Введение в анализ данных: Анализ данных для Интернет рекламы

Юля Киселёва
juliakiseleva@yandex-team.ru
Школа анализа данных

План на сегодня

- Онлайн алгоритмы
- Задача сопоставления
- История рекламы
- Задача adwords

Алгоритмы

- Классическая модель для алгоритма
  - Вы получаете большой объем данных и на основе них строите функцию предсказания
  - Будем называть offline алгоритмами
- Онлайн алгоритмы
  - Вы получаете часть информации во времени и необходимо принять решение, основываясь на этой информации

Интернет реклама

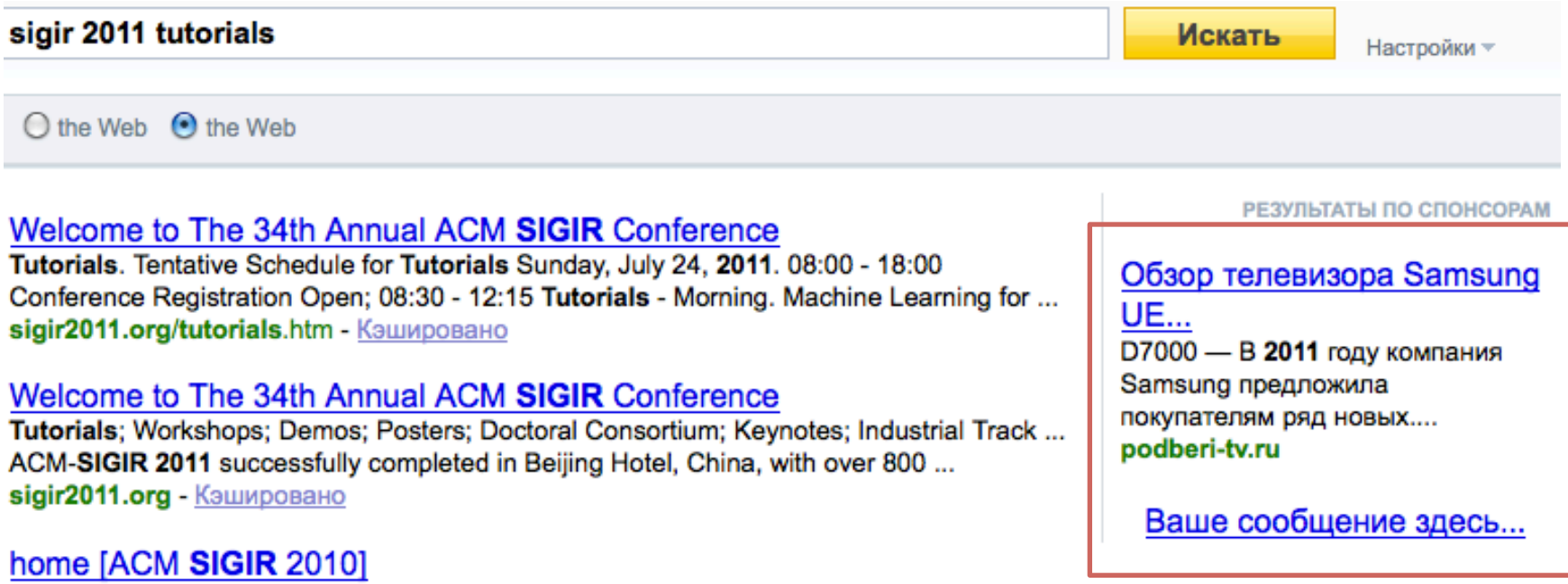
Данные по бюджету интернет-рекламы (US, November 2010)

Год	Online	Online% от всей медиа
2009	$22.7B	13.9%
2010	$25.8B	15.9%
2011	$28.5B	16.7%
2012	$32.6B	18.3%
2013	$36.0B	19.8%
2014	$40.5B	21.5%

Текстовая реклама

- Рекламный поиск (Sponsored Search)
- Контекстная реклама (Context Match)

Пример. Sponsored Search



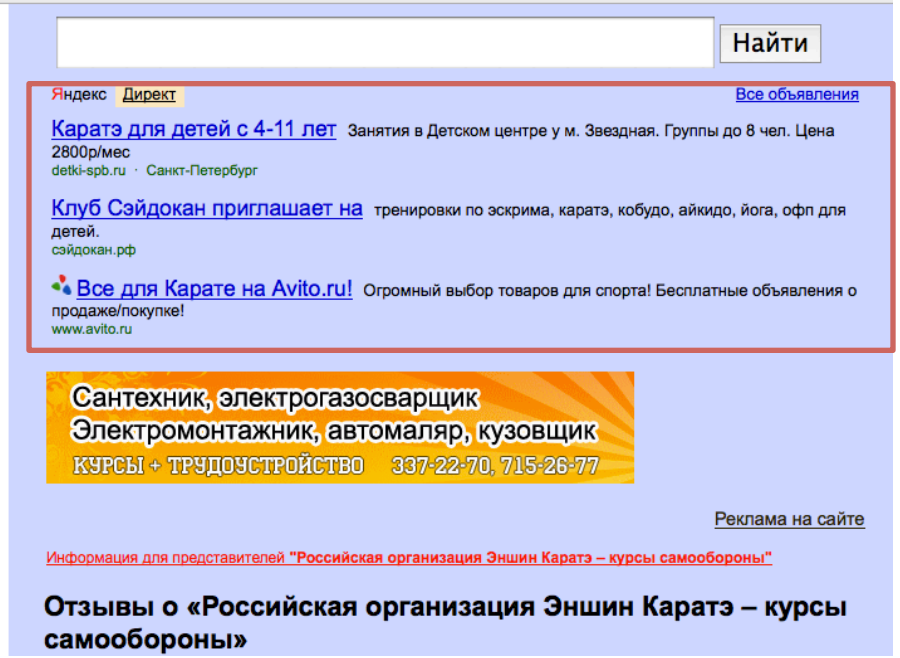
Пример. Content Match

Российская организация Эншин Каратэ – курсы самообороны

Рейтинг: (Еще никто не проголосовал)

Эншин Каратэ это практическая система самозащиты, в основе которой, лежит стратегия циклического перемещения, позволяющая использовать наступательную силу оппонента против него самого.
«Максимальный результат от минимальных физических затрат» – эта стратегия называется «САБАКИ».

Техника Эншин это синтез ударной техники «нокдаун» каратэ и тактико-технических принципов из арсенала дзю-до. Эншин каратэ это рациональное, а главное разумное сочетание ударно-



Оценка

- CPM = cost per thousand impression
  - Обычно используется для баннерной рекламы
- CPC = cost per click
  - Обычно используется для текстовой рекламы
- CPC/CPA = cost per transaction/action

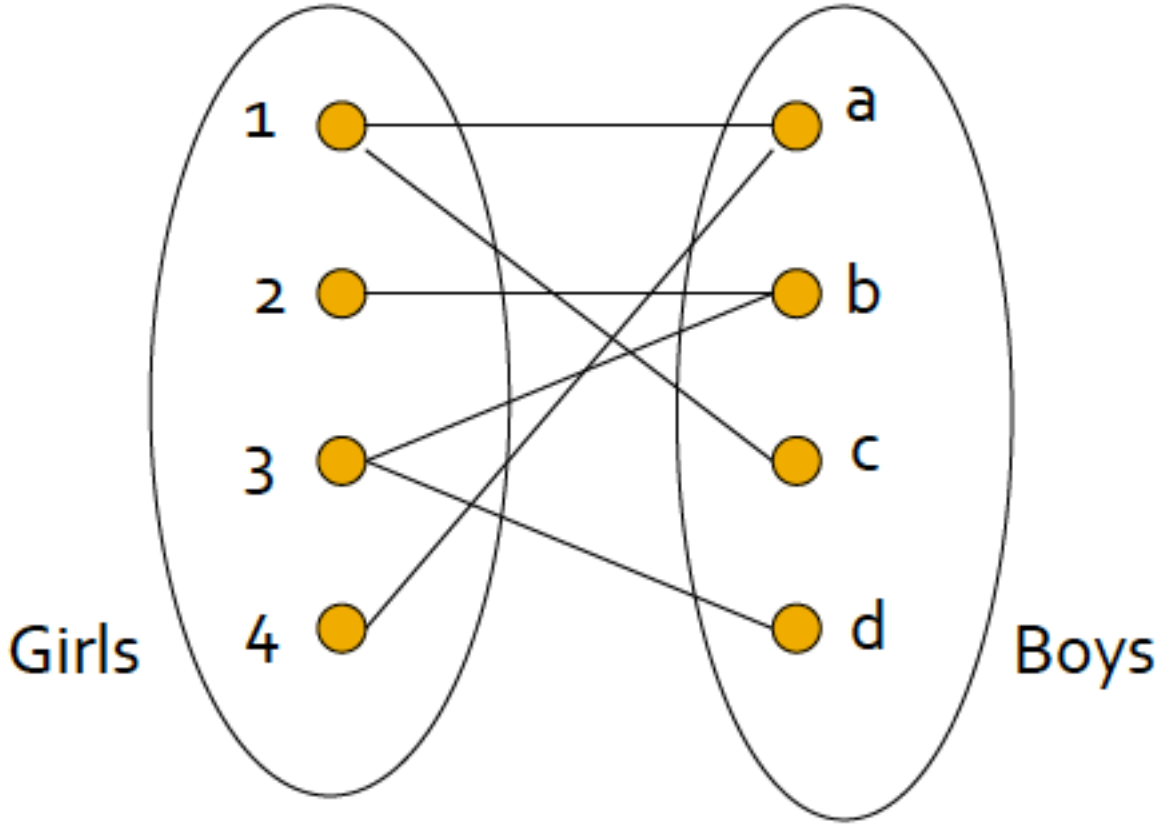
План на сегодня

- Онлайн алгоритмы
- Задача сопоставления
- История рекламы
- Задача adwords

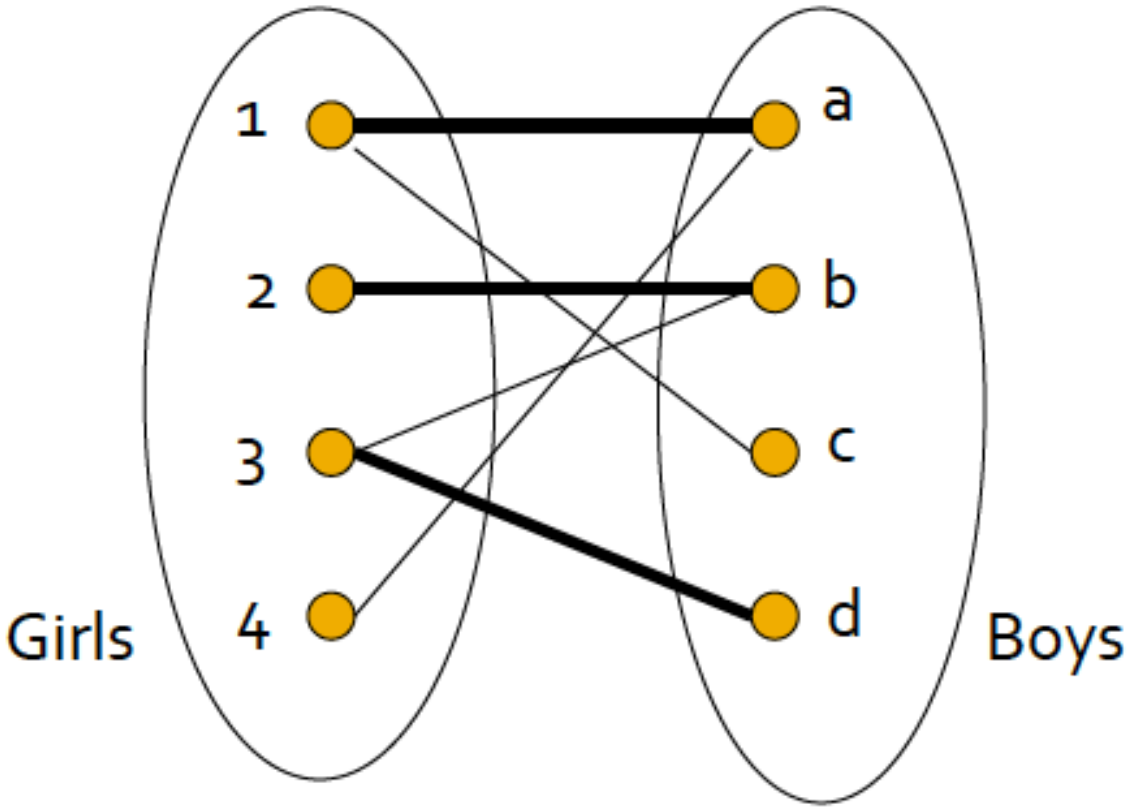
Задача сопоставления (matching problem)

- Пример: определение релевантной рекламы для поисковых запросов
- Можно представить в виде двудольного графа (- граф с двумя наборами узлов: левые и правые)
- Определение: решение задачи сопоставление называется *идеальным*, если все узлы двудольного графа учувствуют в сопоставлении

Пример двудольного графа

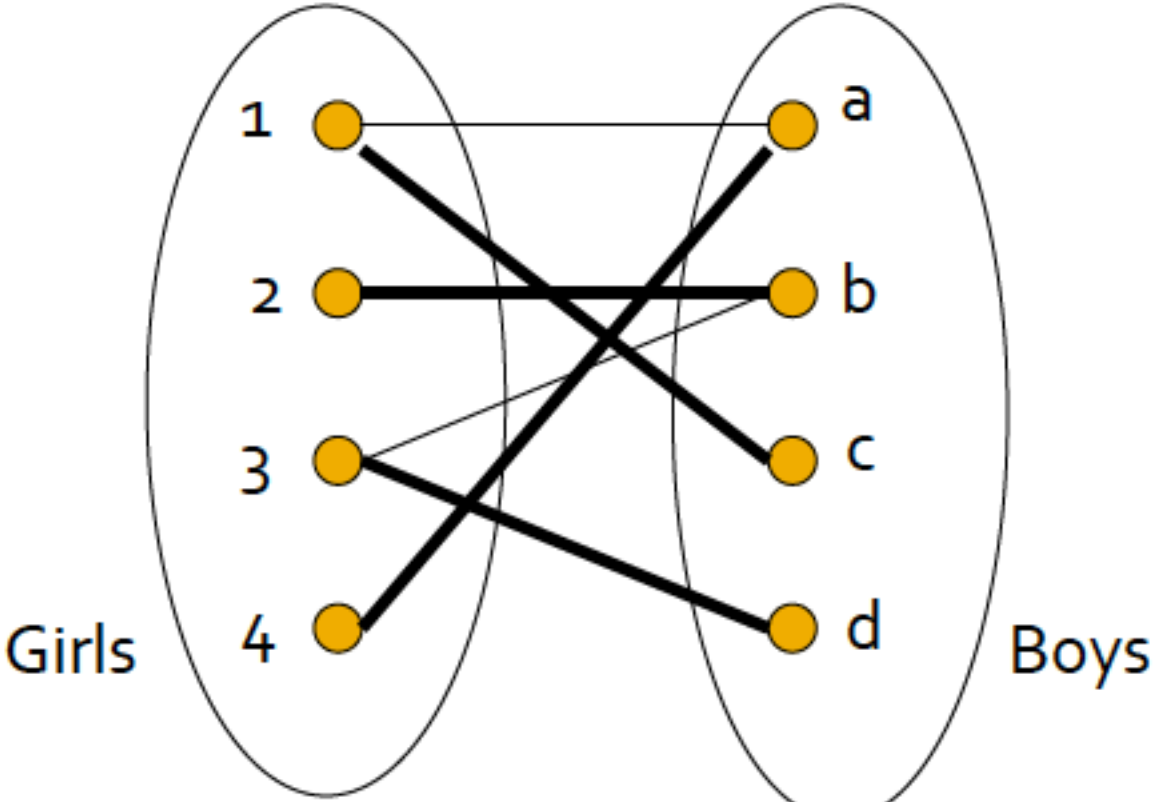


Пример двудольного графа (2)



$M = \{(1,a),(2,b),(3,d)\}$

Матрица переходов A



$$M = \{(1,c),(2,b),(3,d),(4,a)\}$$

Алгоритм сопоставления

- **Задача:** найти максимальное по мощности сопоставление для двудольного графа
  - Найти идеальное если она существует
- Существует оффлайн алгоритм, работающий за полиномиальное время (Hopcroftand Karp 1973)
- Как действовать, если мы не знаем полный граф?

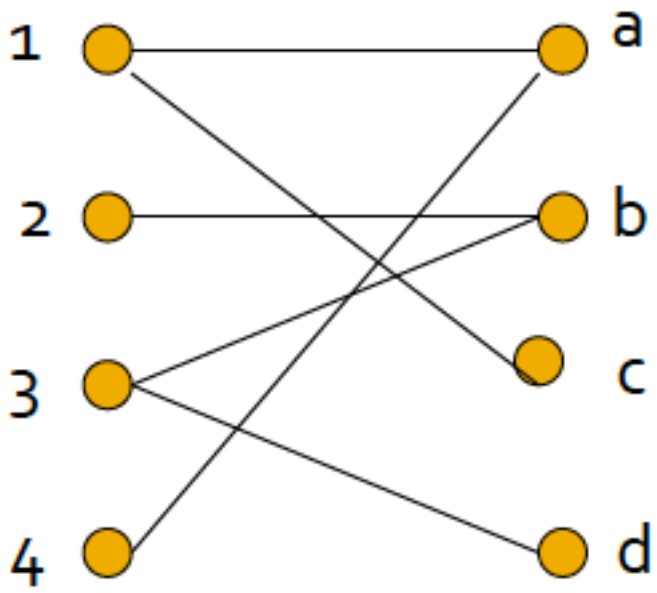
Задача онлайн сопоставления графа

- Изначально мы предполагаем, что есть только мальчики $a = (a_1, a_2, \ldots, a_n)$ и $h = (h_1, h_2, \ldots, h_n)$
- После каждого раунда выявляется выбор каждой девочки
- В этот момент необходимо сделать:
  - Соединить эту пару
  - Отказать

Жадный алгоритм

- Жадный алгоритм для задачи онлайн сопоставления:
  - Составляем пару из девочки и наиболее подходящего мальчика
  - Если такого нет, то оставляем девочку без пары
- Хорош ли данный алгоритм?

Пример



(1,a)
(2,b)
(3,d)

Соотношение для жадного алгоритма

- Для входных данных I, M_{greedy} - это сопоставление, которое получено с помощью жадного алгоритма. M_{opt} - это *оптимальное* сопоставление

$$\text{Competitive ratio} = \min_{\text{all possible inputs } I} (|M_{greedy}|/|M_{opt}|)$$

Анализ жадных алгоритмов

- Рассмотрим множество L – количество сопоставленных левых узлов в M_{opt} , но не в M_{greedy}
- Пусть R – это множество правых узлов, соединенных ребрами с любыми узлами L.

Анализ жадных алгоритмов(2)

1. $|M_o| \leq |M_g| + |L|$, так как среди левых узлов, только L могут быть сопоставлены в M_{opt}, но не в M_{greedy}
2. $|L| \leq |R|$, так как для M_{opt} все узлы в L сопоставлены
3. $|R| \leq |M_g|$
4. (2) и (3) => $|L| \leq |M_g|$
5. (4) и (1) => $|M_o| \leq 2|M_g|$ => $|M_g| \geq \frac{1}{2}|M_o|$

План на сегодня

- Онлайн алгоритмы
- Задача сопоставления
- История рекламы
- Задача adwords

История рекламы

- Баннерная реклама (1995-2001)
  - Типичная форма интернет рекламы
  - Большой риск для рекламодателя (популярные сайты берут x$ за каждые 1000 показов)
- Низкий CRT (Clickthrough rates)
- Не направлено на аудиторию (демографический фактор не присутствует)

История рекламы (2)

- Нововведения в 2000:
  - Рекламодатели устраивают торги (bid) на ключевые слова в поиске
  - Если кто-то ищет ключевое слово, то выигрывает рекламодатель с наибольшей ставкой
  - Рекламодатель сменяется только если на рекламу кликнули
- Подобная модель была адаптирована Google в 2002
  - Была названа adwords

Интересные проблемы

1. Какую рекламу нужно показать для пришедшего поискового запроса?
2. Со стороны рекламодателя: на какие ключевые слова стоит делать ставки, и насколько большие?

План на сегодня

- Онлайн алгоритмы
- Задача сопоставления
- История рекламы
- Задача adwords

Задача adwords

- Поисковые запросы приходят:
  - q1, q2,…
- Несколько рекламодателей поставили на один и тот же запрос
- Когда приходит запрос поисковая машина должна выбрать рекламы, которые будет показывать
- Цель: максимизировать доходы
- Нужен онлайн алгоритм!

Сложности

- Каждый рекламодатель имеет ограниченный бюджет
  - Поисковая компания гарантирует, что рекламодатель не заплатит больше, чем дневной бюджет
- Каждая реклама имеет различную вероятность клика:
  - Рекламодатель 1 ставка $2, вероятность клика = 0.1
  - Рекламодатель 2 ставка $1, вероятность клика = 0.5
  - CTR измеряется исходя из исторических данных
- Простое решение: использовать ожидаемый доход от клика

Adwords

Advertiser	Bid	CTR	Bid * CTR
A	$1.00	1%	1 cent
B	$0.75	2%	1.5 cents
C	$0.50	2.5%	1.125 cents

Жадные алгоритмы

- Входные данные:
  - Только одна реклама может быть показана для одного запроса
  - Все рекламодатели имеют одинаковый бюджет
  - Клики на рекламы равновероятны
- Наиболее простым является жадный алгоритм:
  - Выбрать рекламодателя со ставкой 1

Наиболее плохой сценарий для жадного алгоритма

- 2 рекламодателя: A и B
  - A сделала ставку на ключевое слова X, B сделал ставку на ключевое слово Y и X
  - У обоих бюджет $4
- Пришедший запрос: xxxxyyyy
  - жадный алгоритм: BBBB_ _ _ _
  - Оптимально: AAAABBBB
  - Competitive ration = 1/2

BALANCE алгоритм [MSVV]

- BALANCE by Mehta, Saberi, Vazirani, and Vazirani
- Для каждого запроса выбираем рекламодателя с крупнейшим неизрасходованным бюджетов

Пример. BALANCE

- 2 рекламодателя: A и B
  - A сделала ставку на ключевое слова X, B сделал ставку на ключевое слово Y и X
  - У обоих бюджет $4
- Пришедший запрос: xxxxyyyy
- Результат BALANCE: ABABBB_ _
- Оптимальный: AAAABBBB
- Competitive ration = 3/4

BALANCE: общие результаты

- В общем случае, наиболее плохой результат для competitive ratio ~ 0.63
  - Замечание: ни один другой онлайн алгоритм не имеет лучшего competitive ratio

Общий алгоритм BALANCE

- Запрос q, претендент I
  - Ставка = Xi
  - Бюджет = Bi
  - Уже израсходовано = Mi
  - Доля оставшегося бюджета fi = 1 – Mi/Bi
  - Определяем: $\psi_i(q) = x_i(1-e^{-f_i})$
- Для q => max $\psi_i(q)$
- Competitive ratio = $(1-1/e)$

Резюме

- Узнали про подходы для решения проблем интернет-рекламы
- Поняли задачу сопоставления
- Вспомнили историю интернет рекламы
- И затронули вопрос онлайн-алгоритмов для интернет рекламы
- Нашли оптимальный алгоритм для задачи AdWords