Определение демографических атрибутов пользователей микроблогов[1]

Антон Коршунов, Иван Белобородов, Андрей Гомзин, Кристина Чуприна, Никита Астраханцев, Ярослав Недумов, Денис Турдаков
{korshunov, ivbel, gomzin, chuprina, astrakhantsev, yaroslav.nedumov, turdakov}@ispras.ru

Аннотация. При заполнении полей профиля в различных интернет-сервисах пользователи зачастую по ошибке или преднамеренно не указывают значения некоторых демографических атрибутов, таких как пол, возраст, семейное положение, уровень образования, религиозные и политические взгляды. Вместе с тем, информация об атрибутах пользователей позволяет существенно повысить эффективность систем рекомендации, интернет-маркетинга и других приложений, предполагающих персонализацию результатов. В статье предлагается метод автоматического определения демографических атрибутов пользователей социального сервиса микроблогов Twitter по текстам их сообщений и другой доступной информации из профилей. Метод основан на алгоритме машинного обучения, его отличительными особенностями являются полностью автоматическое построение исходного набора данных для обучения и тестирования, а также поддержка широкого набора языков и демографических атрибутов. Экспериментальные исследования показали высокое качество результатов определения пола, возраста и семейного положения пользователя для наиболее популярных языков: английского, русского, немецкого, французского, итальянского и испанского. Кроме того, для английского языка поддерживается также определение уровня образования, а также религиозных и политических взглядов пользователя.

Ключевые слова: демографические характеристики; демографические атрибуты; социальные сети; микроблоги; обработка текстов на естественном языке; анализ содержимого; компьютерная лингвистика; машинное обучение.

1. Введение

В связи с увеличением количества пользователей интернета, а также появлением новых средств для обмена информацией, количество свободно

[1] Работа выполнена при финансовой поддержке Минобрнауки Российской Федерации по государственному контракту от 10.10.2013 г. № 14.514.11.4111 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы»

доступных персональных данных (включая текстовые сообщения) постоянно растёт. Учитывая склонность пользователей интернета к анонимности, актуальны методы частичной идентификации авторов сообщений по значениям их демографических атрибутов. В частности, в системах интернет-маркетинга и рекомендаций особую важность представляет определение демографических атрибутов пользователя для таргетированного продвижения товаров и услуг в группах пользователей с одинаковыми значениями атрибутов. Помимо интернет-сервисов, такие социо-демографические характеристики находят применение в различных дисциплинах: социология, психология, криминология, экономика, управление персоналом и др.

Демографические атрибуты можно условно разделить на категориальные (пол, национальность, раса, семейное положение, уровень образования, профессия, трудоустроенность, религиозные и политические взгляды) и численные (возраст, уровень доходов). Условность разделения связана с тем, что значения численного атрибута можно отобразить в набор категорий и в дальнейшем рассматривать этот атрибут как категориальный. В частности, значения возраста можно разделить на несколько возрастных категорий, что часто применяется на практике. Набор атрибутов одного пользователя составляет его социо-демографический профиль.

В статье предложен метод определения демографических атрибутов пользователей сети Twitter по текстам их сообщений, обладающий следующими преимуществами:

1. широкий набор поддерживаемых атрибутов: пол, возраст, семейное положение, уровень образования, религиозные и политические взгляды;
2. полностью автоматический метод сбора и разметки корпусов сообщений пользователей интернета для всех поддерживаемых атрибутов;
3. поддерживаемые языки: русский, английский, испанский, немецкий, французский, итальянский;
4. высокое качество результатов.

Дальнейшее изложение строится следующим образом. Раздел 2 содержит обзор литературы. Раздел 3 посвящён деталям предложенного метода. В разделе 4 описываются полученные экспериментальные результаты.

2. Обзор литературы

Задача определения скрытых демографических атрибутов пользователей интернета по текстам их сообщений сводится к классической задаче социолингвистики: определению характерных особенностей языка представителей различных социальных групп, позволяющих производить частичную идентификацию человека по принадлежности к этим группам

(автороведческая экспертиза). Вместе с тем, существует ряд отличий рассматриваемой задачи, связанных с электронной сообщений пользователей:

- ограниченная длина - вследствие ограничения на длину сообщений и/или стремления пользователей сэкономить время длина сообщений редко превышает нескольких сотен символов;
- неформальный стиль - в сообщениях часто встречаются нестандартные аббревиатуры, слэнг, неологизмы, пользователи сознательно пренебрегают классическими нормами орфографии и пунктуации, что можно расценивать как следствие общей тенденции переноса повседневного неформального общения в социальную сеть;
- использование специфических лингвистических конструкций - пользователи активно используют специальные синтаксические конструкции для быстрой категоризации сообщений (теги), добавления ссылок на других пользователей, описания собственных эмоций (эмотиконы) и т.д.;
- нестабильность качества пользовательского контента (большое количество спама и ложных аккаунтов пользователей) обуславливает необходимость в методах фильтрации зашумленных данных.

В связи с перечисленными особенностями и вследствие ограниченной применимости классических методов атрибуции текста к электронным сообщениям пользователей интернета, в последнее десятилетие было предложено множество методов, специализированных для сервисов мгновенных сообщений, электронной почты, форумов, блогов, социальных сетей и других источников текстовых сообщений пользователей. Эти методы, в основном, основаны на применении методов машинного обучения с учителем с целью классификации пользователей по лингвистическим и другим признакам в предопределённые классы, соответствующие различным значениям изучаемых атрибутов. Сообщения пользователя рассматриваются как набор символьных строк, из которых извлекаются признаки, а для разметки обучающей выборки применяются дополнительные источники данных о пользователе.

2.1 Задача определения гендерной принадлежности

Наиболее простым способом определения пола пользователя некоторого Интернет-ресурса является определение пола пользователя по имени, указанному в его профиле. Этот подход используется, например, в работе [1]. Недостатки подхода очевидны: вместо действительных имен пользователи могут указывать псевдонимы, также существует ряд имен, универсальных для обоих полов.

В современных исследованиях, посвященных задаче определения пола, активно используются методы машинного обучения. С точки зрения методов машинного обучения задачу определения пола интернет-пользователя можно

рассматривать как задачу бинарной классификации, где признаки объектов выбираются из размещаемой пользователями информации.

В исследовании [2] для гендерной классификации пользователей социальной сети Facebook помимо словарей имен использовались признаки, являющиеся значениями полей «interested_in» и «relationship_status» профилей пользователей, также учитывалось гендерное распределение друзей каждого пользователя.

Во многих исследованиях используются признаки, являющиеся N-граммами символов и слов сообщений пользователей. Так, исследования [3], [4] используют N-граммы символов длины от 1 до 5, извлекаемые из текстов сообщений, для определения пола пользователей социальной сети Twitter. Следует отметить, что при использовании N-грамм в качестве признаков необходимо осуществлять выбор наиболее репрезентативных N-грамм во избежание получения пространства признаков огромной размерности. Выбор наиболее репрезентативных N-грамм позволяет не только сократить время работы алгоритма, но и улучшить его производительность [4].

В [5] исследуется улучшение качества классификации пользователей сети Twitter при добавлении к признакам – N-граммам символов и слов, извлекаемым из текстов сообщений пользователей, N-грамм, извлекаемых из метаданных (полей «Screen name», «Full name», «Description» профиля пользователя). В [6] для классификации пользователей сети Facebook на основе текстов их статусов в качестве признаков в дополнение к N-граммам слов использовались темы, извлекаемые из текстов статусов при помощи тематического моделирования.

Также можно выделить группу структурных признаков:

1. признаки на основе символов (общее количество символов, общее количество букв, общее количество букв в верхнем регистре, общее количество цифр, общее количество пробелов и т.д.);
2. признаки на основе слов (общее количество слов, средняя длина слова в символах, общее количество различных слов, общее количество коротких слов – слов, длина которых меньше трех символов, количество слов, длина которых превосходит шесть символов, и т.д.);
3. признаки на основе предложений (общее количество различных символов препинания);
4. признаки на основе всего текста (общее количество предложений, общее количество абзацев, среднее число предложений в абзаце, среднее число слов в абзаце, среднее число слов в предложении и т.д.).

В исследовании [7] структурные признаки 1–3 были использованы для классификации пользователей сервиса Youtube. В исследовании [8] структурные признаки 1–4 использовались для гендерной классификации авторов новостных статей и электронных писем. Также, наряду со структурными признаками, в [8] использовались и социолингвистические признаки. В исследованиях [9] и [10] для гендерной классификации

пользователей социальных сетей наряду с признаками – N-граммами использовались социолингвистические признаки. В [11] социолингвистические признаки использовались для гендерной классификации авторов блогов.

В исследованиях [12], [13] для гендерной классификации пользователей сети Twitter использовались признаки, получаемые на основе первых k слов, стемм, хештегов, диграмм и триграмм, встречающихся в текстах пользователей обоих классов. В [12] также исследовалась влияние использования поля имени пользователя на качество работы алгоритма, в [13] – влияние окрестности пользователя в социальном графе.

2.2 Задача определения возраста

Существует три основных подхода к постановке задачи определения возраста. При первом подходе ставится задача определения возраста как непрерывной переменной.

При втором подходе ставится задача определения возрастной категории. Например, можно поставить задачу определения одной из следующих категорий: до 20 лет, от 20 до 40 лет, 40 лет и больше.

Также существует ряд задач, для которых важен не столько возраст пользователя, сколько жизненный этап, на котором он находится. Примерами жизненных этапов могут являться следующие: учащийся школы, студент университета, пенсионер.

Для решения задачи определения возраста широко используются методы машинного обучения. Так, например, в исследовании [14] задача определения возраста была поставлена как задача классификации с двумя классами: 40-, 40+; в исследовании [9] — как задача классификации с двумя классами: 30-, 30+; в исследованиях [15], [16] — как задача классификации с тремя классами: 13–17, 23–27, 33–42; в исследовании [17] — как регрессионная задача. В [18] регрессионные оценки возраста используются для классификации по возрастным категориям. В исследовании [19] проводится сравнение между тремя перечисленными выше подходами для задачи определения возраста пользователей социальной сети Twitter.

Исследования по определению возраста проводились для разнообразных типов данных: блогов ([15], [16]), записей телефонных раговоров ([14]), данных социальных сетей ([9], [19], [20]). Признаки, используемые для решения задачи определения возраста, можно разделить на две основные группы:

1. N-граммы слов и символов, получаемые из текстов сообщений ;
2. стилистические признаки (такие, как части речи, сленг, средняя длина предложения, пунктуация, акронимы, эмотиконы и т.д.).

Также используются признаки, специфические для источника данных. Например, в работе [21] для определения возраста пользователей платформы

LiveJournal использовались такие признаки, как количество друзей пользователя, количество постов, отображаемых на странице пользователя, общее количество постов пользователя, среднее число комментариев к посту пользователя.

2.3 Определение других атрибутов

В [22] описывается определение политической ориентации и этнической принадлежности пользователей сети Twitter с использованием признаков, извлекаемых из профиля пользователя (длина имени, количество букв в имени, количество цифр в имени и т. д.), особенностей поведения (общее количество сообщений, среднее число сообщений в день, среднее время между сообщениями и т. д.), текстового содержимого сообщений (наличие слов, характерных для классов, между которыми проводится классификация), окружения пользователя.

В [23] проводится сравнение двух подходов к определению политической ориентации пользователей сети Twitter. Первый из рассмотренных подходов основывался на классификации пользователей при помощи алгоритма машинного обучения «Метод опорных векторов», использовавшего признаки, извлекаемые из текстов сообщений пользователей; второй подход использовал алгоритм обнаружения сообществ.

В [24] для решения задачи определения географического положения пользователя социальной сети Twitter применяются методы тематического моделирования. В [25] описывается система, определяющая географическое положение пользователя сети Twitter на основе распределения слов по географическим локациям.

2.4 Выводы

Проведен обзор научного направления, связанного с проблематикой задачи, рассматриваемой в рамках данной НИР. Рассмотрены существующие подходы к решению задачи определения таких социо-демографических атрибутов пользователей Интернет-ресурсов, как гендерная принадлежность, возраст, политическая ориентация, этническая принадлежность и географическое положение.

Задача определения социо-демографических атрибутов успешно решается методами машинного обучения. При этом наиболее часто используются признаки, извлекаемые из текстов сообщений пользователей Интернет-ресурсов, т. к. этот тип признаков является универсальным практически для любого ресурса. Добавление специфических для ресурса признаков позволяет улучшить качество определения атрибутов.

Рассмотренные текстовые признаки можно разделить на две основные группы. Признаки первой группы являются независимыми от языка сообщений. В эту группу входят структурные признаки и N-граммы символов

и слов. Признаки второй группы в общем случае зависят от языка. Это различные социолингвистические признаки. Любая система, использующая для определения социо-демографических атрибутов зависящие от языка признаки, также является зависящей от языка.

Недостатки существующих подходов:

1. ограниченный набор поддерживаемых атрибутов пользователя (большинство методов ограничивается определением пола и возраста);
2. отсутствие или недостаточная функциональность автоматических средств для сбора корпусов сообщений пользователей интернета и разметки их с помощью социо-демографических атрибутов пользователей (сообщения собираются и размечаются, в основном, вручную, что накладывает ограничения на размер корпуса и достоверность разметки);
3. отсутствие или недостаточная функциональность методов фильтрации зашумленных недостоверных данных (спама и ложных аккаунтов пользователей);
4. недостаточное использование социолингвистических методов для определения признаков, специфических для отдельных атрибутов и их значений;
5. недостаточное обоснование выбора используемых методов машинного обучения (извлечение признаков, отбор высокоинформативных признаков, обучение, классификация), что накладывает ограничение на качество результатов;
6. отсутствие методов, позволяющих осуществлять моделирование пользовательских атрибутов не по отдельности, а в зависимости от других атрибутов, что накладывает ограничение на качество результатов;
7. немногочисленные открытые программные реализации и веб-сервисы ограничены по функционалу (определяются только пол и/или возраст пользователя), а также в силу недостаточного качества реализации непригодны для интегрирования с промышленными приложениями;
8. отсутствие открытых программных реализаций и/или веб-сервисов для русского языка.

3. Метод

Абсолютное большинство современных методов определения демографических атрибутов пользователей основаны на применении методов машинного обучения с учителем с целью классификации пользователей по лингвистическим и другим признакам в предопределённые классы, соответствующие различным значениям изучаемых атрибутов. Сообщения пользователя рассматриваются как набор символьных строк, из которых извлекаются признаки, а для разметки применяются дополнительные источники данных о пользователе, причём в большинстве случаев разметка производится вручную.

Разработанный метод обладает следующими преимуществами:

- автоматическое построение исходного набора данных;

- извлечение большого количества признаков различных типов из текстов сообщений пользователей Twitter;
- расширяемый набор поддерживаемых атрибутов: все поля Facebook-профиля, а также любая информация о предпочтениях и интересах пользователя могут быть использованы в качестве атрибутов;
- расширяемый набор поддерживаемых языков благодаря использованию автоматической идентификации языка текста сообщений и применению метода построения исходного набора данных, не зависимого от языка.

Метод состоит из следующих этапов:

- построение исходного набора данных;
- предварительная обработка текста;
- построение признакового описания;
- обучение;
- классификация.

Все этапы, за исключением первого, выполняются отдельно для каждого атрибута.

На этапе **построения исходного набора данных** производится сбор данных пользователей из сети Twitter. Для каждого пользователя сначала запрашивается только его профиль в сети Twitter. При наличии в нём ссылки на профиль того же пользователя в сети Facebook (в которой набор пользовательских атрибутов существенно больше, чем в Twitter) запрашиваются и сохраняются все доступные сообщения пользователя из сети Twitter. После чего для текущего пользователя запрашивается и сохраняется его профиль в сети Facebook, из которого извлекаются указанные пользователем значения его атрибутов.

Таким образом, элементом набора данных для каждого атрибута и языка является набор символьных строк, полученных из текстов сообщений и профиля одного пользователя в Twitter, а также значение атрибута у данного пользователя в Facebook.

На этапе **предварительной обработки текста** к текстам полученного на предыдущем этапе набора данных применяется метод определения языковой принадлежности текста (библиотека *language-detection*). После этого данные пользователей распределяются в различные наборы данных в зависимости от языка пользователя.

Предварительно осуществляется фильтрация сообщений, авторство которых не принадлежит пользователю (*ретвиты*). Поскольку цитирование сообщений других пользователей является весьма популярным способом распространения информации в сети Twitter, этот шаг предварительной обработки особенно важен для повышения точности метода.

На этапе **построения признакового описания** из сообщений пользователей извлекаются лингвистические признаки. Из полученных токенов строится набор признаков в виде N-грамм размером от 1 до 3 с учётом порядка токенов. Каждый тип признаков представлен двумя подтипами: с учётом и без учёта регистра символов.

Итоговый вектор признаков для пользователя является бинарным, то есть содержит только информацию о наличии или отсутствии признака в его текстовых данных. Количество экземпляров одного признака игнорируется.

На этапе **обучения** производится построение модели классификации с использованием алгоритма SVM (машины опорных векторов).

На этапе **классификации** в качестве входных данных используются тексты сообщений и поля профиля произвольного пользователя. Выполняется алгоритм классификация для заданного языка и атрибута. Результатом является значение атрибута выбранного пользователя.

4. Результаты экспериментов

Общая схема экспериментальных исследований для каждого демографического атрибута следующая:

1. Получить размеченные данные о пользователях, содержащие корректное значение исследуемого атрибута.
2. Обучить модель классификатора на части размеченных данных, а именно на случайном подмножестве пользователей мощностью 90% от исходного множества.
3. Классифицировать с помощью обученной модели остальную часть размеченных данных.
4. Оценить качество классификации путем сравнения значений исследуемого атрибута в размеченных данных и в результате классификации.

Размеченные данные о пользователях, содержащие корректное значение исследуемого социо-демографического атрибута, извлекаются на этапе сбора и разметки сообщений пользователей. Объем извлекаемых данных: 500 пользователей для каждого исследуемого социо-демографического атрибута и поддерживаемого языка. Таким образом, обучение происходит на 450 случайных пользователях, тестирование – на оставшихся 50 пользователях.

В качестве оценки качества используется общепринятая метрика точности (*Accuracy*).

В таблице 1 приведены результаты экспериментов при использовании N-грамм порядка три.

Таблица 1 Результаты экспериментальных исследований

Тестовая задача	**Язык**	**Точность, %**
Определение пола	английский	84
	русский	86
	испанский	94
	немецкий	88
	французский	94
	итальянский	82
Определение возраста	английский	94
	русский	92
	испанский	92
	немецкий	80
	французский	84
	итальянский	94
Определение семейного положения	английский	98
	русский	96
	испанский	98
	немецкий	94
	французский	98
	итальянский	94
Определение уровня образования	английский	92
Определение религиозных взглядов	английский	94
Определение политических взглядов	английский	82

5. Заключение

Непосредственной областью применения предложенного метода является интернет-маркетинг: повышение точности таргетированного продвижения товаров и услуг в интернете позволит повысить результативность рекламных кампаний и в конечном итоге увеличить прибыль производителей и посредников.

Использование метода в политических целях позволит собирать дополнительную информацию об избирателях и более эффективно расходовать средства рекламных кампаний.

Применение метода органами правопорядка позволит производить частичную де-анонимизацию преступников путём анализа их сообщений.

Ряд задач прикладного характера может быть решён напрямую с использованием разработанного метода, в частности:

- система рекомендаций товаров и услуг для интернет-магазина с учётом демографических профилей их пользователей, построенных путём анализа текстов их сообщений/отзывов;
- система рекомендаций пользователей для установления связей дружбы/следования в социальных сетях с учётом похожести демографических профилей различных пользователей, построенных путём анализа текстов их сообщений;
- система рекомендаций телепередач с учётом демографический профилей пользователей средств массовой информации, построенных путём анализа их отзывов и сообщений о различных СМИ.

Также разработанный задел открывает возможность для проведения исследований по различным направлениям в области обработки персональных данных пользователей сети Интернет, например, в ходе анализа социальных сетей (поиск сообществ и кластеризация социального графа), персонализации информационного поиска (вывод более точного поискового результата), упрощение решения задач компьютерной лингвистики (устранение многозначности и поиск омонимов во время исследования сообщений пользователей).

Список литературы

[1]. **Sloan L.** Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. [Текст] / L. Sloan [et al.] – Sociological Research Online. – 2013. – Т. 18. – №. 3. – p. 7.

[2]. **Tang C.** What's in a name: A study of names, gender inference, and gender behavior in facebook. [Текст] / C. Tang [et al.] – Database Systems for Adanced Applications. – Springer Berlin Heidelberg, 2011. – pp. 344–356.

[3]. **Miller Z.** Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. [Текст] / Z. Miller, B. Dickinson, W. Hu – International Journal. – 2012. – Т. 2.

[4]. **Deitrick W.** Gender identification on twitter using the modified balanced winnow. [Текст] / W. Deitrick [et al.] – Communications and Network. – 2012. – Т. 4. – №. 3. – pp. 189–195.

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

[6]. **Schwartz H. A.** Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. [Текст] / H. A. Schwartz [et al.] – PloS one. – 2013. – Т. 8. – №. 9. – p. 73791.

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

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

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

[10]. **Rao D.** Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. [Текст] / D. Rao [et al.] – ICWSM. – 2011.

[11]. **Mukherjee A.** Improving gender classification of blog authors. [Текст] / A. Mukherjee, B. Liu – Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 207–217.

[12]. **Liu W.** What's in a Name? Using First Names as Features for Gender Inference in Twitter. [Текст] / W. Liu, D. Ruths – 2013 AAAI Spring Symposium Series. – 2013.

[13]. **Al Zamal F.** Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. [Текст] / F. Al Zamal, W. Liu, D. Ruths – ICWSM. – 2012.

[14]. **Garera N.** Modeling latent biographic attributes in conversational genres. [Текст] / N. Garera, D. Yarowsky – Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – Association for Computational Linguistics, 2009. – Vol. 2, pp. 710–718.

[15]. **Schler J.** Effects of Age and Gender on Blogging. [Текст] / J. Schler [et al.] – AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. – 2006. – pp. 199–205.

[16]. **Goswami S.** Stylometric analysis of bloggers' age and gender. [Текст] / S. Goswami, S. Sarkar, M. Rustagi – Third International AAAI Conference on Weblogs and Social Media. – 2009.

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

[18]. **van Heerden C.** Combining regression and classification methods for improving automatic speaker age recognition. [Текст] / C. van Heerden [et al.] – Acoustics Speech

and Signal Processing (ICASSP), 2010 IEEE International Conference on. – IEEE, 2010. – pp. 5174–5177.
[19]. Nguyen D. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. [Текст] / D. Nguyen [et al.] – Seventh International AAAI Conference on Weblogs and Social Media. – 2013.
[20]. Peersman C. Predicting age and gender in online social networks. [Текст] / C. Peersman, W. Daelemans, L. Van Vaerenbergh – Proceedings of the 3rd international workshop on Search and mining user-generated contents. – ACM, 2011. – pp. 37–44.
[21]. Rosenthal S. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. [Текст] / S. Rosenthal, K. McKeown. – Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Association for Computational Linguistics, 2011. – Vol. 1, pp. 763–772.
[22]. Pennacchiotti M. A Machine Learning Approach to Twitter User Classification. [Текст] / M. Pennacchiotti, A. M. Popescu – ICWSM. – 2011.
[23]. Conover M. D. Predicting the political alignment of twitter users. [Текст] / M. D. Conover [et al.] – Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). – IEEE, 2011. – pp. 192–199.
[24]. Eisenstein J. A latent variable model for geographic lexical variation [Текст] / J. Eisenstein [et al.] – Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 1277–1287.
[25]. **Cheng Z.** You are where you tweet: a content-based approach to geo-locating twitter users. [Текст] / Z. Cheng, J. Caverlee, K. Lee – Proceedings of the 19th ACM international conference on Information and knowledge management. – ACM, 2010. – pp. 759–768.
[26]. **Al Zamal F., Liu W., Ruths D.** Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors //ICWSM. – 2012.
[27]. **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.
[28]. **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.
[29]. **Kótyuk G., Buttyán L**. A machine learning based approach for predicting undisclosed attributes in social networks //Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on. – IEEE, 2012. – С. 361-366.
[30]. **Caruana, G**. A survey of emerging approaches to spam filtering [Текст] / G. Caruana, M. Li // ACM Computing Surveys (CSUR), Vol. 44, No.2, February 2012. pp. 1-27.
[31]. **Stafford, G.** An Evaluation of the Effect of Spam on Twitter Trending Topics [Электронный ресурс] — Электрон. дан. - США, [2013] — Режим доступа: http://homepages.gac.edu/~lyu/Grant_paper.pdf, свободный. — Англ.
[32]. **Martinez-Romo, J.** Detecting malicious tweets in trending topics using a statistical analysis of language [Текст] / J. Martinez-Romo, L. Araujo // Expert Systems with Applications, Vol. 40, No.8, June 2013. pp. 2992-3000.

[33]. **Almeida, T. A.** Advances in spam filtering techniques. In Computational Intelligence for Privacy and Security [Текст] / T. A. Almeida, A.Yamakami // Computational Intelligence for Privacy and Security, Vol. 394, 2012. pp. 199-214.
[34]. **Wang, A. H.** Machine Learning for the Detection of Spam in Twitter Networks [Текст] / A. H. Wang // e-Business and Telecommunications, Vol. 222, 2012. pp. 319-333.
[35]. **Ahmed, F.** Generic Statistical Approach for Spam Detection [Текст] / F.Ahmed, M. A. Abulaish // Computer Communications, Vol. 36, June 2013. pp. 1120-1129.
[36]. **Thomas, K.** Suspended Accounts in Retrospect: An Analysis of Twitter Spam [Текст] / K. Thomas, C. Grier, V. Paxson, D. Song // Proceedings of the Internet Measurement Conference 2011 (IMC 2011) , Berlin, Germany, November 2-4. 2011. pp. 243-258.
[37]. **Sridharan, V.** Twitter games: how successful spammers pick targets [Текст] / V. Sridharan, V. Shankar, M. Gupta // Proceedings of the 28th Annual Computer Security Applications Conference, Orlando, Florida, USA, December 3-7. 2012. pp. 389-398.
[38]. **Левенштейн, В.И.** Двоичные коды с исправлением выпадений, вставок и замещений символов [Текст] / В.И. Левенштейн // Доклады Академий Наук СССР, 1965, Т. 163, №4. С. 845-848.
[39]. **Lin P.C.** A study of effective features for detecting long-surviving Twitter spam accounts [Текст] / P.C. Lin, P.M. Huang // The 15th International Conference on Advanced Communications Technology, Phoenix Park, PyeongChang, South Korea, January 27-30. 2013. pp. 841 - 846
[40]. **Романов А.С., Мещеряков Р.В.** *Определение пола автора короткого электронного сообщения* // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 25–29 мая 2011 г.). М. : Изд-во РГГУ, 2011. Вып. 10 (17). С. 620–626

Detection of demographic attributes of microblog users

Anton Korshunov, Ivan Beloborodov, Andrey Gomzin, Christina Chuprina, Nikita Astrakhantsev, Yaroslav Nedumod, Denis Turdakov
{korshunov, ivbel, gomzin, chuprina, astrakhantsev, yaroslav.nedumov, turdakov}@ispras.ru
ISP RAS, Moscow, Russia

Abstract. Users of internet services often make errors or intentionally provide misleading information about their demographic attributes, including gender, age, marital status, education, religious and political views. At the same time, knowing values of user attributes allows to enhance the performance of recommender systems, internet marketing solutions, and other applications based on personalized results. In the paper, a method is proposed for automatic detection of demographic attributes of Twitter users by analyzing their textual messages and other data from their profiles. The method is based on a machine learning algorithm trained with binary vectors of token N-grams extracted from user posts. Its distinctive features are fully automatic compilation of training and testing data sets as well as support for a broad and extendable range of languages and demographic attributes. This is achieved by exploiting Facebook accounts associated with user profiles in Twitter. Additional steps are detecting language of posts and filtering borrowed content. Experimental study showed high accuracy of gender, age, and marital status detection for the most popular languages: English, Russian, German, French, Italian, and Spanish. Besides, detection of education, religious and political views is also supported for English.

Keywords: demographic characteristics; demographic attributes; social networks; microblogs; natural language processing; content analysis; computational linguistics; machine learning.

References

[1]. Sloan L. et al. Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. Sociological Research Online. – 2013. – Т. 18. – №. 3. – p. 7.
[2]. Tang C. et al. What's in a name: A study of names, gender inference, and gender behavior in facebook. Database Systems for Adanced Applications. – Springer Berlin Heidelberg, 2011. – pp. 344–356.
[3]. Miller Z., Dickinson B., Hu W. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. International Journal. – 2012. – Т. 2.
[4]. Deitrick W. Gender identification on twitter using the modified balanced winnow. Communications and Network. – 2012. – Т. 4. – №. 3. – pp. 189–195.
[5]. 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. – pp. 1301–1309.
[6]. Schwartz H. A. et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PloS one. – 2013. – Т. 8. – №. 9. – p. 73791.
[7]. 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. – pp. 1478–1488.
[8]. Cheng N., Chandramouli R., Subbalakshmi K. P. Author gender identification from text. Digital Investigation. – 2011. – Т. 8. – №. 1. – pp. 78–88.
[9]. 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. – pp. 37–44.
[10]. Rao D. et al. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. ICWSM. – 2011.
[11]. Mukherjee A., Liu B. Improving gender classification of blog authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 207–217.
[12]. Liu W. Ruths D. What's in a Name? Using First Names as Features for Gender Inference in Twitter. 2013 AAAI Spring Symposium Series. – 2013.
[13]. Al Zamal F., Liu W., Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. ICWSM. – 2012.
[14]. Garera N., Yarowsky D. Modeling latent biographic attributes in conversational genres. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – Association for Computational Linguistics, 2009. – Vol. 2, pp. 710–718.
[15]. Schler J. et al. Effects of Age and Gender on Blogging. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. – 2006. – pp. 199–205.
[16]. Goswami S., Sarkar S., Rustagi M. Stylometric analysis of bloggers' age and gender. Third International AAAI Conference on Weblogs and Social Media. – 2009.
[17]. 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. – pp. 115–123.
[18]. van Heerden C. et al. Combining regression and classification methods for improving automatic speaker age recognition. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. – IEEE, 2010. – pp. 5174–5177.
[19]. Nguyen D. et al. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. Seventh International AAAI Conference on Weblogs and Social Media. – 2013.
[20]. 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. – pp. 37–44.
[21]. Rosenthal S. McKeown K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Association for Computational Linguistics, 2011. – Vol. 1, pp. 763–772.
[22]. Pennacchiotti M., Popescu A.M. A Machine Learning Approach to Twitter User Classification. ICWSM. – 2011.
[23]. Conover M. D. Predicting the political alignment of twitter users. Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). – IEEE, 2011. – pp. 192–199.

[24]. Eisenstein J. et al. A latent variable model for geographic lexical variation Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 1277–1287.
[25]. Cheng Z., Caverlee J., Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. Proceedings of the 19th ACM international conference on Information and knowledge management. – ACM, 2010. – pp. 759–768.
[26]. Al Zamal F., Liu W., Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. ICWSM. – 2012.
[27]. 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. – S. 37-44.
[28]. 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. – S. 1301-1309.
[29]. Kótyuk G., Buttyán L. A machine learning based approach for predicting undisclosed attributes in social networks. Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on. – IEEE, 2012. – S. 361-366.
[30]. Caruana, G., Li M. A survey of emerging approaches to spam filtering. ACM Computing Surveys (CSUR), Vol. 44, No.2, February 2012. pp. 1-27.
[31]. Stafford G., Yu L. L. An Evaluation of the Effect of Spam on Twitter Trending Topics. Social Computing (SocialCom), 2013 International Conference on. – IEEE, 2013. – C. 373-378.
[32]. Martinez-Romo, J., Araujo L. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, Vol. 40, No.8, June 2013. pp. 2992-3000.
[33]. Almeida T. A., Yamakami A. Advances in spam filtering techniques. Computational Intelligence for Privacy and Security. – Springer Berlin Heidelberg, 2012. – C. 199-214.
[34]. 34. Wang, A. H. Machine Learning for the Detection of Spam in Twitter Networks. e-Business and Telecommunications, Vol. 222, 2012. pp. 319-333.
[35]. Ahmed, F., Abulaish M.A. Generic Statistical Approach for Spam Detection. Computer Communications, Vol. 36, June 2013. pp. 1120-1129.
[36]. Thomas, K., Grier C., Paxson V., Song D. Suspended Accounts in Retrospect: An Analysis of Twitter Spam. Proceedings of the Internet Measurement Conference 2011 (IMC 2011) , Berlin, Germany, November 2-4. 2011. pp. 243-258.
[37]. Sridharan V., Shankar V., Gupta M. Twitter games: how successful spammers pick targets. Proceedings of the 28th Annual Computer Security Applications Conference, Orlando, Florida, USA, December 3-7. 2012. pp. 389-398.
[38]. Levenshtejn V.I. Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshhenij simvolov [Binary codes with correction for deletions, insertions, and substitutions of characters]. Doklady Akademij Nauk SSSR [The Proceedings of the USSR Academy of Sciences], 1965, T. 163, №4. C. 845-848. (in Russian)
[39]. Lin P.C., Huang P.M. A study of effective features for detecting long-surviving Twitter spam accounts. The 15th International Conference on Advanced Communications Technology, Phoenix Park, PyeongChang, South Korea, January 27-30. 2013. pp. 841 — 846
[40]. Romanov A.S., Meshheryakov R.V. Opredelenie pola avtora korotkogo ehlektronnogo soobshheniya [Gender identification of the author of a short message]. Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoj Mezhdunar. konf. «Dialog» (Bekasovo, 25–29 maya 2011 g.). [Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue” (Bekasovo, May 25-29, 2011)] Moscow.: RGGU, 2011.Issue 10 (17) – Moscow, RGGU, 2011. pp. 620–626. (in Russian)