

A decorative graphic in the top-left corner of the slide, consisting of a 4x4 grid of squares. The squares are colored in a pattern: the top row has one teal square; the second row has one orange and one brown square; the third row has one orange, one teal, and one light brown square; the bottom row has one light brown, one orange, one orange, and one brown square.

Ранжирование

Повторение

Рекомендательные системы

- Рекомендательные системы сокращают объём информации, необходимый для принятия решения
- Не нужно читать отзывы на 1000 фильмов — модель сама выберет лучший
- Netflix: 2/3 просмотренных фильмов найдены через рекомендательную систему
- Amazon: 35% продаж через полки рекомендаций
- Youtube: 60% просмотров благодаря рекомендациям

Типичная рекомендательная система

- Объект: пара «user-item»
- Целевая переменная: клики, длинные клики, досмотры, покупки, дослушивания, лайки и т.д.
- Решаем задачу классификации/регрессии/ранжирования

Особенности:

- Выбор целевой переменной
- Выбор метрики качества
- Факторы для модели
- Слишком много товаров/видео/песен/...

Отбор кандидатов

- Простая и быстрая модель, которая отбирает тысячи товаров для данного пользователя
- Сложная модель применяется только к отобранным кандидатам

Основные подходы

- Есть методы, разработанные напрямую для рекомендаций
- Коллаборативная фильтрация
 - Рекомендации на основе сходства действий пользователей
- Контентные рекомендации

Обозначения

- Множество товаров:
- Множество пользователей:
- Множество пар «пользователь-товар», для которых известны оценки:
- Если для пары известен рейтинг, то будем писать
- Оценки — рейтинги фильмов, индикаторы покупки товара и т.д.

Оценки

- Оценки (или фидбэк) бывают явные и неявные
- Явные оценки
 - Пользователь поставил оценку фильму/товару
 - Пользователь написал отзыв
 - Пользователь поставил лайк
- Неявные оценки
 - Пользователь посмотрел фильм
 - Пользователь добавил товар в корзину
 - Пользователь долго смотрел на запись в социальной сети
- Неявные оценки более шумные, но их больше

Сходство пользователей

- $I_{uv} = \{i \in I \mid \exists r_{ui} \text{ и } \exists r_{vi}\}$ — множество товаров, которые оценили и пользователь u , и пользователь v
- Сходство пользователей (корреляция):

$$w_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}},$$

где \bar{r}_u и \bar{r}_v — средние рейтинги пользователей

User-based collaborative filtering

- Дан пользователь u_0
- Найдём пользователей, которые похожи на него:

$$U(u_0) = \{v \in U \mid w_{u_0 v} > \alpha\}$$

- Порекомендуем те товары, которые часто покупались пользователями из $U(u_0)$

User-based collaborative filtering

Недостатки:

- Много параметров, которые сложно выбирать
 - Какой порог сходства для пользователей?
 - Сколько похожих пользователей должны были купить товар, чтобы мы его порекомендовали?
- Требуется хранить всю матрицу оценок

Есть и другие методы, основанные на сходствах, но все обладают теми же недостатками.

Модели со скрытыми переменными

- Обучим вектор p_u для каждого пользователя u
- Обучим вектор q_i для каждого товара i
- Оценка приближается их скалярным произведением:

$$r_{ui} \approx \langle p_u, q_i \rangle$$

- Находим векторы только по известным оценкам
- После этого можем предсказать оценку для любой пары «пользователь-товар»

Модели со скрытыми переменными

- Оптимизационная задача:

$$\sum_{(u,i) \in R} (r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle)^2 \rightarrow \min_{P, Q}$$

- Решение: градиентный спуск, Alternating Least Squares (ALS) и другие методы

SVD для построения рекомендаций

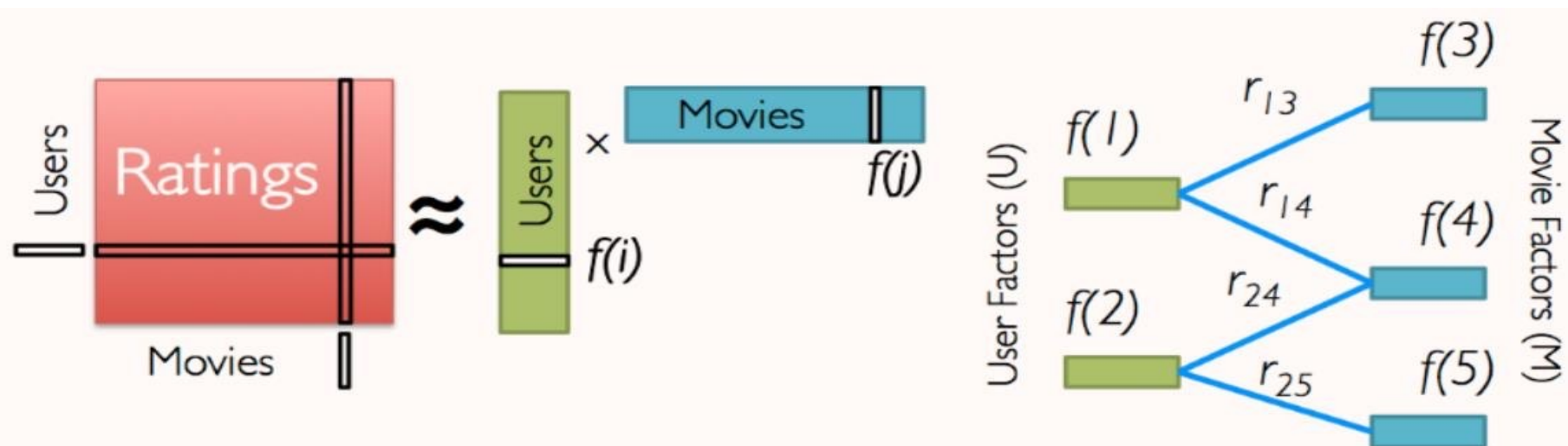
- Матрица товарных предпочтений (матрица, где строки это пользователи, а столбцы это продукты, с которыми пользователи взаимодействовали) представляется произведением трех матриц:

$$\begin{matrix} & \overbrace{\hspace{1cm}}^n \\ \underbrace{\hspace{1cm}}_m \begin{matrix} \vdots \\ \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots & \dots & \vdots \\ \vdots & & \vdots \\ \vdots & & \vdots \end{matrix} & \approx & \begin{matrix} \overbrace{\hspace{1cm}}^k \\ U \end{matrix} & \times & \begin{matrix} \Sigma \end{matrix} & \times & \begin{matrix} V^T \\ \underbrace{\hspace{1cm}}_k \end{matrix} \end{matrix}$$

U – описание характеристик пользователя

V – описание характеристик продукта

Матричная факторизация



Iterate:

$$f[i] = \arg \min_{w \in \mathbb{R}^d} \sum_{j \in \text{Nbrs}(i)} (r_{ij} - w^T f[j])^2 + \lambda ||w||_2^2$$

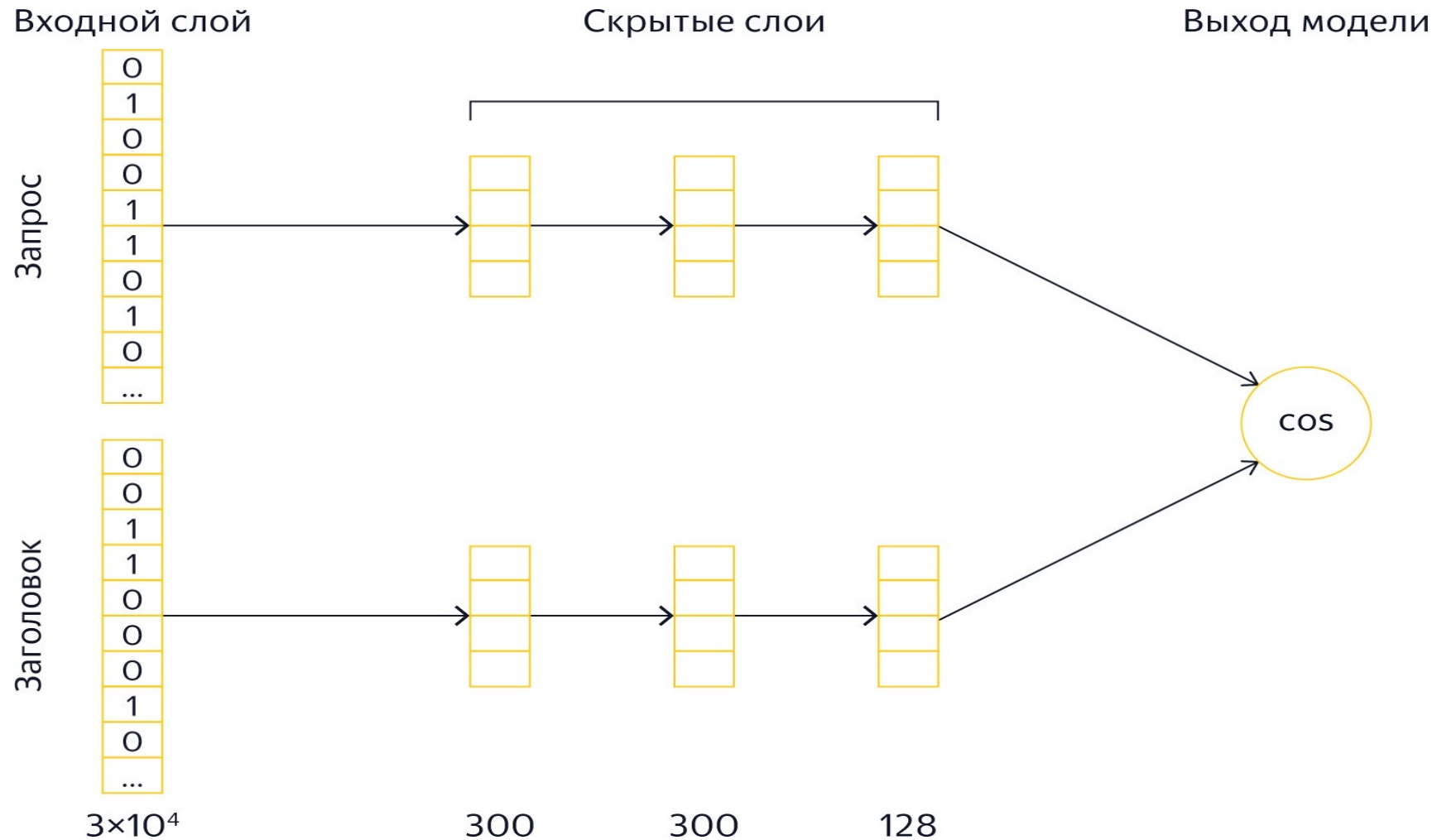
Taken from the BerkeleyX Course Big Data Analysis with Spark

<https://habr.com/ru/company/lanit/blog/421401/>

Контентные рекомендации

- Сведём задачу к обычному обучению с учителем
- Объект: пара «пользователь-товар»
- Ответ: отклик пользователя
- Факторы: информация про пользователя и про товар
- Обучаем любую модель на этих данных
- Среди факторов могут быть и прогнозы коллаборативных моделей

Deep Structured Semantic Model



Качество предсказаний

В зависимости от целевой переменной:

- MSE, MAE, R^2
- Accuracy, HitRate, precision/recall, AUC-ROC
- Метрики качества ранжирования (дальше в курсе)

Другие метрики

- Покрытие
 - Какая доля товаров рекомендовалась хотя бы раз?
 - Какой доле пользователей хотя бы раз показаны рекомендации?
- Новизна
 - Как много рекомендованных товаров пользователь встречал раньше?
- Прозорливость (serendipity)
 - Способность предлагать товары, которые отличаются от купленных ранее
- Разнообразие


Ранжирование


Пример


Яндекс


Поиск Картинки Видео Карты Маркет Новости Переводчик Кью Услуги Музыка


Результаты поиска


 **Анатомия рекомендательных систем. Часть первая / Хабр**
[habr.com](#) > [ru/company/lanit/blog/420499/](#) ...
Задача **рекомендательной системы** – проинформировать пользователя о товаре ...
Рекомендательные системы – это про то, что предложить клиенту, чтобы сделать его счастливым. Читать ещё


 **Рекомендательная система — Википедия**
[ru.wikipedia.org](#) > Рекомендательная система ...
Рекомендательные системы — программы, которые пытаются предсказать, какие объекты (фильмы, музыка, книги, новости, веб-сайты) будут интересны пользователю...

 **Что такое и как работают рекомендательные системы**
[skillbox.ru](#) > [media/code...rekomendatelnye-sistemy_i...](#) ...
Рекомендательная система, основанная на контенте, посоветует ещё 25... Гибридные **рекомендательные системы** сочетают разные подходы. Читать ещё

 **Рекомендательные системы: как помочь пользователю...**
[vc.ru](#) > [marketing/152926-rekomendatelnye-sistemy...to...](#) ...
Что такое **рекомендательные системы**? **Рекомендательная система** — комплекс алгоритмов, программ и сервисов, задача которого предсказать, что может заинтересовать того или иного пользователя. Читать ещё

 **Рекомендательные системы — Викиконспекты**
[neerc.ifmo.ru](#) > [wiki/index.php?...Рекомендательные...](#) ...
Рекомендательные системы — программы, которые пытаются предсказать, какие объекты будут интересны пользователю, имея определённую информацию о его профиле. Читать ещё

 **Как работают рекомендательные системы**
[neurohive.io](#) > [...osnovy-data...rekomendatelnye-sistemy...](#) ...
Рекомендательные системы, основанные на машинном обучении, получили широкое распространение для бизнеса в последние годы. Читать ещё

 **Как устроены современные рекомендательные системы?**
[proglib.io](#) > [...sovremennye-rekomendatelnye-sistemy...02](#) ...
Современные **рекомендательные системы**. Не далее чем в мае 2019 Facebook выложил в открытый доступ исходный код некоторых своих подходов к **рекомендациям** и представил DLRM (Deep-Learning... Читать ещё



Ранжирование

- Дан набор запросов $\{q_1, \dots, q_m\}$
- Дан набор документов $\{d_1, \dots, d_n\}$
- Рассматриваем пары «запрос-документ» (q, d)
- Для некоторых троек (q, d_1, d_2) известно, что для запроса q документ d_1 должен стоять раньше, чем d_2
- Обозначение: R — множество троек (q, d_1, d_2) , для которых известен такой порядок

Ранжирование

- Раньше: строим модель $a(x)$, которая приближает ответы
- Сейчас: строим модель $a(q, d)$, которая правильно упорядочивает документы для запросов

$$(q, d_1, d_2) \in R \Rightarrow a(q, d_1) > a(q, d_2)$$

Пример

- Для запроса q известны пары (d_3, d_1) , (d_3, d_2) , (d_1, d_4)
- Какие наборы прогнозов модели лучше?
- $(3, 2, 4, 1)$
- $(2, 3, 4, 1)$
- $(3, 4, 2, 1)$
- **$(13, 10, 20, 7)$**
- Важен порядок, а не абсолютные значения!

Метрики качества ранжирования

Целевая переменная

- Определение задачи через пары — правильно, но сложно
- Упростим постановку:
 - Объекты — пары «запрос-документ» $x_i = (q, d)$
 - Ответы — числа y_i
 - Требование — если есть объекты (q, d_1) и (q, d_2) , такие что $y_1 > y_2$, то должно быть $a(q, d_1) > a(q, d_2)$

Целевая переменная, пример

- $(q_1, d_1), 1$
- $(q_1, d_2), 0.7$
- $(q_1, d_3), 0$
- $(q_2, d_1), 0$
- $(q_2, d_2), 1$
- Для q_1 должны получить ранжирование (d_1, d_2, d_3)
- Для q_2 должны получить ранжирование (d_2, d_1)

Качество ранжирования

Машинное обучение — Википедия

ru.wikipedia.org > Машинное обучение ▾

Машинное обучение (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](https://ccrp.ru/rabochie/mashinist_burilno-kranovoy...) > rabochie/mashinist_burilno-kranovoy... ▾

Обучение машиниста бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](https://ngpedia.ru/id201843p1.html) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](https://ccrp.ru/rabochie/mashinist_burilno-kranovoy...) > rabochie/mashinist_burilno-kranovoy... ▾

Обучение машиниста бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](https://ngpedia.ru/id201843p1.html) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

Машинное обучение — Википедия

ru.wikipedia.org > Машинное обучение ▾

Машинное обучение (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

Обучение машиниста бурильно-крановых машин — АНО...

[ccrp.ru](https://ccrp.ru/rabochie/mashinist_burilno-kranovoy...) > rabochie/mashinist_burilno-kranovoy... ▾

Обучение машиниста бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

Машинное обучение — Википедия

ru.wikipedia.org > Машинное обучение ▾

Машинное обучение (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

Обучение - машина - Большая Энциклопедия Нефти...

[ngpedia.ru](https://ngpedia.ru/id201843p1.html) > id201843p1.html ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

- Какое ранжирование лучше?
- Какое хуже всех?

DCG (Discounted cumulative gain)

$$\text{DCG}@k(q) = \sum_{i=1}^k \frac{2^{y_i} - 1}{\log(i + 1)}$$

- Вычисляется по первым k документам из выдачи для запроса q
- y_i — истинный ответ для документа на i -й позиции
- Чтобы получить итоговую оценку, DCG усредняется по всем запросам

DCG (Discounted cumulative gain)

W **Машинное обучение** — Википедия

ru.wikipedia.org > Машинное обучение ▾

Машинное обучение (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

📄 **Обучение машиниста бурильно-крановых машин** — АНО...

ccrp.ru > [rabochie/mashinist_burilno-kranovoy...](https://ccrp.ru/rabochie/mashinist_burilno-kranovoy...) ▾

Обучение машиниста бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

📖 **Обучение - машина** - Большая Энциклопедия Нефти...

ngpedia.ru > [id201843p1.html](https://ngpedia.ru/id201843p1.html) ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

$$DCG = \frac{2^1 - 1}{\log(2)} + \frac{2^0 - 1}{\log(3)} + \frac{2^0 - 1}{\log(4)} \approx 1.44$$

📄 **Обучение машиниста бурильно-крановых машин** — АНО...

ccrp.ru > [rabochie/mashinist_burilno-kranovoy...](https://ccrp.ru/rabochie/mashinist_burilno-kranovoy...) ▾

Обучение машиниста бурильно-крановой самоходной машины регламентировано Приказом Минтруда России № 208н от 01.03.2017 г...

W **Машинное обучение** — Википедия

ru.wikipedia.org > Машинное обучение ▾

Машинное обучение (англ. Machine Learning) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи...

📖 **Обучение - машина** - Большая Энциклопедия Нефти...

ngpedia.ru > [id201843p1.html](https://ngpedia.ru/id201843p1.html) ▾

После **обучения машины** или в ходе его, смотря по алгоритму, проводится прогнозирование новых катализаторов...

$$DCG = \frac{2^0 - 1}{\log(2)} + \frac{2^1 - 1}{\log(3)} + \frac{2^0 - 1}{\log(4)} \approx 0.91$$

Доля дефектных пар

$$DP@k(q) = \frac{2}{k(k-1)} \sum_{i < j}^k [y_i < y_j]$$

- Число инверсий порядка среди первых k документов

pFound

- Вероятностная модель поведения пользователя
- При неуспехе с очередным документом выдачи пользователь разочаруется и уйдет с вероятностью P_{out}
- P_i — вероятность дойти до i -ого документа, y_i — вероятность того, что пользователь удовлетворится i -ым документом

$$P_1 = 1, \quad P_{i+1} = P_i(1 - y_i)(1 - P_{\text{out}})$$

$$\text{pFound}@k(q) = \sum_{i=1}^k P_i y_i$$

pFound



Разнообразие поисковой выдачи

- Неоднозначные запросы
- Пример: «ягуар»
 - Животное?
 - Марка автомобиля?
 - Танк? (немецкий или китайский?)
 - Напиток?

Разнообразие поисковой выдачи

- Неоднозначные запросы
- С точки зрения обычных метрик, весь топ выдачи нужно замостить одинаковыми релевантными документами
- Разнообразие позволяет собрать разнородную выдачу, чтобы удовлетворить в среднем всех

Wide pFound

- Предполагается, что пользователь, делая запрос, мог иметь в виду один из интенгов $I = \{I_1, \dots, I_m\}$
- Примеры интенгов: автомобили, картинки, новости, животные, ...
- Каждый интенг имеет некоторую вероятность $p(I_i)$ и порождает собственное распределение релевантностей на документах

$$\text{wide pFound} = \sum_{i=1}^m p(I_i) \text{pFound}(I_i)$$

Wide pFound

- Как вычислить вероятности интенгов?
- Интент пользователя определяется по продолжениям введенного запроса
- Продолжения классифицируются по различным тематикам
- Тематики являются интенгами
- Вероятности определяются по частоте соответствующих продолжений запросов

Качество ранжирования

- Также можно сформулировать задачу классификации ($Y = \{0, 1\}$):

$$\text{precision} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{retrieved}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant}\} \cap \{\text{retrieved}\}|}{|\{\text{relevant}\}|}$$

Методы ранжирования

Поточечный (pointwise) подход

- Обучим модель $a(q, d)$, чтобы она как можно точнее приближала ответы y_i
- Например, линейная регрессия:

$$\sum_{(q,d,y) \in R} (\langle w, x(q, d) \rangle - y_i)^2 \rightarrow \min_w$$

- $x(q, d)$ — признаки для пары «запрос-документ»

Поточечный (pointwise) подход

- Простой в реализации
- Можно использовать любую из известных моделей (линейные, деревья, случайные леса, нейронные сети...)
- Восстанавливает точные значения y_i , хотя нас интересует порядок

Попарный (pairwise) подход

- В ранжировании требуется правильно располагать пары документов — формализуем это

$$\sum_{(q, d_i, d_j) \in R} [a(q, d_i) - a(q, d_j) < 0]$$

- Штрафуем, если второй документ из пары оказался раньше

Попарный (pairwise) подход

- Получили разрывный функционал — сложно оптимизировать
- Перейдём к гладкой верхней оценке (как в линейных классификаторах):

$$\sum_{(q, d_i, d_j) \in R} [a(q, x_i) - a(q, x_j) < 0] \leq \sum_{(q, d_i, d_j) \in R} L(a(q, x_i) - a(q, x_j))$$

- Пример: $L(z) = \log(1 + e^{-z})$

Попарный (pairwise) подход

- Сложнее поточечного (больше слагаемых в функционале)
- Обычно даёт качество выше, чем поточечный
- Реализации: SVM^{light} , xgboost (rank:pairwise)

Признаки в задачах ранжирования

Типы признаков

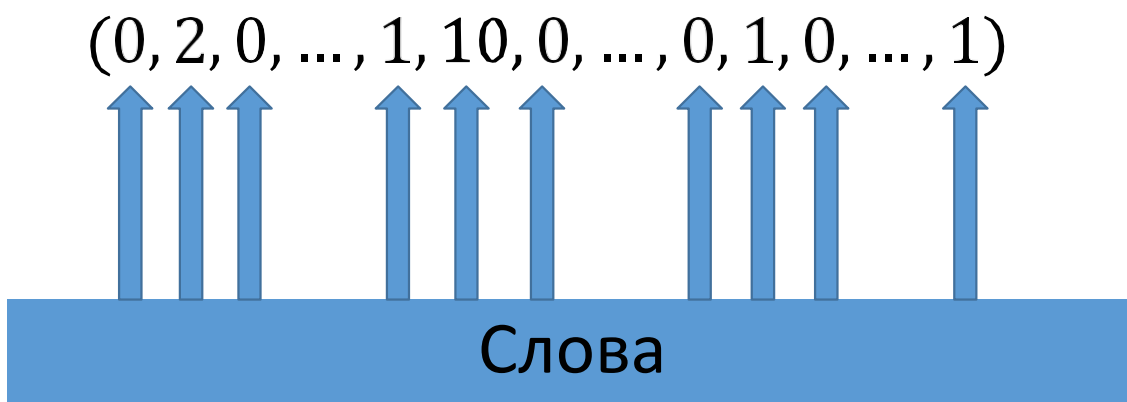
- Запросные
 - Популярность запроса
 - Тип запроса (навигационный, товарный и т.д.)
- Статические — зависят только от документа
 - Популярность документа
 - Тематика
 - Распределение слов
- Динамические — зависят от документа и от запроса
 - Расстояния между запросом и документом

Мешок слов

- $v(\text{большое}) = (1, 0, 0, 0, \dots, 0)$
- $v(\text{спасибо}) = (0, 1, 0, 0, \dots, 0)$
- $v(\text{минус}) = (0, 0, 1, 0, \dots, 0)$
- $v(\text{зарубежный}) = (0, 0, 0, 1, \dots, 0)$
- ...
- $v(\text{инквизиция}) = (0, 0, 0, 0, \dots, 1)$

Мешок слов

- Текст — это вектор x , содержащий счётчики слов



Косинусное расстояние

- Пусть \vec{q} — вектор запроса, \vec{d} — вектор документа
- Мера сходства:

$$s(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n q_i d_i}{\|\vec{q}\| \|\vec{d}\|}$$

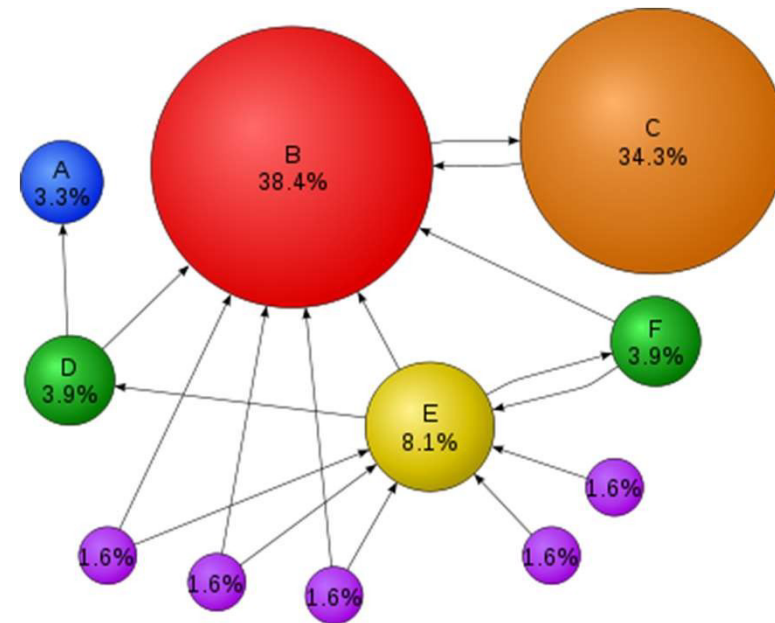
- Чем больше, тем сильнее тексты похожи по долям слов

Продвинутое расстояние: BM25

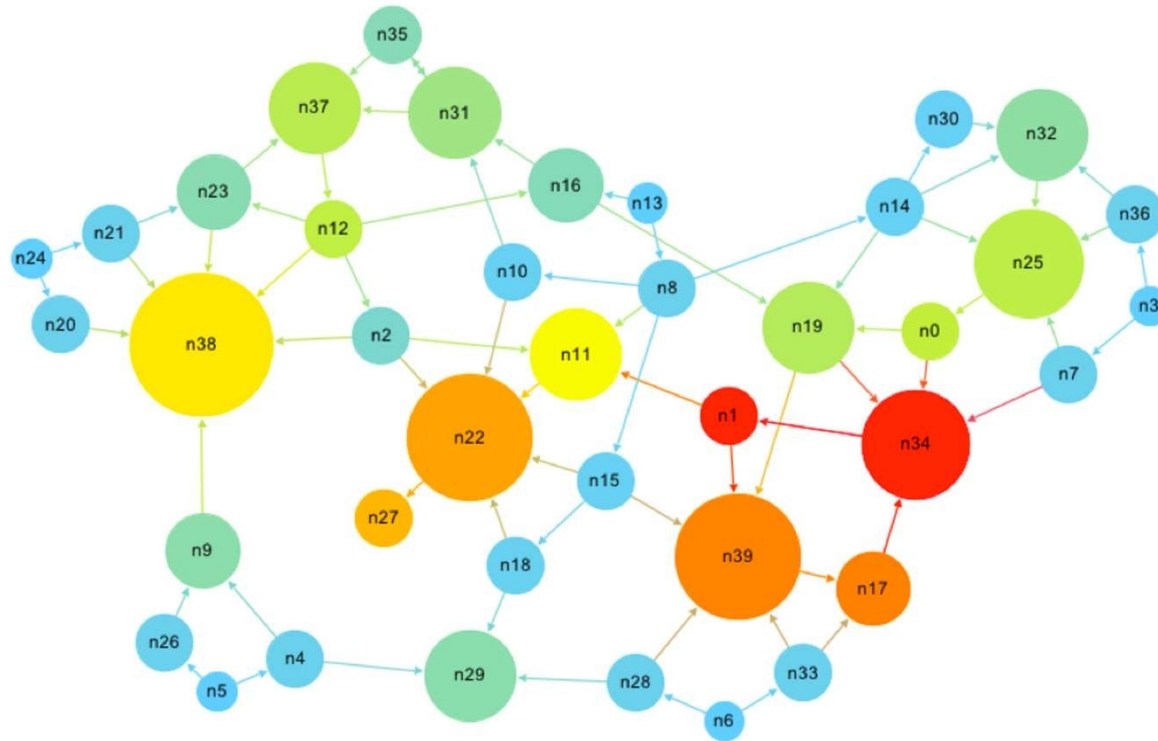
$$\text{BM25}(q, d) = \sum_{i=1}^n \text{IDF}(q_i) \frac{\text{tf}(q_i, d)(k_1 + 1)}{\text{tf}(q_i, d) + k_1 \left(1 - b + b \frac{|D|}{\bar{n}_d}\right)}$$

PageRank

- Документы в сети ссылаются друг на друга
- Если документ A ссылается на документ B, то он «голосует» за B
- Чем меньше голосов отдаёт A, тем сильнее его голос
- Документ B важен, если за него отдано много сильных голосов



PageRank



PageRank

- Пусть пользователь бродит по сети
- Стартует из случайного документа
- С вероятностью $(1 - \delta)$ переходит по одной из ссылок с равными вероятностями
- С вероятностью δ переходит на случайный документ из всей сети
- PageRank — вероятность при таком случайном блуждании попасть в данный документ

PageRank

- PageRank страницы u зависит от PageRank страниц v из множества B_u (страниц, которые ссылаются на u), поделенного на число исходящих ссылок $L(v)$ из страницы v :

$$\text{PR}(u) = \sum_{v \in B_u} \frac{\text{PR}(v)}{L(v)}$$

PageRank

- Учтем, что пользователь может остановиться в какой-то момент
- Установим damping factor (фактор затухания) – обычно $d \approx 0.85$
- N – число рассматриваемых страниц

$$\text{PR}(u) = \frac{1 - d}{N} + d \sum_{v \in B_u} \frac{\text{PR}(v)}{L(v)}$$

Резюме

- Ранжирование — задача сортировки документов по релевантности
- Метрика должна учитывать позиции, а не абсолютные значения прогнозов — например, DCG
- Поточечный и попарный подходы
- Отдельная задача — разработка признаков

Спасибо за внимание!



Ildar Safilo

@Ildar_Saf

irsafilo@gmail.com

<https://www.linkedin.com/in/isafilo/>