 Обновление \mathbf{W} на всех узлах
- 5 $\mathbf{H} \leftarrow$ вычисление при фиксированном \mathbf{W} распределено на d узлах
- 6 Обновление \mathbf{H} на всех узлах



Сравнение методов

В эксперименте [R.Gemulla и др, 2013] использовался R-кластер:

- 16 узлов, каждый
 - Intel Xeon E5530 (2.4GHz, 8 ядер)
 - 48ГБ ОЗУ

Для сравнения методов:

- 8 узлов, каждый с 8 параллельными задачами

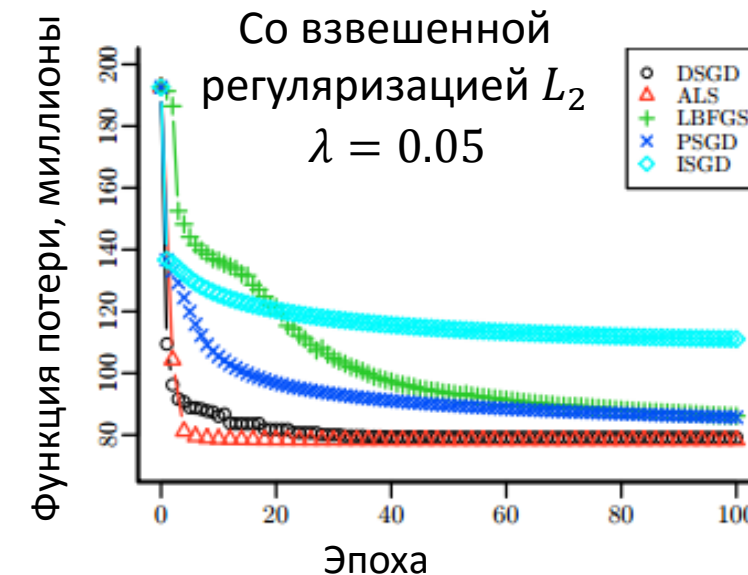
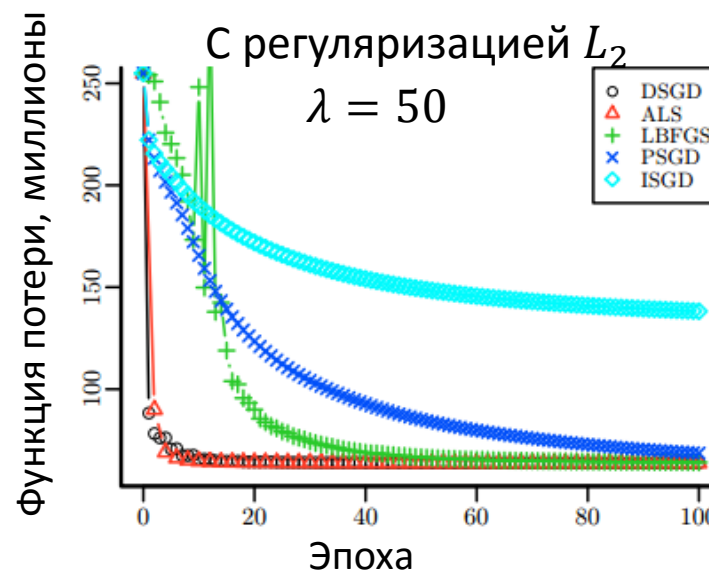
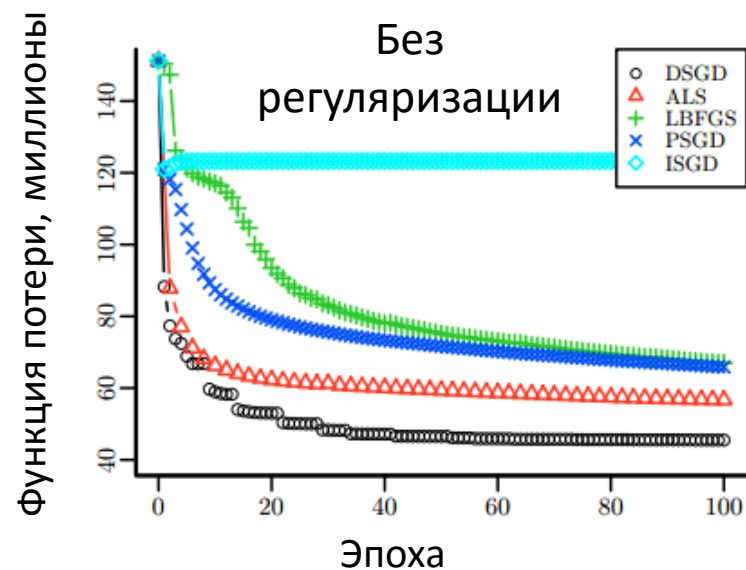
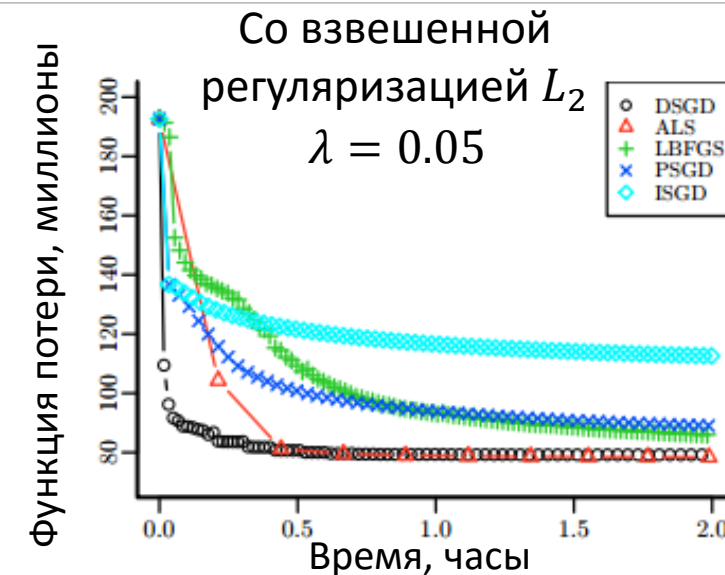
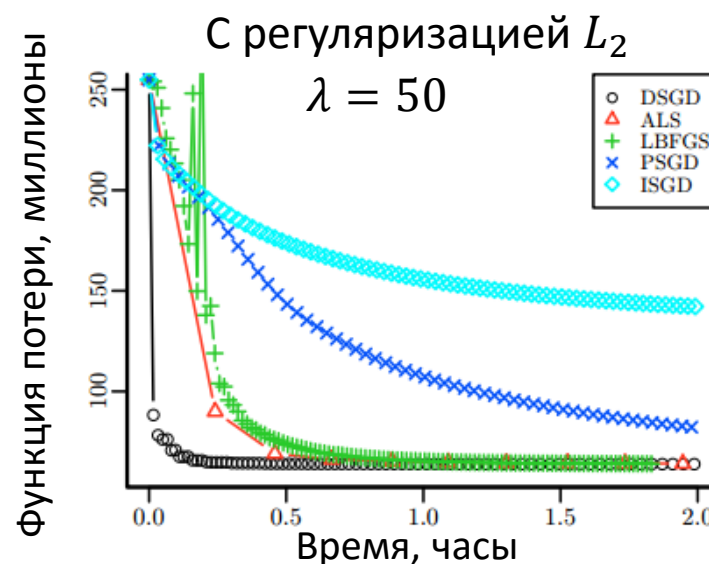
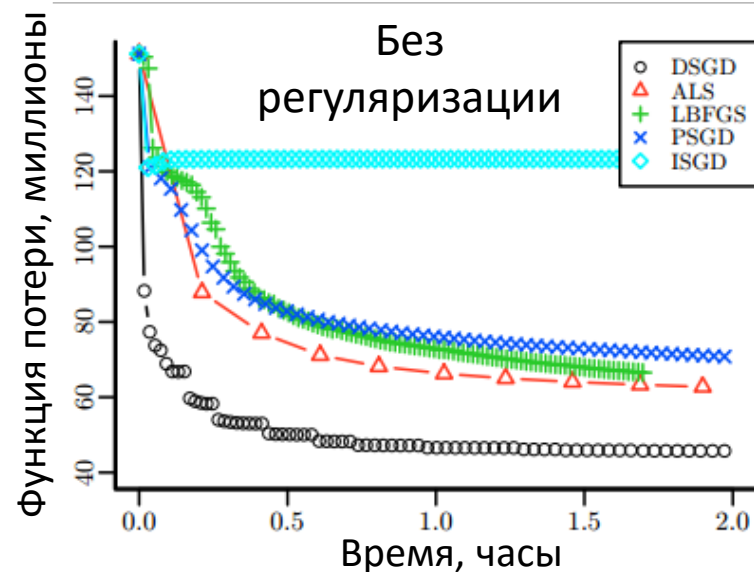
Набор данных Netflix:

- 100 миллионов рейтингов
- 480 тысяч пользователей
- 18 тысяч фильмов



Сравнение методов

[R.Gemulla и др, 2013]



Рекомендательные системы на базе Spark

Матрица рейтингов R

	1	2	3	4
1		2	2	3
2			5	
3	4		3	5
4		3	3	

Выбор модели (параметров)

$$\left\{ \begin{array}{c} \text{количество факторов,} \\ \text{коэф. регуляризации} \\ \text{и пр} \end{array} \right\}$$

Матрица объектов O

The diagram shows a data matrix with 3 rows (objects) and 3 columns (parameters). The columns are labeled 'id', 'title', and 'img'. The rows are labeled '1', '2', and '3'. The first column, corresponding to the 'id' parameter, is highlighted in green.

Обучение лучшей модели

The diagram illustrates the multiplication of two matrices, W and H , resulting in a product matrix (indicated by a dot \cdot between them).

Matrix W (Left):

- Rows are labeled W_{1*} , W_{2*} , W_{3*} , W_{4*} , W_{5*} , and W_{6*} .
- Columns are labeled "Пользователи" (Users) and "Факторы" (Factors).
- Row W_{3*} is highlighted in yellow.
- Row W_{4*} is highlighted in blue.

Matrix H (Right):

- Columns are labeled H_{*1} , H_{*2} , H_{*3} , H_{*4} , H_{*5} , and H_{*6} .
- Rows are labeled "Факторы" (Factors) and "Объекты" (Objects).
- Column H_{*3} is highlighted in green.
- Column H_{*4} is highlighted in orange.

Список для рекомендации

<i>user_id</i>						
<i>object_id</i>						

Рекомендации (прогнозирование)

(   и др.)

Список новых рейтингов

<i>user_id</i>					
<i>object_id</i>					
рейтинг	5	1	4	2	5

Добавление новых рейтингов

Повторное обучение



ALS в MLlib

Библиотека Mllib включает модуль Recommendation, который содержит класс для реализации распределенного ALS

Метод train класса ALS

```
train(ratings, rank, iterations=5, lambda_=0.01, blocks=-1, nonnegative=False, seed=None)
```

```
trainImplicit(...)
```

Основные параметры обучения ALS

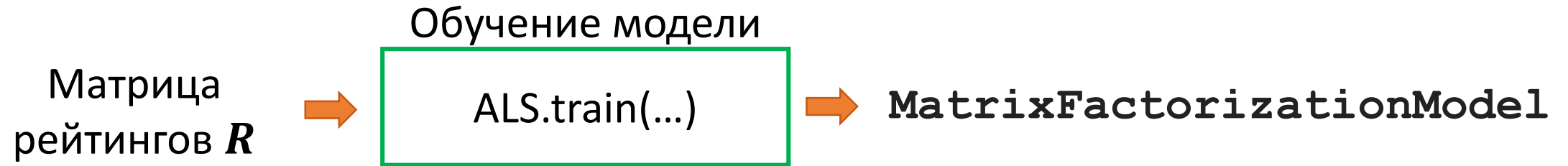
- *numBlocks* – число блоков
- *rank* - количество факторов
- *iterations* - количество итераций
- *lambda* - параметр регуляризации





Рекомендации на базе Spark и MLlib

Модель факторизованной матрицы



Рекомендации:

- `predict(user, product)`
- `predictAll(user_product)`
- `recommendProducts(user, num)`
- `recommendProductsForUsers(num)`
- `recommendUsers(product, num)`
- `recommendUsersForProducts(num)`

Другое:

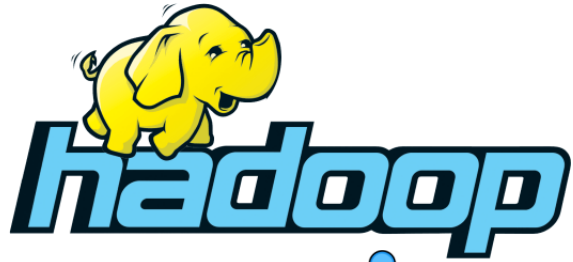
- `load(sc, path)`
- `productFeatures()`
- `userFeatures()`



MLlib



Другие решения

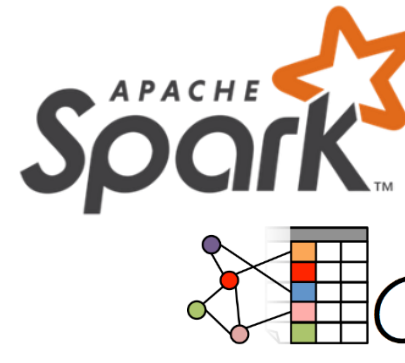


Frequent
Itemsets

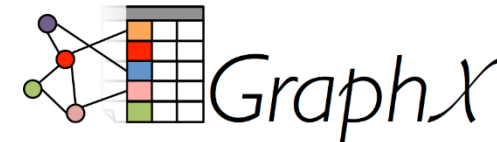


ALS

Item-Based и User-Based
коллаборативная
фильтрация



SVD++





ИСТОЧНИКИ

Efthalia Karydi and Konstantinos G. Margaritis «Parallel and Distributed Collaborative Filtering: A Survey», University of Macedonia, Department of Applied Informatics Parallel and Distributed Processing Laboratory

Qun Liu, Xiaobing Li «A New Parallel Item-Based Collaborative Filtering Algorithm Based on Hadoop» School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. Journal of Software, Volume 10, Number 4, April 2015

Simon Doms, Pieter Audenaert, Jan Fostier, Toon De Pessemier, Luc Martens «In-Memory, Distributed Content-Based Recommender System». Agency for Innovation by Science and Technology (IWT Vlaanderen)

Boduo Li, Sandeep Tata, Yannis Sismanis «Sparkler: supporting large-scale matrix factorization». EDBT '13 Proceedings of the 16th International Conference on Extending Database Technology. PP 625-636, 2013

Yehuda Koren «Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model». AT&T Labs, August 24–27, 2008, Las Vegas, Nevada, USA

Xavier Amatriain «Recommender Systems». Machine Learning Summer School, Published on Jul 21, 2014

Xavier Amatriain «Big & Personal: data and models behind Netflix recommendations». BigMine'13 August 2013. Chicago, Illinois, USA



ИСТОЧНИКИ

R. Gemulla, P. J. Haas, E. Nijkamp, Y. Sismanis «Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent», IBM Research Report RJ10481, March 2011 Revised February, 2013

Yunhong Zhou, Dennis Wilkins, Robert Schreiber, Rong Pan «Large-Scale Parallel Collaborative Filtering for the Netflix Prize» AAIM '08 Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management. PP 337–348, Springer-Verlag Berlin, Heidelberg, 2008

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, Ciyu Zhu «A limited memory algorithm for bound constrained optimization». SIAM Journal on Scientific Computing, Volume 16 Issue 5, Sept. PP 1190-1208, 1995

Boduo Li, Sandeep Tata, Yannis Sismanis «Sparkler: supporting large-scale matrix factorization». EDBT '13 Proceedings of the 16th International Conference on Extending Database Technology. PP 625-636, 2013

Trevor Hastie, Rahul Mazumder, Jason D. Lee, Reza Zadeh «Matrix completion and low-rank SVD via fast alternating least squares». The Journal of Machine Learning Research, Volume 16, Issue 1, PP 3367-3402, 2015

Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, Inderjit Dhillon «Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems». ICDM '12 Proceedings of the 2012 IEEE 12th International Conference on Data Mining, PP 765-774, 2012



Спасибо за внимание!