Чем старше человек,
тем больше ему лет

Методы построения социо-демографических профилей пользователей сети Интернет

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*

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

Введение. Гомофилия и поведение

Для определения демографических атрибутов используются поведенческие признаки и свойство гомофилии.

Гомофилия – тенденция индивидов создавать и поддерживать связи с похожими на них (is the tendency of individuals to associate and bond with similar others).
Пример:

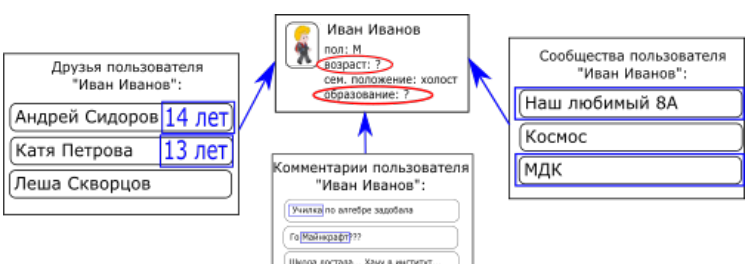
- Отношение дружбы в соцсетях

Поведенческие признаки:

- Сообщения
- Подписка на сообщества
- оценки (нравится / не нравится)

Постановка задачи (1)

Дано: социальная сеть
Найти: неуказанные значения демографических атрибутов пользователей

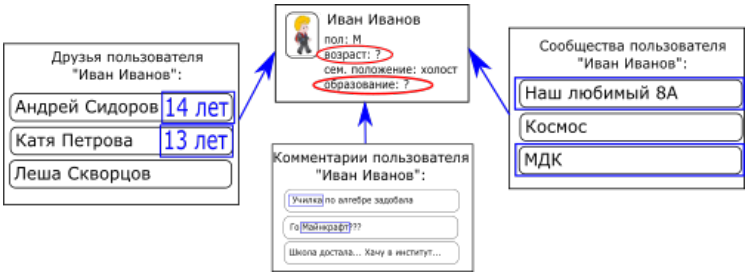


Демографические атрибуты:

- пол
- возраст
- семейное положение
- уровень образования
- ...

Постановка задачи (2)

Дано: социальная сеть
Найти: неуказанные значения демографических атрибутов пользователей



Анализируемые данные:

- отношение дружбы
- подписка на сообщества
- тексты сообщений
- значения атрибутов
- ...

Краткий обзор существующих решений

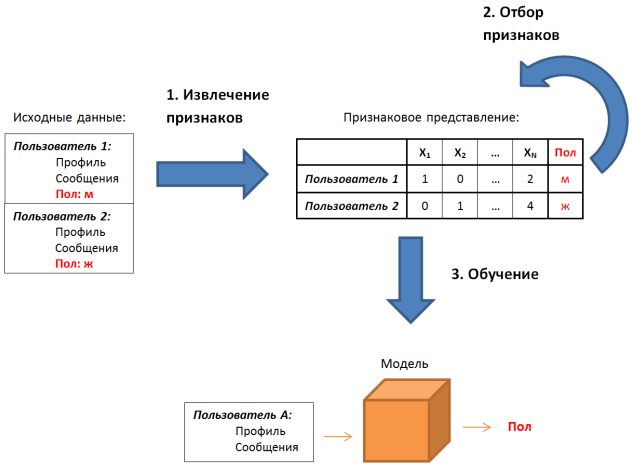
Данные:

- Twitter
- Facebook
- Youtube
- e-mail

Подходы:

- Эвристики (пол по имени)
- Однородные даные -> Машинное обучение с учителем
- Графы -> Кластеризация графов, векторное представление вершин
- Подходы, использующие разнородные данные

Машинное обучение с учителем



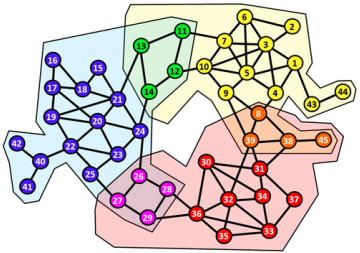
Машинное обучение с учителем. Признаки

Текстовые признаки:

- *n*-граммы:
  слова, символы, части речи
- статистические
  - средняя длина сообщения
  - LIWC – Linguistic Inquiry and Word Count:
    - доля эмоций (+/-)
    - доля местоимений
    - доля длинных слов ($|\cdot| > 6$)
- ресурсозависимые:
  - #хештеги
  - @упоминания

Графовые признаки:

- найденные (нечеткие) кластеры:



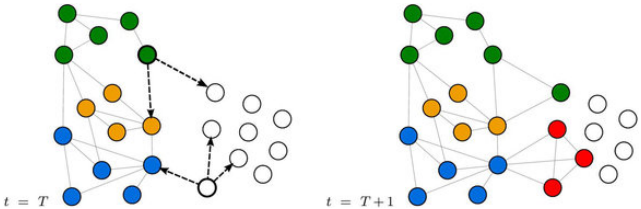
Прочие ресурсозависимые признаки:

- Цвет фона/текста/ссылок (Twitter)

Кластеризация социального графа

Распространение меток в графе:

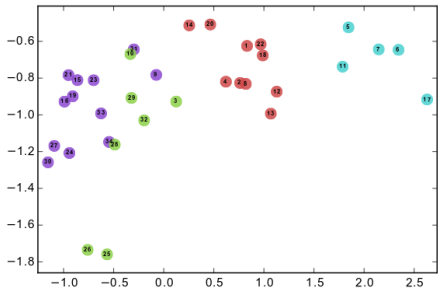
1. инициализация узлов метками
2. несколько итераций распространения меток соседям:
   - Задается стратегия отправки меток
   - Задается стратегия приема меток



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")

Векторное представление вершин графа

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

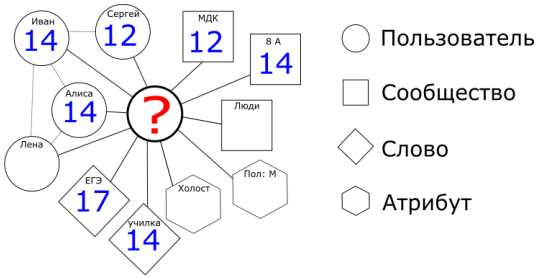


Разнородные данные

- Сообщества в графе + текстовые признаки -> машинное обучение
- Тексты сообщений + тексты сообщений друзей -> машинное обучение
- $\min ||XW - Y|| + ||W|| + E$ Сообщество объединяет пользователей с одинаковыми значениями атрибута

Предлагаемый подход

- Построение социо-лингвистического графа
- Распространение меток



Социо-лингвистический граф

Узлы:

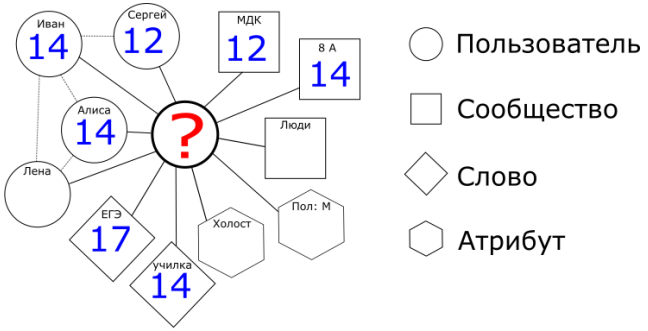
- пользователи
- сообщества
- текстовые признаки (слова)
- значения атрибутов

Ребра:

- отношение дружбы
- подписка на сообщества
- употребление слов
- явно указанные значения

Метки:

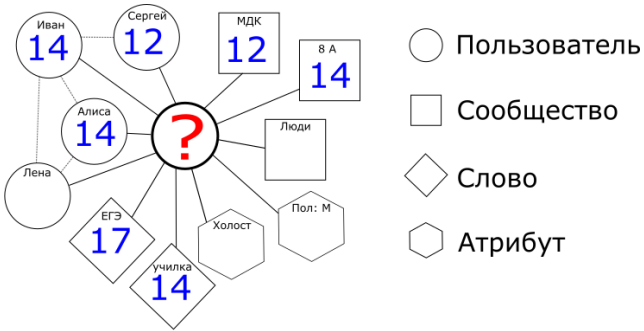
- значения атрибута



Алгоритм определнения значений атрибутов

Схема алгоритма:

1. Инициализация (разметка)
2. Распространение меток от узлов-пользователей к другим узлам (сообщества, слова, атрибуты)
3. Распространение меток от всех узлов к узлам-пользователям



Разметка (Вконтакте). Возраст, семейное положение

- Возраст
  - 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
- Семейное положение
  - Поле “семейное положение” пользователя Вконтакте принимает следующие значения:
    - **1 — не женат/не замужем; => «не в официальном браке»**
    - **2 — есть друг/есть подруга; => «не в официальном браке»**
    - 3 — помолвлен/помолвлена;
    - **4 — женат/замужем; => «в официальном браке»**
    - 5 — всё сложно;
    - **6 — в активном поиске; => «не в официальном браке»**
    - 7 — влюблён/влюблена;
    - 0 — не указано.

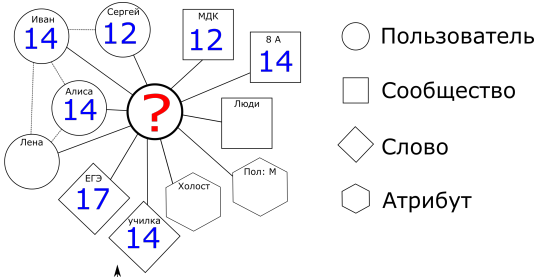
Разметка (Вконтакте). Уровень образования

- Уровень образования
  - Максимальный год окончания школы => среднее образование
  - Минимальный год окончания университета => высшее образование
  - Атрибут «Образование» по таблице:

		окончание вуза		
		неизвестно	отсутствует	имеется
окончание средней школы	неизвестно	неизвестно	среднее	высшее
	отсутствует	отсутствует	отсутствует	неизвестно
	имеется	среднее	среднее	высшее

Стратегия приема меток. Идея

1. Вычисляем распределение значений атрибута соседей.
2. Зашумляем его. Это фильтрует узлы-признаки с малым количеством соседей (например, редкие слова, которые используются 2-3 пользователями)
3. Вычисляем, насколько в распределении одно из значений преобладает; если сильно преобладает, то присваиваем метку узлу, если нет — то узел остается без метки.



Стратегия приема меток (1)

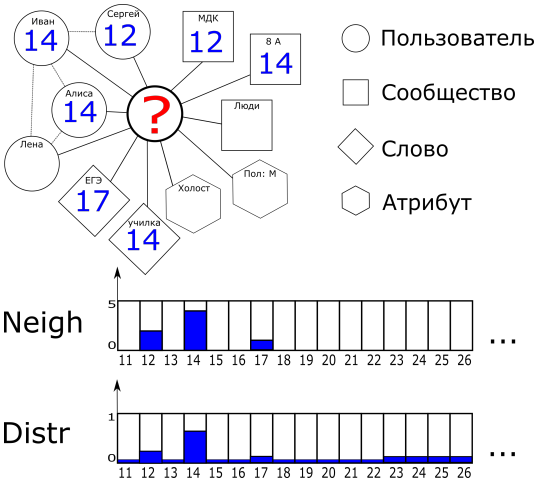
Gen – распределение значений атрибута на всех данных. $||Gen|| = 1$

1. *Neigh* – распределение значений атрибута соседей. $Neigh(v)$ – количество соседей
2.

$$Distr(v) = \frac{Neigh(v) + w \times Gen(v)}{N}$$

где N подбирается таким образом, что:

$$|Distr| = \sqrt{\sum_{v \in V} Distr^2(v)} = 1$$



Стратегия приема меток (2)

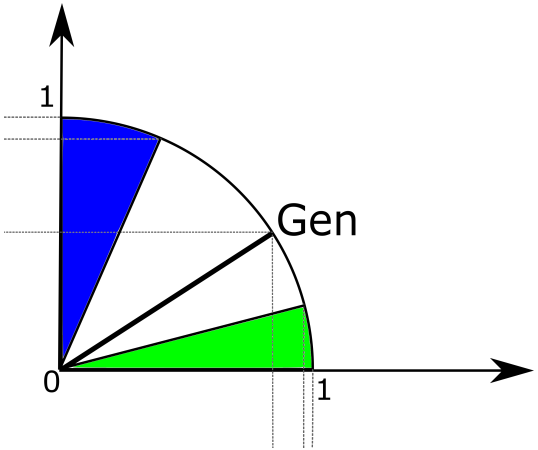
Gen – распределение значений атрибута на всех данных. $||Gen|| = 1$

3. Вычисляем метку *val* и вес *S*:

$$s = \max_v \frac{Distr(v) - Gen(v)}{1 - Gen(v)}$$

$$val = \arg\max_v \frac{Distr(v) - Gen(v)}{1 - Gen(v)}$$

4. Узел принимает соответствующую метку, если $s > t$.



Эксперименты. Данные

Социальная сеть: Вконтакте.
Данные:

- 50M профилей
- 50M связей типа "дружба"
- 1M связей типа "подписка на сообщество"
- тексты комментариев за 0.5 года из 1M активных сообществ

Количество пользователей с указанными атрибутами:

- Пол: 34521070
- Возраст: 10183309
- Семейное положение: 4720535
- Образование: 8818867

Эксперименты. Гомофилия vs Поведение

- Измерялись значения достоверности (accuracy) и F1-меры
- Точность для атрибута "Возраст". Возраст определен точно, если:

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

- Эксперименты:
  - Только узлы-**пользователи**:
    (>20 друзей, хотя бы у одного указан атрибут)
  - Только узлы-**сообщества**:
    (>20 сообществ)
  - Только узлы-**слова**:
    (>20 слов)
  - **Пользователи, сообщества, слова**:
    (пересечение всех предыдущих)
- Параметры:
  - Распространение от пользователей к другим узлам: $T = 0.8, W = 2.0$
  - Распространение к пользователям (предсказание): $T = 0.0, W = 0.0$

Эксперименты. Гомофилия vs Поведение. Возраст

	Пользователи	Сообщества	Слова	Все
MAE	1.66	1.13	5.4	
Acc (15%)	0.90	0.91	0.53	
Доля пользователей	0.77	0.01	<0.01	

Эксперименты. Гомофилия vs Поведение. Семейное положение

	Пользователи	Сообщества	Слова	Все
Acc	0.70	0.72	0.64	
F1macro	0.64	0.66	0.63	
Доля пользователей	0.42	0.20	<0.01	

Эксперименты. Гомофилия vs Поведение. Образование

	Пользователи	Сообщества	Слова	Все
Acc	0.59	0.63	0.47	
F1macro	0.47	0.44	0.36	
Доля пользователей	0.31	0.07	<0.01	

Эксперименты. Планы

- **Сравнение с методами, использующими разнородные данные**
- Национальность, доход, дети, языки
- Лайки, репосты
- Стены пользователей и комментарии (вместо комментариев в группах)
- Определение неверно указанных значений

Спасибо за внимание

Заключение

- Гомофилия, поведение, социо-демографические профили пользователей
- Метод построения социо-демографического профиля
- Выводы:
  - Качество – пойдет
  - Метод шустрый и масштабируемый
  - Гомофилия – возраст, поведение – везде
  - Нужно фильтровать данные

Эксперименты. Сравнение

Эксперименты:

- SVM классификатор. Текстовые признаки (юниграммы)
- Распространение меток. Только текстовые признаки
- main class

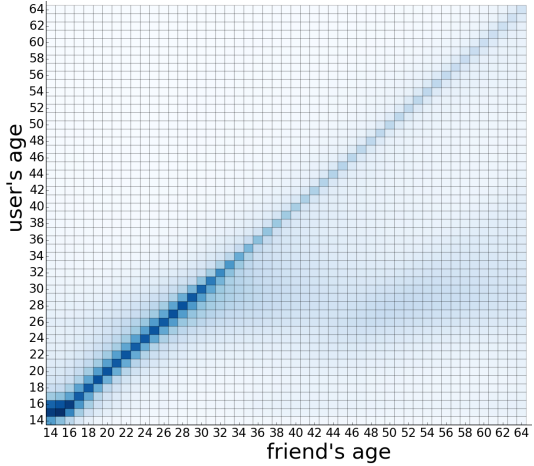
Результаты CrossVal 10 fold:

- Все примерно одинаково ($\pm 2\%$)

Вывод:

- Тексты комментариев – плохой источник

Определение возраста. "Векторная модель"



Определение возраста. Результаты

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

"Векторная модель". Резульаты на всех данных для vk

Из отчета одного проекта:

- возраст: точность = **75,1** %
- семейное положение: F1-мера = **76,3** %
- образование: F1-мера = 49,9 %
  - среднее образование: F1-мера = 49,0 %, достоверность: 60,0 %
  - высшее образование: F1-мера = 45,3 %, достоверность: 84,3 %

Из автореферата (1)

Целью данной работы является разработка подхода к построению социо-демографических профилей пользователей сети Интернет с использованием модели социо-лингвистического графа. Разработанный подход должен обладать следующими свойствами:

- Для определения атрибутов должны использоваться разнородные данные: тексты публичных сообщений, атрибуты профиля, социальные связи, при этом для определения значений атрибутов пользователя должно быть достаточно наличия хотя бы одного из перечисленных типов данных;
- Характеристики качества должны соответствовать соответствующим характеристикам методов, представленных в современной литературе;
- Время работы должно линейно зависеть от количества связей в социо-лингвистическом графе, решение должно быть масштабируемо;

Из автореферата (2)

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Исследовать существующие методы определения демографических характеристик пользователей;
2. Разработать модель социо-лингвистического графа, объединяющая в себя как социальные, так и лингвистические сущности и связи;
3. Разработать метод, определяющий значения демографических атрибутов, обладающий описанными выше свойствами;
4. Разработать прототип системы построения социо-демографических профилей пользователей;
5. Провести экспериментальное сравнение разработанного метода с аналогичными методами по точности и производительности.

Из автореферата (3)

Новизна

Разработанная модель социо-лингвистического графа объединяет в себя различные виды сущностей и позволяет определять значения демографических атрибутов пользователя даже при наличии лишь частичного набора связанных с ним сущностей. Математически доказана оценка вычислительной сложности разработанного метода. Экспериментально показано, что качество разработанного метода сответствует качеству аналогичных методов, описанных в современной литературе.

Разработанная модель универсальна. Социо-лингвистический граф может быть расширен новыми видами данных (сущностей и связей между ними). Кроме того, модель может применяться не только для определения значений демографических атрибутов, но и для решения других задач классификации пользователей, сообщений или других сущностей.