

Рекомендательные системы на основе вложенных тегов

Исполнил:

Чеботаев А. П.

Научный руководитель:

Бухановский А. В., д.т.н

Кафедра Компьютерных Технологий, ФИТиП

СПб 2011

Рекомендательные системы

- C – пользователи, S – объекты
- $u : C \times S \rightarrow \mathbb{R}$ – функция «полезности»

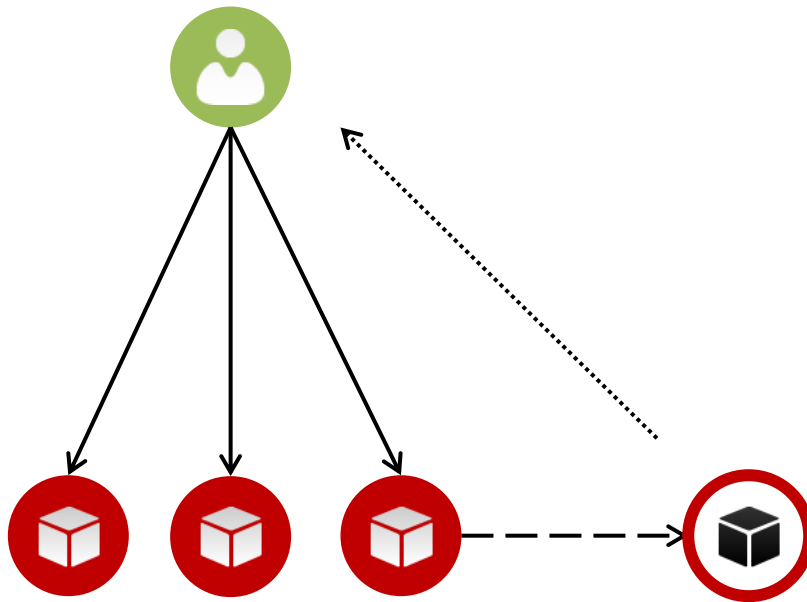
Современные РС

- Упорядочивание

Направления развития

- Обоснование
- Фильтрация
- Поиск по критериям

Контентные методы

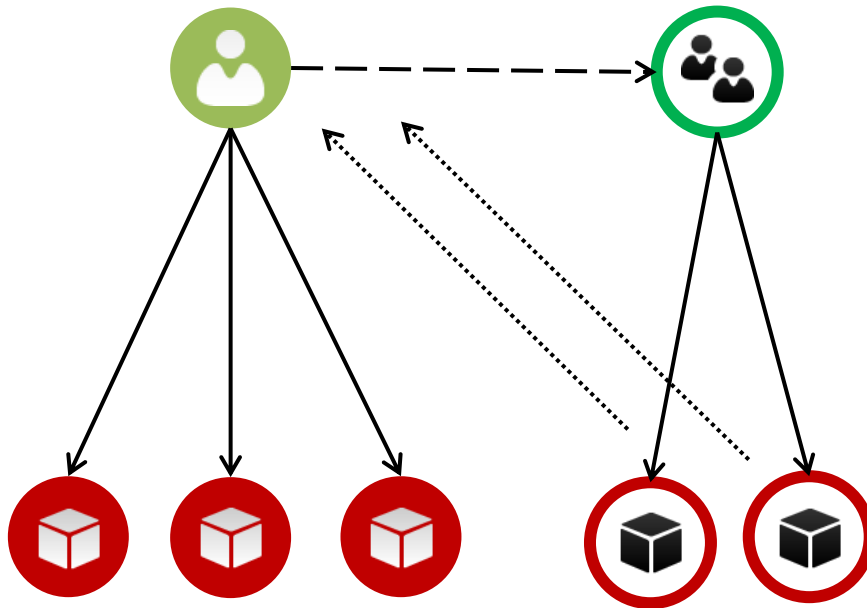


- ✗ Ограниченность анализа
- ✗ Фиксированная предметная область
- ✗ Узкие рекомендации

amazon.com



Совместная фильтрация



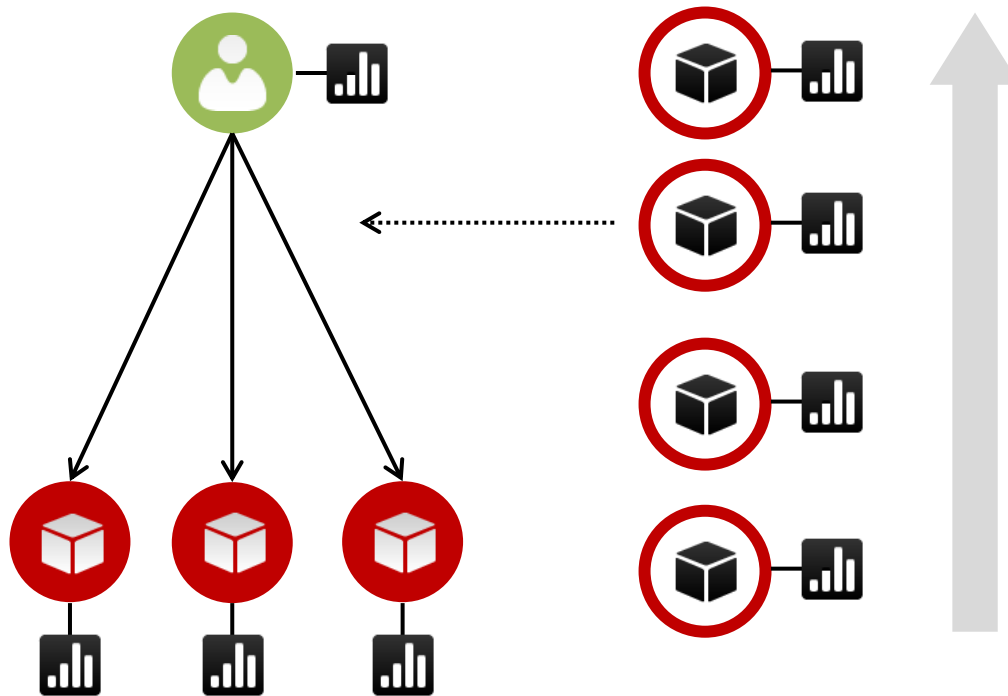
- ✗ Новый объект/пользователь
- ✗ Избирательность внимания
- ✗ Ресурсоемкость

last.fm

NETFLIX

imhonet
рекомендательный
сервис

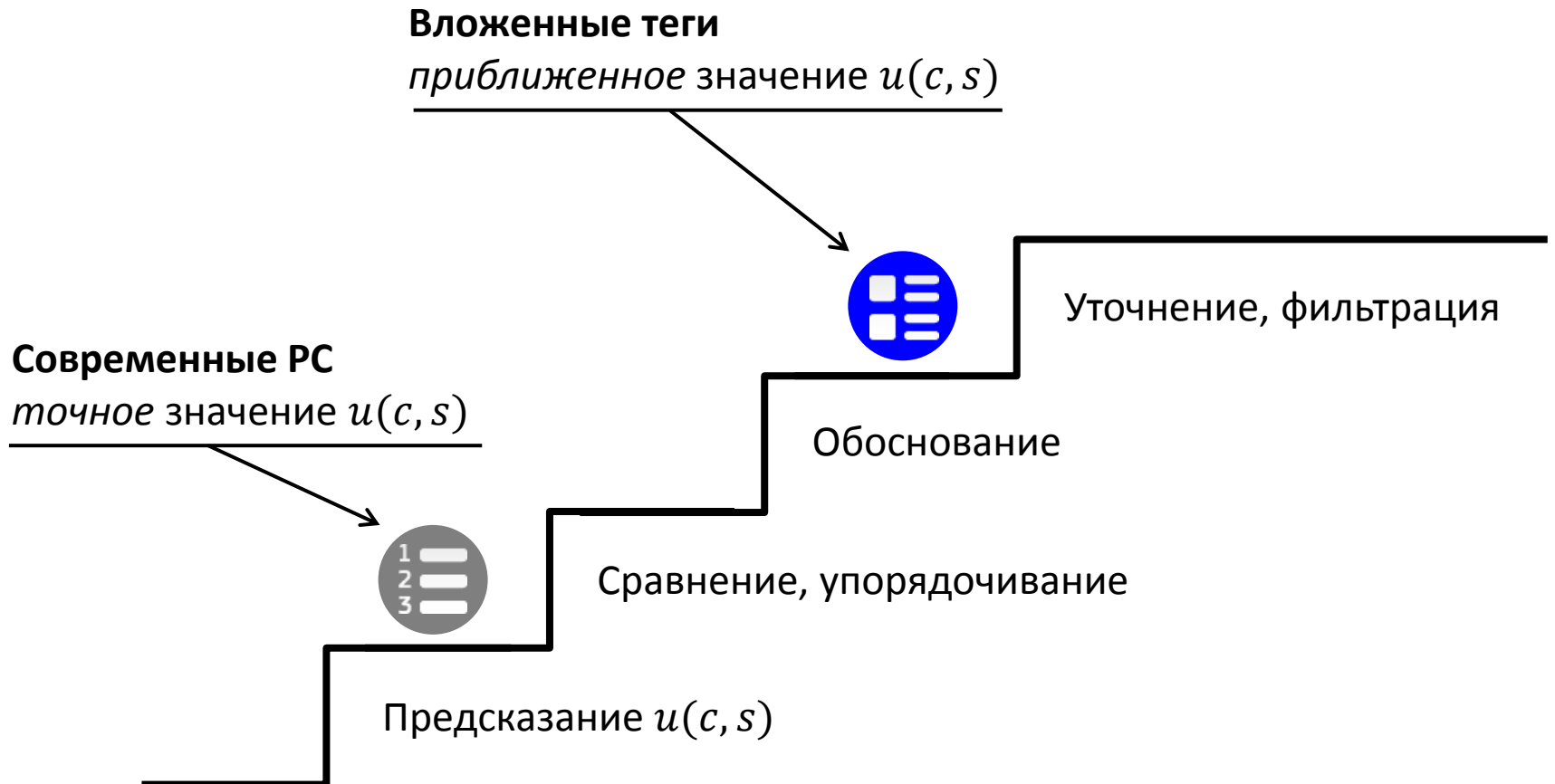
Скрытые факторы



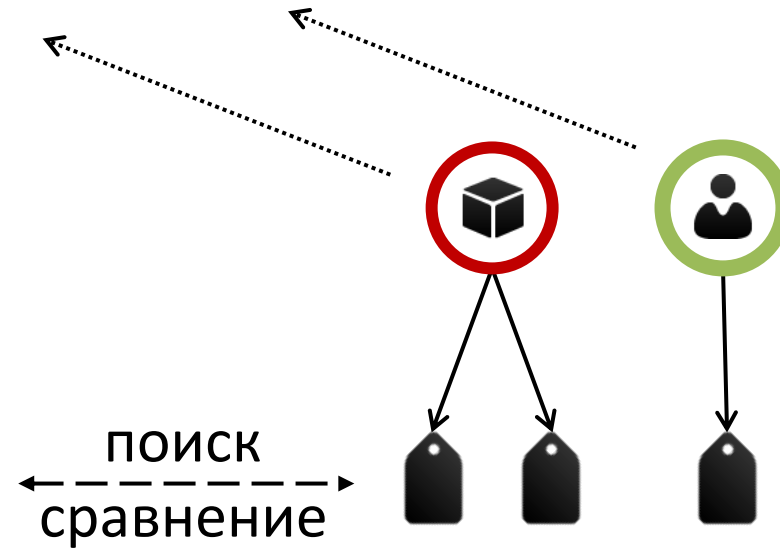
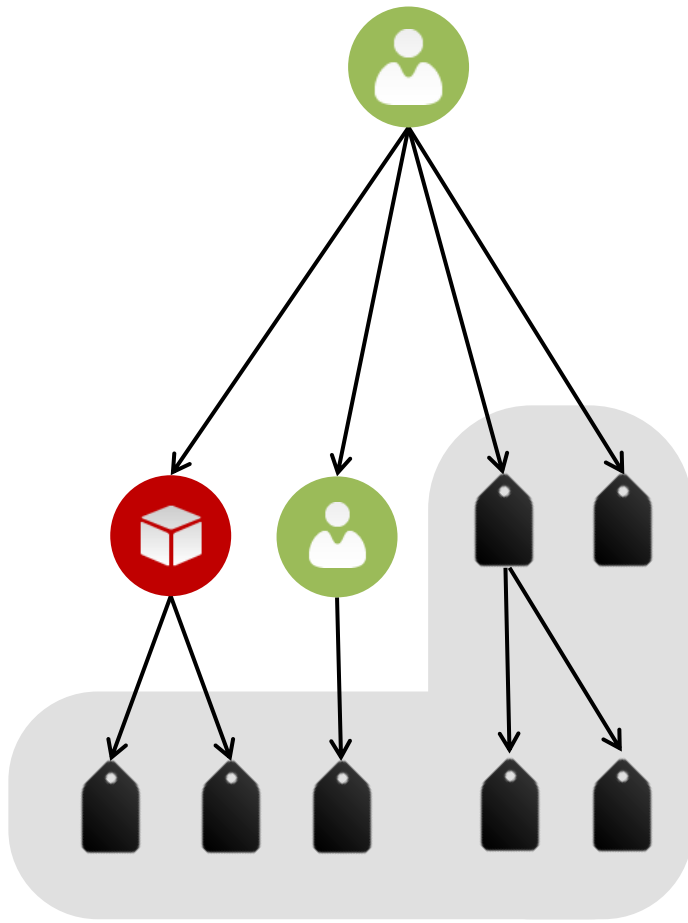
- ✗ Ресурсоемкость
- ✗ Невозможность обоснования
- ✗ Ручной подбор параметров



Рекомендательные системы



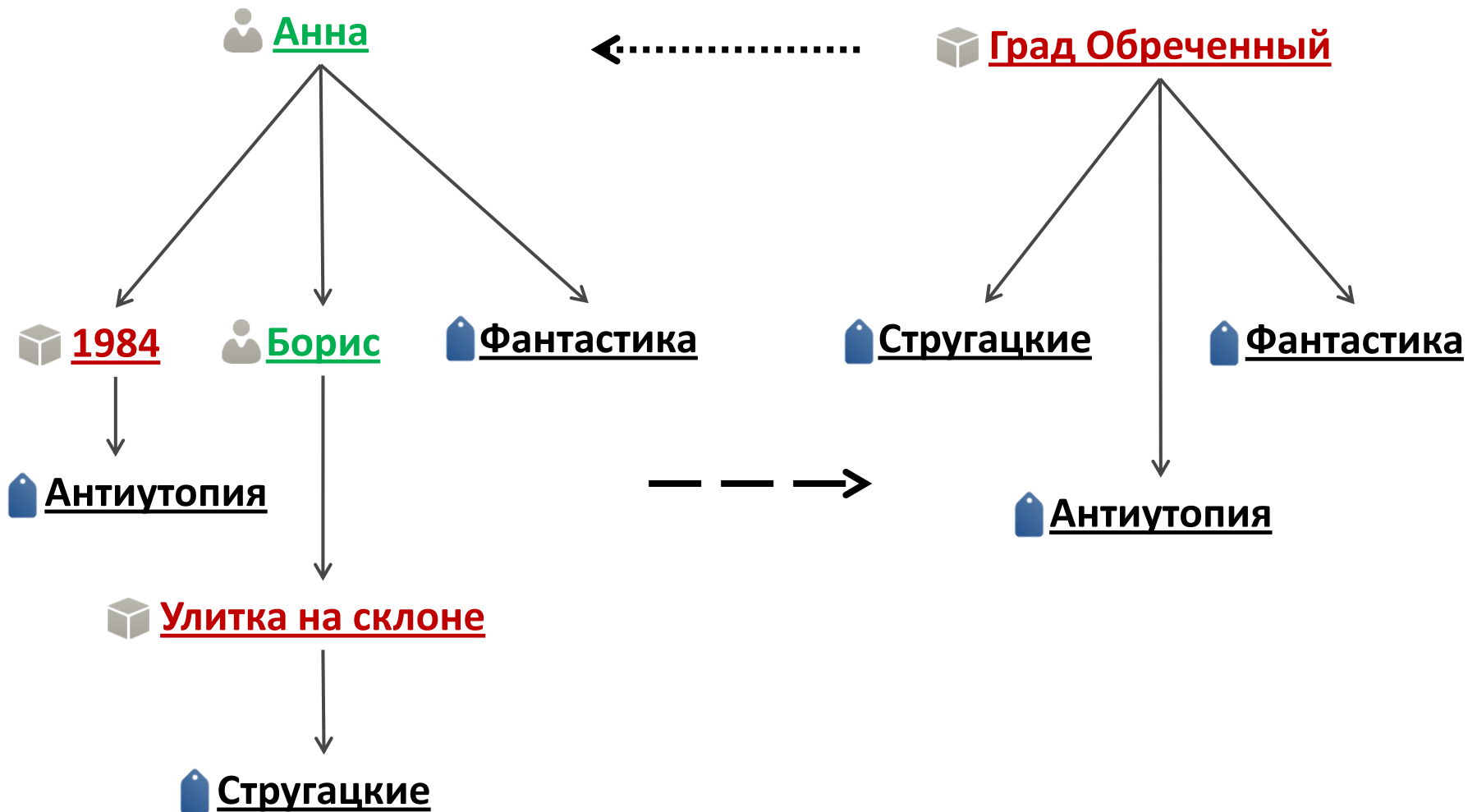
Вложенные теги



$$u(\text{green person}, \text{red cube}) \quad u(\text{red cube}, \text{red cube})$$

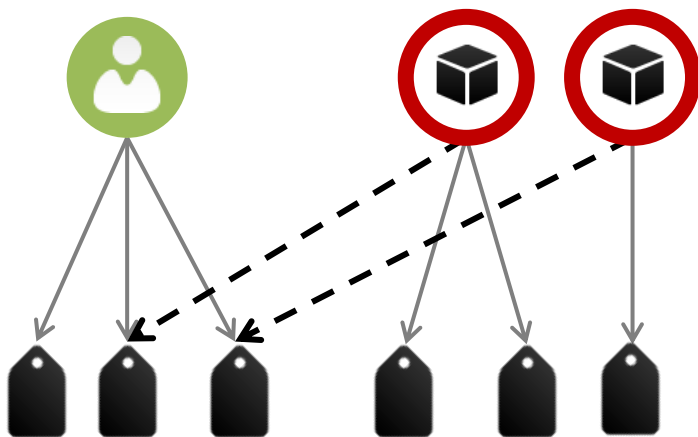
$$u(\text{green person}, \text{green person}) \quad u(\text{green person}, \text{black tag})$$

Вложенные теги



Поиск объектов

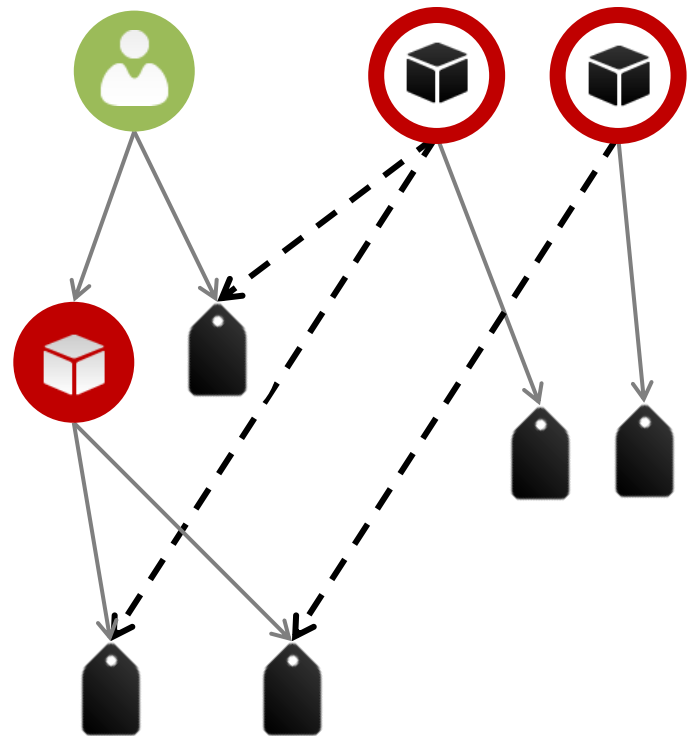
1. Поиск компонент
2. Поиск их родителей
3. Сравнение



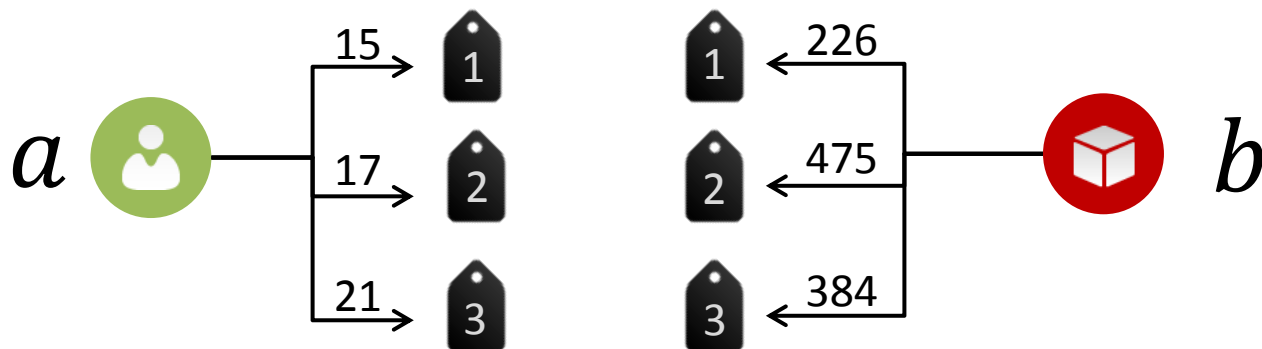
Время работы поиска — $O(k^2)$

Сравнение — $O(k)$

k — среднее количество компонент

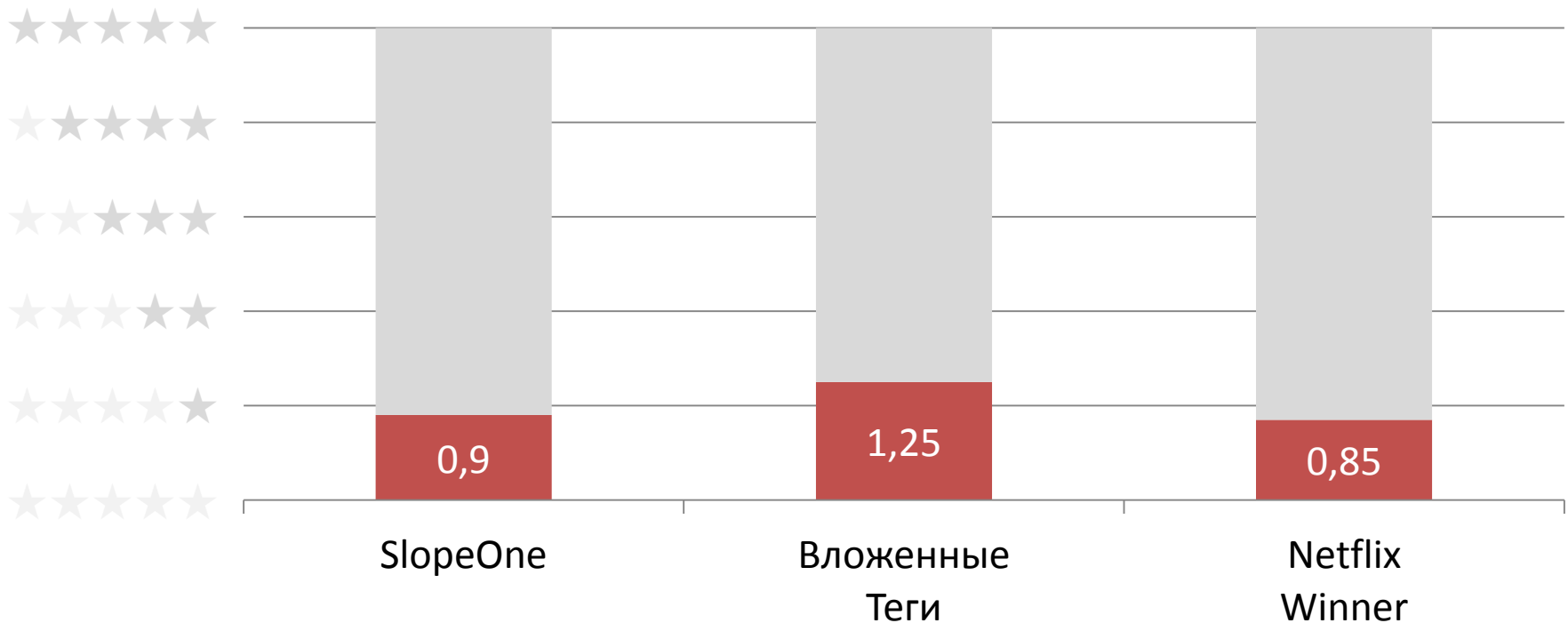


Сравнение объектов



- Объект a связан с $S_a = \{o_{a_1}, o_{a_2}, \dots, o_{a_n}\}$
Вес связи $o_1 \rightarrow o_2$ обозначим $w(o_1, o_2)$
- $$pw(a, o) = \frac{w(a, o_i)}{\sum_i^{o_i \in S_a} w(a, o_i)}$$
- $$u(a, b) = \sum_i^{o_i \in S_b \cup S_a} \min(pw(o_i, a), pw(o_i, a),)$$

Падение точности

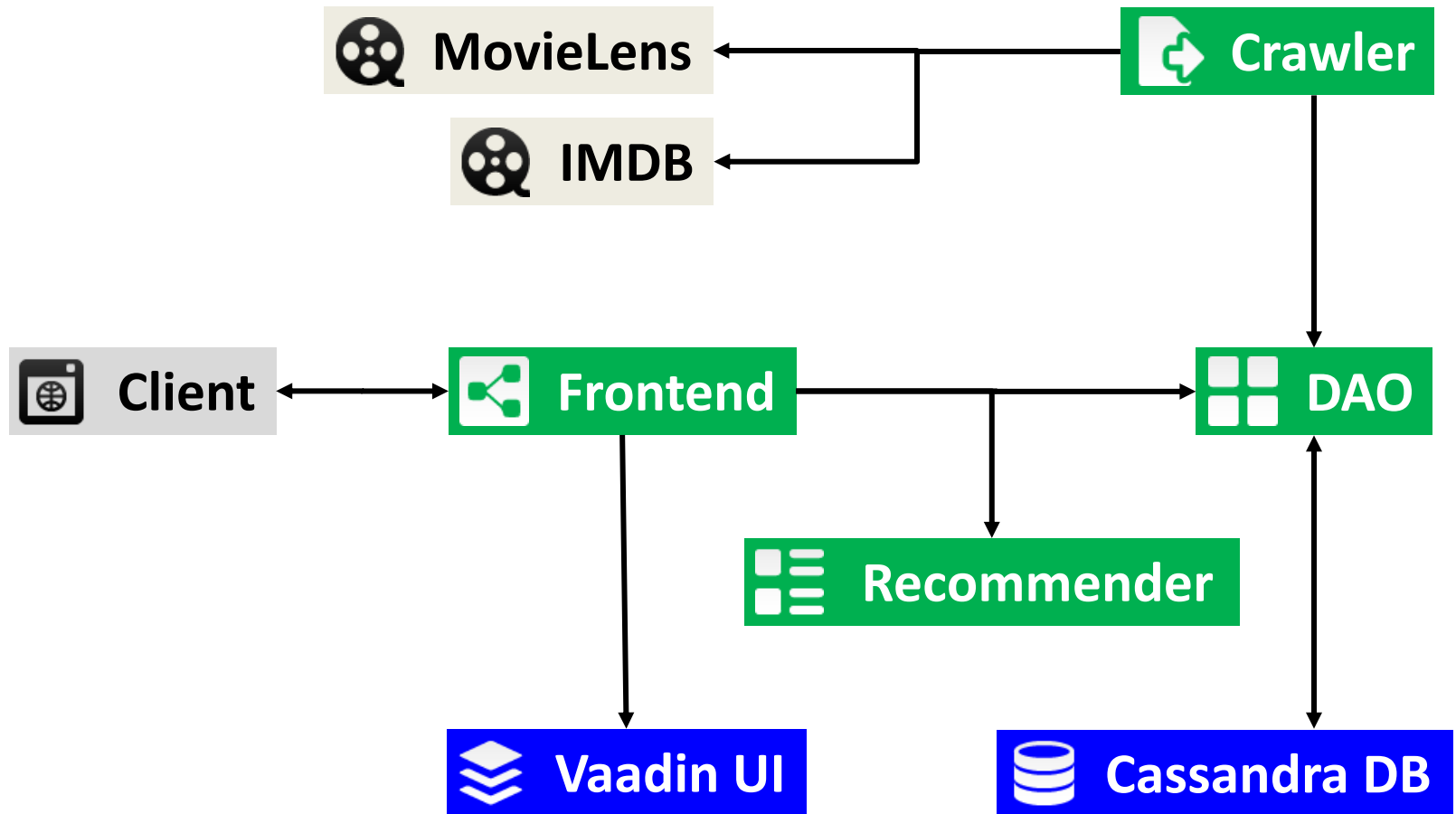


- 1 000 000 рейтингов
- 6 000 пользователей
- 100 000 ключевых слов
- 3 900 фильмов




- Обучение на 800 000 рейтингах
- Проверка на 200 000 рейтингах
- Измерение среднеквадратичной ошибки

Прототип рекомендательной системы по методу вложенных тегов



Прототип рекомендательной системы по методу вложенных тегов

Add New Tag Load Tag Current Tag... Search Monitoring User...

 **Антон Чеботаев**

jedi-knight 10,0 | star wars 10,0 | Robert (I) Rodriguez 10,0

George (I) Lucas 10,0 | Samuel L. Jackson 10,0 | nonlinear 10,0





death-star 10,0 | vampire-hunter 10,0 | Quentin Tarantino 10,0

blood 10,0 | stop-action 10,0 | gangster 10,0 | Four Rooms 10,0

Bruce Willis 10,0 | coffee 10,0 | neo-noir 10,0

slow-motion-action-scene 10,0 | lightsaber 10,0 | vampire-slayer 10,0

violence 10,0 | space-battle 10,0 | George Clooney 10,0

Илья Пименов 10,7

Bruce Willis Jason Statham Pulp Fiction Gary Oldman Johnny Depp
Evil Dead, The Brad Pitt

Snatch 5,2




slow-motion-action-scene stop-action Guy Ritchie Tim Fal Day
Jason Statham neo-noir blood narrated-by-character
nonlinear-timeline gangster violence rhyming-slang
stuffed-in-a-trunk yardie gangster-comedy hasidic-jew
man-eating-pig jewish-mafia head-brace attacked-by-a-d

Star Wars: Episode II - Attack of the Clones 5,0

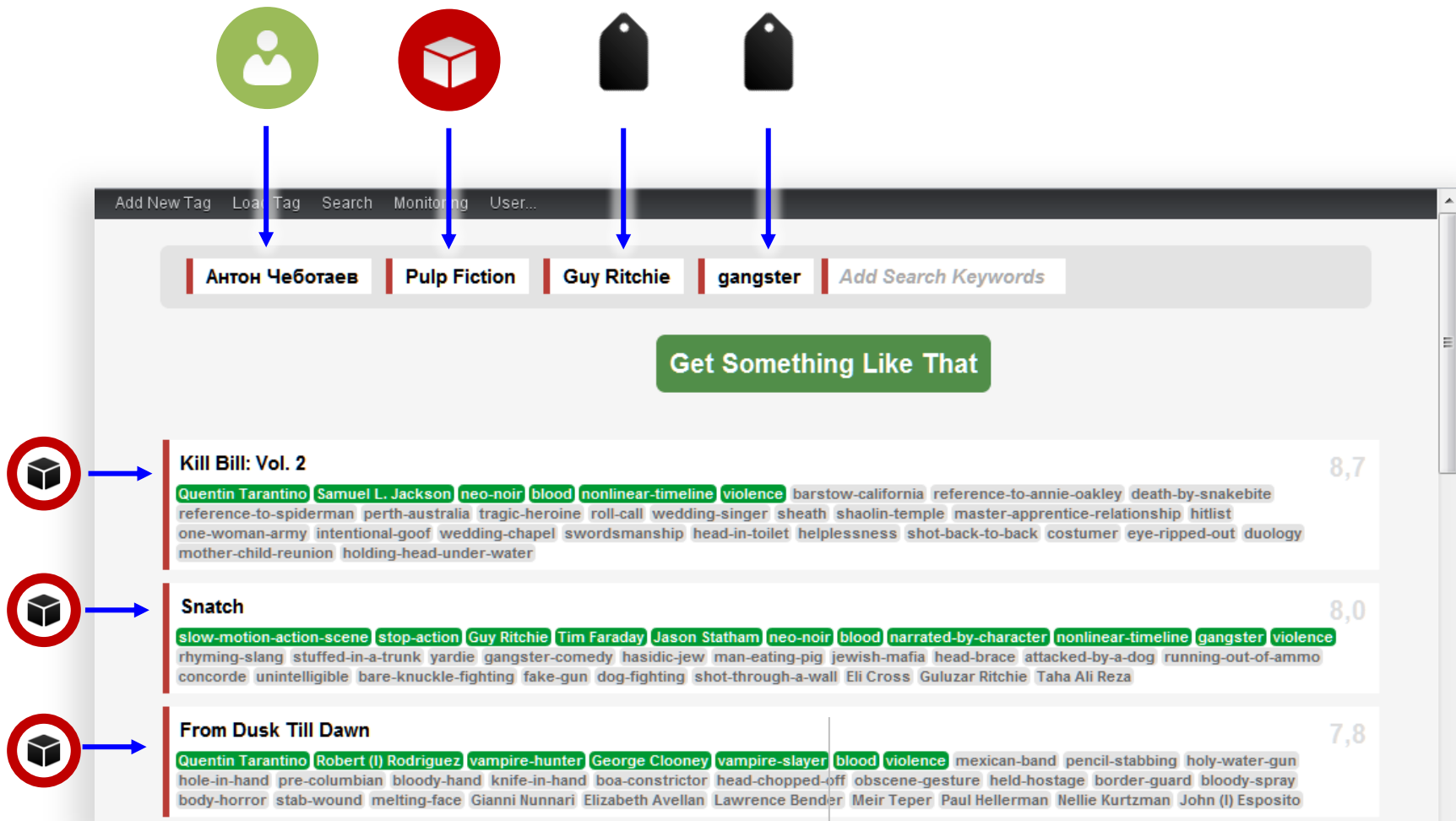
death-star jedi-knight George Lucas lightsaber space battle
Samuel L. Jackson star wars
murdered-before-giving-protagonist-information prequel-cult-film
mistaking-reality-for-dream god-becoming-evil galactic-war
character-feels-around-for-missing-head foreshadow
mechanical-hand closing-the-eyes-of-a-dead-person
fate-of-the-universe changeling decapitation ends-with-a-wedding

From Dusk Till Dawn 4,5

Quentin Tarantino Robert (I) Rodriguez vampire-hunter
George Clooney vampire-slayer blood violence mexican-band
pencil-stabbing holy-water-gun hole-in-hand pre-columbian
bloody-hand knife-in-hand boss-constrictor head-chopped-off
obscene-gestures held-hostage border-guard bloody-spr

Прототип рекомендательной системы по методу вложенных тегов



Развитие

- Функции сравнения
- Алгоритмы поиска
- Качество описания объектов
 - Анализ рецензий фильмов
 - Фильтрация пользовательских данных

Вывод

Преимущества

- ✓ Обоснование рекомендаций
- ✓ Поиск с учетом состояния пользователя
- ✓ Требуется меньше ресурсов

Недостатки

- ✗ Качество зависит от собранных данных
- ✗ На данный момент точное значение $u(c, s)$ ниже