рамблер

Рекомендательные алгоритмы на больших данных

Павел Клеменков
p.klemenkov@rambler-co.ru

Рекомендательные системы неформально

Рекомендательная система – программа, задачей которой является предсказать оценку, которую пользователь поставит объекту, с которым он еще не встречался, на основании характеристик этого объекта и/или профиля пользователя

Рекомендательные системы формально

U – множество пользователей

I – множество объектов

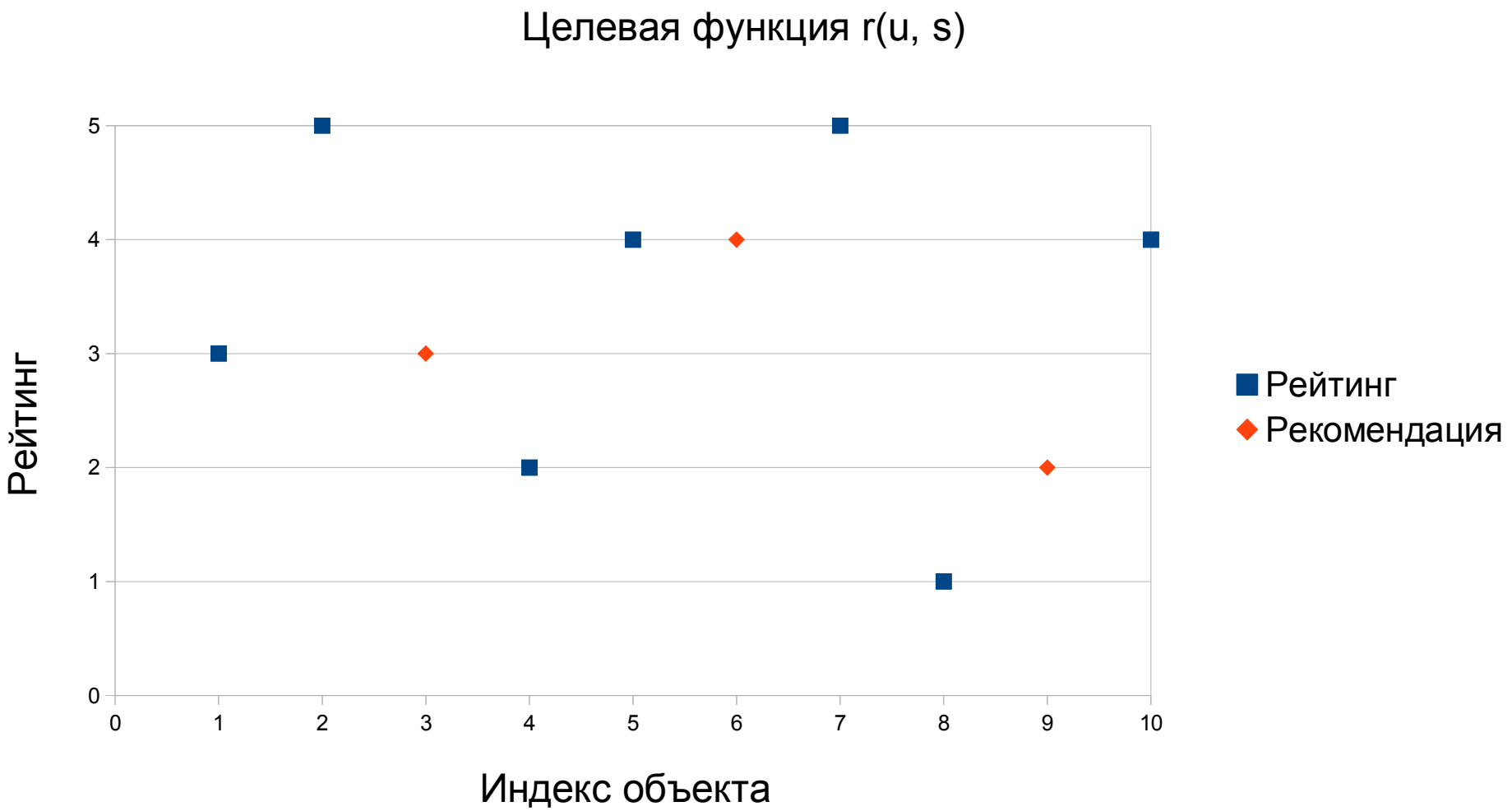
$$r : U \times I \rightarrow R$$

$$\forall u \in U, i'_u = \underset{i \in I}{argmax}\, r(u, i)$$

Задача рекомендательной системы

	Солярис	Возвращение	Легенда 17	12 стульев
Василий	4	3	None	4
Петр	None	4	5	5
Иван	2	2	4	None
Мария	3	None	5	2

Задача рекомендательной системы



Классификация рекомендательных систем

Коллаборативные

Контентные

Гибридные

Коллаборативные системы формально

$$r_{u,i} = \underset{u' \in \hat{U}}{aggr}\, r_{u',i}$$

$$r_{u,i} = \frac{1}{N} \sum_{u' \in \hat{U}} r_{u',s}$$

$$r_{u,i} = k \sum_{u' \in \hat{U}} \mathrm{sim}(u, u') * r_{u',i}$$

$$r_{u,i} = \bar{r}_u + k \sum_{u' \in \hat{U}} \mathrm{sim}(u, u') * (r_{u',i} - \bar{r}_{u'})$$

Коллаборативные системы формально

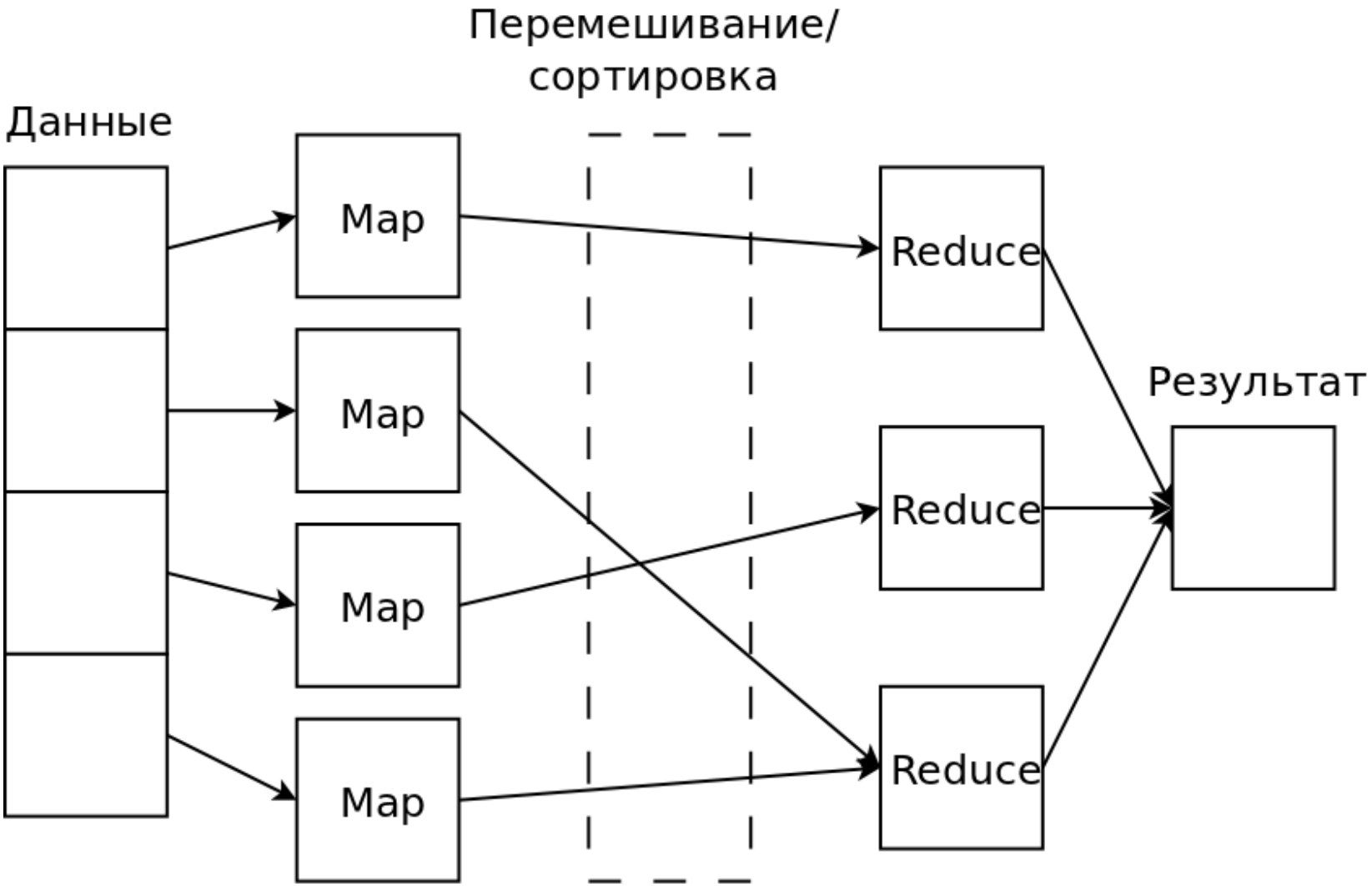
$$\mathrm{sim}(a,b)=\frac{\sum_{i\in I_{ab}}\left(r_{a,i}-\bar{r}_a\right)\left(r_{b,i}-\bar{r}_b\right)}{\sqrt{\sum_{i\in I_{ab}}\left(r_{a,i}-\bar{r}_a\right)^2}\sqrt{\sum_{i\in I_{ab}}\left(r_{b,i}-\bar{r}_b\right)^2}}$$

$$\mathrm{sim}(a,b)=\cos(a,b)=\frac{\sum_{i\in I_{ab}} r_{a,i} r_{b,i}}{\sqrt{\sum_{i\in I_{ab}} r_{a,i}^{\,2}}\sqrt{\sum_{i\in I_{ab}} r_{b,i}^{\,2}}}$$

При чем здесь Big Data?


hadoop

MapReduce


Перемешивание/
сортировка
Данные
Map
Map
Map
Map
Reduce
Reduce
Reduce
Результат

Ключевые особенности Hadoop

- Данные пишутся один раз, читаются много
- Изменить часть данных невозможно
- Файловая система оптимизирована для работы с очень большими файлами
- Локальность вычислений
- Большие задержки


mahout

Apache Mahout

- Коллаборативная фильтрация
- K-Means, Fuzzy K-Means
- LDA
- SVD
- Наивный Байес
- Random Forest
- и многое другое...

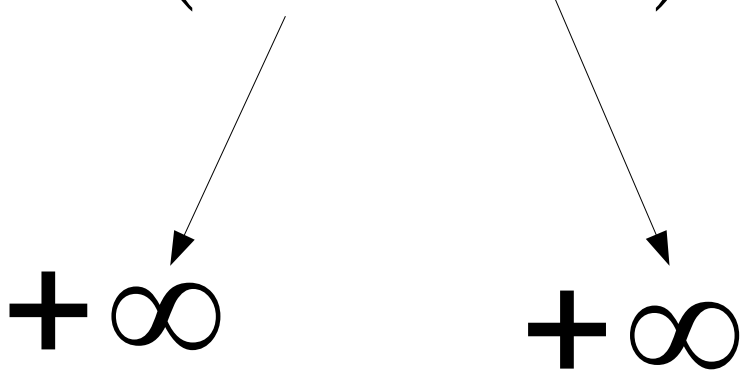
Фильтрация по близости пользователей

```
1 DataModel model = new FileDataModel(new File("data.txt"));

2 UserSimilarity userSimilarity = new PearsonCorrelationSimilarity(model);

3 UserNeighborhood neighborhood =
        new NearestNUserNeighborhood(3, userSimilarity, model);

4 Recommender recommender =
        new GenericUserBasedRecommender(model, neighborhood, userSimilarity);

5 Recommender cachingRecommender = new CachingRecommender(recommender);

6 List<RecommendedItem> recommendations =
        cachingRecommender.recommend(1234, 10);
```

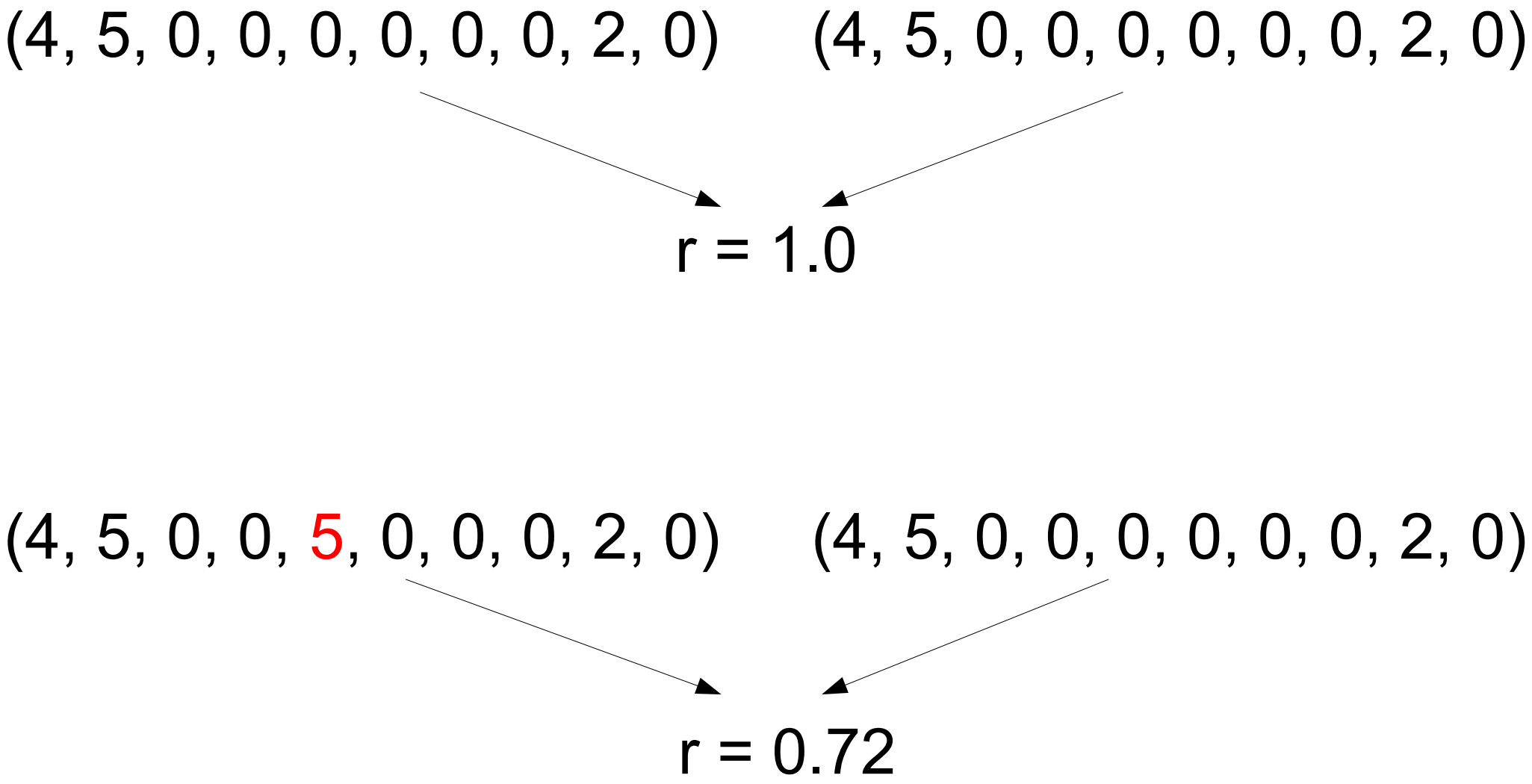
Вычислительная сложность

$O(MN)$ – в худшем случае

$O(M+N)$ – в реальности



В Mahout **нет** распределенной реализации фильтрации по близости пользователей


(4, 5, 0, 0, 0, 0, 0, 0, 2, 0)
(4, 5, 0, 0, 0, 0, 0, 0, 2, 0)
r = 1.0
(4, 5, 0, 0, 5, 0, 0, 0, 2, 0)
(4, 5, 0, 0, 0, 0, 0, 0, 2, 0)
r = 0.72

Фильтрация по близости объектов

$$\mathrm{sim}(a,b)=\frac{\sum_{u\in U_{ab}}\left(r_{u,a}-\bar{r}_u\right)\left(r_{u,b}-\bar{r}_u\right)}{\sqrt{\sum_{u\in U_{ab}}\left(r_{u,a}-\bar{r}_u\right)^2}\sqrt{\sum_{u\in U_{ab}}\left(r_{u,b}-\bar{r}_u\right)^2}}$$

$$r_{u,i}=\frac{\sum_{i'\in\hat{I}}\mathrm{sim}(i,i')*r_{u,i}}{\sum_{i'\in\hat{I}}\mathrm{sim}(i,i')}$$

Фильтрация по близости объектов

- Матрицу близости объектов можно посчитать заранее
- Векторы объектов, обычно, менее разреженные, чем векторы пользователей

$O(N^2 M)$ – в худшем случае

$O(NM)$ – в реальности

Фильтрация по близости объектов

```
1 DataModel model = new FileDataModel(new File("data.txt"));

2 Collection<GenericItemSimilarity.ItemItemSimilarity> correlations;

3 ItemSimilarity itemSimilarity = new GenericItemSimilarity(correlations);

4 Recommender recommender =
        new GenericItemBasedRecommender(model, itemSimilarity);;

5 Recommender cachingRecommender = new CachingRecommender(recommender);

6 List<RecommendedItem> recommendations =
        cachingRecommender.recommend(1234, 10);
```

Проблемы коллаборативных систем

- Сколько ближайших соседей выбирать?
- Какую меру близости использовать?
- Как бороться с “холодным стартом”?
- Что делать, когда векторы сильно разрежены?
- Как лучше использовать неявные рейтинги?

Ваш город: Москва | +5° | $ 32,27 | € 43,77

рамблер

Найти

Билеты на лучшие места

Спорт | Авто | Кино | Музыка | Игры | Путешествия | Здоровье | Бизнес | Дети | Светская хроника | Еда | Мода и красота | Ещё

Полонский объявлен в международный розыск

Российский предприниматель Сергей Полонский объявлен в международный розыск по линии Интерпола.

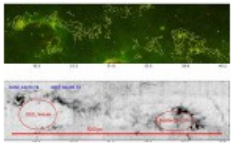
ЧАС НАЗАД

Руководство овощебазы в Бирюлево проведет проверку после ЧП

4 ЧАСА НАЗАД

Военный напророчил Израилю землетрясение «разрушительнее, чем война»

40 МИНУТ НАЗАД



Астрономы нашли в Млечном пути крупнейший газовый поток

5 ЧАСОВ НАЗАД

Телеведущий Алексей Пиманов покинет Совет Федерации

2 ЧАСА НАЗАД

Навальный открыл сбор подписей за визовый режим для мигрантов

3 ЧАСА НАЗАД

«Бюджетник» Peugeot получил новый «автомат»

Авторамблер

7 часов назад на **Снобе**

«Люди требуют справедливости и главоубийства»

Контентные рекомендательные системы

$$w_{ij} = TF_{ij} * IDF_i$$

$$\text{Content}(d_j) = (w_{1j}, \ldots, w_{kj})$$

$$r(u, i) = \text{Score}(\text{ContentProfile}(u), \text{Content}(i))$$

$$r(u, i) = \cos(\vec{w}_u, \vec{w}_i) = \frac{\vec{w}_u \cdot \vec{w}_i}{\|\vec{w}_u\| \|\vec{w}_i\|}$$

Mahout K-means

```
mahout seq2sparse \
    -i /data/input/ \
    -o /data/sparse-kmeans \
    –maxDFPercent 85 \
    –namedVector

mahout kmeans\
    -i /data/sparse-kmeans/tfidf-vectors/ \
    -c /data/kmeans-clusters \
    -o /data/kmeans \
    -dm CosineDistanceMeasure \
    -x 10 -k 20 -ow –clustering
```

Mahout Dirichlet Process Clustering

```
mahout seq2sparse \
    -i /data/input/ \
    -o /data/sparse-dirichlet \
    –maxDFPercent 85 \
    –namedVector

mahout dirichlet\
    -i /data/sparse-dirichlet/tfidf-vectors/ \
    -o /data/dirichlet \
    -dm CosineDistanceMeasure
    -md DistanceMeasureClusterDistribution
    -mp DenseVector
    -x 10 -k 20 -ow –clustering
```

Контентные рекомендательные системы

+ Рекомендации строятся на основе одного профиля
+ Легко показать, откуда взялась рекомендация
+ Нет проблемы “холодного старта” нового объета
- Ограниченные возможности анализа (эксперты, отсутствие специфических фич)
- Сужение кругозора
- “Холодный старт” нового пользователя

Не забывайте использовать экосистему!

- **Pig** - среда исполнения и язык программирования вычислений
- **Hive** - распределенное хранилище с SQL-подобным языком запросов
- **HBase** – распределенное колоночное хранилище

Вопросы?