Восстановление пропущенных значений

Лекция №2

Лектор: Шевляков Артём

Проблема!

- В таблице могут быть незаполненные ячейки, или же содержимое некоторых ячеек заведомо некорректное и противоречит здравому смыслу

Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	3
Запеканка	?	185	64	-4
Ватрушкина	0	168	666	2
Ололоева	0	?	85	1

Что нужно делать?

- Удалить объект (т.е. строку).
- Удалить столбец, если в нём очень много пропусков.
- Заменить значение в ячейке на среднее (медиану, моду…) из значений столбца.

Например, для столбца-признака (0,1,2,4,4,5,?) пропущенное значение можно заменить на среднее 2.67, медиану 3, моду 4.

А что делать с номинальными признаками?

Есть признак «пол человека» (0,0,0,1,1,?). На что заменить пропущенное значение???

Ваши предложения?

А что делать с номинальными признаками?

Есть признак «пол человека» (0,0,0,1,1,?). На что заменить пропущенное значение???

Ваши предложения?

1. Заменить на моду (то есть на 0).
2. Сгенерировать 0 с вероятностью 3/5 или 1 с вероятностью 2/5.
3. Волевым решением объявить «пол» числовым признаком (со всеми вытекающими отсюда последствиями), и применить к нему методы восстановления для числовых признаков.

Восстановление данных с помощью метрики

Появление метрики

А что такое метрика? Это обобщение понятия расстояния из геометрии. Но метрика (в отличие от расстояния)…

1) может быть вычислено для объектов произвольной природы;
2) не обязана вычисляться по известной формуле из школьного учебника геометрии.

Известные метрики

Если даны два набора

$$P = (p_1, p_2, \dots, p_n) \quad Q = (q_1, q_2, \dots, q_n)$$

то расстояние между *P* и *Q* может быть найдено, например, такими способами:

1. как в учебнике геометрии (евклидова метрика)

$$\rho(P,Q) = \sqrt{\begin{matrix}(p_1 - q_1)^2 + (p_2 - q_2)^2 + \\ \dots + (p_n - q_n)^2\end{matrix}}$$

2. Метрика Манхеттен

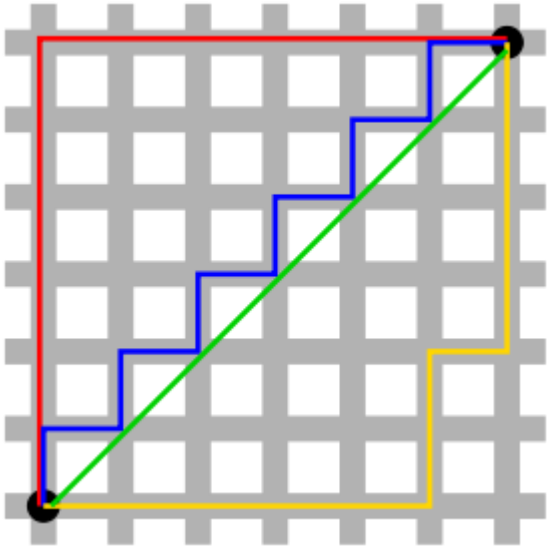
$$\rho(P,Q) = |p_1 - q_1| + |p_2 - q_2| + \ldots + |p_n - q_n|$$

3. max-метрика

$$\rho(P,Q) = \max\{|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_n - q_n|\}$$

В общем, метрик очень много, вы можете придумать свои собственные.

При чём тут Манхеттен?



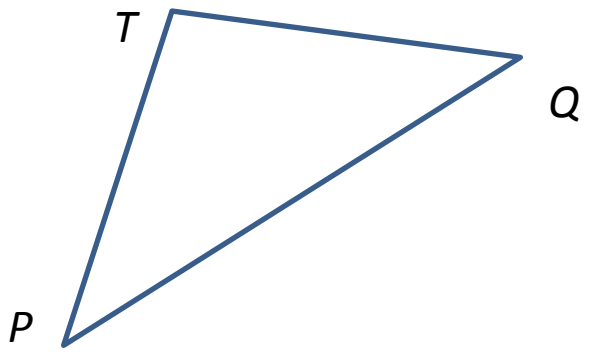
Свойства метрики

1. $\rho(P,P) = 0$ (в принципе, логично)
2. $\rho(P,Q) = \rho(Q,P)$ (разумно). Но можете ли привести пример из жизни, когда расстояние от *P* до *Q* не равно расстоянию от *Q* до *P*?
3. $\rho(P,Q) \leq \rho(P,T) + \rho(T,Q)$ (неравенство треугольника).

А есть ситуации, когда

это неравенство

не выполняется?



Восстановление данных с помощью метрики

Пусть у объекта *A* значение признака *P* не корректно. Как восстановить *P* для *A*?

1. Исключим пока из таблицы столбец с признаком P (мы предполагаем, что остальные ячейки в таблице нормальные).
2. Найдем расстояния (с помощью некоторой метрики) от строки A до остальных объектов таблицы. Получим числа

$$\rho(A, A_1), \rho(A, A_2), \dots, \rho(A, A_n)$$

Восстановление данных с помощью метрики (продолжение)

3. Пусть значения признака *P* для объектов $A_1, A_2, \ldots, A_n$ равны $P(A_1), P(A_2), \ldots, P(A_n)$.

Итак, у нас есть числа $P(A_1), P(A_2), \ldots, P(A_n)$ и $\rho(A, A_1), \rho(A, A_2), \ldots, \rho(A, A_n)$ как их собрать в одну формулу?

Щас прервёмся на примерчик…

Пример

Объекты	P1	P2	P3	P4	P
A1	3	4	5	3	4
A2	5	5	5	4	3
A3	4	3	3	2	5
A	5	4	3	3	?

Если пропуск заменить на среднее или медиану по столбцу (здесь они равны друг другу), то нужно писать 4. Попытаемся заполнить пропуск с помощью различных метрик.

Объекты	P1	P2	P3	P4	P
A1	3	4	5	3	4
A2	5	5	5	4	3
A3	4	3	3	2	5
A	5	4	3	3	?

Вид метрики	От A до A1	От A до A2	От A до A3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

Кто догадается, что с этими числами делать дальше?

Ключевая идея: признак *P* для объекта *A* должен быть близок к значению признака *P* у близких к *A* объектов.

Нужно взять их лин. комбинацию со значениями признака *P*

То есть по евклидовой метрике надо так:

Вид метрики	От A до A1	От A до A2	От A до A3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

Объекты	P
A1	4
A2	3
A3	5
A	?

$$\frac{1}{\frac{1}{2.83}+\frac{1}{2.45}+\frac{1}{1.73}}\left(\frac{4}{2.83}+\frac{3}{2.45}+\frac{5}{1.73}\right)$$
$$= 4.15$$

Аналогично дается ответ с помощью других метрик

Вид метрики	От А до А1	От А до А2	От А до А3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

Например,

макс-метрика даёт:

Объекты	Р
А1	4
А2	3
А3	5
А	?

$$\frac{1}{1/2 + 1/2 + 1/1}(4/2 + 3/2 + 5/1) = 4.25$$

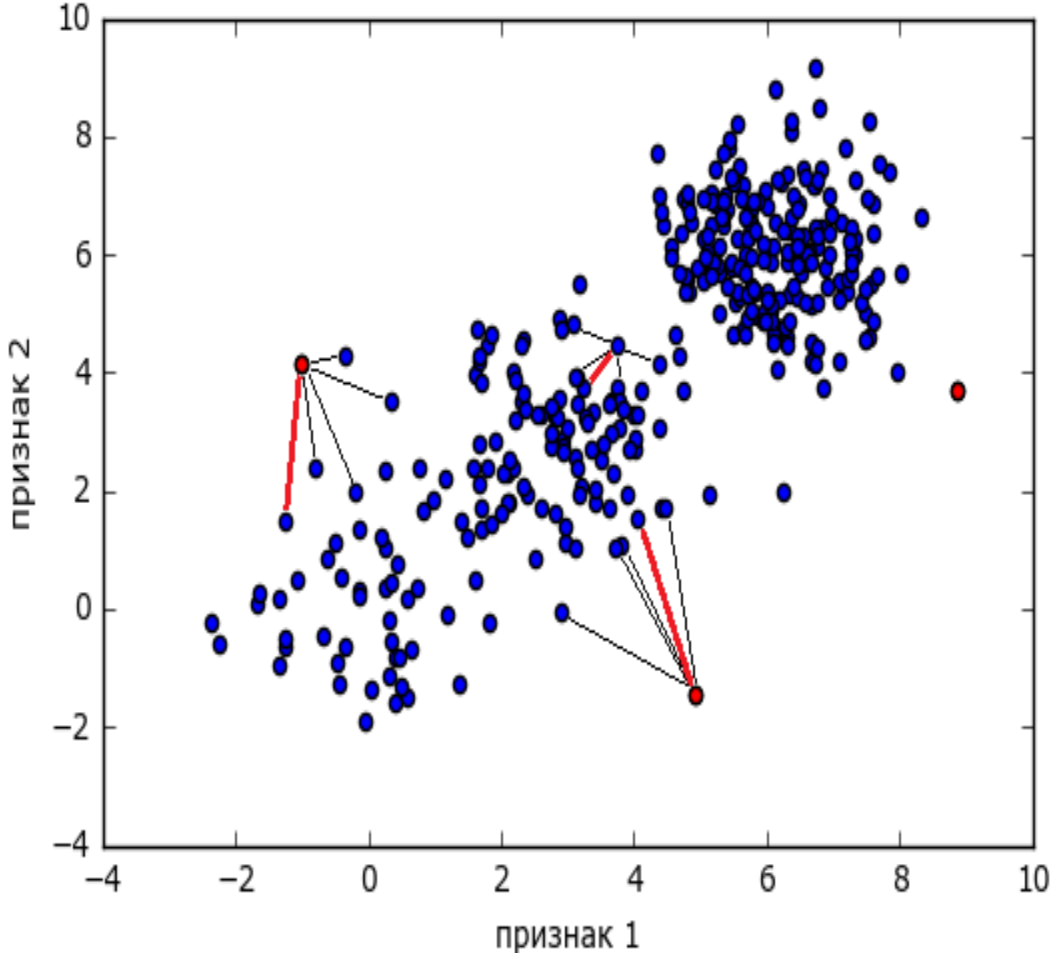
Умная формула для восстановления данных с помощью метрики

$$P(A) = \frac{1}{\sum_{i=1}^{n} \frac{1}{\rho(A, A_i)}} \left(\sum_{j=1}^{n} \frac{P(A_j)}{\rho(A, A_j)} \right)$$

Замечание об использовании метрики

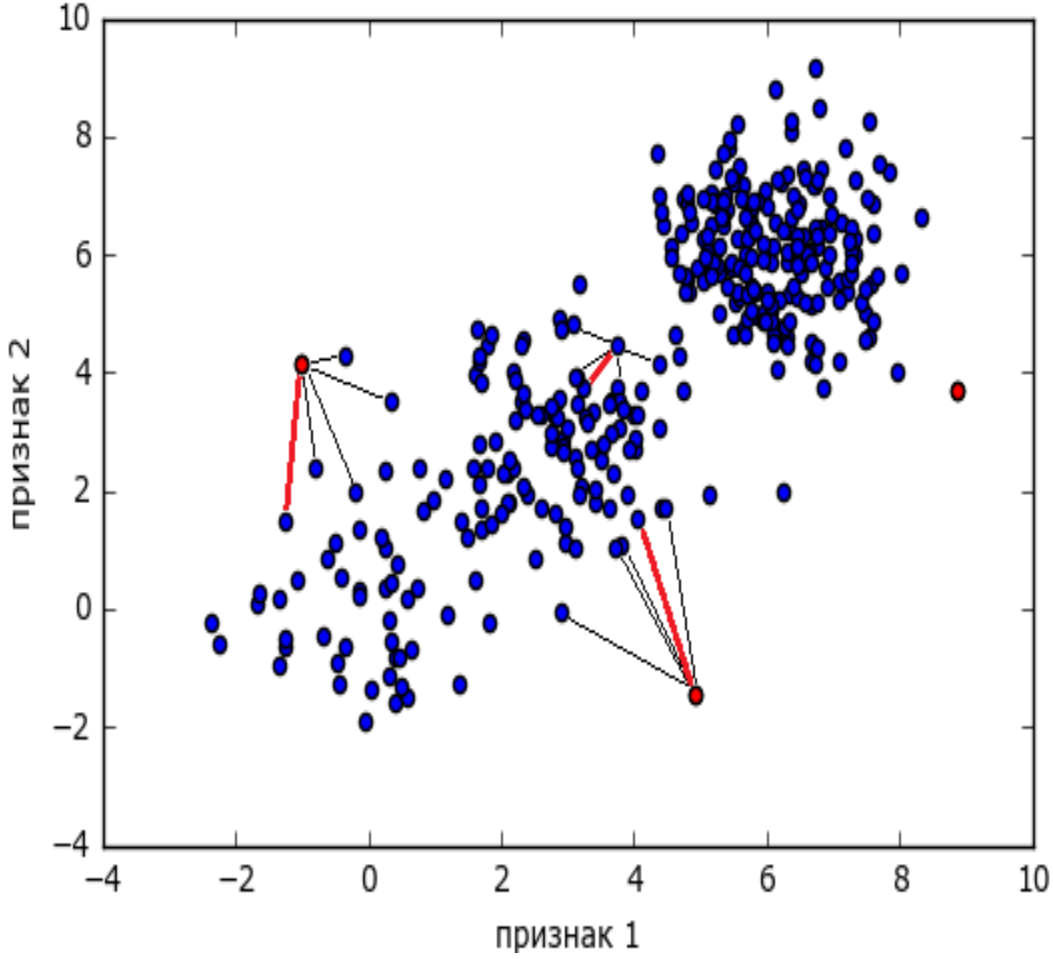
Замечание об использовании метрики

Часто (см. будущие лекции) объекты представляются точками в пространстве признаков и между ними считаются расстояние (метрика).



Замечание об использовании метрики

Для адекватной работы необходимо, чтобы все признаки (значения по осям) имели одинаковый масштаб. Иначе…



Замечание об использовании метрики

… некоторые признаки фактически будут проигнорированы (в след. примере различия в росте и различия в весе имеют ОЧЕНЬ разную ценность – это из-за разных единиц измерения).

Студент	Вес, кг	Рост, м
Иванов	61	1,76
Сидорова	56	1,50
Петров	100	1,98

При вычислении метрики все признаки приводить к единой шкале (нормировать)

Признак: $P = (p_1, p_2, \dots, p_n)$

Далее используем обозначения:

$\bar{p}$ - среднее значение, S_P - отклонение

Способы нормировки признака:

1. Перевести все значения признака в интервал [0,1]:

$$p_i' = \frac{p_i - \min\{p_i\}}{\max\{p_i\} - \min\{p_i\}}$$

При вычислении метрики все признаки приводить к единой шкале (нормировать)

2. Выполнить преобразование

$$p_i' = \frac{p_i - \bar{p}}{s}$$

после этого у признака *P* среднее значение и отклонение будут равны…

Нужно все признаки приводить к единой шкале (нормировать)

2. Выполнить преобразование

$$p_i' = \frac{p_i - \bar{p}}{s}$$

после этого у признака *P* среднее значение и отклонение будут равны 0 и 1 соответственно.

3. Помимо формул из пп.1-2 к признакам можно предварительно применять разные функции (например, *log*)

По-хорошему в последнем примере признаки тоже нужно нормировать

Объекты	P1	P2	P3	P4	P
A1	3	4	5	3	4
A2	5	5	5	4	3
A3	4	3	3	2	5
A	5	4	3	3	?

То есть лучше работать с такой таблицей:

Объекты	P1	P2	P3	P4	P
A1	0	0.5	1	0.5	4
A2	1	1	1	1	3
A3	0.5	0	0	0	5
A	1	0.5	0	0.5	?

Использование коэффициента корреляции для восстановления данных

Восстановление данных с помощью других столбцов

В отличие от предыдущего метода (восстановления с помощью строк) тут есть трудность: значения в разных столбцах могут иметь разный масштаб.

	Грудь	Талия	Бёдра	Рост	Вес
Ж1	99	56	91	160	58
Ж2	89	58	89	157	48
Ж3	91	64	91	165	54
Ж4	91	51	91	170	54
Ж5	86	56	84	157	44
Ж6	97	53	86	175	56
Ж7	?	51	91	165	54

Восстановление данных с помощью других столбцов

Нужна величина, которая показывает, как значения одного признака определяют значения другого признака. Эта величина должна иметь смысл и для признаков с разными единицами измерения.

В статистике для таких задач используют коэффициент корреляции (КК).

Пример зависимости между столбцами

Здесь сильная зависимость между признаком *P1* и признаками *P3,P4,P5.*

P1	P2	P3	P4	P5
0	1	0	10	4
1	0	100	11	3
2	3	200	12	2
3	2	300	13	1

Так как их пары их значений ложатся на прямую.

Формула для КК

Пусть $P = (p_1, p_2, \dots, p_n)$ $Q = (q_1, q_2, \dots, q_n)$

столбцы (строки) из таблицы.

Тогда КК считается по формуле

$$r(P,Q) = \frac{\sum_{i=1}^{n} p_i q_i - n\bar{p}\bar{q}}{(n-1)s_P s_Q}$$

Свойства КК

Коэффициент корреляции (КК) – это число из отрезка [–1,1], которое имеет следующий смысл:

1. Если КК=0 (или близок к нему), то очевидной зависимости между признаками *P,Q* нет.
2. Если КК>0, то бОльшим значениям признака *P*, как правило, соответствуют бОльшие значения признака Q.
3. Если КК<0, то бОльшим значениям признака *P*, как правило, соответствуют меньшие значения признака Q.
4. Чем ближе значение модуля КК к единице, тем сильнее зависимость между признаками *P,Q* (то есть по значению одного признака легче предсказать значение второго).
5. Если модуль КК равен 1, то между признаками *P,Q* существует линейная зависимость.

Задача на понимание

Однажды я попросил, чтобы студенты ответили на 2 вопроса анкеты «ваш год рождения» и «ваш возраст». Из их ответов я сформировал таблицу, в которой был столбец P=«год рождения студента» и Q=«возраст студента».

Вопрос 1: оцените (приближенно) КК $r(P,Q)$. Он больше или меньше нуля?

Задача на понимание

Однажды я попросил, чтобы студенты ответили на 2 вопроса анкеты «ваш год рождения» и «ваш возраст». Из их ответов я сформировал таблицу, в которой был столбец P=«год рождения студента» и Q=«возраст студента».

Вопрос 1: оцените (приближенно) КК $r(P,Q)$.

Вопрос 2: как зависит $r(P,Q)$ от месяца, в котором проводится опрос (я не шучу)?

Восстановление данных с помощью других столбцов (вернулись к задаче)

Нужно для каждого столбца (без использования последней строки) подсчитать его среднее и КК всех столбцов с 1-м столбцом.

	Грудь	Талия	Бёдра	Рост	Вес
Ж1	99	56	91	160	58
Ж2	89	58	89	157	48
Ж3	91	64	91	165	54
Ж4	91	51	91	170	54
Ж5	86	56	84	157	44
Ж6	97	53	86	175	56
А	?	51	91	165	54

	Среднее
Грудь	92,17
Талия	56,33
Бёдра	88,67
Рост	164
Вес	52,33

Коэффициенты корреляции:

r(Грудь,Талия)= -0,22

r(Грудь,Бедра)= 0,34

r(Грудь,Рост)= 0,46

r(Грудь,Вес)= 0,91

Смысл в том, признак с большим (по модулю) КК имеет большее влияние на размер груди.

Осталось изобрести подходящую формулу…

Умная формула

Пусть *P(A)* - значение признака *P* объекта *A*.

$\bar{P}$ - среднее значение признака *P.*

Требуется определить P(A) по столбцам-признакам $P_1, P_2, \ldots, P_m$

$$P(A) = \bar{P} + \frac{\sum_{j=1}^{n} r(P, P_i)(P_i(A) - \bar{P}_i)}{\sum_{i=1}^{m} |r(P, P_i)|}$$

Примечание: в формуле все величины вычисляются без учета строки объекта *A*

В наше примере имеем

A	?	51	91	165	54

	Среднее
Грудь	92,17
Талия	56,33
Бёдра	88,67
Рост	164
Вес	52,33

r(Грудь,Талия)= -0,22

r(Грудь,Бедра)= 0,34

r(Грудь,Рост)= 0,46

r(Грудь,Вес)= 0,91

Формула дает

$$P(A) = 92.17 + \frac{1}{0.22 + 0.34 + 0.46 + 0.91} \cdot$$

$$(-0.22(51 - 56.33) + 0.34(91 - 88{,}67) + 0.46(165 - 164) + 0.91(54 - 52.33))$$

$$= 92.17 + 0.52(1{,}17 + 0.79 + 0.46 + 1.52) = 94.22$$

КК можно применять и для строк

Для этого нужно вычислить КК строки с пропуском и остальных строк таблицы.

И по аналогичной формуле восстановить пропущенное значение.

	Грудь	Талия	Бёдра	Рост	Вес
Ж1	99	56	91	160	58
Ж2	89	58	89	157	48
Ж3	91	64	91	165	54
Ж4	91	51	91	170	54
Ж5	86	56	84	157	44
Ж6	97	53	86	175	56
А	?	51	91	165	54

КК и нормирование

При использовании КК в работе с данными признаки можно не нормировать (нормировка «зашита» в формулу для вычисления КК).

$$r(P,Q) = \frac{\sum_{i=1}^{n} p_i q_i - n\bar{p}\bar{q}}{(n-1)s_P s_Q}$$

КК также не изменится при изменении масштаба признаков и переводе признаков в другие единицы измерения.

Применение метрик и КК в рекомендательных системах (РС)

Применение метрик и КК в рекомендательных системах (РС)

Формально РС - это таблица: строки соответствуют пользователям, столбцы – товарам, а в *(i,j)*-ячейке стоит оценка, которую поставил i-й пользователь j-му товару (в ячейке может быть пустой, если оценивания не было).

	Чупачупс	Доширак	Боярышник	ВАЗ-2101	Семечки
Вася	3	4	5	3	4
Петя	5	5	5	4	3
Маша	4	3	3	2	5
Саша	5	4	3	3	?

Предсказать оценку в РС – это фактически восстановить данные в таблице!!!

А выше было показано, как это делается с помощью метрики и КК.

Например, при использовании КК для строк этой таблицы получится ответ 4.12

Результаты восстановления с помощью метрики были выше (эта таблица повторялась на предыдущих слайдах).

	Чупачупс	Доширак	Боярышник	ВАЗ-2101	Семечки
Вася	3	4	5	3	4
Петя	5	5	5	4	3
Маша	4	3	3	2	5
Саша	5	4	3	3	4.12

РС с анонимными пользователями

Предыдущий пример РС существенно опирался на знание историй пользователей РС. А как быть, когда сервис (сайт) посещается анонимными пользователями?

Как в это случае определять расстояние между пользователями (товарами)?

РС с анонимными пользователями

Можно исхитриться так (на примере интернет-магазина): пользователи, как правило, составляют заказ из нескольких товаров. Мерой близости товаров может стать величина «как часто они попадают в один заказ»

РС с анонимными пользователями

Более формально: признак $P_i=1$, если товар попал в заказ с номером i (иначе $P_i=0$). Должна получиться примерно такая таблица

Товары	Номера заказов								
	1	2	3	4	5	6	7	8	9
A	1	1	0	0	1	1	0	0	0
B	0	1	0	0	1	0	1	1	0
C	0	0	0	1	1	0	1	0	1

Теперь можно считать расстояния (или КК) между строками таблицы. Это и будет мерой близости товаров друг к другу.

Другие способы восстановления данных

Восстановление данных с помощью моделей предсказания

(Более подробно идеи этой темы будут обсуждаться на лекциях по моделям предсказания)

Идея: если мы сумеем научить предсказывать значение признака *P* используя для этого все другие признаки, то автоматически будет решена проблема восстановления данных для признака *P*.

В этой таблице восстановление данных эквивалентно предсказанию места на олимпиаде по остальным данным студента

Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	?
Запеканка	1	185	64	?
Ватрушкина	0	168	61	?
Ололоева	0	201	85	?

Используемая литература

- Статья habrahabr.ru/company/infopulse/blog/283168/
- Обзорная статья: «Collaborative Filtering Recommender Systems» by M.Ekstrand, J.Riedl, J.Konstan
- Книга Т.Сегаран «Программируем коллективный разум» (глава про рекомендательные системы)
- Гуглить фразу «Коллаборативная фильтрация»