

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

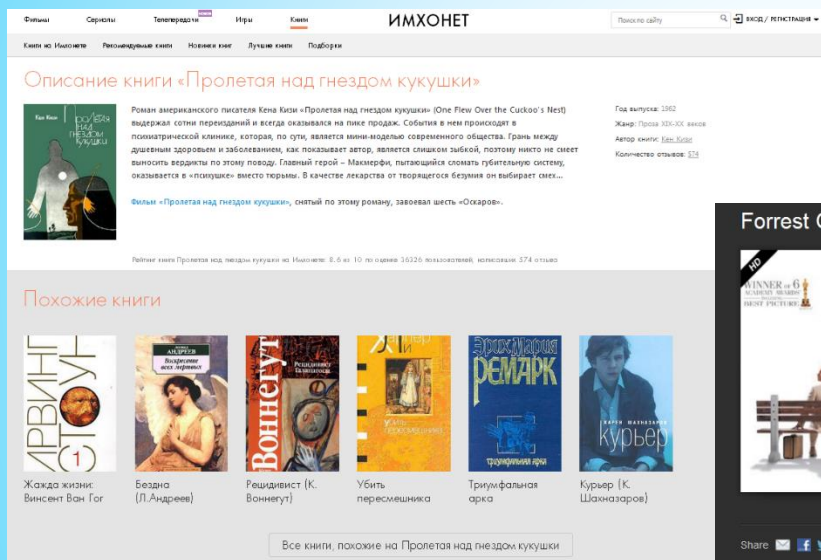
ЛЕКЦИЯ 8. РЕКОМЕНДУЮЩИЕ СИСТЕМЫ

к.т.н., Кашевник Алексей Михайлович,
alexey@iias.spb.su

к.т.н., Пономарёв Андрей Васильевич
ponomarev@iias.spb.su

РЕКОМЕНДУЮЩИЕ СИСТЕМЫ

Рекомендующие системы (РС) – это класс систем поддержки принятия решений, предназначенных для облегчения процесса выбора из множества вариантов.



ИМХОНЕТ

Описание книги «Пролетая над гнездом кукушки»

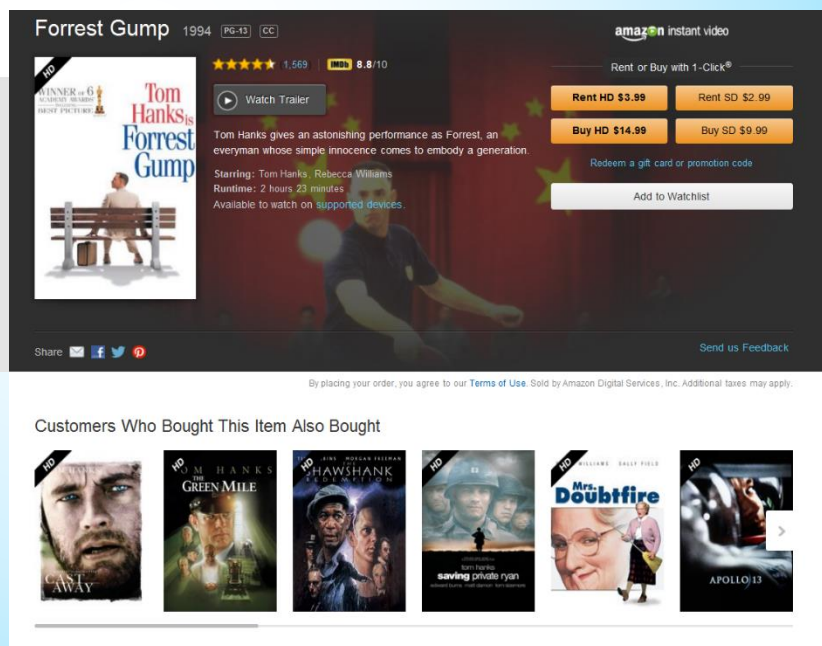
Роман американского писателя Кена Kesey «Пролетая над гнездом кукушки» (One Flew Over the Cuckoo's Nest) выдержал сотни переизданий и всегда оказывался на пике продаж. События в нем происходят в психиатрической клинике, которая, по сути, является мини-моделью современного общества. Грань между душевным здоровьем и заболеванием, как показывает автор, является слишком тонкой, поэтому никто не смеет выносить вердикты по этому поводу. Главный герой – Макмерфи, пытающийся сломать губительную систему, оказывается в «пинжике» вместо тюрьмы. В качестве лекарства от творческого безумия он выбирает смех...

Фильм «Пролетая над гнездом кукушки», снятый по этому роману, завоевал шесть «Оскаров».

Похожие книги

- Жагда жизни: Винсент Ван Гог
- Бездна (Л. Андреев)
- Решение (К. Воннегут)
- Убить пересмешника
- Триумфальная арка
- Курьер (К. Шамазаров)

Все книги, похожие на Пролетая над гнездом кукушки



Forrest Gump 1994 (PG-13) CC

amazon instant video

Watch Trailer

Tom Hanks gives an astonishing performance as Forrest, an everyman whose simple innocence comes to embody a generation.

Starring: Tom Hanks, Rebecca Williams
Runtime: 2 hours 23 minutes
Available to watch on supported devices.

Rent or Buy with 1-Click®

- Rent HD \$3.99
- Rent SD \$2.99
- Buy HD \$14.99
- Buy SD \$9.99

Redeem a gift card or promotion code

Add to Watchlist

Share

Send us Feedback

By placing your order, you agree to our [Terms of Use](#). Sold by Amazon Digital Services, Inc. Additional taxes may apply.

Customers Who Bought This Item Also Bought

- AWAY
- GREEN MILE
- HAWSHANK REED
- SAVING PRIVATE RYAN
- Mrs. Doubtfire
- APOLLO 13

ОСНОВНЫЕ СВОЙСТВА РС

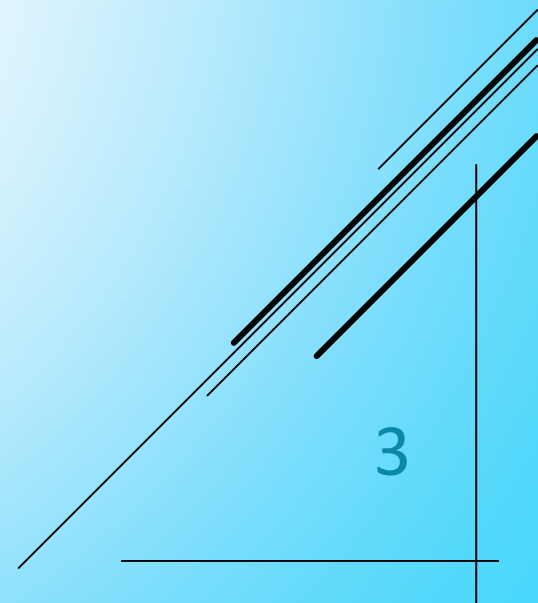


- 1) РС предназначены для облегчения выбора из *однородного множества вариантов*;
- 2) Характер этого выбора и критерии его осуществления являются, во многом, *субъективными*.

Рекомендующая система, таким образом, нацелена на обобщение и формализацию процесса *субъективного выбора*.

Рекомендующие системы vs. Экспертные системы:

- 1) Назначение
- 2) Источник информации
- 3) Представление знаний



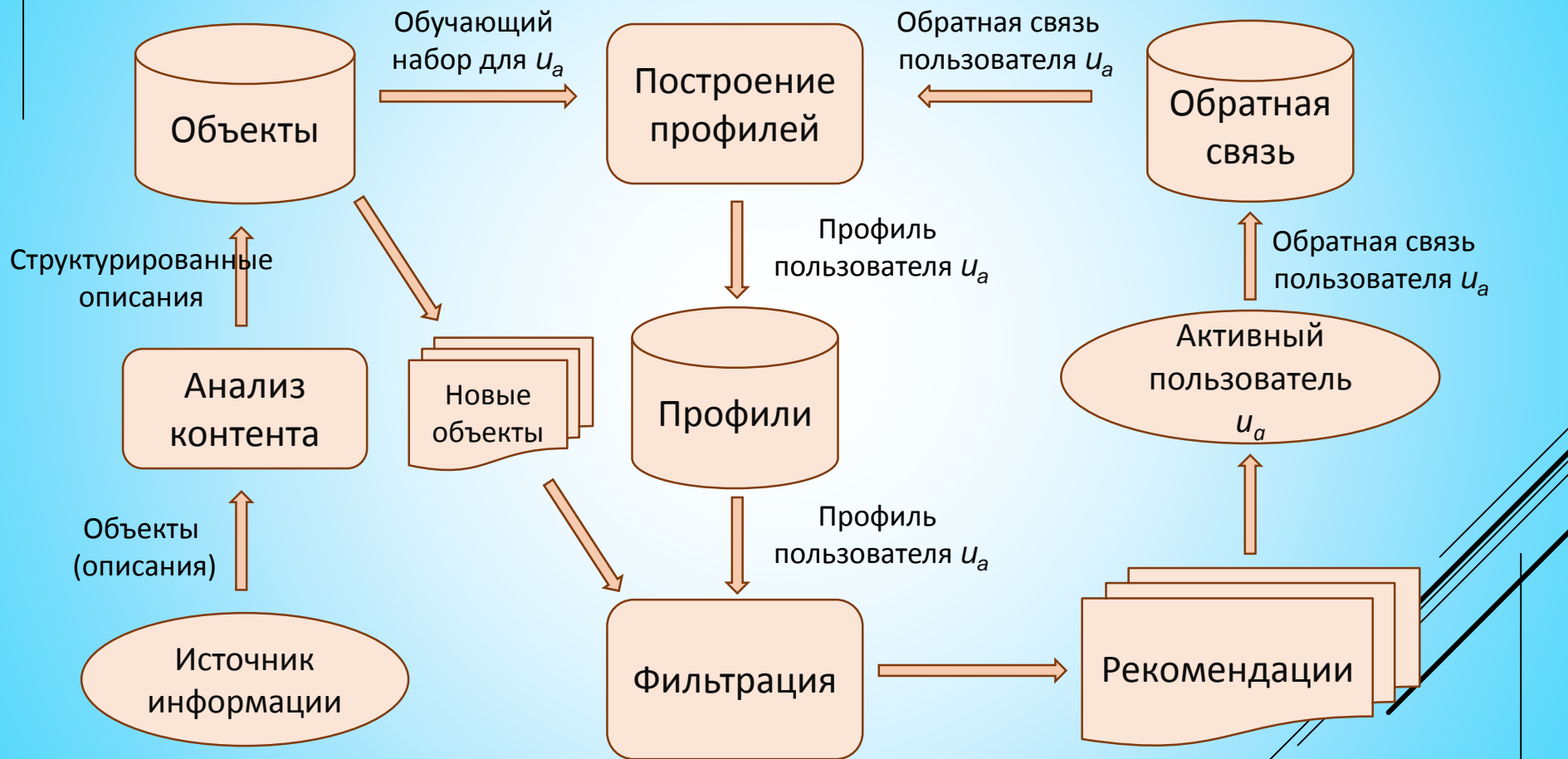
КЛАССИФИКАЦИЯ РС



Критерий классификации: вид информации, используемой для формирования рекомендаций.

- 1) Контентные РС (content-based).
- 2) Системы коллаборативной фильтрации (collaborative filtering).
- 3) Демографические (demographic).
- 4) Основанные на знаниях (knowledge-based).
- 5) Социальные (community-based).
- 6) Гибридные (hybrid).

КОНТЕНТНЫЕ РЕКОМЕНДУЮЩИЕ СИСТЕМЫ. ОБЩАЯ АРХИТЕКТУРА



КОНТЕНТНЫЕ РЕКОМЕНДУЮЩИЕ СИСТЕМЫ. ПРЕДСТАВЛЕНИЕ ОБЪЕКТОВ (1)

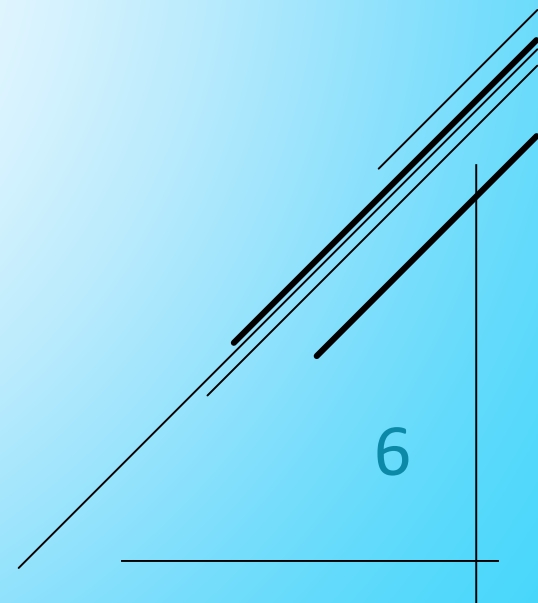


Зависит от природы объектов, которые собираемся рекомендовать!

Упрощение: работать не с самим содержимым, а с его *текстовым описанием*.

Широко используются методы информационного поиска (information retrieval):

- сайты (по содержимому);
- новостные сообщения;
- сообщения электронной почты;
- продукты (по описанию);
- ...



КОНТЕНТНЫЕ РЕКОМЕНДУЮЩИЕ СИСТЕМЫ. ПРЕДСТАВЛЕНИЕ ОБЪЕКТОВ (2)



Для текстовых объектов/описаний: **векторная модель документа, взвешенная по TF-IDF (и для профилей, и для объектов)**

$D = \{d_1, \dots, d_N\}$ – корпус документов;

$T = \{t_1, \dots, t_n\}$ – словарь (множество слов из корпуса);

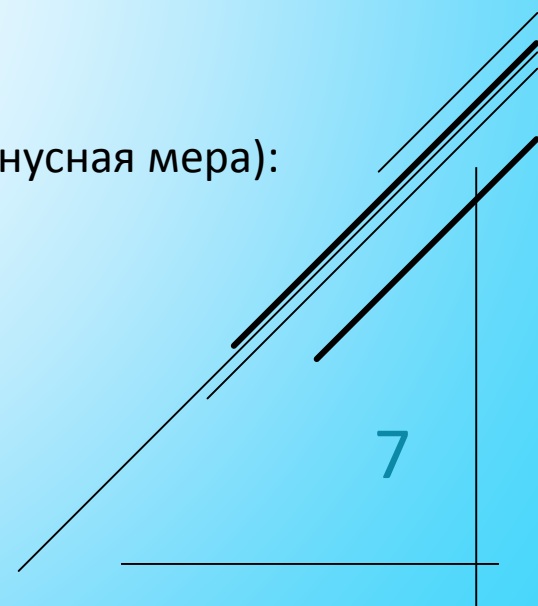
Документ представляется в виде $d_j = (w_{1j}, w_{2j}, \dots, w_{nj})$.

$$TFIDF(t_k, d_j) = TF(t_k, d_j) \cdot \log \frac{N}{n_k}$$

$$w_{kj} = \frac{TFIDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} TFIDF(t_s, d_j)^2}}$$

Вычисление степени сходства между двумя документами (косинусная мера):

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} w_{kj}}{\sqrt{\sum_k w_{ki}^2} \sqrt{\sum_k w_{kj}^2}}$$



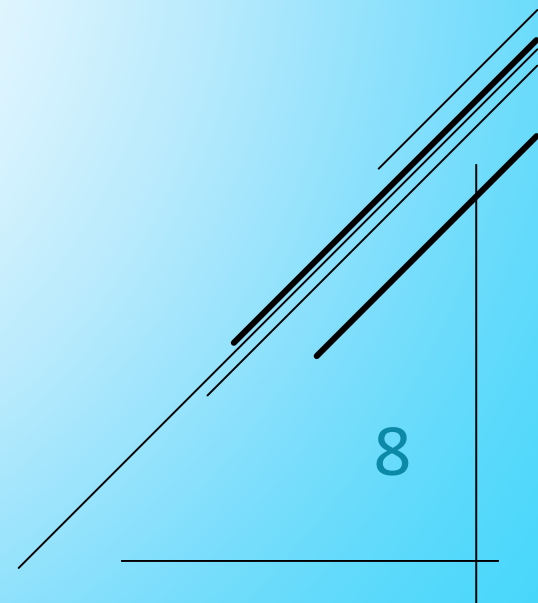
КОНТЕНТНЫЕ РЕКОМЕНДУЮЩИЕ СИСТЕМЫ. ПРОФИЛЬ ПОЛЬЗОВАТЕЛЯ



Множество объектов T_u , для каждого из которых известно: понравился ли (был ли полезен) пользователю u данный объект: $T_u = \{(X^{(1)}, y^{(1)}), \dots, (X^{(l_u)}, y^{(l_u)})\}$, $y^{(l_u)} \in \{0, 1\}$.

Варианты:

- дерево решений;
- наивный байесов классификатор;
- метод опорных векторов (support vector machine);
- ...



ДОСТОИНСТВА И НЕДОСТАТКИ КОНТЕНТНЫХ РЕКОМЕНДУЮЩИХ СИСТЕМ



Достоинства:

- отсутствие зависимости от других пользователей
- прозрачность получаемых рекомендаций («похожие объекты»)
- легкость добавления новых объектов

Недостатки:

- ограничения, связанные с анализом контента, сложность при построении моделей объектов
- излишняя специализация (низкие шансы на «неожиданную находку»)
- сложность формирования рекомендаций для нового пользователя

СИСТЕМЫ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ



Общей особенностью систем коллаборативной фильтрации является то, что *единственным* источником информации при формировании рекомендаций является информация об *оценках*, присвоенных пользователями различным объектам.

					
Alice		5	4	5	
Bob	5		5	?	5
Roger			5	5	

Предсказание
оценки (рейтинга)

Основные подходы:

- 1) Оценка сходства (neighborhood approach).
- 2) Модельные методы (например, моделирование латентных факторов).

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ПОСТАНОВКА ЗАДАЧИ



m – количество пользователей;

n – количество объектов;

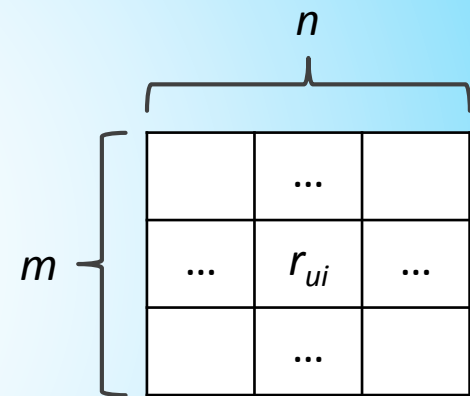
r_{ui} – оценка, присвоенная пользователем u объекту i , выражающая степень привлекательности (или полезности) объекта для пользователя;

K – множество известных оценок;

$R(u)$ – множество объектов, оцененных пользователем u ;

$R(i)$ – множество пользователей, оценивших объект i ;

\hat{r}_{ui} – «предсказанная» оценка.



КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. МЕТОДЫ, ОСНОВАННЫЕ НА СХОДСТВЕ

...пользователей
(user-based CF)

w_{uv}

Alice	5	3	4	3
Bob	5	3	5	?
Roger	2	5	3	5

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}$$

...объектов
(item-based CF)

w_{ij}

Alice	5	3	4	4
Bob	5	3	5	?
Roger	2	5	3	5

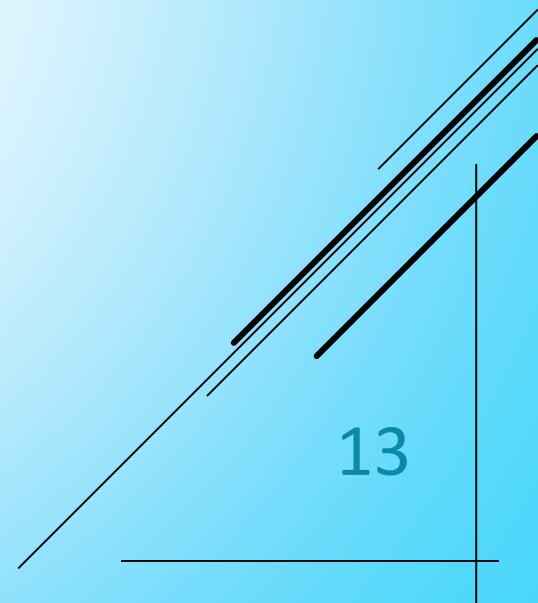
$$\hat{r}_{ui} = \frac{\sum_{j \in N_u(i)} w_{ij} r_{uj}}{\sum_{j \in N_u(i)} |w_{ij}|}$$

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. МЕТОДЫ, ОСНОВАННЫЕ НА СХОДСТВЕ (2)



Помимо выбора отношения вида отношения соседства (между объектами или между пользователями), для определения рекомендуемой процедуры необходимо задать:

- 1) Способ нормализации оценок.
- 2) Способ вычисления сходства (объектов или пользователей).
- 3) Способ выбора «ближайших соседей».



КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. НОРМАЛИЗАЦИЯ ОЦЕНОК



Проблема: пользователи могут по-разному использовать шкалу оценок.

Например:

Оценки пользователя 1: 10, 10, 8, 9, 10, 10.

Оценки пользователя 2: 8, 8, 5, 6, 7, 8.

Если использовать оценки пользователя 2 в качестве основы для вычисления оценок пользователя 1, то предсказания будут занижены.

$$h(r_{ui}) = r_{ui} - \bar{r}_u$$

Например:

Оценки пользователя 1: 10, 10, 8, 9, 10, 10 => 0.5, 0.5, -1.5, -0.5, 0.5, 0.5.

Оценки пользователя 2: 8, 8, 5, 6, 7, 8 => 1, 1, -2, -1, 0, 1.

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in N_i(u)} w_{uv} (r_{vi} - \bar{r}_v)}{\sum_{v \in N_i(u)} |w_{uv}|}$$

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ВЫЧИСЛЕНИЕ СТЕПЕНИ СХОДСТВА (1)

Косинусная мера:

$$cv(u, v) = \frac{\sum_{i \in R(u) \cap R(v)} r_{ui} r_{vi}}{\sqrt{\sum_{i \in R(u)} r_{ui}^2 \sum_{i \in R(v)} r_{vi}^2}}$$

				
Alice	5	3	4	3
Bob	5	3	5	?
Roger	2	5	3	5




$$cv(Alice, Bob) = \frac{5 \cdot 5 + 3 \cdot 3 + 4 \cdot 5}{\sqrt{(5^2 + 3^2 + 4^2 + 3^2) (5^2 + 3^2 + 4^2)}} \approx 0.91525$$

1	0.915	0.853
0.915	1	0.656
0.853	0.656	1

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ВЫЧИСЛЕНИЕ СТЕПЕНИ СХОДСТВА (2)

Коэффициент корреляции (Пирсона):

$$pc(u, v) = \frac{\sum_{i \in R(u) \cap R(v)} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in R(u) \cap R(v)} (r_{ui} - \bar{r}_u)^2 \sum_{i \in R(u) \cap R(v)} (r_{vi} - \bar{r}_v)^2}}$$

				
Alice	5	3	4	3
Bob	5	3	5	?
Roger	2	5	3	5

$$cv(Alice, Bob) \approx \frac{2 \cdot 0.67 + 0 \cdot (-1.33) + 1 \cdot 0.67}{\sqrt{(2^2 + 0^2 + 1^2) (0.67^2 + (-1.33)^2 + 0.67^2)}} \approx 0.828$$

1	0.828	-0.986
0.828	1	-0.896
-0.986	-0.896	1

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ВЫБОР «БЛИЖАЙШИХ СОСЕДЕЙ»



Реализация выбора включает ответ на два вопроса:

- 1) определение того, какие степени сходства, которые будут сохраняться рекомендующей системой (user-based – m^2 , item-based – n^2);
- 2) фактический выбор «ближайших соседей» среди сохраненных.

Эвристики, применяемые при хранении «соседей»:

- 1) Лучшие N ($N \geq K$).
- 2) Превышающие определенный заранее заданный порог.
- 3) Сигнализирующие о сходстве пользователей (а не о различии).
(или их комбинация)

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. МОДЕЛИРОВАНИЕ ЛАТЕНТНЫХ ФАКТОРОВ

Отображение и пользователей, и объектов в единое пространство латентных факторов размерности f таким образом, чтобы исходные данные (оценки, полученные от пользователей), можно было восстановить с помощью операции скалярного произведения соответствующих векторов в пространстве латентных факторов.

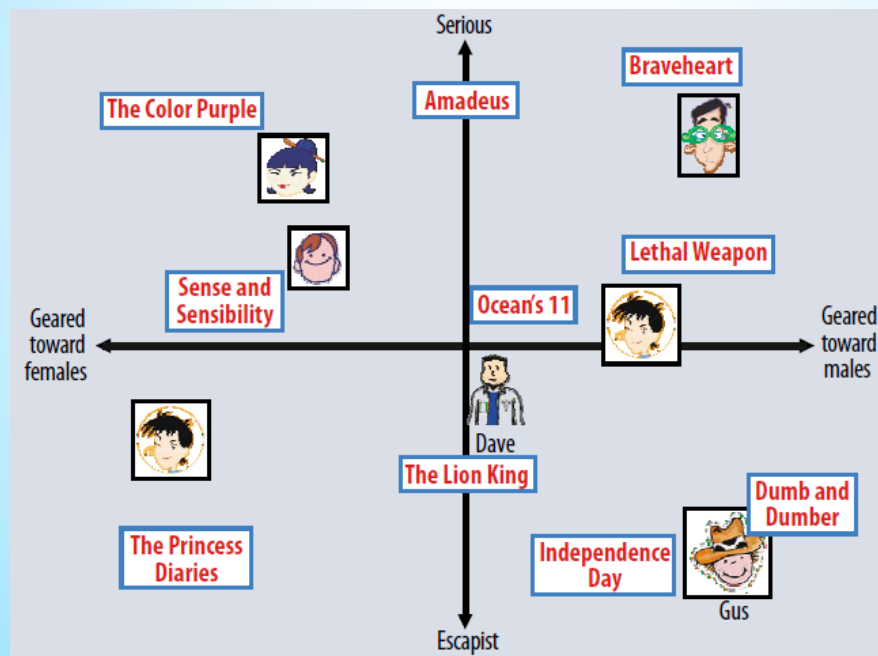


Иллюстрация из Koren Y., Bell R., Volinsky C. Matrix factorization techniques for recommender systems

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. МОДЕЛИРОВАНИЕ ЛАТЕНТНЫХ ФАКТОРОВ



Т.е. переход от множества рейтингов, к множеству векторов следующего вида:

- пользователи – $p_u \in \mathbb{R}^f$ (в какой мере пользователю близки (интересны) объекты, обладающие каждым из факторов);
- объекты – $q_i \in \mathbb{R}^f$ (в какой мере объект обладает каждым из факторов).

Оценка объекта i пользователем u представляется как скалярное произведение $q_i^T p_u$.

Идея метода:

- 1) Найти значения векторов p_u и q_i , используя для этого известные оценки r_{ui} .
- 2) Использовать найденные значения этих векторов для вычисления предсказаний неизвестных оценок \hat{r}_{ui} .

(Вариант интерпретации: $R = PQ^T$ - разложение (факторизация) матрицы.)

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ПОИСК РАЗЛОЖЕНИЯ



Функция стоимости:

$$J(p, q) = \frac{1}{2} \sum_{(u,i) \in K} (r_{ui} - q_i^T p_u)^2$$

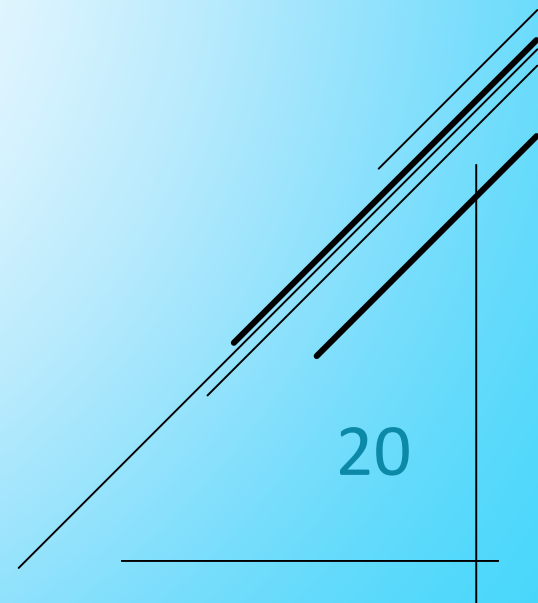
Для всех известных
оценок

Задача параметрического обучения:

$$(p^*, q^*) = \arg \min_{p, q} \frac{1}{2} \sum_{(u,i) \in K} (r_{ui} - q_i^T p_u)^2$$

Как искать?

- Градиентный спуск
- Стохастический градиентный спуск
- Метод чередующихся наименьших квадратов



КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ПОИСК РАЗЛОЖЕНИЯ



(Метод стохастического градиентного спуска)

Основное отличие от классического метода градиентного спуска заключается в том, что *вместо точной оценки значения градиента* в точке, вычисляется лишь *оценка градиента с использованием только одного обучающего примера*, выбранного случайным образом.

Вычислительная схема нахождения векторов p^* и q^* :

Параметры: α (коэффициент обучения)

1. <random initialization of p and q >
2. repeat
3. shuffle(K);
4. for (u, i, r) in K :
5. $\text{old_p} = p[u]$; $\text{old_q} = q[i]$;
6. $e = r - \text{predict}(u, i)$;
7. $p[u] = \text{old_p} + \alpha * e * \text{old_q}$; // vectorized
8. $q[i] = \text{old_q} + \alpha * e * \text{old_p}$; // vectorized
9. until <converges>;

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ПОИСК РАЗЛОЖЕНИЯ



(улучшенная модель)

Пусть μ – средняя оценка по всем обучающим примерам, b_u соответствует базовому смещению оценки в соответствии с особенностями пользователя u , а b_i – смещению оценки в соответствии с особенностями объекта i , и пусть предсказание оценки с учётом этих параметров вычисляется по следующей формуле:

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

Тогда функция стоимости может быть записана следующим образом:

$$J(b_i, b_u, p, q) = \sum_{(u,i) \in K} (r_{ui} - \mu - b_i - b_u - q_i^T p_u)^2 + \lambda(b_i^2 + b_u^2 + \|q_i\|^2 + \|p_u\|^2).$$

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ. ПОИСК РАЗЛОЖЕНИЯ



(улучшенная модель, продолжение)

Процедура обучения модели схожая (стохастический градиентный спуск), отличается только формулами пересчёта переменных:

$$b_u \leftarrow b_u + \alpha(e_{ui} - \lambda b_u)$$

$$b_i \leftarrow b_i + \alpha(e_{ui} - \lambda b_i)$$

$$q_i \leftarrow q_i + \alpha(e_{ui}p_u - \lambda q_i)$$

$$p_u \leftarrow p_u + \alpha(e_{ui}q_i - \lambda p_u)$$

ОРГАНИЗАЦИЯ ПРОЦЕССА



Простая схема:

- 1) Разбиваем множество известных оценок на два: тренировочное (80%) и проверочное (20%).
- 2) Обучаем модель на тренировочном.
- 3) Измеряем функцию ошибки на проверочном.

Более правильная схема (например, для подбора коэффициента регуляризации):

- 1) Разбиваем множество известных оценок на три: тренировочное (60%), множество валидации (20%) и проверочное (20%).
- 2) Обучаем модель на тренировочном.
- 3) Измеряем функцию ошибки на множестве валидации.
- 4) Повторяя шаги 2-3, настраиваем параметры так, чтобы добиться минимальной ошибки на примерах из множества валидации.
- 5) Измеряем функцию ошибки на проверочном.

ДОСТОИНСТВА И НЕДОСТАТКИ СИСТЕМ КОЛЛАБОРАТИВНОЙ ФИЛЬТРАЦИИ



Достоинства:

- универсальность, отсутствие необходимости анализа и моделирования контента
- высокая точность рекомендаций
- шансы найти неожиданный незнакомый объект выше, чем в контентных системах

Недостатки:

- варианты проблемы «холодного старта» (новый пользователь, новый объект)
- сложности с объяснением рекомендаций
- возможность мошенничества/атак

УЧЁТ КОНТЕКСТА В РЕКОМЕНДУЮЩИХ СИСТЕМАХ. ВИДЫ КОНТЕКСТА



Во многих приложениях условия, в которых пользователь оценивает объект, существенным образом влияют оценку. Подобные условия в исследованиях по рекомендующим системам получили название *контекста*.

Виды контекста, актуальные для рекомендующих систем (в первую очередь, мобильных):

- Физический контекст (время, положение, вид деятельности пользователя, погода, освещенность и т.п.).
- Социальный контекст (наличие и роль окружающих людей).
- Контекст устройства (вид и характеристики устройства, с которого осуществляется доступ к информации).
- Модальный контекст (настроение пользователя, цель, опыт, когнитивные способности).

УЧЁТ КОНТЕКСТА В РЕКОМЕНДУЮЩИХ СИСТЕМАХ. ПЕРЕХОД К МНОГОМЕРНОЙ ЗАДАЧЕ



Классические рекомендующие системы

Известные оценки:
 $K = \{(user, item, rating)\}$

	...	
...	r_{ui}	...
	...	

Контекстно-зависимые рекомендующие системы

Известные оценки:
 $K = \{(user, item, C_1, \dots, C_k, rating)\}$

	...	
...	r_{uic_1}	...
	...	

УЧЁТ КОНТЕКСТА В РЕКОМЕНДУЮЩИХ СИСТЕМАХ. ПОДХОДЫ



В области использования контекста в рекомендующих системах наметилось три подхода:

- *предварительная контекстная фильтрация* (contextual pre-filtering)
- *контекстная постфильтрация* (contextual post-filtering)
- *моделирование контекста* (contextual modeling)

ОЦЕНКА КАЧЕСТВА РЕКОМЕНДАЦИЙ (1)

Классический способ для систем, предсказывающих оценки:

$$RMSE = \sqrt{\frac{1}{|K|} \sum_{(u,i) \in K} (\hat{r}_{ui} - r_{ui})^2}$$

Классический способ для систем, формирующих список рекомендаций (не предсказывающих оценки):

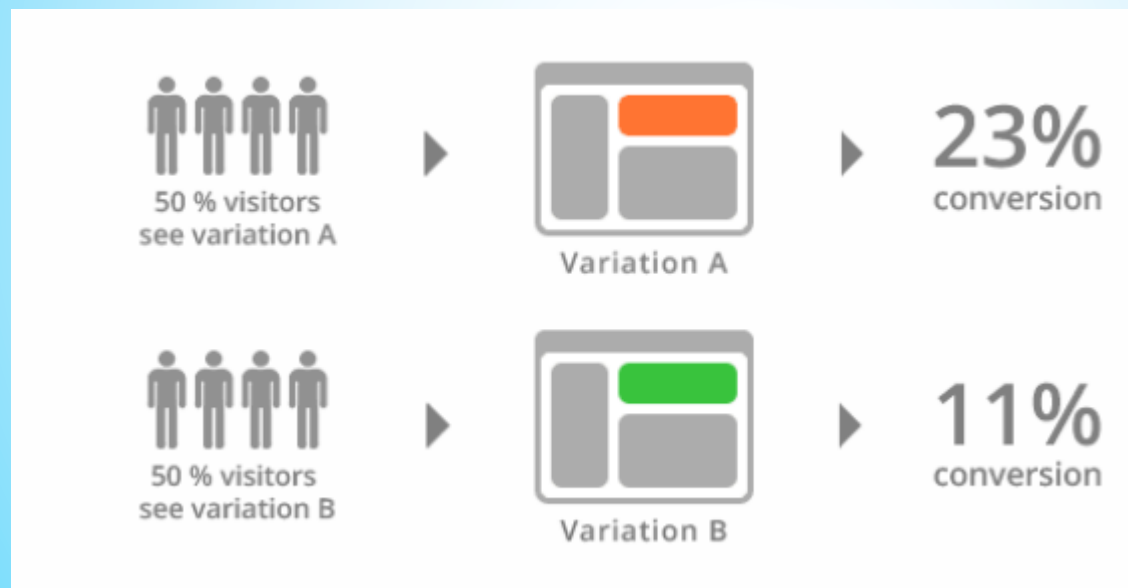
	Рекомендован	Не рекомендован
Полезен	tp	fn
Бесполезен	fp	tn

$$Precision = \frac{\#tp}{\#tp + \#fp}$$

$$Recall = \frac{\#tp}{\#tp + \#fn}$$

ОЦЕНКА КАЧЕСТВА РЕКОМЕНДАЦИЙ (2)

- Измерять нужно то, что соответствует цели
- Цели внедрения рекомендующих систем... только ли удобство пользователей?
- Более сложные и «жизненные» меры качества в новых (экспериментальных) поколениях РС



Только
онлайн

30

Иллюстрация с <https://vwo.com/ab-testing/>

ЛИТЕРАТУРА



- 1) Recommender Systems Handbook, Ricci F., Rokach L., Shapira B., Kantor P.B. (Eds.), 2011
- 2) Koren Y., Bell R., Volinsky C. Matrix Factorization Techniques for Recommender Systems // Computer, vol. 42, iss. 8, 2009, pp. 30-37
- 3) Adomavicius G., Tuzhilin A., Context-aware recommender systems // AI Magazine, vol. 32, no 3.

Материалы конференций:

ACM Conference on Recommender Systems