Товарные рекомендации

2017.08.31 @d_key, Дмитрий Колодезев, Промсофт

0. План

- Постановка задачи
- Обзор подходов к решению
- Готовые библиотеки
- Данные и методики оценки
- Велосипеды

1. Постановка задачи

- У клиента возникает потребность.
- Заходит на сайт из контекстной рекламы
- Смотрит товары, добавляет в корзину.
- Ассортимент большой (десятки тысяч)
- Есть блок «Добавьте в корзину»
- Аналог прикассовой зоны супермаркета
- Нужно продать больше, чем сейчас

2. Данные

- 0.5 M покупок.
  - Список просмотренных страниц из 170k
  - Список товаров в корзине из 60k
- Нет негативных «отзывов»
- Зашумлены от рождения
- Confidence vs Preference
- Сложно оценивать

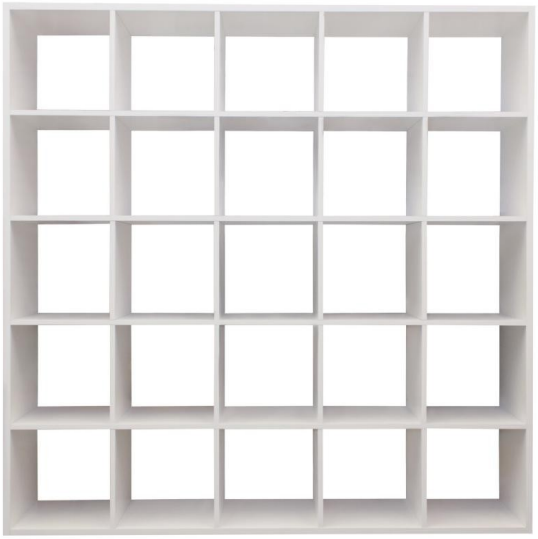
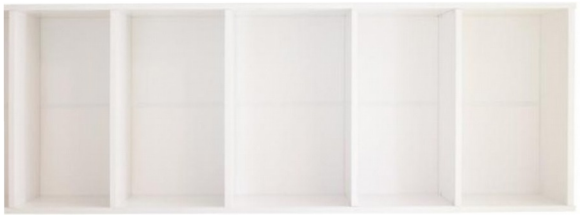
3. Варианты

- Первая покупка (конверсия)
- Помощь в подборе (конверсия, чек)
- Последняя покупка (средний чек)
- 404 страница (конверсия)
- Удержание при уходе (конверсия)
- Много рекомендаторов
- Рассмотрим «прикассовую зону»

4. Модели поведения

- Модель касаний. Ждем, когда откроется
  Никуда не денется.
  Яндекс-директ etc
- Модель встречи. Уточняет ожидания и порог.
  Не встретит — уйдет.
  Встретит — сразу купит.
  Товарные рекомендации

5. Факторизуй и в продакшен


- У нас тут вам не кинофильмы
- Неявные оценки
- Нет негативных оценок
- Короткая история, мало данных о клиенте
- Интеншен (номинализация, жаргон)

6. Предположения

- Клиент похож на других клиентов
- По просмотрам, демографии, корзине
- Давайте продадим ему то, что купили другие
- Нет. Меняется интеншен
- Люди меняются
- Один заказчик — несколько человек
- Один визит — один интеншен (с натяжкой)

7. Предположение -2

- Товар похож на другие товары
- Эти отношения устойчивы
- Предложим такое же
- Нет. Это не проблема прикассовой зоны
- Если купили шурупы 4*50, рекомендуем шурупы 4*40 4*45 4*55 4*60
- Он это все видел, когда выбирал
- Уголь + шашлык + пиво, а не 3 вида угля

8. Предположение -3

- Интеншен — спросить или вычислить
- «Помощник в подборе»
- По поведению
- HMM — в планах
- LDA — работает, но не в планах (скорость)
- Товары устаревают (айфон2 больше не шик)
- Интеншены меняются и заменяются

9. Предположение -4

- Товары образуют устойчивые группы
- Дополняющие товары, руками
- Ассоциативные правила
- Apriori
- Не ловит негативное влияние
- Не ловит длинный хвост
- Сложно считать и использовать
- Префиксное дерево

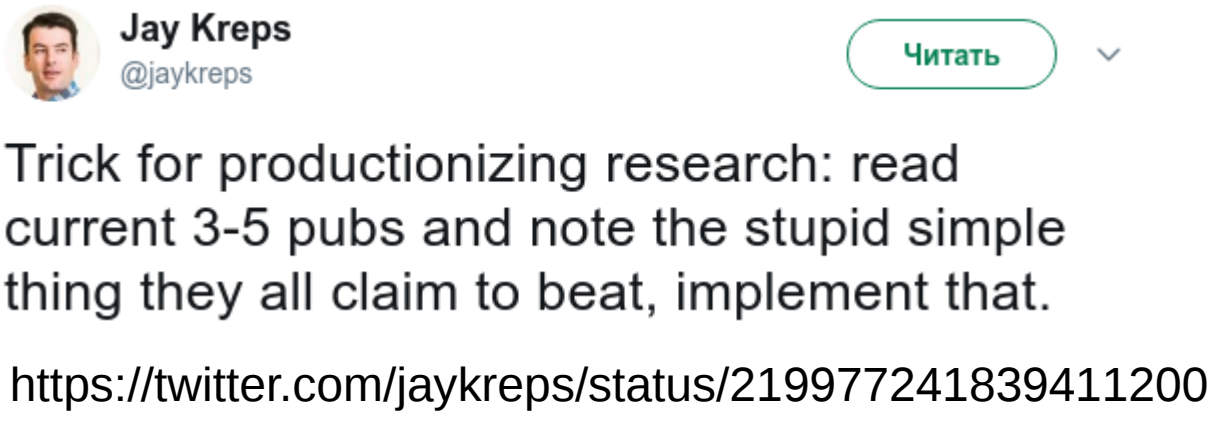
10. Предположение -5

- Наивный, наивный Б.
- Товары в корзине и просмотры страниц
- Предположение о независимости (неверное)
- Упрощаем цепочное правило
- Произведения попарной встречаемости
- «правдоподобие»

$$\mathrm{P}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right)$$

11. И в продакшен!

- Внезапно хорошо работает
- Интерпретируемый



12. А как же?

- Многорукие бандиты слишком многорукие
- Факторизация украсит любой интерьер
  - https://github.com/benfred/implicit
  - http://yifanhu.net/PUB/cf.pdf
- Строй классификаторов (если товаров мало)

13. Велосипед детектед

- Surprise Python scikit
- PredictionIO Spark + Scala
  - Personalized recommendations—user-based
  - Similar items—item-based
  - Viewed this bought that—item-based cross-action
  - Popular Items and User-defined ranking
  - Item-set recommendations for complimentarty purchases or shopping carts—item-set-based
  - Hybrid collaborative filtering and content based recommendations—limited content-based

14. Кстати, иерархия

- Узлы товарного дерева как фичи
- Холодный старт для товаров

15. Оценка

- Нет негативных примеров
- На продакшене — A/B тест
- В разработке — CV на 5 фолдах
  - Перемешиваем товары
  - Убираем левый товар и его страницу
  - По оставшимся предсказываем
  - Смотрим, попал ли это товар в ТОП-5
  - Baseline — 5 самых популярных, 2%
  - Есть куда расти

16. Типа формула

- Строим матрицу M признаки-товары
- Признаки — товары + страницы
- Значение ячейки — P(C|R)
- Произведение = «правдоподобие»
- Ранжируем, выкидываем те, что уже есть
- Profit

17. Детали

- Считаем попарную частоту Т-П
- Нулевую частоту смягчаем (+0.1)
- Делим на частоту товара
- Логарифмируем
- Умножаем на вектор признаков

18. Верните мне мою память

- 60k * 170k * 4 = 41Gb плотная матрица
- Не все товары в продаже, выкидываем часть
- Частоты переносим, или нет.
- Умножение на константу не меняет ранг
- Умножим все частоты на 10.
- log(M*10/F*10) = log(M*10) — log(F*10)
- Log(0.1*10) в большинстве ячеек
- Разреженная матрица и вектор частот, < 1Gb

19. И это работает?

- Ранее использовали Apriori
- Переход на ЭТО дал увеличение размера корзины
- Для новых работает лучше, чем для старых

20. Холодный старт

- У клиента есть хотя бы одна страница
- Товару — бонус новичка
  - Наиболее похожие товары
  - По времени или по продажам отменяем

21. Дальше-то что?

- Категориальные признаки
  - Город
  - Источник трафика
  - История заказов
- Устаревание корзин (вес от времени)

22. Вопросы?

dk.promsoft@mail.ru