Автоматическое определение возрастной принадлежности пользователя в социальных сетях

Аннотация. В работе проводилось исследование методов определения возраста пользователей социальных сетей. При регистрации предполагается заполнение его профиля. Анкета содержит множество полей, к которым относится и возрастная принадлежность. Зачастую не указываются точные данные или профиль заполняется не полностью. В этой связи возникает необходимость уточнения неизвестных характеристик.

Точно предоставленный и установленный методологическим путем возраст интересен рекомендательным и маркетинговым системам. С помощью этой информации удается отбирать целевую аудиторию продвигаемых товаров и услуг. Помимо этого, такие данные могут быть применены для более точного выявления демографического профиля интернет-сообществ. Также появляется возможность определить целевую аудиторию для рекламных кампаний, проводимых в Интернете.

В статье рассматривается метод, который предсказывает неизвестные значения возрастного атрибута. Он использует заполненные поля профиля – указанный возраст и сведения о социальных связях, предполагая распространение меток по числу друзей и подписок участников социальной сети.

Ключевые слова: интернет сообщества, возраст пользователя, профиль пользователя, возрастная принадлежность, демографические характеристики, целевая аудитория

Введение

Профили (страницы) участников социальных сетей обычно состоят из следующих демографических атрибутов: возраст, пол, семейное положение, образование, политические, религиозные взгляды и прочее. Данную информацию используют в рекомендательных и маркетинговых системах. Значения демографических характеристик позволяют отобрать целевую аудиторию, соответствующую продвигаемым товарам и услугам.

Пользователи по определённым причинам заполняют далеко не все атрибуты. Достаточно часто участники социальных сетей указывают недостоверные сведения. Представленная работа посвящена алгоритмам, с помощью которых возможно предсказать незаполненные или неточные возрастные показатели. Данный метод позволяет установить искомые значения посредством анализа социальных связей участника и точного возраста, указанного другими пользователями сети. Исследование проводилось на базе открытой информации социальной сети Вконтакте. Использовались следующие данные: профили участников, их подписки на сообщества, а также профили их друзей. Способы поиска недостоверно указанных атрибутов нами не рассматривались.

Предложенный в статье метод выявления демографических характеристик базируется на применении социального графа. При этом, участники социальной сети и сообщества – это узлы графа, а взаимоотношения между юзерами (их дружба), подписки на паблики (сообщества) – это ребра. Сообщество – это некая страница в социальной сети, которая объединяет ее участников по интересам: пользователи подписываются на интересующие их паблики, чтобы регулярно получать релевантную информацию. Значения искомых признаков предугадываются путем распределения меток в данном графе. В нашем случае, метки – это значения возраста.

Для начала рассмотрим имеющиеся на данный момент методы решения задачи выявления демографических признаков и смежных задач. Далее приведем разработанный нами способ. В конце работы отражены результаты экспериментального исследования рекомендуемого метода.

Возможные решения

В данном разделе статьи приводится краткий обзор решений задач выявления демографических признаков участников социальных сетей. Для исследователей данного вопроса наибольший интерес представляют такие социальные сети, как Facebook и Twitter. Помимо этих ресурсов, в некоторых работах для анализа применяются комментарии на Youtube [5], а также новостные сообщения и электронные письма [2].

Среди ученых при установлении значений демографических атрибутов наиболее популярен подход, который предполагает выявление признаков из текстовых сообщений пользователей посредством метода машинного обучения. Для начала опишем признаки, применяемые авторами, далее причислим основные алгоритмы.

Определяя пол участников Youtube, применяют способ распространения пола в графе пользователи-видео, где ребро между видео и участником свидетельствует о факте просмотра видеоматериала. Далее в качестве характеристик изучаются следующие статистические признаки: возраст, словесные n-граммы, средняя длина комментария в символах/словах/предложениях и распределение пола, которое получено на основе алгоритма распространения атрибута «пол» в графе пользователи-видео [5].

По текстам сообщений пользователей социальной сети Twitter проводится определение пола участников посредством символьных и словесных n-грамм [1]. Также проводилось исследование на базе пользователей, общающихся письменно на голландском языке [8]. При этом возраст участников разделен на интервалы. За признаки применяются символьные и словесные 1,2 и 3-граммы. Кроме решений, в которых атрибуты возраста участников подразделяются на интервалы, есть алгоритмы, которые определяют числовое значение возраста [7].

Определяются политические взгляды участников Twitter [3]. К рассмотрению принимается 3 группы: республиканцы, демократы и неопределенная политическая позиция. Признаками выступают словесные юниграммы, хэштеги, сообщества, которые получены посредством приема, базирующегося на распространении меток в социальном графе участников.

Наиболее простым методом считается Наивный байесовский классификатор [1]. При распределении признаков на 2 группы зачастую применяют линейный классификатор. Одним из широко используемых алгоритмов обучения линейного классификатора является способ опорных векторов [1], [2], [7], [8]. В научных статьях также можно встретить решающие деревья и логистическую регрессию [2]. Для выявления числового параметра возраста применяется линейная регрессия [7]. Имеются также более основательный обзор способов установления демографических атрибутов участников социальных сетей, предполагающих анализ текстовых сообщений пользователей [10].

Для установления демографических атрибутов помимо текстовых данных используются также социальные связи, например, университетская сеть [6]. Атрибуты устанавливаются посредством алгоритма кластеризации социального графа способом распространения меток. В качестве источника информации может выступать мобильная социальная сеть, в которой связи между участниками формируются на основе звонков и коротких текстовых сообщений между пользователями [4].

Могут применяться сразу два вида данных: текстовых сообщений и социальных связей пользователей [9]. Устанавливается тональность сообщений участников Twitter путем построения графа, состоящего из профилей пользователей, сообщений, слов, эмотиконов, а далее применяется способ распространения меток в полученном графе. Теперь переходим к рассмотрению метода определения возраста участников путем распространения меток в графе, состоящего из профилей, сообществ и взаимосвязей между ними.

Содержание методики

Для реализации метода выявления демографических атрибутов требуются такие сведения:

- страницы участников социальных сетей (профили), которые содержат соответствующие значения демографических характеристик;
- социальные связи (перечень друзей участников или подписчиков сообществ).

В первую очередь отбираются значения возраста, что называется разметкой. Далее для всех участников неуказанные возрастные значения устанавливаются на базе социальных связей. Для начала рассмотрим описание краулера – сборщика информации. После этого выполним описание разметки атрибутов, далее – алгоритма выявления неуказанных значений атрибутов.

Информация была взята из социальной сети ВКонтакте. Аккумулирование данных осуществлялось методом VK API для разработчиков приложений. Важно, что выборка имеет отношение ко всем пользователям, но не ко всем сообществам. При сборе профилей участников и при скачивании графа дружбы, краулер заблаговременно получает перечень идентификаторов всех участников из каталога (https://vk.com/catalog.php). Скачивание графа подписок на сообщества проводится для 1 млн. «самых активных» групп VK. Перечень групп был предварительно подготовлен при помощи ранжирования доступных на то время сообществ по дате самой поздней публикации.

Для отбора профилей участников социальной сети применяются методы API users.get и groups.getById, принимающие на вход перечни идентификаторов участников или групп и возвращающие перечни их профилей в формате JSON. Во время одного запроса к каждому из способов скачивается порядка 200 профилей. В свою очередь, для сбора графов дружбы и подписки полезны методы API friends.get и groups.getMembers, принимающие идентификатор одного участника или сообщества и возвращающие перечни идентификаторов его подписчиков или друзей.

Перечисленные способы сбора информации применяют версию API 5.52. Реализация краулера данных осуществлена на базе фреймворка MODIS Crawler (дает возможность одновременно осуществлять множество запросов).

Предложенный алгоритм установления возраста участников социальной сети для решения поставленной задачи использует значения возраста, указанные другими пользователями, которые берутся из даты рождения, указанной в профиле.

Три варианта возможного вида поля «дата рождения»:

1) DD-MM – доступны число и месяц рождения;
2) YYYY – указан только год;
3) DD-MM-YYYY – дата отображена полностью.

Там, где год рождения известен, возраст определяется следующим образом:

$$Y_c - Y_u, \tag{1}$$

где Yc – это текущий год,
Yu – указанный год.

Неизвестные значения атрибутов система выявления возраста определяет исходя из данных о размеченных характеристиках и социальных связей (граф подписчиков и граф друзей). Как указывалось, социальный граф включает в себя узлы и связи между ними. Первые представляют собой сообщества и участники, а вторые бывают следующими:

- граф подписок на сообщества – предполагает взаимосвязь между участником и сообществом;
- граф друзей – это связи между участниками (между пользователем и его друзьями).

Каждому узлу в графе присваивается набор меток. Соответствующая метка – это определенное значение атрибута (к примеру, «возраст=23»).

Порядок алгоритма следующий:

а) инициализация – узлы пользователя получают свои метки;

б) создание векторной модели;

в) определение весов участников и сообществ, распределение меток на последние;

г) создание векторной модели на базе весов;

д) распределение меток на узлы-пользователей на базе весов – метки узлов участников и узлов-сообществ распространяются на узлы-пользователей, которые не имеют их (возраст не указан).

Рассмотрим более подробно алгоритм вычисления метки.

В первую очередь узлы-пользователи инициализируются метками согласно разметке. Для соответствующего участника осуществляется распределение значений признака между его соседями. Пример распределения приведен на рисунке 1.

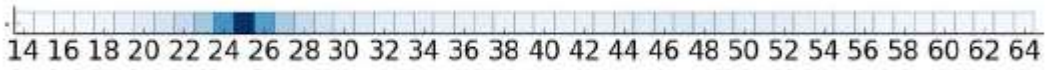

Рисунок 1 – Распределение значений возраста соседей участника социальной сети[1]

Далее все распределения сгруппировываются по значению атрибута участника, после чего для каждого определяется среднее распределение возраста соседей. В результате получаем векторную модель для выявления возрастной принадлежности пользователей. К примеру, на рисунке 2 приводится векторная модель, в которой для каждого атрибута возраста задается распределение возрастов соседей (модель $Model_{avg}$).

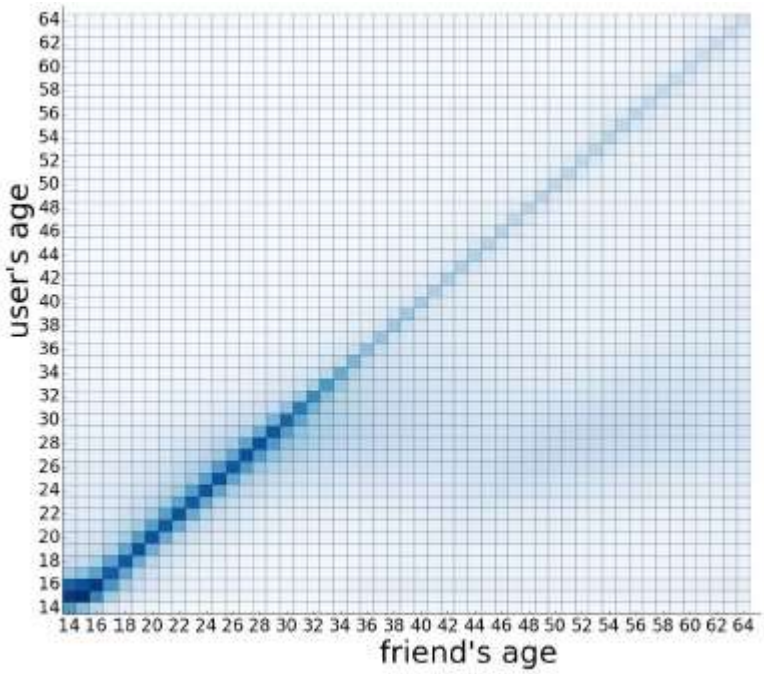

Рисунок 2 – Векторная модель для атрибута «возраст»[2]

Векторная модель, в которой осуществляется усреднение распределений по всем критериям, используется при распространении меток в сторону участников. Для узлов-сообществ применяется модель, в которой распределение атрибутов возраста соседей для отдельного значения обретает вид согласно формуле:

[1] Вероятность отражена интенсивностью цвета (чем цвет темнее, тем больше вероятность)

[2] В каждой отдельной строке отражены усредненные значения возраста соседей

$$p(val_n|val_c) = \begin{cases} 1, & \text{если } val_n = val_c \\ 0, & \text{если } val_n \neq val_c \end{cases}, \quad (2)$$

где $p(val_n|val_c)$ – это вероятность того, что значения возраста соседа равно val_n, только в том случае, если свое значение атрибута равно val_c (модель $\text{Model}_{\text{max}}$).

Векторные модели зачастую применяются для оценки близости распределения соседей узла, для которого определяется метка к определенному распределению из модели. Наибольшая близость в этом случае достигается, когда все метки соседей узла имеют одинаковое значение.

Переходим к моделированию распространения меток по социальным связям. На основе значений характеристик соседей для каждого сообщества отдельно вычисляются метка (атрибутное значение) и вес. В свою очередь, вес – это вещественное число, которое определяет соответствие между меткой конкретного пользователя или сообщества и векторной моделью. Данный термин также можно толковать как убежденность алгоритма в собственном решении.

Для определения метки (атрибутного значения) узла создается распределение значений соответствующего атрибута у соседей Distr (рис. 1), далее для каждого значения рассчитывается близость этого распределения к характерному распределению векторной модели. За меру близости применяется косинусная мера. Следовательно, расчёт значения метки можно представить в виде следующей формулы:

$$L = \arg\max_{val}\big(sim(Model_*(val), Distr)\big) \quad (3)$$

$$S = \max_{val}\big(sim(Model_*(val), Distr)\big) = sim(Model_*(L), Distr) \quad (4)$$

$$sim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \quad (5)$$

где L – метка (значение атрибута),

S – близость соответствует максимально подходящему значению атрибута val,

Distr – распределение значений атрибута соседей,

$Model_*(val)$ – модельное распределение для значения атрибута val (модель $\text{Model}_{\text{avg}}$ используется для пользователей, а $\text{Model}_{\text{max}}$ – для сообществ).

Показатель S и L рассчитываются для узлов, имеющих множество узлов-соседей, у которых значение атрибута заполнено. Показатель S применяют для выявления весов узлов W(node), которые рассчитываются для соответствующего вида отдельно. Найденные значения S группируют по возрастанию и помещают в массив. После этого для расчета применяют следующую формулу:

$$W(node) = \left(\frac{pos(S_{node})}{N}\right)^2, \quad (6)$$

где $pos(S_{node})$ – порядковый номер значения S (от 1 до N) в полученном массиве,

N – число узлов с выявленным значением S.

При этом для тех узлов, у которых значение S не определено, вес приравнивается 0. Далее переходим к ручному установлению веса для каждого типа соседа (сообщества или участники социальной сети), задающего вклад источнику информации (WComm, WUser).

При тестировании они подбираются для всех атрибутов. После этого приступают к распространению меток для пользователей, на основе выявленных на предшествующем этапе весов $W(node)$.

Таким образом, вложение каждого сообщества-соседа участника в распределение Distr приравнивается к его весу. Для каждой группы соседей (сообщество или участник) по-отдельности рассчитывается распределение их меток соответствующей группы и умножается на соответствующий вес. Полученная сумма нормализуется, а для распространения значений атрибутов применяется модель $\text{Model}_{\text{avg}}$. Незаполненные признаки указываются по меткам.

Проведение тестирования

Для качественного выявления демографических признаков применяют кросс-валидацию с группированием информации на 10 сегментов, запускаемую при отличии параметров WUser и WComm. Все параметры принимают значения 1, 10 или 100.

Рассмотрим выборку, метрики качества и полученные результаты. Сначала из всех сообществ отбираются те, у которых насчитывается хотя бы K подписчиков с верно внесенным в профиле значением возраста, а потом участники с критериями:

- значение возраста размечено;
- имеется хотя бы K социальных связей (друзья с точным значением атрибута).

Для эксперимента в выборке насчитывалось 28940134 участников, а в тестах K = 10. Для возрастного атрибута устанавливается точность, с возрастанием которого абсолютная ошибка при угадывании становится менее критичной, следовательно, применяем величину относительной ошибки. Принято считать, что значение возраста определено верно, если:

$$\left|age_u - age_p\right| \leq 0{,}15 \times age_u, \tag{7}$$

где age_u – возраст участника из разметки,
age_p – предположенное значение возраста.

Для атрибута возраст средняя абсолютная ошибка (MAE) рассчитывается по формуле:

$$\frac{\sum|age_u - age_p|}{N}, \tag{8}$$

где N – число предположенных значений.

Эксперименты проводились при разных значениях параметров W_{Comm} и W_{User}, конфигурации которых изучаются при равенстве и когда один из них превалирует. Значения средней абсолютной ошибки и точности рассмотрим в таблице 1.

Таблица 1 – Результаты эксперимента

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

Таким образом, из таблицы 1 видно, что значительный результат дает граф друзей.

Заключение

В данной статье была изучена задача выявления возраста участников социальных сетей. В ходе работы нами был предложен подход, который позволяет установить значения возраста участников, у которых указывается хотя бы один тип из следующей информации: перечень друзей или подписок на сообщества (группы). Алгоритм базируется на распределении меток в социальном графе, где участники и сообщества – это его узлы, а взаимоотношения между участниками (отношения дружбы, подписки) – ребра. В свою очередь, метки – это конкретные значения возраста. В результате эксперимента получили достаточно приемлемые итоги при установлении возраста пользователей социальных сетей. Планируется использовать данный подход и к остальным атрибутам профилей, применять генерируемый текстовый контент и зависимости между значениями атрибутов различных категорий (к примеру, между уровнем образования и возрастом).

Список литературы

1. John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1301-1309. Association for Computational Linguistics, 2011.
2. Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. Digital Investigation, 8(1):78-88, 2011.
3. Michael D Conover, Bruno Gon¸calves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pages 192-199. IEEE, 2011.
4. Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 15-24. ACM, 2014.
5. Katja Filippova. User demographics and language in an implicit social network. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1478-1488. Association for Computational Linguistics, 2012.
6. Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You are who you know: inferring user profiles in online social networks. In Proceedings of the third ACM international conference on Web search and data mining, pages 251-260. ACM, 2010.
7. Dong Nguyen, Noah A Smith, and Carolyn P Ros´e. Author age prediction from text using linear regression. In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 115-123. Association for Computational Linguistics, 2011.
8. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In Proceedings of the 3rd international workshop on Search and mining user-generated contents, pages 37-44. ACM, 2011.
9. Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In Proceedings of the First workshop on Unsupervised Learning in NLP, pages 53-63. Association for Computational Linguistics, 2011.
10. Гомзин А.Г., Кузнецов С.Д. Методы построения социо-демографических профилей пользователей сети Интернет. Труды ИСП РАН, том 27, вып. 4, 2015, стр. 129-144. DOI: 10.15514/ISPRAS-2015-27(4)-7