Метод автоматического определения возраста пользователей с помощью социальных связей

Андрей Гомзин

Институт Системного программирования РАН
Факультет ВМК МГУ

Содержание

Введение. Социо-демографический профиль

Социо-демографические характеристики пользователей: пол, возраст, семейное положение, уровень образования.

- Не все поля заполняются пользователями
- Указываются неверные значения

атрибуты не указаны *ошибки* *дубликаты профилей*

В данной работе рассматривается атрибут "Возраст", решается задача предсказания неуказанных значений.

Введение. Целевая аудитория

Целевая аудитория – группа людей, объединённых общими признаками, или объединённой ради какой-либо цели или задачи[1]

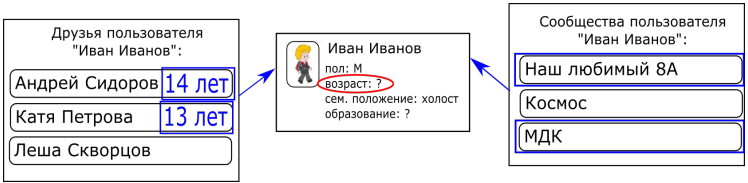
Явно указанные и извлеченные социо-демографические атрибуты используются в рекомендательных и маркетинговых системах для:

- определения целевой аудитории продукта
- поиска потенциальных потребителей

[1] Википедия: https://ru.wikipedia.org/wiki/Целевая_группа

Постановка задачи

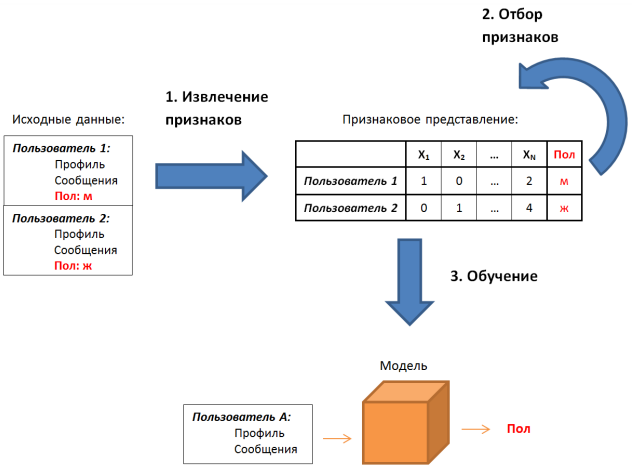
Дано: социальная сеть (социальный граф и частично указанные значения возраста)
Найти: неуказанные значения **возраста** пользователей



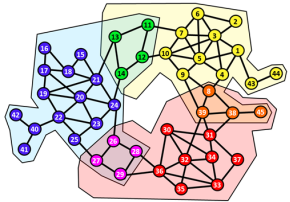
Анализируемые данные:

- отношение дружбы
- подписка на сообщества
- ...

Обзор. Машинное обучение с учителем



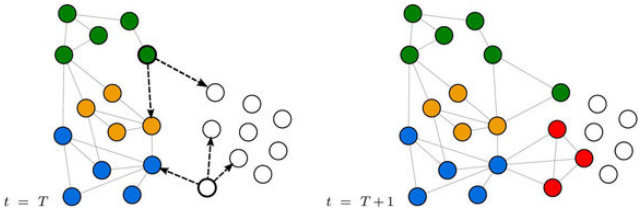
Признаки – найденные (нечеткие) кластеры:



Обзор. Кластеризация социального графа

Распространение меток в графе:

1. инициализация узлов метками
2. несколько итераций распространения меток соседям:
   - Задается стратегия отправки меток
   - Задается стратегия приема меток



Определение демографических атрибутов с помощью кластеризации:
Кластер объединяет в себя пользователей с одинаковым набором атрибутов.
Для этого инициализируем метки согласно известным значениям атрибута (например, "М" и "Ж")

Предлагаемый подход. Социальный граф

Узлы:

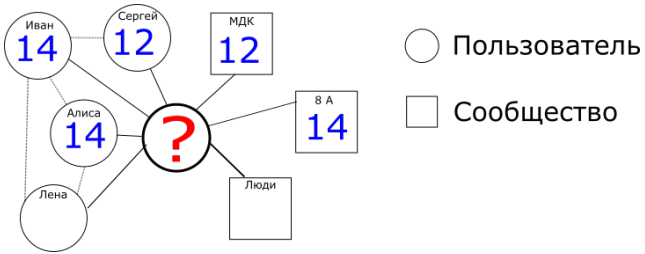
- пользователи
- сообщества

Ребра:

- отношение дружбы
- подписка на сообщества

Метки:

- значения возраста

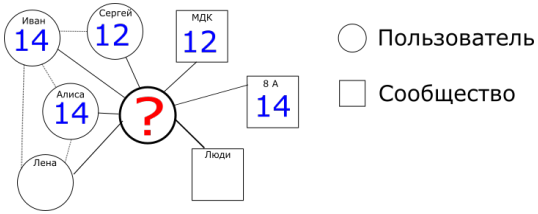


Возраст определяется путем распространение меток в социальном графе.

Алгоритм определения значений атрибутов

Схема алгоритма:

1. Инициализация (разметка)
2. Построение векторной модели
3. Вычисление весов пользователей и сообществ, распространение меток на узлы-сообщества
4. Построение векторной модели с учетом весов
5. Распространение меток на узлы-пользователей с учетом весов



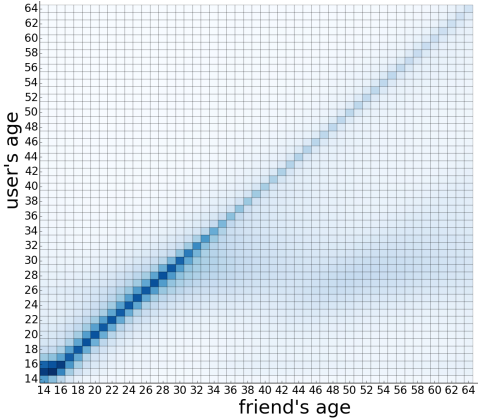
Разметка (Вконтакте)

- Возраст пользователя извлекается из даты его рождения. Поле «дата рождения» может быть представлено в трех вариантах:
  1. **DD-MM-YYYY** - доступна полная дата
  2. **YYYY** - доступен год рождения
  3. DD-MM - доступна дата без года

Определение возраста. "Векторная модель"

Распределения значений возраста соседей узла, сгруппированные по возрасту узла:

Распространение к узлам-**пользователям**:



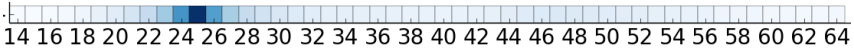
Распространение к узлам-**сообществам**:

$$p(val_n|val_c) = \begin{cases} 1 & \text{если } val_n = val_c \\ 0 & \text{если } val_n \neq val_c \end{cases} \quad (1)$$

Здесь $p(val_n|val_c)$ — вероятность того, что значения атрибута соседа равно val_n, при условии что свое значение атрибута равно val_c.

Вычисление весов, распространение меток на сообщества (1)

1. Строится распределение значений возраста у соседей *Distr*



2. для каждого значения возраста вычисляется близость *Distr* к соответствующему распределению из векторной модели:

$$L = \arg\max_{val}(sim(Model_*(val), Distr))$$

$$S = \max_{val}(sim(Model_*(val), Distr)) = sim(Model_*(L), Distr)$$

где:

$$sim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Вычисление весов сообществ, распространение меток на сообщества (2)

3. Значения S используются для определения весов узлов $W(node)$. Веса вычисляются отдельно для каждого типа узла. Вычисленные значения S сортируются по возрастанию и помещаются в массив. Затем вес узла определяется по формуле:

$$W(node) = \left(\frac{pos(S_{node})}{N} \right)^2$$

Здесь $pos(S_{node})$ — порядковый номер значения S (от 1 до N) в отсортированном массиве, N — количество узлов, для которых вычислено значение S.
Для узлов, у которых не вычислено значение S вес равен 0.

Распространение меток на пользователей с учетом весов (1)

- Параметры:
  Для каждого типа соседей (пользователи, сообщества) вручную задается вес, задающий вклад каждого источника данных (W_{User}, W_{Comm}, соответственно).
- Распространение меток к узлам-пользователям:
  - от пользователей (дружба)
  - от сообществ (подписка)
- Учет весов $W(node)$, полученных на предыдущем шаге:
  вклад каждого соседа пользователя в распределение $Distr$ равен его весу.

Распространение меток на пользователей с учетом весов (2)

1. Распределение значений атрибута соседей вычисляется отдельно для каждого типа узлов (пользователь, сообщество)
2. Распределение домножается на соответствующий вес (W_{User}, W_{Comm})
3. Полученная сумма распределений нормализуется
4. Вычисляется значение атрибута с использованием векторной модели

После распространения меток незаполненные атрибуты заполняются в соответствие с распространенными метками.

Эксперименты. Данные и метрики

Социальная сеть: Вконтакте.
Данные:

- 300M профилей
- 300M связей типа "дружба"
- 1M связей типа "подписка на сообщество" (1M активных сообществ)

Метрики:

- MAE – средняя абсолютная ошибка:

$$\frac{\sum |act - pred|}{N}$$

- Точность для атрибута "Возраст". Возраст определен точно, если:

$$|act - pred| < 0.15act$$

Определение возраста. Результаты

Скользящий контроль 10-fold

Значения весов	Метрика	Значение
$W_{User} = 1,\ W_{Comm} = 1$	точность	81,3 %
	MAE	2,79 года
$W_{User} = 1,\ W_{Comm} = 10$	точность	77,6 %
	MAE	3,28 года
$W_{User} = 10,\ W_{Comm} = 1$	точность	81,1 %
	MAE	2,81 года

Вывод: наибольший вклад в качество определения возраста приносит граф друзей.

Заключение. Планы

- Заключение
  - Социо-демографические атрибуты используются в системах маркетинга и рекомендаций
  - Метод предсказания возраста с использованием социального графа (друзья и подписки)
  - Для атрибута возраст наилучшее качество достигается при использовании графа дружбы
- Планы
  - Добавление новых типов узлов в граф (лайки, репосты, текстовые признаки, ...)
  - Другие атрибуты (семейное положение, уровень образования, ...)
  - **Сравнение с методами, использующими разнородные данные**
  - Определение неверно указанных значений

Спасибо за внимание

Обзор. Векторное представление вершин графа

- Каждая вершина v графа $G(V, E)$ отображается в R^n
- Близость между узлами графа сохраняется в R^n.