Методы построения социо-демографических профилей пользователей сети Интернет

[1, 2]А.Г. Гомзин <gomzin@ispras.ru>
[1, 2, 3]С.Д. Кузнецов <kuzloc@ispras.ru>
[1] Институт системного программирования РАН, 109004, Россия, г. Москва, ул. А. Солженицына, дом 25
[2] Московский государственный университет имени М.В. Ломоносова, 119991 ГСП-1 Москва, Ленинские горы, МГУ имени М.В. Ломоносова, 2-й учебный корпус, факультет ВМК
[3]Московский физико-технический институт (государственный университет), 141700, Московская область, г. Долгопрудный, Институтский пер., 9

Аннотация. Работа посвящена методам построения социально-демографического профиля пользователей Интернета. Примерами демографических атрибутов являются пол, возраст, политические и религиозные взгляды, район проживания, состояние отношений с другими людьми. Эта работа представляет собой обзор методов, которые обнаруживают демографические атрибуты из профиля пользователя и сообщений. Большинство известных работ посвящены выявлению пола. Возраст, политические взгляды и области также интересуют исследователей.

Самыми популярными источниками данных для извлечения демографических атрибутов являются социальные сети, такие как Facebook, Twitter, Youtube. Большинство решений основано на машинном обучении с учителем. Машинное обучение позволяет найти целевые значения (демографические атрибуты) в зависимости от входных данных и использовать их, чтобы предсказать значение целевого атрибута для новых данных. в работе анализируются следующие шаги решения задачи: сбор данных, извлечение признаков, отбор информативных признаков, методы обучения классификаторов, оценка качества.

Исследования используют различные виды данных, чтобы предсказать демографические атрибуты. Самым популярным источником данных является текст. Последовательности слов (n-граммы), части речи, смайлики, особенности относящиеся к конкретным ресурсам (например, @ и # в Twitter) извлекаются и используются в качестве входных данных для алгоритмов машинного обучения. Социальные графы также используются в качестве исходных данных. Сообщества пользователей, которые автоматически извлекаются из социального графа пользователем в качестве признаков для прогнозирования атрибутов. Текстовые данные дает много возможностей. Алгоритмы выбора признаков необходимы для снижения признакового пространства.

В статье исследуются функции выбора, классификации и регрессии алгоритмы, показатели оценки.

Ключевые слова: демографические атрибуты; демографические характеристики; социальные сети; обработка текстов на естественном языке; машинное обучение

DOI: 10.15514/ISPRAS-2015-27(4)-7

Для цитирования: Гомзин А.Г., Кузнецов С.Д. Методы построения социо-демографических профилей пользователей сети Интернет. Труды ИСП РАН, том 27, вып. 4, 2015 г., стр. 129-144. DOI: 10.15514/ISPRAS-2015-27(4)-7.

1. Введение

Многие ресурсы в сети Интернет позволяют своим пользователям принимать активное участие в создании контента. К таким ресурсам относятся блоги, форумы, социальные сети. Кроме того, многие интернет-магазины, новостные сайты и другие подобные сервисы позволяют пользователям оставлять комментарии, отзывы. Как правило, помимо сообщений, комментариев, отзывов, оценок пользователь оставляет на ресурсе некоторую информацию о себе. Эта информация включает в себя имя, пол, возраст, интересы, контактные и другие данные. Такая информация о пользователе, как правило, помещается на отдельную страницу ресурса. Набор таких атрибутов называется профилем пользователя.

Контент, генерируемый пользователями, отражает его интересы, взгляды. Так, например, лексика, используемая в социальных сетях подростками и взрослыми людьми, различается. В качестве контента, генерируемого пользователями, можно рассматривать тексты, изображения, аудио- и видеоконтент. В данном обзоре рассматриваются только работы, посвященные анализу текстового контента пользователей.

В статье рассматривается задача составления социо-демографических профилей пользователей сети Интернет. Далеко не все пользователи полностью заполняют свой профиль. Кроме того, в некоторых случаях пользователи преднамеренно указывают неверные данные. В связи с этим возникает задача предсказания неизвестных социо-демографических атрибутов, таких как пол, возраст, политические предпочтения, по имеющейся информации о пользователе. Обычно анализируется только находящийся в публичном доступе контент пользователя, т.е. его сообщения, комментарии, отзывы.

Методы автоматического определения демографических атрибутов пользователей могут использоваться для исследования определенных групп пользователей, даже если не все пользователи указывают значения атрибутов.

Полученные с помощью таких методов значения атрибутов могут применяться в рекомендательных системах [1], для таргетированной рекламы [2], а также в других приложениях.

Абсолютное большинство работ по определению пола, возраста и других атрибутов основано на методах машинного обучения. Решение задачи разбивается на несколько этапов:

1. сбор данных для построения модели
2. построение (обучение) модели
3. классификация с использованием полученной модели и оценка качества модели

В первом разделе рассматриваются исходные данные и особенности их сбора. Второй раздел описывает решения, где не используется машинное обучение. Третий раздел посвящен решению задачи с использованием методов машинного обучения. Затем следуют выводы и заключение.

2. Данные

Наибольший интерес для исследователей представляют активно развивающиеся социальные сети, такие как Facebook, Twitter и другие.

Facebook – самая крупная по количеству зарегистрированных пользователей социальная сеть [3]. В ней зарегистрировано более 1,2 млн пользователей. В профилях пользователей можно встретить различные демографические атрибуты: пол, возраст, семейное положение, политические и религиозные взгляды и т.д. Классификаторы, полученные с использованием контента пользователей и их профилей, позволяют с высокой точностью предсказывать значения атрибутов, которые не указаны в профиле других пользователей.

Другой ресурс, не менее популярный у исследователей демографических атрибутов, – социальная сеть Twitter. Это сервис микроблоггинга, в котором длина каждого сообщения не превышает 140 символов. Такие тексты имеют свои особенности, которые описаны в работе [4]. На данном ресурсе в профилях пользователей отсутствуют демографические атрибуты, что усложняет этап сбора обучающей выборки. Об этом подробнее будет написано в разделе 2.1.

Кроме данных ресурсов, в некоторых исследованиях анализируются комментарии на Youtube [5], новости и электронные письма [6].

В нашем обзоре представлены работы, в которых решаются задачи определения следующих атрибутов: пол, возраст, политические взгляды, регион проживания.

2.1 Сбор данных

Первым этапом решения задачи определения демографических атрибутов является сбор данных. Данные содержат текстовый контент пользователей, а также истинные значения атрибутов. Значения атрибутов используются алгоритмами машинного обучения с учителем и оценки качества.

Социальная сеть представляется в виде графа, в котором вершины соответствуют пользователям, а ребра – наличию социальной связи между пользователями (отношения дружбы, подписки и т.д.). Возникает задача обхода вершин графа с целью получения репрезентативной выборки. Такой процесс обхода называется сэмплингом. Исследования [7, 8] показали, что наиболее репрезентативные выборки получаются при использовании алгоритмов сэмплирования «Лесного пожара» (Forest Fire) и Метрополиса-Гастингса (Metropolis–Hasting).

При решении задач определения демографических атрибутов в обучающую выборку попадают только пользователи с явно указанными атрибутами. Такие данные можно собрать, например, с использованием сервиса поиска друзей в социальных сетях. Например, в социальной сети Вконтакте сервис поиска пользователей[1] позволяет явно указать значения атрибутов интересующих пользователей. Встречаются также работы, в которых значения целевых атрибутов определяются экспертами [9, 10, 11, 12].

Не всегда на анализируемом ресурсе профиль пользователей содержит нужный атрибут. Примером такого ресурса со скудными профилями является сервис микроблогинга Twitter. В профиле Twitter не указывается пол, возраст, семейное положение и другие атрибуты. Найти значение нужного атрибута можно, например, если знать где находится профиль этого же человека на другом ресурсе. В профиле Twitter имеется поле URL, в котором пользователи часто указывают гиперссылку на свой профиль в другой социальной сети. Перейдя по гиперссылке, можно найти значения целевых атрибутов. Такой подход используется в работах [13, 14].

Даже если данных с атрибутами достаточно для построения решений и реализации методов, могут возникать технические проблемы, связанные с ограничениями на ресурсах. Например, API Twitter позволяет выдавать не более 300 запросов на скачивание страницы сообщений пользователя в течение 15 минут для одного приложения[2]. Кроме того, на многих ресурсах имеются неявные ограничения. Например, при слишком частом обращении к сервису, он может заблокировать запросы и не возвращать данные. Частоту, с которой нужно делать запросы, не получая блокировку, можно определить только эмпирически.

3. Методы, не использующие машинное обучение

Как правило, задачи определения демографических атрибутов пользователей решаются с использованием методов машинного обучения. Но в некоторых задачах высокая точность достигается и при использовании более простых решений.

[1] https://vk.com/people

[2] https://dev.twitter.com/rest/reference/get/statuses/user_timeline

Например, пол человека с высокой точностью определяется по его имени. В работе [15] использовались различные словари имен. Такие методы предполагают, что пользователи указывают свое настоящее имя.

Для определения остальных атрибутов требуются более сложные методы. В большинстве исследований, посвященных определению демографических атрибутов пользователей, используется машинное обучение с учителем.

4. Машинное обучение

Машинное обучение позволяет найти зависимость целевых значений от исходных данных и использовать ее для предсказания значения целевого атрибута для новых данных. В нашем случае целевые данные – это демографические атрибуты, а исходные данные – информация о пользователе. Информация о пользователе варьируется от ресурса к ресурсу, но, как правило, включает в себя сообщения, автором которых он является, и профиль – набор атрибутов и их значений. В социальных сетях также доступна информация об отношениях между пользователями, а также отношениях между пользователями и объектами: так называемые социальные связи. В некоторых работах [5, 9] эти связи используются для определения демографических атрибутов.

Различаются машинное обучение без учителя и машинное обучение с учителем. При обучении без учителя алгоритм находит закономерности в исходных данных, с помощью которых разбивает эти данные на группы (кластеры). При обучении с учителем, помимо исходных данных, для алгоритма требуются значения целевых атрибутов. В процессе обучения строится модель, с помощью которой предсказываются целевые значения для новых исходных данных, в которых значения целевых атрибутов неизвестны. В задачах определения демографических атрибутов, как правило, имеется выборка, в которой значения атрибутов указаны явно. Соответственно, применяется обучение с учителем.

При использовании машинного обучения процесс решения задачи включает в себя четыре этапа:

1. извлечение признаков;
2. отбор признаков (опционально);
3. обучение модели;
4. оценка качества алгоритма.

На рис. 1 изображена схема обучения с учителем. В качестве целевого атрибута рассматривается пол пользователя.

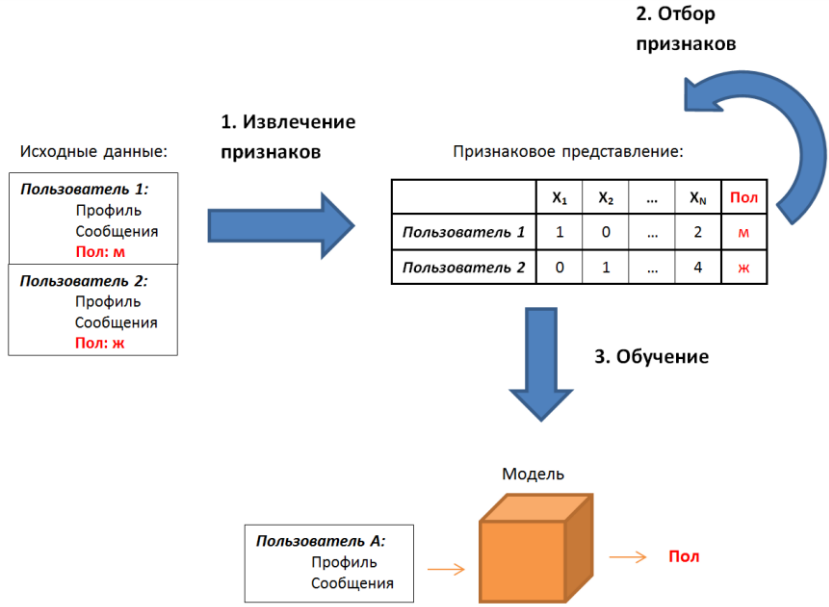

Рис. 1. Машинное обучение с учителем

4.1 Извлечение признаков

Данный подраздел посвящен признакам, используемым при обучении классификаторов пользователей по демографическим атрибутам. Сначала опишем признаки, а затем исследования, в которых используются эти признаки.

Исходные данные в задачах определения атрибутов представляют собой профиль пользователя и набор текстовых сообщений, автором которого он является. Но алгоритмы машинного обучения работают с признаковыми представлениями объектов (рис. 1), где каждый объект представляет собой набор векторов в пространстве признаков. Если применяются методы машинного обучения с учителем, значение целевого атрибута также входит в объект. Соответствующий этап машинного обучения называется извлечением признаков (рис. 1).

Для начала необходимо определить, какие признаки можно выделить из текстов и профилей. Текст состоит из слов. В качестве признаков используются слова и последовательности подряд идущих слов. Такие последовательности называются n-граммами. Здесь n – длина последовательности. В качестве значений признаков можно использовать бинарное значение: 1, если слово встретилось в тексте, 0 – иначе. Кроме того,

значением признака может быть частота встречаемости слова в текст, мера TF-IDF и т.д.

Часто в определении пола, возраста и других атрибутов помогают части речи в последовательности слов. Части речи последовательностей подряд идущих слов могут рассматриваются как признаки (POS n-граммы).

С другой стороны, текст можно рассматривать как последовательность символов. Соответственно, в качестве признаков могут использоваться символьные n-граммы, т.е. последовательности подряд идущих символов.

Существует также подход, при котором в качестве признаков рассматриваются последовательности фонем – единиц звука речи. В этом случае ключевую роль играет не написание слов, а их произношение.

В некоторых работах используются такие признаки, как сокращения, эмотиконы, признаки, специфичные для определенных ресурсов (например, @упоминания и #хэштэги в Твиттере).

Еще одна группа признаков – статистические. Числовые значения таких признаков – некоторые статистические значения, полученные из тестов. Примерами статистических признаков являются средняя длина одного сообщения пользователя, частота эмотиконов, частота знаков препинания в сообщениях и т.д.

Помимо текстов сообщений пользователя, анализируется его профиль. Значения атрибутов профиля могут использоваться как признаки. В некоторых случаях из значений атрибутов извлекаются описанные выше признаки (например, n-граммы). Например, для определения пола пользователя полезным бывает имя. Встречаются также работы, где используется цвет страницы пользователя, заданный им в личных настройках.

В социальных сетях, помимо текстов и профилей пользователей, существуют социальные связи. Социальный граф, в котором вершины соответствуют пользователям, а ребра – наличию отношений между пользователями, также используется для предсказания атрибутов. Например, в таком графе с помощью специальных алгоритмов [16] могут выделяться сообщества. Набор сообществ, в которые входит пользователь, используется в качестве признаков.

Как правило, из исходных данных извлекаются комбинации различных типов признаков. Но важно помнить, что не всегда использование большего числа признаков улучшает качество получаемых классификаторов. Может возникнуть переобучение. Способы избежать переобучения описываются в следующем разделе.

Универсального рецепта по выбору признаков нет. Выбор признаков существенно зависит от задачи (т.е. какой атрибут определяется) и исходных данных. Далее описываются некоторые работы и признаки, используемые авторами этих работ.

В работе [5] определяется пол пользователей Youtube по комментариям и графу пользователи-видео, где ребро между пользователем и видео означает факт просмотра видео пользователем. В качестве признаков рассматриваются статистические признаки, такие как средняя длина комментария в символах/словах/предложениях, словесные n-граммы, возраст пользователя, распределение пола, полученное с помощью модели распространения атрибута «пол» в графе пользователи-видео.

Авторы работы [14] определяют пол пользователей Twitter. В этом исследовании извлекаются признаки из профиля, в том числе цвета фона/текста/ссылок, которые пользователь указал в настройках своей страницы. Также выделяются признаки из имен пользователей. При этом имя преобразовывается в последовательность фонем.

В работе [13] пол пользователей Twitter определяется по текстам их твитов. Используются символьные и словесные n-граммы. Для решения той же задачи авторы [12] использовали символьные n-граммы.

В работе [17] рассматривалась задача определения возраста пользователей, пишущих на голландском языке. Возраст пользователей разбивался на 2 или 4 интервала разными способами (до 16, после 16 лет; до 16, после 18 лет; до 16, после 25 лет; то же самое, но для каждого значения пола). В качестве признаков использовались символьные и словесные 1,2 и 3-граммы.

Помимо решений, в которых множество значений возраста пользователей разбивается на несколько интервалов, существуют методы, которые предсказывают числовое значение возраста [18]. В [18] методы тестировались на нескольких наборах исходных данных. При этом в качестве признаков использовались словесные юниграммы (1-граммы), юниграммы и биграммы (2-граммы) частей речи слов, статистические признаки.

Пол и возраст – атрибуты, наиболее популярные среди исследователей. Существуют также и работы, в которых рассматриваются другие атрибуты, такие как политические взгляды и регион проживания.

Авторы [9] определяют политические предпочтения пользователей социальной сети Twitter. Рассматриваются три класса: демократы, республиканцы, неявная политическая позиция. В качестве признаков используются словесные юниграммы, хэштеги, сообщества пользователей (полученные с помощью алгоритма, основанного на распространении меток в социальном графе пользователей).

В работе [10] решаются задачи определения пола, возраста (до 30, после 30), региона (юг и северо-восток США) и политических взглядов (республиканцы, демократы). Авторы рассматривали в качестве признаков юниграммы и биграммы слов, а также социолингвистические признаки – эмотиконы, аббревиатуры, повторяющиеся знаки препинания и др.

Мы затронули лишь некоторые исследования, посвященные определению демографических атрибутов. В большинстве работ в качестве признаков присутствуют n-граммы, извлеченные из текстов сообщений пользователей.

При использовании n-грамм пространство признаков велико, в связи с чем возникает задача отбора информативных признаков.

4.2 Отбор признаков, уменьшение размерности

В процессе извлечения признаков из текста каждый пользователь представляется большим количеством признаков. Если рассматривать все различные признаки, встречающиеся у всех пользователей, то их получается на порядки больше, чем самих пользователей. Например, в работе [13] обучающая выборка содержит 180000 пользователей, при этом в ней 15000000 различных признаков. При таких данных высока вероятность возникновения переобучения. При переобучении модели классификаторов получаются очень сложными, так как в них присутствуют все признаки. В этом случае классификатор правильно предсказывает результат для тех данных, на которых он обучился, и неправильно – для новых данных. Чтобы избавиться от переобучения, нужно уменьшить количество признаков.

Один из способов – избегать большого числа признаков на этапе выбора признаков. Например, в работе [14] рассматриваются последовательности фонем, извлеченных из имени пользователя. В данной работе используется до 16000 различных признаков при том, что набор данных содержит около 180000 пользователей.

В случае, когда анализируются тексты сообщений, нельзя обойтись без этапа отбора признаков. С помощью специальных методов, о которых будет сказано далее, выбираются наиболее информативные признаки для анализируемого набора данных и у каждого объекта оставляются только те признаки, которые считаются информативными.

Среди методов отбора признаков существуют такие, которые не рассматривают значения целевых атрибутов. Одним из простых примеров является фильтрация признаков по частоте. Для каждого признака вычисляется количество объектов, в которых данный признак присутствует. Выбираются признаки с наибольшим значением частоты, а редкие признаки не рассматриваются. Другой способ – оставлять признаки с высокой дисперсией. В этом случае удаляются признаки, значение которых несущественно варьируется у объектов.

В этих методах не учитываются значения целевых атрибутов у объектов обучающей выборки. Информативность признака должна оцениваться в контексте целевого атрибута. Например, значение признака «окончание имени пользователя» коррелирует с полом соответствующего пользователя: как правило, имена женщин заканчиваются на гласную букву, имена мужчин – на согласную.

В работе [12] используются несколько методов отбора признаков: Хи-квадрат (Chi-Square), прирост информации (Information Gain), отношение прироста информации (Information Gain Ratio), Relief, симметричная неопределенность (Symmetrical Uncertainty), Filtered Attribute Evaluation. Все эти методы

оценивают, насколько хорошо значения признаков разделяют выборку по классам.

Некоторые алгоритмы машинного обучения имеют встроенную возможность отбора признаков. Как правило, алгоритмы подбирают параметры модели, чтобы минимизировалась ошибка на обучающей выборке. Для уменьшения количества признаков в модели используется регуляризация. Суть ее заключается в том, что сложность модели (пропорциональная количеству признаков с ненулевыми весами) минимизируется одновременно с ошибкой на тренировочных данных. Примером такой регуляризации, уменьшающей размерность данных, является регуляризации LASSO. К функционалу ошибки прибавляется сумма модулей весов признаков. Таким образом, если признак используется в модели, функционал ошибки увеличивается на модуль его веса.

Рассмотренные способы уменьшают размерность исходных данных путем удаления неинформативных признаков. В этом случае в качестве признаков на вход обучающему алгоритму приходит некоторое подмножество исходных признаков. Существуют другие подходы уменьшения размерности данных, в которых признаковое пространство полностью меняется.

Для текстовых данных применимы методы тематического моделирования [19]. В результате применения таких методов каждый текст (сообщение пользователя) представляется в виде распределения над темами.

Исходные данные можно представить в виде матрицы (объекты × признаки). Существуют различные методы матричных разложений, где исходная матрица представляется в виде произведения двух других: $A=W\times H$. Здесь A – исходная матрица (размерности $n\times m$) признаковых представлений объектов (n объектов, m признаков) раскладывается на произведение двух других матриц: W (размерности $n\times t$) и H (размерности $t\times m$). Матрицу W можно рассматривать как новое признаковое представление объектов в пространстве t неявных признаков.

Отбору признаков как отдельной проблеме посвящены работы [20, 21].

4.3 Используемые алгоритмы машинного обучения

Множество значений исследуемых демографических атрибутов, как правило, состоит из нескольких элементов. Например, пол принимает два значения: мужской, женский. В некоторых случаях атрибут принимает числовое значение (например, возраст). Таким образом, исходная задача сводится к задаче классификации или регрессии.

Одним из самых простых алгоритмов классификации является Наивный байесовский классификатор. Он предполагает, что все признаки независимы. В его основе лежит теорема Байеса. Данный алгоритм используется в работах [12, 13, 14] для определения пола. Преимущество данного метода состоит в том, что он поддерживает онлайн-обучение, т.е. при добавлении в обучающую

выборку новых объектов модель классификатора не пересчитывается заново, а обновляется согласно новым данным.

При классификации на два класса часто используется линейный классификатор. Одним из популярных алгоритмов обучения линейного классификатора является метод опорных векторов (в англоязычной литературе также известный как SVM – Support Vector Machine). Основная идея метода – поиск разделяющей гиперплоскости с максимальным зазором до объектов классов. Метод опорных векторов использовался в работах [6, 9, 10, 13, 17, 18].

Метод опорных векторов не является онлайновым. При необходимости «дообучения» линейного классификатора нужно использовать другие алгоритмы, такие как Перцептрон и Balanced Winnow. Такие алгоритмы используются в работах [11, 12, 13].

В работах также встречаются решающие деревья [6, 14], логистическая регрессия [6].

4.4 Оценка качества

Оценка качества нужна, чтобы иметь представление о том, насколько хорошо работает полученный алгоритм. Для этого измеряются такие значения, как точность (accuracy), достоверность (precision), полнота (recall) и F-1 мера.

Точность определяется как доля объектов, по которым классификатор принял правильное решение:

$$Accuracy = \frac{P}{N}$$

Достоверность и полнота рассматриваются в пределах одного класса, который обозначается *положительным*. Сначала составляется таблица (табл. 1)

Таблица 1. Обозначение результатов классификатора по отношению к истинным значениями

		Истинное значение	
		положительное	*отрицательное*
Результат классификатора	*Положительный*	TP	FP
	Отрицательный	FN	TN

Тогда достоверность определяется как:

$$Precision = \frac{TP}{TP + FP}$$

Полнота:

$$Recall = \frac{TP}{TP + FN}$$

F-мера представляет собой гармоническое среднее между достоверностью и полнотой:

$$F = 2\frac{Precision \times Recall}{Precision + Recall}$$

Как правило, качество алгоритмов оценивается с использованием кросс-валидации. Набор размеченных данных разбивается на несколько частей. Затем для каждой части происходит обучение на оставшихся частях и проверка на выбранной части данных. Измерения значений точности или других параметров усредняются по всему набору данных.

Для задач регрессии используется, как правило, метрика MAE – средняя абсолютная ошибка:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Здесь y_i – предсказанное значение, $\hat{y}_i$ – истинное значение.

Результаты исследований показывают большой разброс в зависимости от данных, извлекаемых признаков и алгоритмов машинного обучения. Здесь приводятся интервалы значений точности для различных атрибутов:

Пол: 75-95 %;

Политические взгляды: 79-95%;

Возраст (MAE): 5-17.

5. Заключение

В данной статье рассматриваются методы построения социо-демографичского профиля пользователей Интернета. Большинство обозреваемых работ посвящено анализу сообщений социальных сетей. Особый интерес представляет социальная сеть Twitter, так как профили пользователей не содержат явно указанных демографических атрибутов. Кроме того, сообщения имеют свои стилистические особенности.

Много работ посвящено определению пола. Также встречаются статьи, посвященные определению возраста, политических взглядов, региона проживания.

Абсолютное большинство решений основано на использовании методов машинного обучения с учителем. В статье рассмотрен каждый этап решения: сбор данных, извлечение признаков, отбор информативных признаков, методы обучения классификаторов, оценка качества.

В заключение можно выделить направления дальнейших улучшений работы алгоритмов. Первое направление – определять все атрибуты вместе, учитывая зависимость между этими атрибутами. Второе направление – исследовать

возможность построения классификаторов, не зависящих от источника исходных данных.

Список литературы

[1]. Li Q., Kim B. M. Constructing user profiles for collaborative recommender system //Advanced Web Technologies and Applications. – Springer Berlin Heidelberg, 2004. – С. 100-110.

[2]. Bharat K., Lawrence S., Sahami M. Generating user information for use in targeted advertising : заяв. пат. 10/750,363 США. – 2003.

[3]. Список социальных сетей. [электронный ресурс] https://ru.wikipedia.org/wiki/Список_социальных_сетей

[4]. Коршунов А. и др. Определение демографических атрибутов пользователей микроблогов //Труды Института системного программирования РАН. – 2013. – Т. 25, стр. 179-194. DOI: 10.15514/ISPRAS-2013-25-10

[5]. Filippova K. User demographics and language in an implicit social network //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. – С. 1478-1488.

[6]. Cheng N., Chandramouli R., Subbalakshmi K. P. Author gender identification from text //Digital Investigation. – 2011. – Т. 8. – №. 1. – С. 78-88.

[7]. Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

[8]. Gjoka M. et al. Walking in Facebook: A case study of unbiased sampling of OSNs //INFOCOM, 2010 Proceedings IEEE. – IEEE, 2010. – С. 1-9.

[9]. Conover M. D. et al. Predicting the political alignment of twitter users //Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. – IEEE, 2011. – С. 192-199.

[10]. Rao D. et al. Classifying latent user attributes in twitter //Proceedings of the 2nd international workshop on Search and mining user-generated contents. – ACM, 2010. – С. 37-44.

[11]. Deitrick W. et al. Gender identification on Twitter using the modified balanced winnow. – 2012

[12]. Miller Z., Dickinson B., Hu W. Gender prediction on twitter using stream algorithms with N-gram character features. – 2012.

[13]. Burger J. D. et al. Discriminating gender on Twitter //Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2011. – С. 1301-1309.

[14]. Alowibdi J. S., Buy U. A., Yu P. Empirical evaluation of profile characteristics for gender classification on twitter //Machine Learning and Applications (ICMLA), 2013 12th International Conference on. – IEEE, 2013. – Т. 1. – с. 365-369.

[15]. Sloan L. et al. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter //Sociological Research Online. – 2013. – Т. 18. – №. 3. – С. 7.

[16]. Fortunato S. Community detection in graphs //Physics Reports. – 2010. – Т. 486. – №. 3. – С. 75-174.

[17]. Peersman C., Daelemans W., Van Vaerenbergh L. Predicting age and gender in online social networks //Proceedings of the 3rd international workshop on Search and mining user-generated contents. – ACM, 2011. – С. 37-44.
[18]. Nguyen D., Smith N. A., Rosé C. P. Author age prediction from text using linear regression //Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. – Association for Computational Linguistics, 2011. – С. 115-123.
[19]. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке //Труды Института системного программирования РАН. – 2012. – Т. 23, стр. 215-244. DOI: 10.15514/ISPRAS-2012-23-13
[20]. Molina L. C., Belanche L., Nebot À. Feature selection algorithms: A survey and experimental evaluation //Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. – IEEE, 2002. – С. 306-313.
[21]. Zheng Z., Wu X., Srihari R. Feature selection for text categorization on imbalanced data //ACM Sigkdd Explorations Newsletter. – 2004. – Т. 6. – №. 1. – С. 80-89.

Methods for Construction of Socio-Demographic Profile of Internet Users

[1,2]A.G. Gomzin <gomzin@ispras.ru>
[1, 2, 3]S.D. Kuznetsov <kuzloc@ispras.ru>
[1] Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., 109004, Moscow, Russia
[2]Lomonosov Moscow State University, 2nd Education Building, Faculty CMC, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation
[3]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russia

Abstract. The paper is devoted to methods for construction of socio-demographic profile of Internet users. Gender, age, political and religion views, region, relationship status are examples of demographic attributes. This work is a survey of methods that detect demographic attributes from user's profile and messages. The most of surveyed works are devoted to gender detection. Age, political views and region are also interested researches. The most popular data sources for demographic attributes extraction are social networks, such as Facebook, Twitter, Youtube.

The most of solutions are based on supervised machine learning. Machine learning allows to find target values (demographic attributes) dependencies from input data and use them to predict the value of the target attribute for the new data. The following problem solving steps are surveyed in the paper: feature extraction, feature selection, model training, evaluation. Researches use different kind of data to predict demographic attributes. The most popular data source is text. Words sequences (n-grams), parts of speech, emoticons, features specific to particular resources (eg, @ mentions and # Hashtags on Twitter) are extracted and used as input for machine learning algorithms. Social graphs are also used as source data.

Communities of users that are automatically extracted from social graph are user as features for attributes prediction.
Text data produces a lot of features. Feature selection algorithms are needed to reduce feature space.
The paper surveys feature selection, classification and regression algorithms, evaluation metrics.

Keywords: demographic attributes; social networks; text processing; machine learning

DOI: 10.15514/ISPRAS-2015-27(4)-7

For citation: Gomzin A.G., Kuznetsov S.D. Methods for Construction of Socio-Demographic Profile of Internet Users. *Trudy ISP RAN/Proc. ISP RAS*, vol. 27, issue 4, 2015, pp. 129-144 (in Russian). DOI: 10.15514/ISPRAS-2015-27(4)-7.

References

[1]. Li Q., Kim B. M. Constructing user profiles for collaborative recommender system. Advanced Web Technologies and Applications. – Springer Berlin Heidelberg, 2004. – С. 100-110.
[2]. Bharat K., Lawrence S., Sahami M. Generating user information for use in targeted advertising : patent. 10/750,363 США. – 2003.
[3]. Spisok social'nyh setej. Wikipedia. List of social networks https://ru.wikipedia.org/wiki/Список_социальных_сетей
[4]. Koeshunov A. et al., Opredelenie demograficheskih atributov pol'zovatelej mikroblogov [Microblogs' users' demographic attributes detection]. *Trudy ISP RAN* [*The Proceedings of ISP RAS*], 2013. – T. 25, pp. 179-194. DOI: 10.15514/ISPRAS-2013-25-10
[5]. Filippova K. User demographics and language in an implicit social network. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. – P. 1478-1488.
[6]. Cheng N., Chandramouli R., Subbalakshmi K. P. Author gender identification from text. Digital Investigation. – 2011. – Т. 8. – №. 1. – P. 78-88.
[7]. Leskovec J., Faloutsos C. Sampling from large graphs. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – P. 631-636.
[8]. Gjoka M. et al. Walking in Facebook: A case study of unbiased sampling of OSNs. INFOCOM, 2010 Proceedings IEEE. – IEEE, 2010. – P. 1-9.
[9]. Conover M. D. et al. Predicting the political alignment of twitter users. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. – IEEE, 2011. – P. 192-199.
[10]. Rao D. et al. Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents. – ACM, 2010. – P. 37-44.
[11]. Deitrick W. et al. Gender identification on Twitter using the modified balanced winnow. – 2012
[12]. Miller Z., Dickinson B., Hu W. Gender prediction on twitter using stream algorithms with N-gram character features. – 2012.

[13]. Burger J. D. et al. Discriminating gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2011. – P. 1301-1309.

[14]. Alowibdi J. S., Buy U. A., Yu P. Empirical evaluation of profile characteristics for gender classification on twitter. Machine Learning and Applications (ICMLA), 2013 12th International Conference on. – IEEE, 2013. – T. 1. – P. 365-369.

[15]. Sloan L. et al. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online. – 2013. – T. 18. – №. 3. – P. 7.

[16]. Fortunato S. Community detection in graphs. Physics Reports. – 2010. – T. 486. – №. 3. – P. 75-174.

[17]. Peersman C., Daelemans W., Van Vaerenbergh L. Predicting age and gender in online social networks. Proceedings of the 3rd international workshop on Search and mining user-generated contents. – ACM, 2011. – P. 37-44.

[18]. Nguyen D., Smith N. A., Rosé C. P. Author age prediction from text using linear regression. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. – Association for Computational Linguistics, 2011. – P. 115-123.

[19]. Korshunov A., Gomzin A. Tematicheskoe modelirovanie tekstov na estestvennom yazyke [Topic modeling of natural language texts]. *Trudy ISP RAN* [*The Proceedings of ISP RAS*] – 2012. – T. 23. pp. 215-244. DOI: 10.15514/ISPRAS-2012-23-13

[20]. Molina L. C., Belanche L., Nebot À. Feature selection algorithms: A survey and experimental evaluation. Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. – IEEE, 2002. – P. 306-313.

[21]. Zheng Z., Wu X., Srihari R. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter. – 2004. – T. 6. – №. 1. – P. 80-89.