Построение рекомендательных систем с использованием нейронных сетей

Николай Карпов
Доцент каф. ПМИ, НИУ ВШЭ, Нижний Новгород
New Tang Dynasty Television (ntd.tv), Нью Йорк, США

Научно-учебная группа
«Анализ мультимедийных данных»
НИУ ВШЭ, г. Нижний Новгород

План

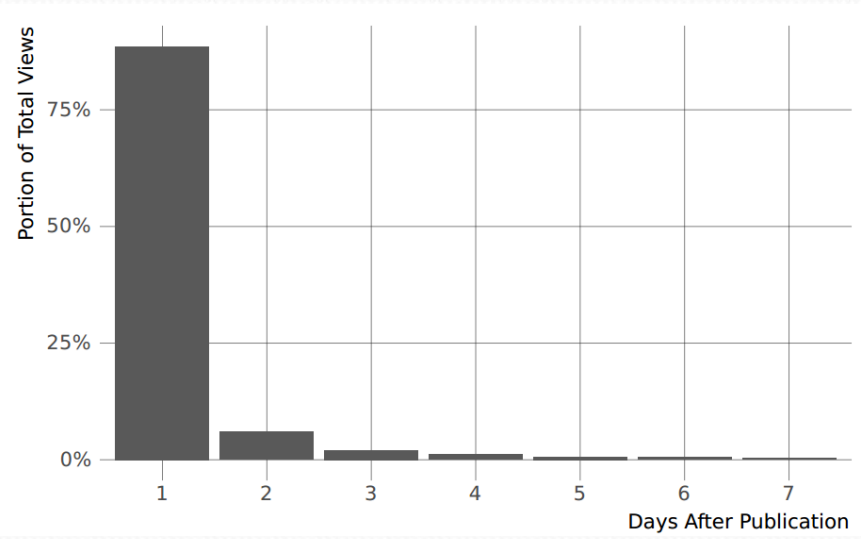
- **Мотивация**
- Модель матричной факторизации
- Модель для решения проблемы «холодного старта»
- Модели на основе нейронных сетей
- Развертывание
- Заключение

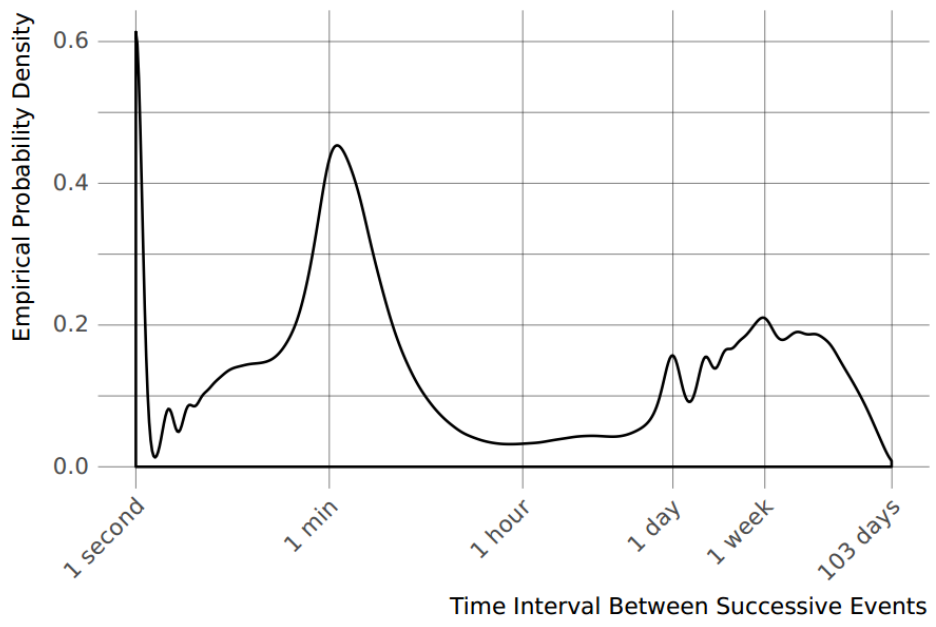
Мотивация

- Большенство сайтов имеют только неявную обратную связь
- Современные модели (BRP, WARP) не принимают во внимание время
- Нужно решать проблему холодного старта для новых пользователей и новых публикаций
- Алгоритм должен поддреживать GPU для быстрого дообучения

Новостной набор данных

- 600-800 тысяч уникальных пользователей
- 2000-3000 новостных стей в день
- 10 – 20 миллионов взаимодействий с пользоватлями
- 86% просмотров в первые 24 часа
- 90% новых пользователей

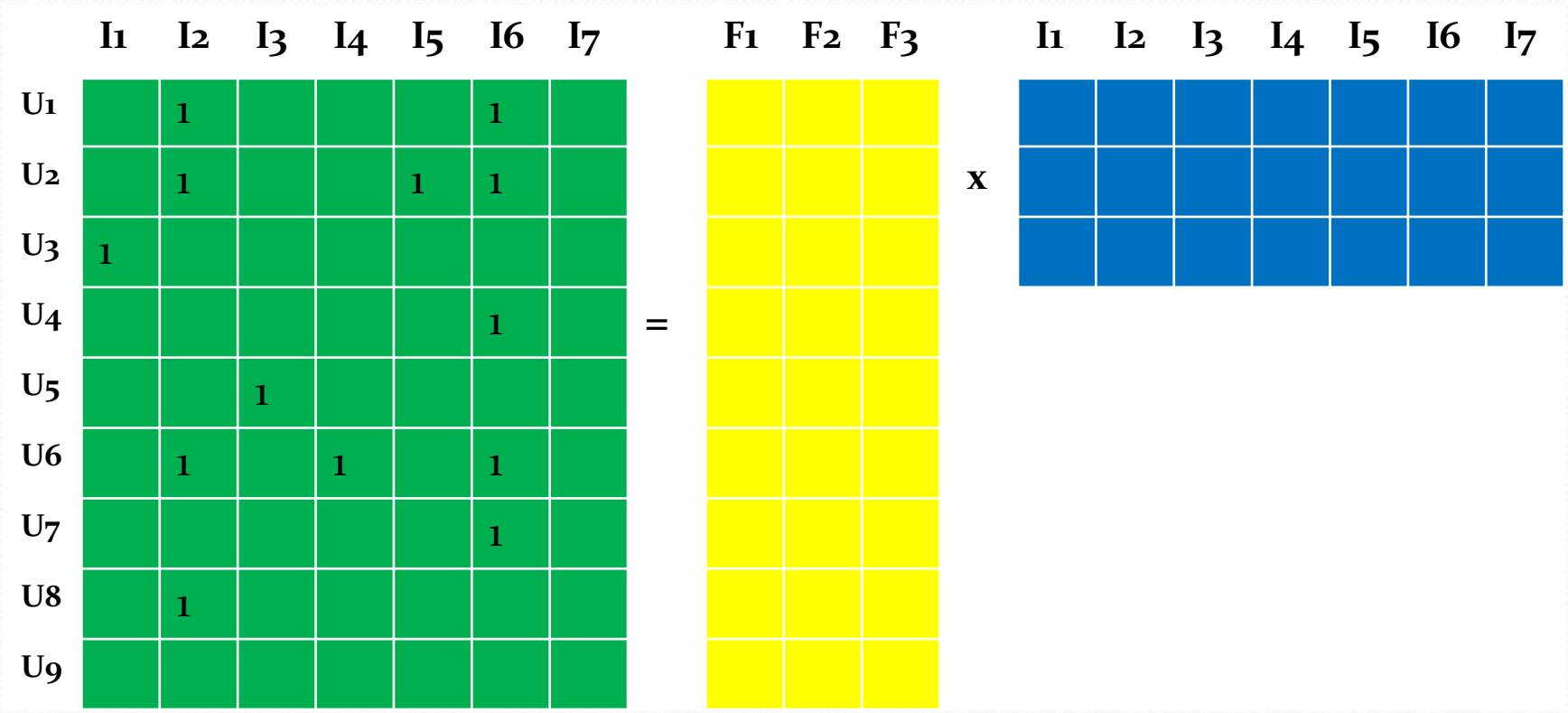




План

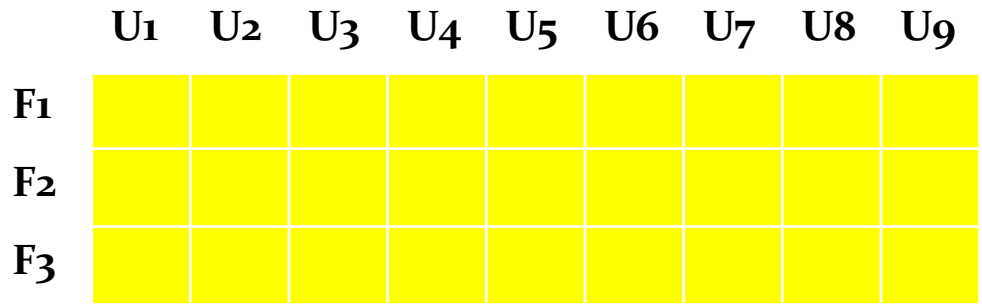
- Мотивация
- **Модель матричной факторизации**
- Модель для решения проблемы «холодного старта»
- Модели на основе нейронных сетей
- Развертывание
- Заключение

Матричная факторизация

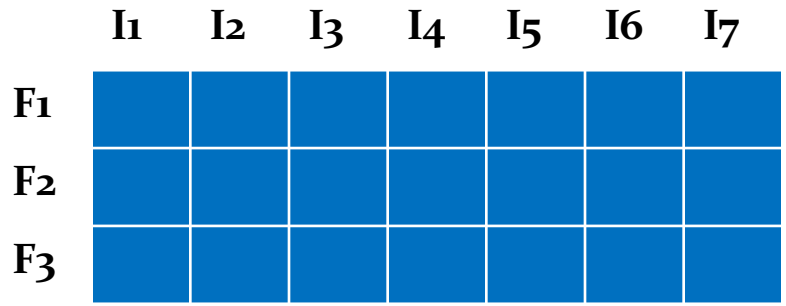


Матричная факторизация

$\Phi_U(x)$



$\Phi_I(y)$



$$f_y(x) = \Phi_U^T(x) \times \Phi_I(y)$$

Целевые функции для неявной обратой связи

- Bayesian Personalized Ranking (BPR) [1]

Парное семплирование, оптимизируем AUC

- Weighted Approximate-Rank Pairwise (WARP) [2]

Оптимизируем только топ в ранцированном списке

$$L_{\text{BPR}} = \sum_{i,j \in D_u} \ln \sigma(f_u(i) - f_u(j))$$

$$L_{\text{WARP}} = \sum_{i,j \in D_u} \Phi(rank_u(i)) |1 - f_u(i) + f_u(j)|_+$$

1. S.Rendle, C.Freudenthaler and Z.Gantner 2009
2. J. Weston, H. Yee, and R. J. Weiss 2013

Позитивное и негативное семплирование

	I1	I2	I3	I4	I5	I6	I7
U1		1				1	
U2		1			1	1	
U3	1						
U4						1	
U5			1				
U6		1		1		1	
U7						1	
U8		1					
U9							

Positive	
U1	I2
U2	I2
U3	I1
U1	I6
U2	I5
U2	I6

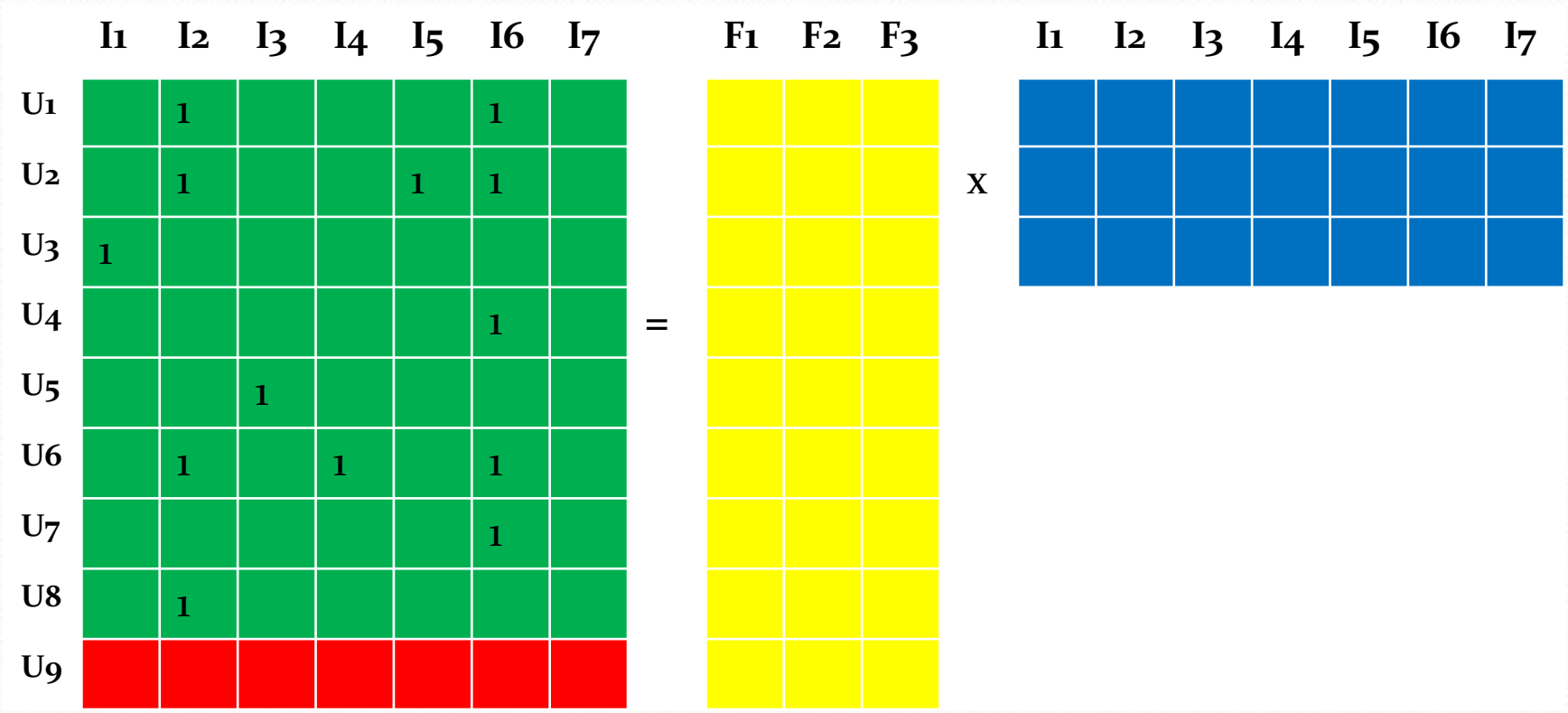
Negative	
U1	I4
U2	I1
U3	I6
U1	I1
U2	I3
U2	I7

$f_u(i)$ $f_u(j)$

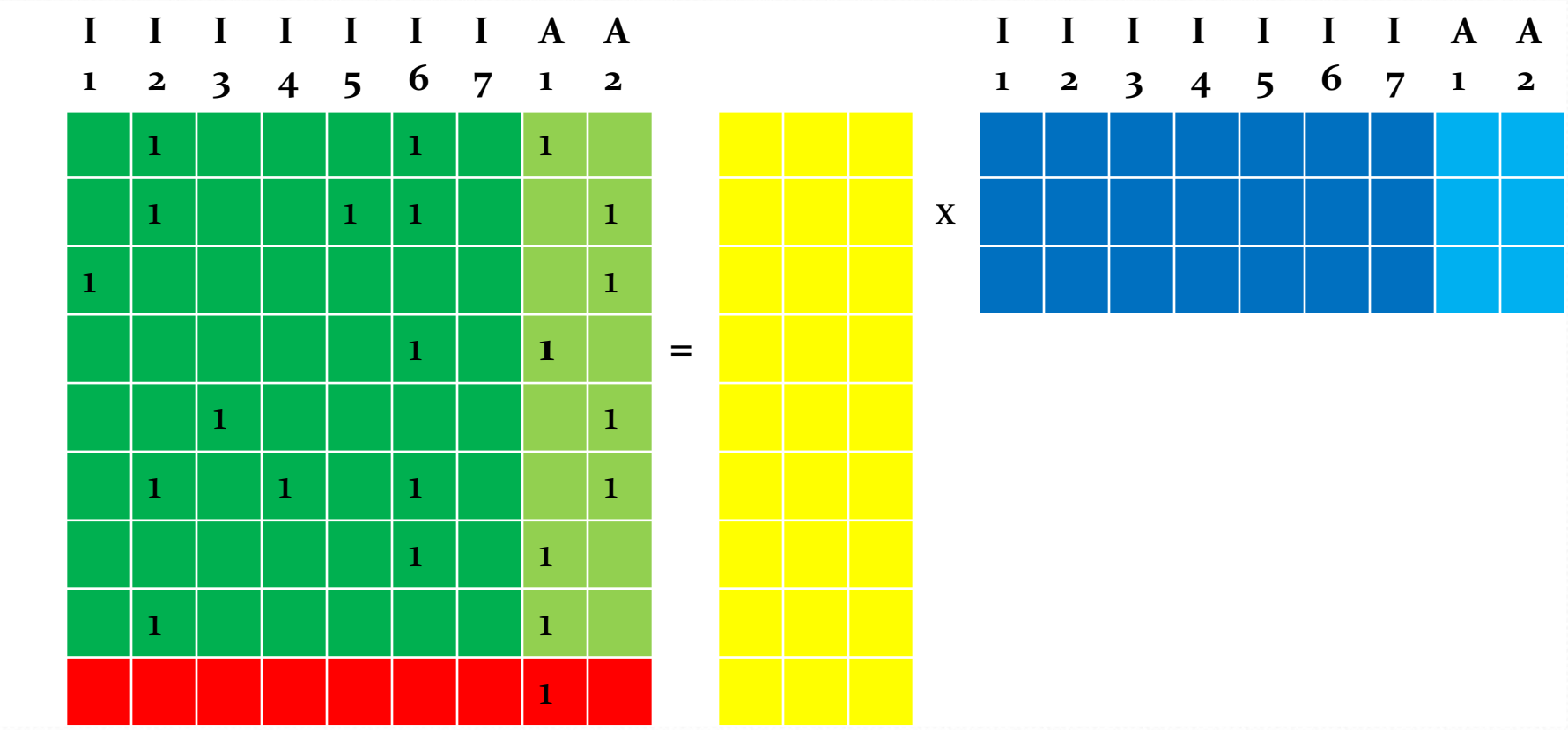
План

- Мотивация
- Модель матричной факторизации
- **Модель для решения проблемы «холодного старта»**
- Модели на основе нейронных сетей
- Развертывание
- Заключение

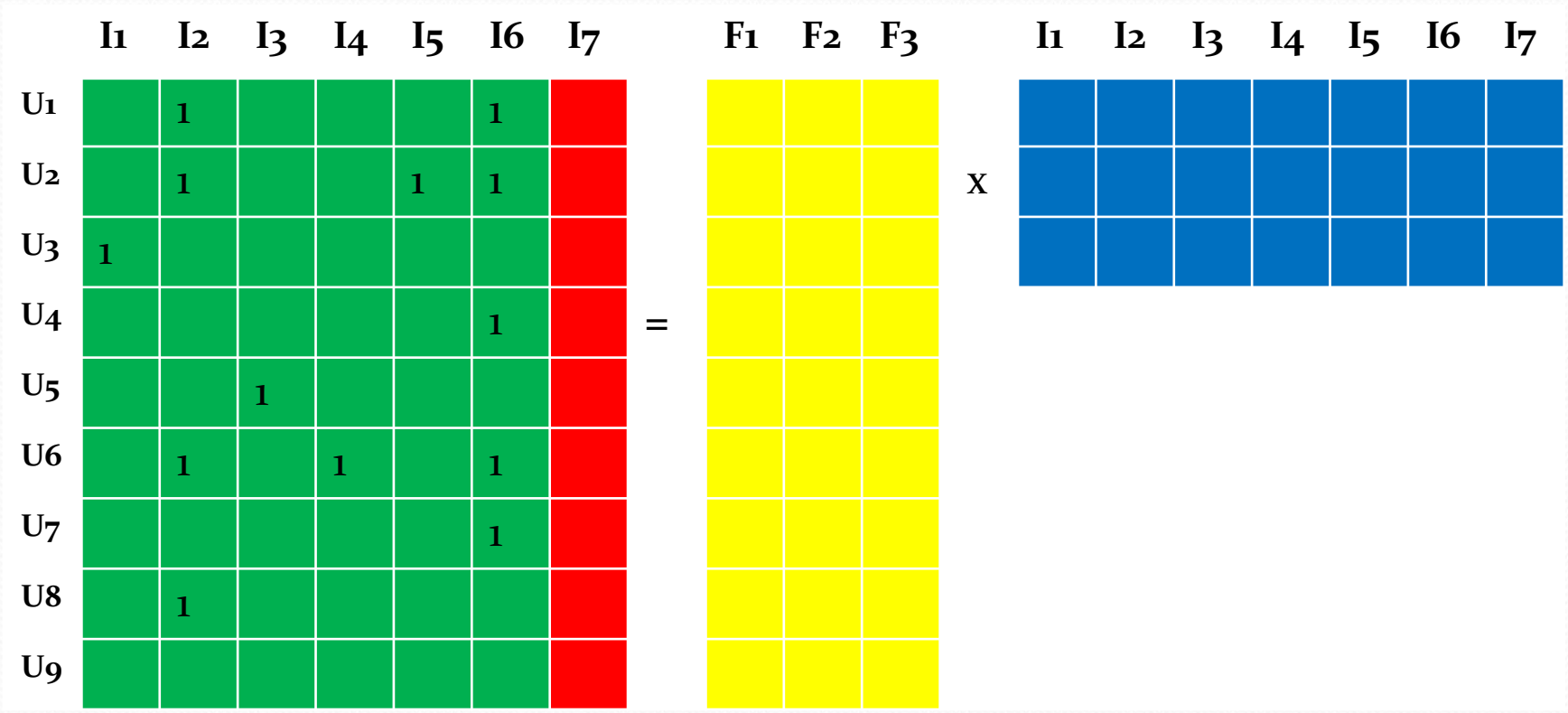
Проблема холодного старта для нового пользователя



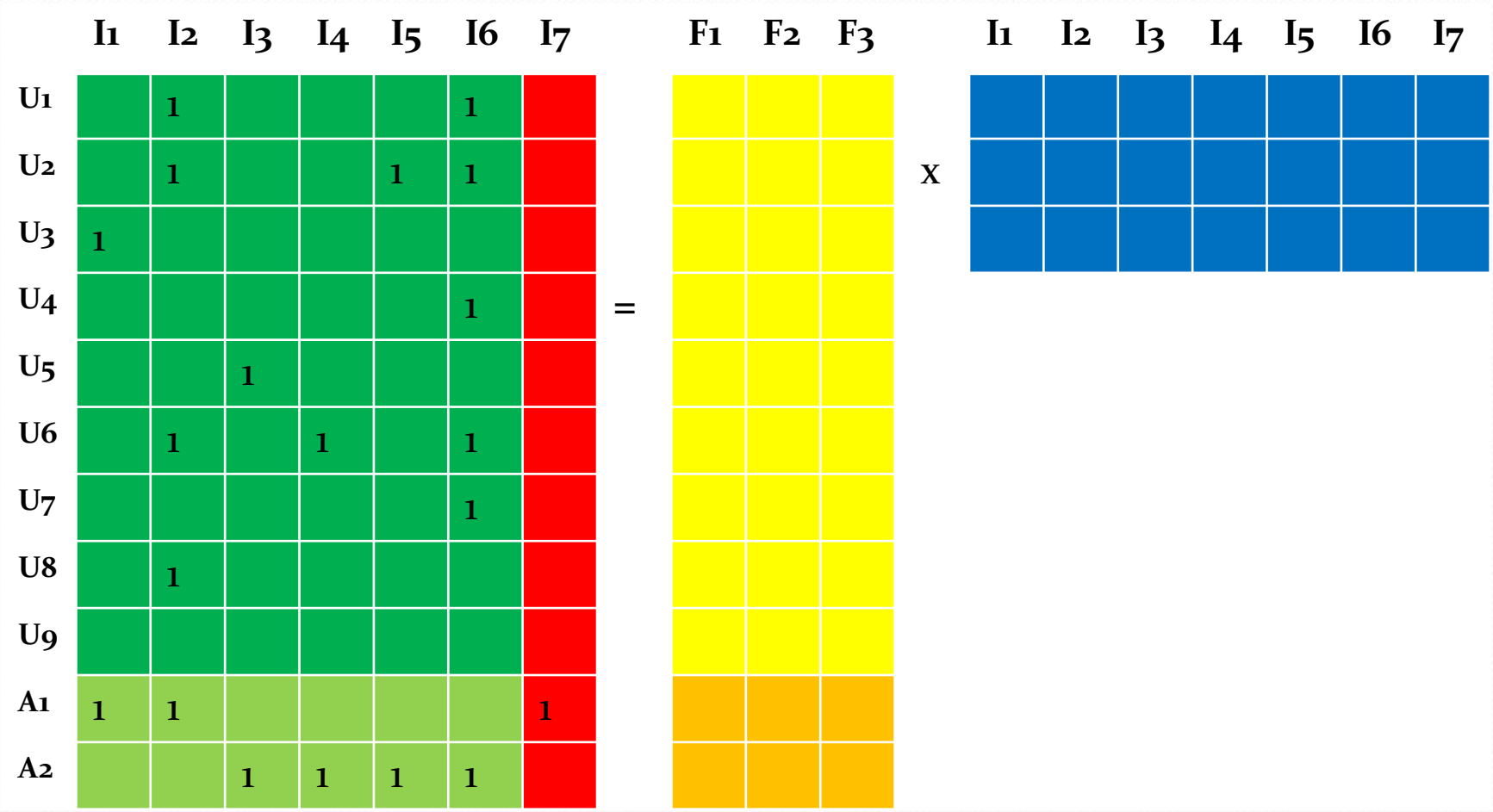
Дополнительные признаки для пользователя



Проблема холодного старта для новой публикации



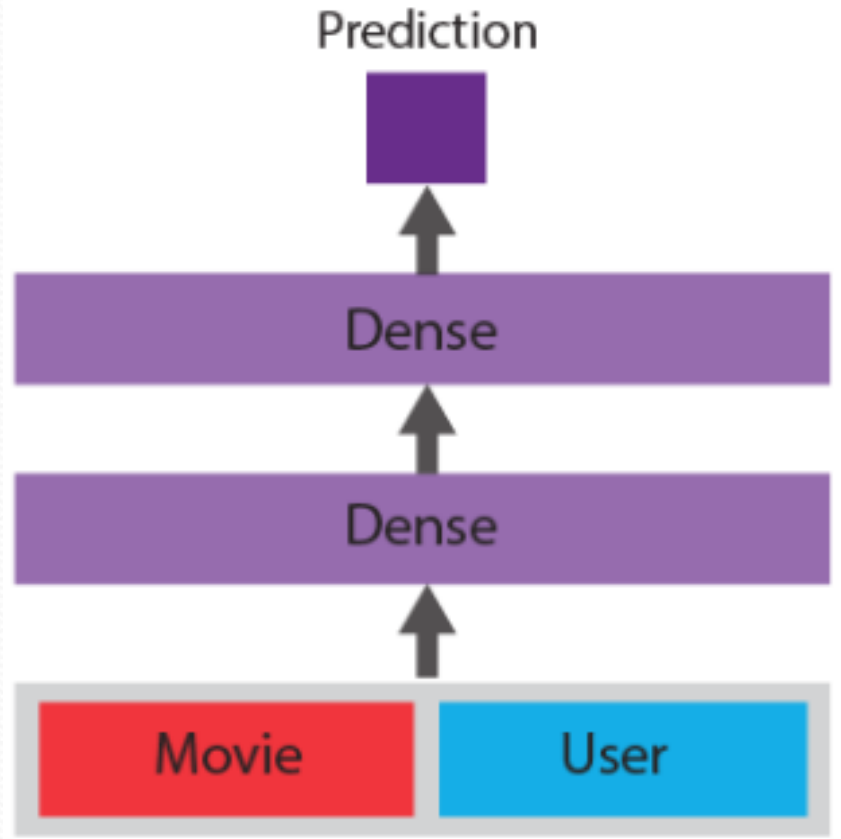
Проблема холодного старта для новой публикации



План

- Мотивация
- Модель матричной факторизации
- Модель для решения проблемы «холодного старта»
- **Модели на основе нейронных сетей**
- Развертывание
- Заключение

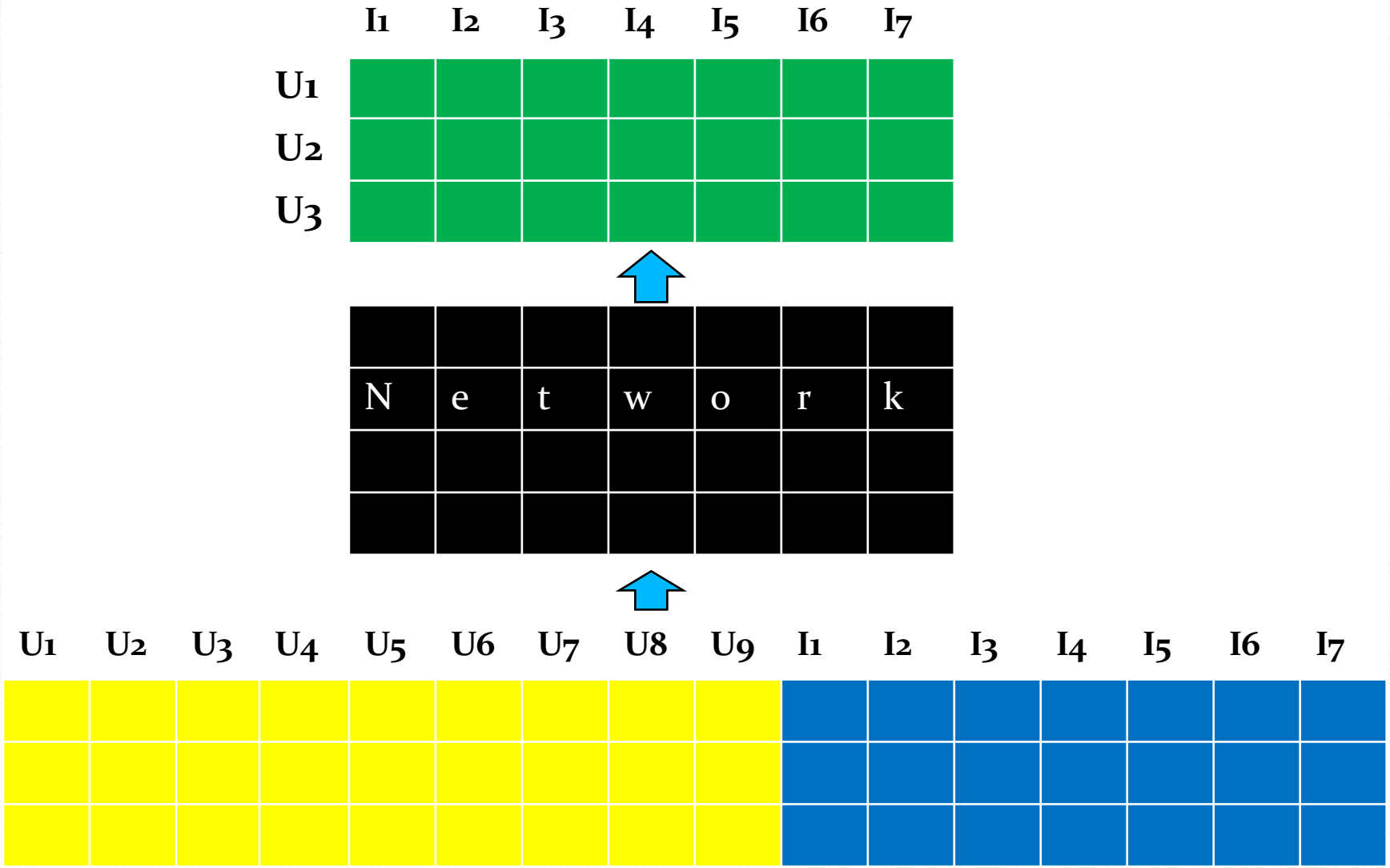
Глубокая матричная факторизация



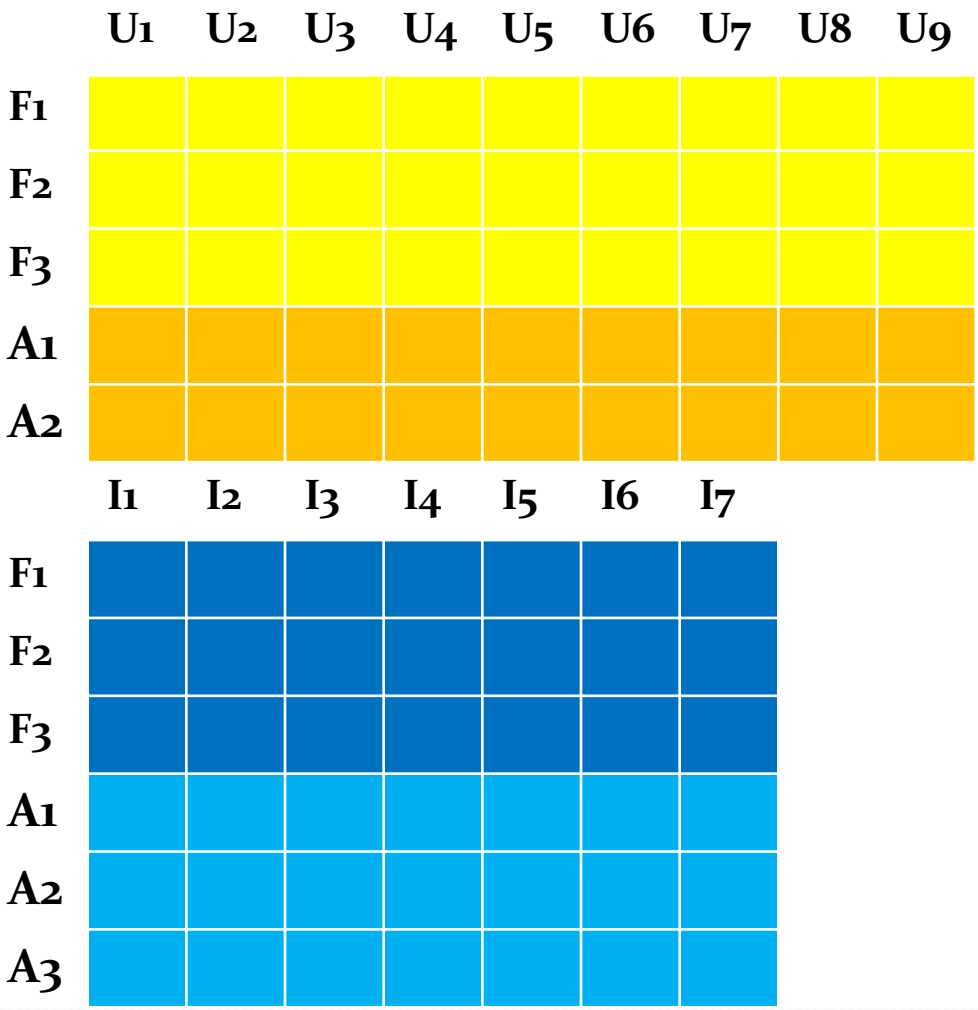
Нейронная сеть может задавать нелинейную функуию в отличии от матричной факторизации

https://www.oreilly.com/ideas/deep-matrix-factorization-using-apache-mxnet

Глубокая матричная факторизация

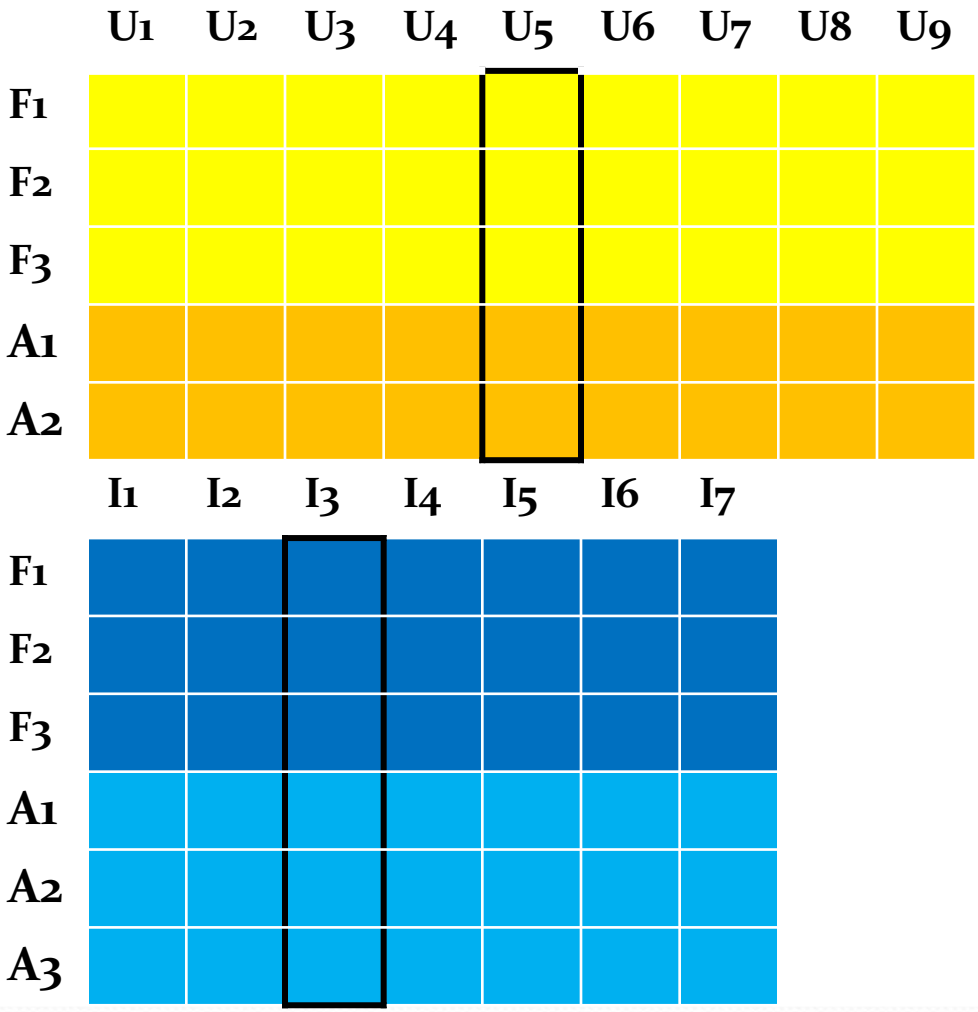


Глубокая матричная факторизация



Т.к. нейронная сеть может задавать нелинейную функуию мы можем подавать на ее вход склеенную матрицу признаков пользователя и статьи

Глубокая матричная факторизация

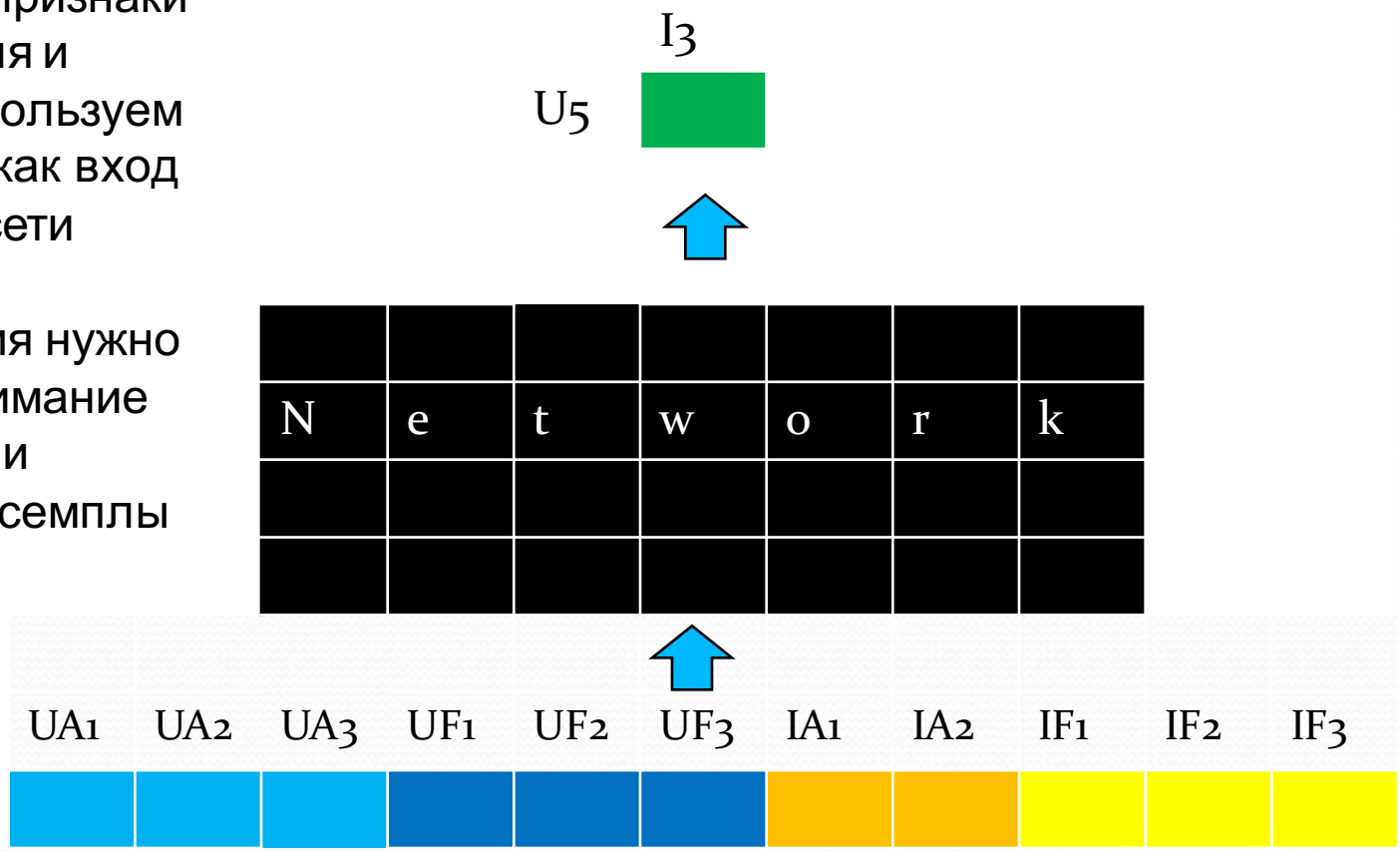


Слой Lookup делает выборку необходымы колонок в матрице

Глубокая матричная факторизация

Склееваем признаки пользователя и статьи и используем этот вектор как вход нейронной сети

Для обучения нужно брать во внимание позитивные и негативные семплы



Наши данные это последовательность событий

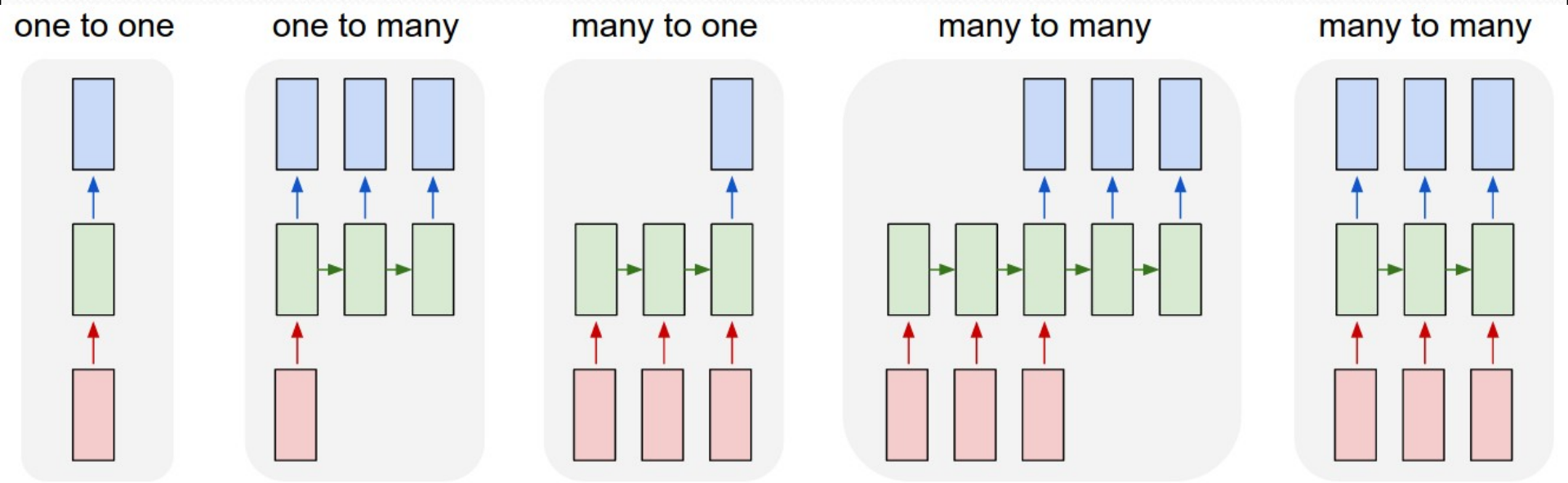
	I1	I2	I3	I4	I5	I6	I7
U1		1				1	
U2		1			1	1	
U3	1						
U4						1	
U5			1				
U6		1		1		1	
U7						1	
U8		1					
U9							

Positive

U1	I2
U2	**I2**
U3	I1
U1	I6
U2	**I5**
U2	**I6**

$$f_u(i)$$

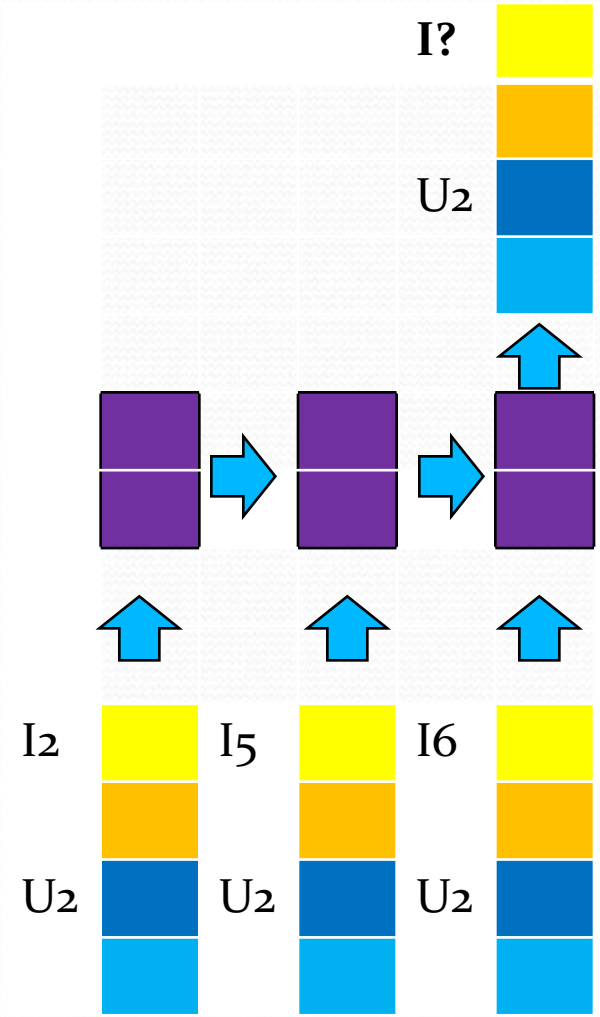
Нейронные сети могут давать болле сложные модели



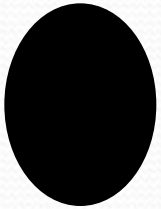
Используя рекурентную нейронную сеть мы можем учесть вермя

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Модель на основе LSTM



Рекурентная нейронная сеть (RNN) позволяет предсказать следующее событие на основе последовательности предыдущих. Кажное событие задано в виде вектора.



Как разворачивать рекомендательную систему?

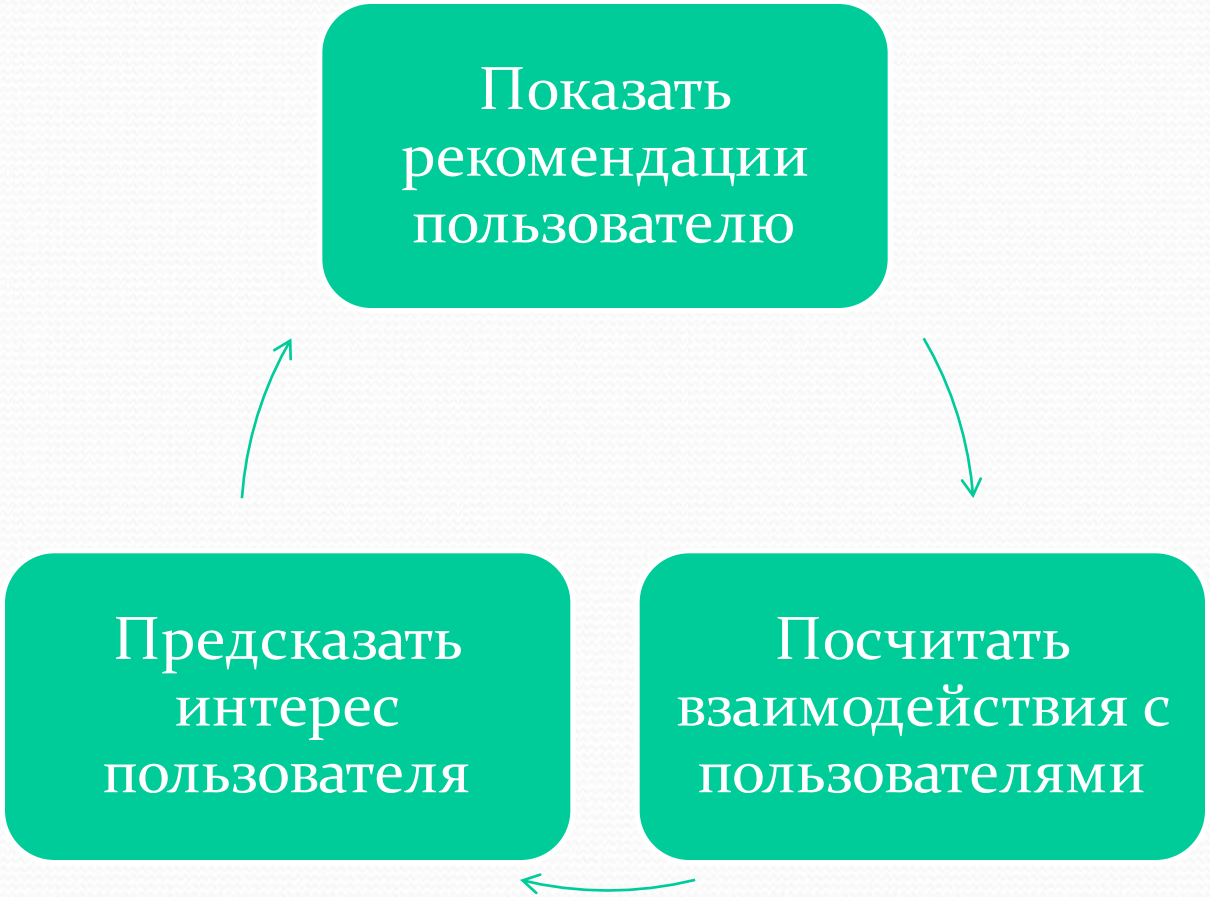
Развертывание

- Обучаем модель на мошьном сервере с GPU
- Используем модель для предсказания на другом сервере
- Обеспечиваем API для взаимодестивия с UI
- Переодически обновляем модель

Технологии: TensorFlow, Docker, NGINX, Flask, uWSGI

Для TensorFlow существует TensorFlow Serving. Но мы испоьзуем другой подход который применим не только для TensorFlow моделей.

Цикл анализа данных

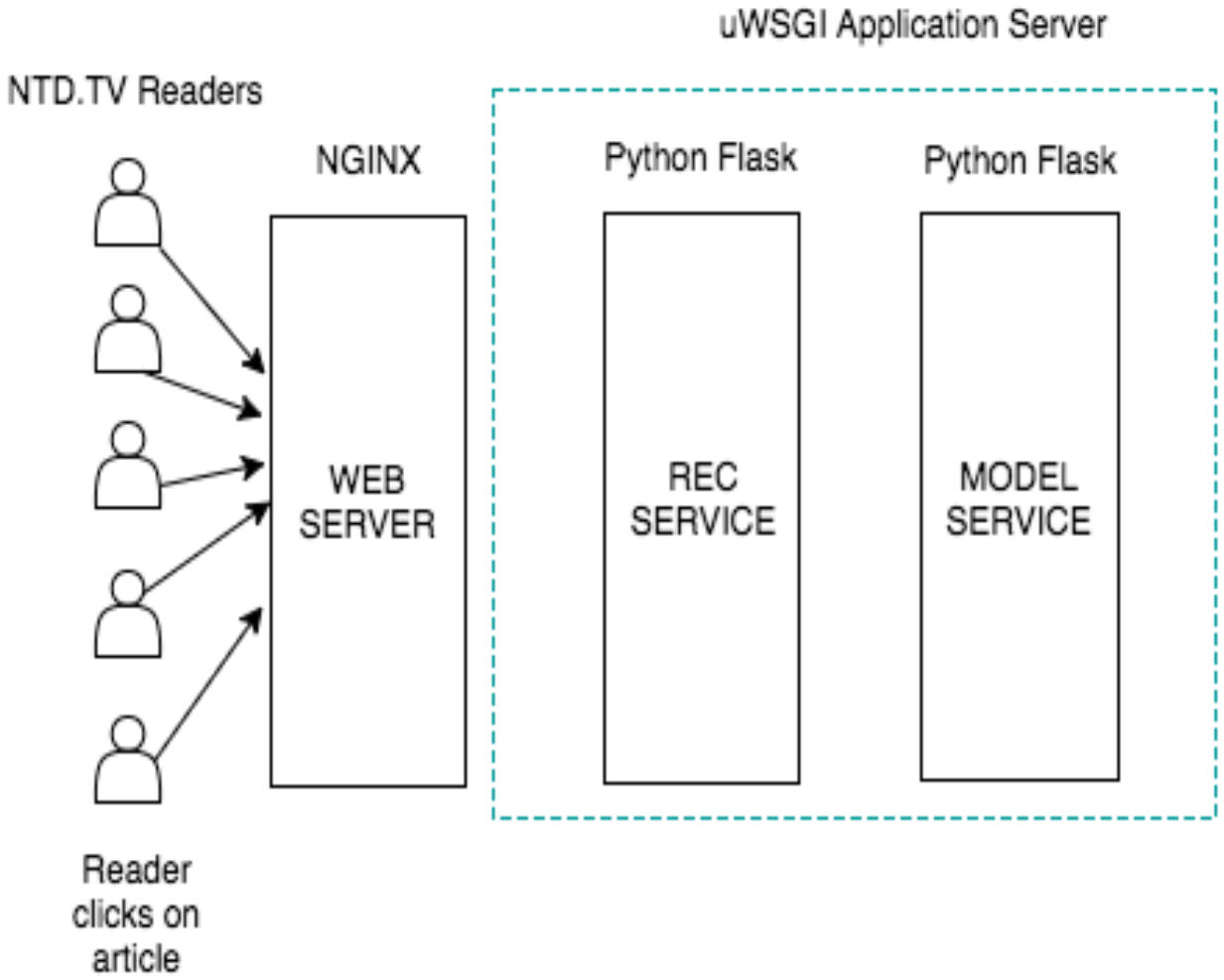


Технологический процесс рекомендательной системы

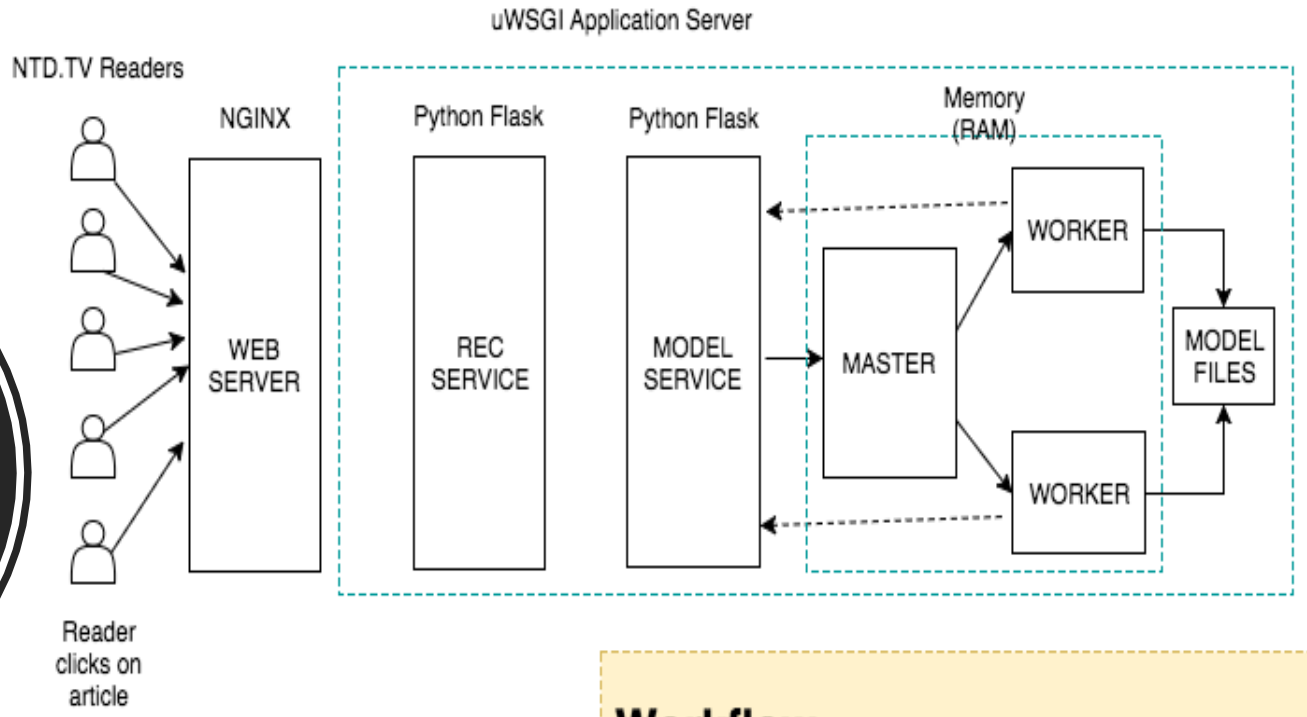
Техническая реализация

THANK YOU KANNAN SANKARAN FOR THIS SECTION

Technical Implementation- Initial Setup



Technical Implementation – How services are started



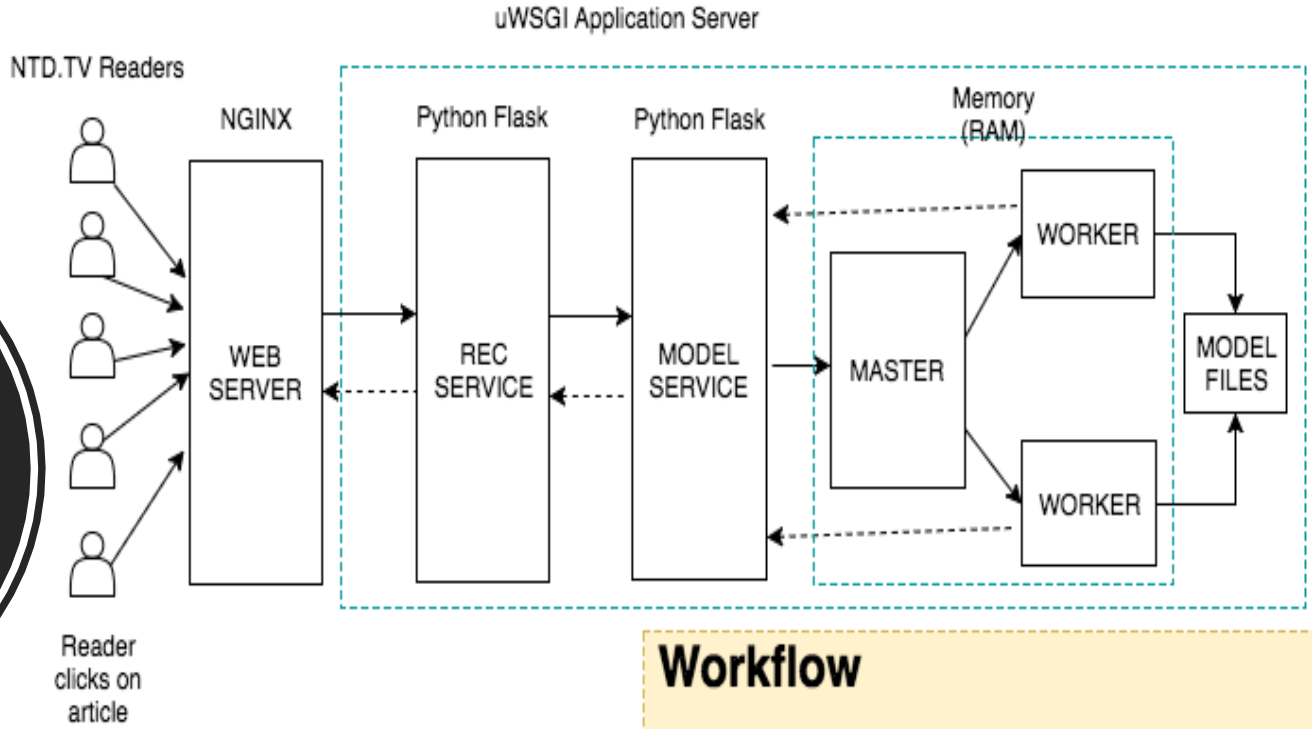
Workflow

1) uWSGI application server starts the Rec service and Model service

2) Model service creates a master, and master creates 2 workers

3) Each worker loads the model files that will be used for recommendation

Technical Implementation – How user requests are handled



Workflow

1) Reader clicks on article

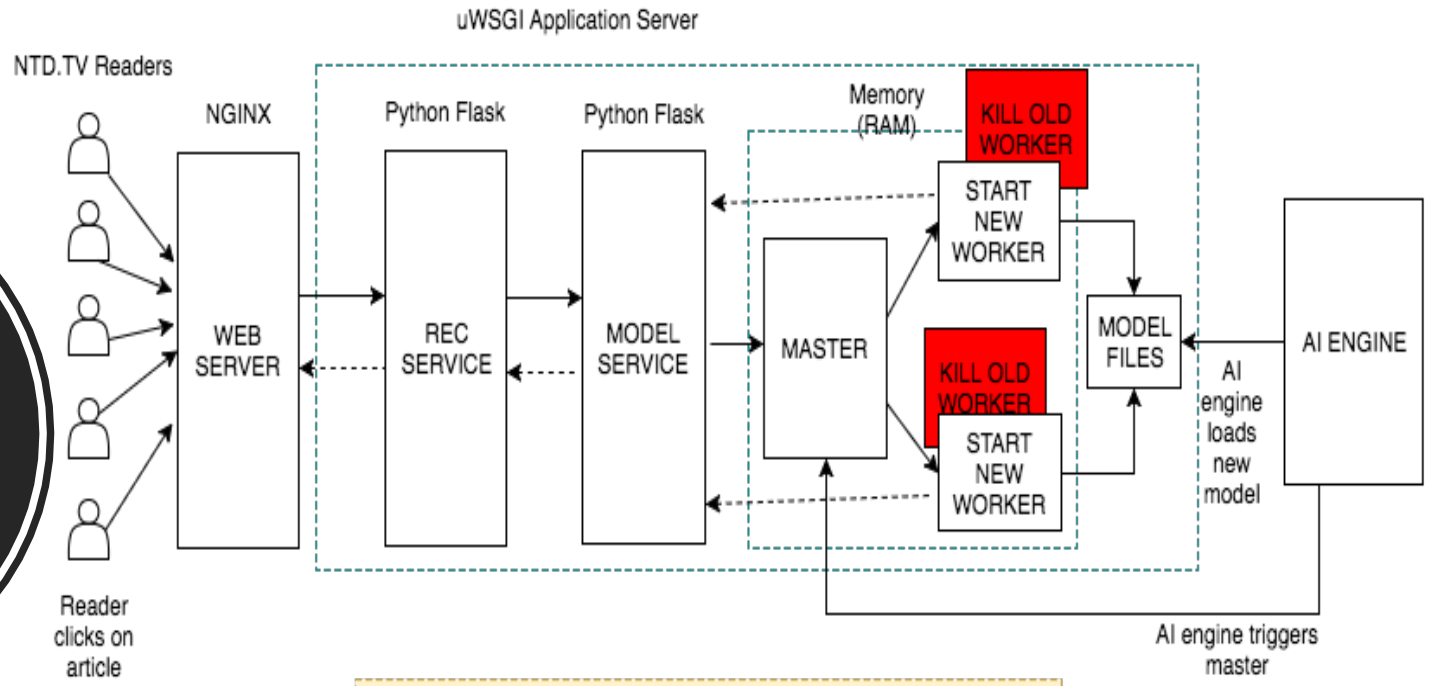
2) Web server forwards the request to Rec service

3) Rec service makes a call to the model service, and passes user_id and top N items as parameters

4) Model service forwards the request to one of the workers that uses the model to predict, and returns article URLS as recommendations to model service

5) Rec service then calls SQL database to get title and image for each article, and sends back the result to the reader's webpage

Technical Implementation – How Model Reloading Works



Workflow

1) AI Engine loads new model, and triggers master

2) Master kills workers, one at a time. This is called chain-reloading.

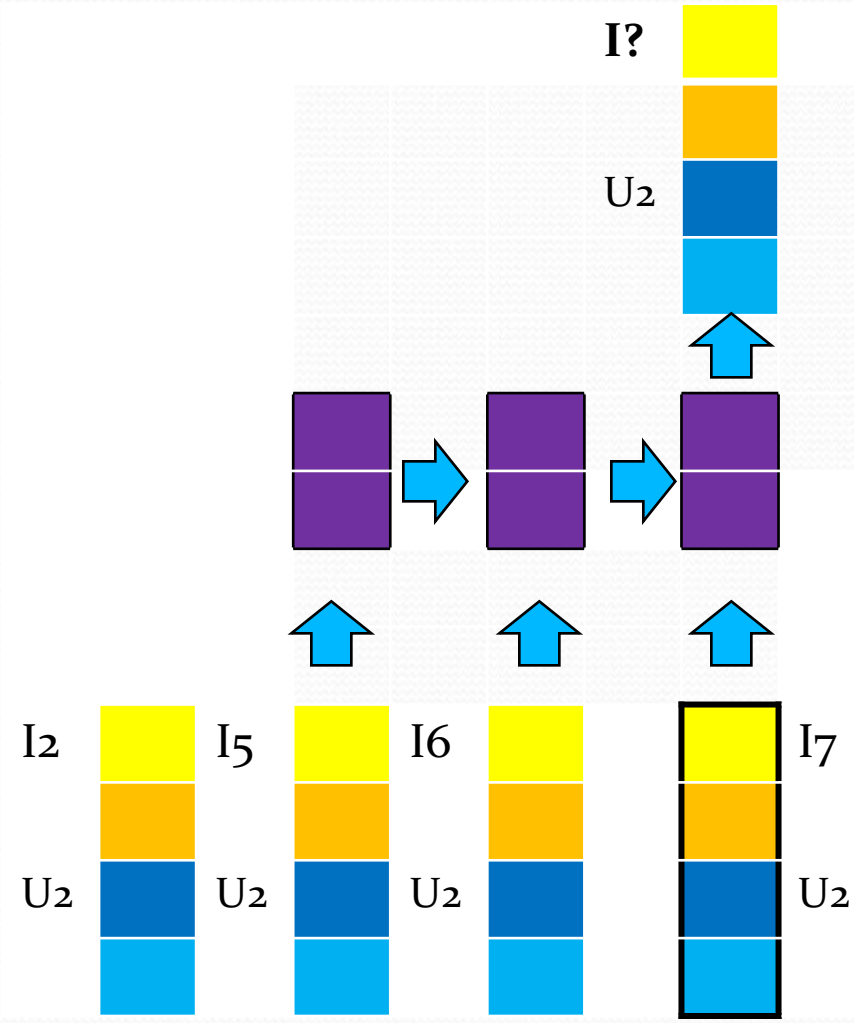
3) Master creates new workers, one at a time

4) Each worker reloads the model, and responds to requests as before

Свойства построенной системы

- Предложенная архитектура обеспечивает насколько сотен API вызовов в секунду.
- При необходимости есть возможность масштабировать горизонтально
- Мы принимаем во внимание дополнительные признаки пользователя и стать для решения холодного старта
- Есть возможность быстрого обновления рекомандаций без дообучения модели

Модель на основе LSTM



Когда пользователь совершает новое заимодействие с контентом мы можем получать новые рекомендации подавая на вход системы новую последовательность без дообучения модели

Новое взаимодействие пользователя

План

- Мотивация
- Модель матричной факторизации
- Модель для решения проблемы «холодного старта»
- Модели на основе нейронных сетей
- Развертывание
- **Заключение**

Заключение

- Мы предложили модель на основе нейронной сети для неявной обратной связи.
- Для учета динамики во времени мы использовали несколько реализаций рекурентных рейронных сетей с памятью.
- Был реализовон алгоритм с целевой функцией WARP на TensorFlow.
- Алгоритмы были опробаваны на собственном наборе данных который мы выложили в открытый доступ.
- Алгоритм позволяет решить проблему холодного старта для нового пользоватлея и новой статьи.
- Система была развернута в облаке и поддерживает несколько сотен запросов в секунду.

Благодарности

- Научно-учебная группа «Анализ мультимедийных данных»
- New Tang Dynasty Television (**ntd.tv**) за возможность работать с данными
- Kannan Sankaran за техническую поддержку при разработке проекта

Спасибо за внимание!

Николай Карпов

Пишите мне:

nkarpov@hse.ru