Методология построения социально-демографических профилей пользователей в Интернете

Аннотация. В статье проводится анализ методов построения социально-демографических профилей пользователей сети Интернет, которые распределяются по группам в зависимости от соответствующих характеристик. К подобным атрибутам относят пол, возраст, семейное положение, фактическое место проживания, а также религиозные и политические взгляды юзеров. Обзор и классификация были проведены на основании сведений, полученных из профилей и переписок пользователей социальных сетей и прочих Интернет-ресурсов. Значительное число исследований проводилось для определения пола. Помимо данного критерия, особое внимание уделялось установлению возраста, региональной принадлежности, а также религиозным мировоззрениям. Основная масса решений основывается на применении методов машинного обучения с учителем. В работе детально разобран каждый этап решения: сбор информации, установление признаков, вычленение информативных характеристик, способы обучения классификаторов и качественная оценка.

Ключевые слова: демографические параметры; демографические атрибуты, социальные сети; обработка текстовых сообщений; анализ текстов на естественном языке; машинное обучение

Введение

В Интернете существует огромное количество возможностей для создания пользователями персонального контента – ресурсы различного рода: социальные сети, блоги и форумы. Помимо этого, основная масса интернет-магазинов, а также различные новостные сайты и прочие аналогичные сервисы своим посетителям предоставляют возможность делиться отзывами либо оставлять комментарии. Через подобные сообщения и оценки пользователь распространяет определенные личные сведения, а именно разглашает свое имя, пол, возраст, предпочтения и даже контакты. Такого рода информация является набором демографических атрибутов и вносится в соответствующие поля, предусмотренные отдельной страницей ресурса, именуемой профилем пользователя.

Таким образом, контент, который был сформирован юзером, напрямую отражает его предпочтения и интересы. К примеру, по характеру лексики возможно определить статус автора будь то подросток или человек в возрасте. К контенту относят не только анкетные данные, но и тексты, картинки, аудио- и видеоматериалы, размещенные на странице пользователя. В данной статье для анализа используется лишь текстовая информация контента.

При написании работы преследовалась главная задача – составить социально-демографический профиль интернет-пользователей. Однако главная трудность заключается в том, что не все юзеры достоверно заполняют анкетные данные, либо вносят их частично. В этой связи возникает задача угадать скрытые социально-демографические характеристики (пол, возраст) по указанной в профиле информации. Зачастую решить задачу позволяют публичные сведения, почерпнутые из текста комментариев, отзывов и прочих сообщений, предоставленных в контенте.

Для выделения конкретных групп пользователей широко применяются методы автоматического определения демографических признаков юзеров, что эффективно даже в тех случаях, когда указаны не все значения. Полученные таким путем данные можно использовать в рекомендательных системах, а также для целей таргетированной рекламы и в прочих приложениях [1, 2].

Основная масса исследований в данном направлении основывается на методах машинного обучения. При этом цель достигается посредством реализации следующих шагов:

- сбора информации для моделирования – изучается исходный материал, рассматриваются особенности его получения;
- создания (обучения) модели – описания решения, где не применяется машинное обучение;
- классификацией на базе составленной модели – решением задачи с использованием методов машинного обучения;
- оцениванием качества модели – формируются заключительные выводы.

Сведения

Больше всего исследователей интересует стремительно развивающиеся социальные сети (Фейсбук, Твиттер и прочие). Самой крупной из них считается Facebook – насчитывает свыше 1,2 млн. зарегистрированных пользователей, в профилях которых содержится информация о различных демографических признаках [3]. Классификаторы, сформированные на основе общедоступного контента, позволяют наиболее точно предсказать значения характеристик, неуказанных в профиле остальных юзеров.

Такой же популярностью у исследователей пользуется вторая по популярности социальная сеть – Twitter. Представляет собой сервис микроблоггинга (длина сообщений менее 140 знаков). Однако сбор информации на данном ресурсе усложняется тем, что профиль пользователя этой социальной сети не содержит демографические атрибуты [4].

Помимо данных площадок, в некоторых работах исследователи анализируют сведения, полученные из электронных писем, а также оценивают информацию, собранную на новостных сайтах и YouTube [5, 6]. В данной статье представлен обзор работ, в которых происходит определение таких признаков, как пол, возраст, место фактического проживания, политические убеждения.

Сбор информации

Сбор информации – это первый этап выявления демографических атрибутов. Информация содержит истинное значение характеристик, а также текстовый контент юзеров. Значения признаков удается определить при помощи алгоритмов машинного обучения и качественного оценивания.

Социальная сеть рассматривается в качестве графа, вершины которого – пользователи, а ребра – связь между ними (дружба или подписка). Ставится задача обойти его вершины для получения показательной подборки. Данный процесс именуется семплингом. Как свидетельствуют проведенные исследования [7, 8], наиболее информативные выборки удается получить при задействовании алгоритмов семплирования «Лесного пожара», а также Метрополиса-Гастингса.

При установлении характерных признаков в подборку попадают лишь те пользователи, в профиле которых приведена информация о демографических атрибутах. Получить сведения данного рода в социальных сетях можно при помощи сервиса поиска друзей. К примеру, в такой социальной сети, как Вконтакте, можно указать интересующее значение при поиске. В некоторых случаях демографические признаки удается установить лишь экспертам [9-12].

Ни на каждом Интернет-ресурсе удается получить нужный атрибут из профиля юзера (например, Twitter). Определить интересующее значение можно только в том случае, если известно наличие профиля данного человека в другой социальной сети. Это возможно благодаря полю URL в профиле, содержащего ссылку на страницу этого пользователя с

другого ресурса. Перейдя по гиперссылке удается отыскать значения целевых характеристик. Данный способ изучается в работах [13 и 14].

Однако может возникнуть и такая ситуация, при которой получить желаемую информацию практически невозможно – это связано с техническими ограничениями, которые существуют на тех или иных площадках. К примеру, API Twitter одному приложению разрешает осуществлять до 300 запросов на получение страницы сообщений юзеров в течение 15 минут. Некоторые платформы имеют скрытые ограничения. Такого плана как блокировка запросов в случае превышения установленного лимита. При этом допустимую частоту можно определить лишь экспериментальным путем.

Методы, при которых не применяется машинное обучение

Зачастую для выявления демографических признаков прибегают к методу машинного обучения, однако, в определенных случаях цель достигается и при задействовании более простых подходов. К примеру, пол юзера легко установить исходя из его имени. В работе [15] были использованы специальные словари, содержащие все возможные имена. Однако результативность данного способа главным образом зависит от достоверности информации, полученной из контента – не все указывают свое подлинное имя. Чтобы определить другие характеристики, нужны более основательные подходы. В основной массе таких исследований применяется машинное обучение с учителем.

Машинное обучение с учителем и без

Данный способ дает возможность отследить взаимосвязь целевых значений и исходных сведений, результаты чего в дальнейшем используются для предположения целевого признака для новых сведений. В данном случае к исходной информации относят сведения о юзере, а к целевой – демографические критерии. В большинстве социальных сетей можно получить информацию о характере взаимоотношений ее пользователей – именуется социальной связью. В исследованиях [5, 9] социальные связи применяются также для установления классификационных атрибутов.

Существует машинное обучение с учителем и без него. Во втором случае закономерности отыскиваются в исходной информации, полученные сведения в дальнейшем классифицируются по группам. Что касается обучения с учителем, то в этом случае для построения алгоритма нужны также и характеристики целевых признаков. Сформированная модель позволяет определять целевые значения для тех исходных сведений, в которых характеристики целевых показателей невозможно установить. Зачастую имеется подборка с теми атрибутами, значения которых непременно представлены.

Применение машинного обучения заключается в следующих этапах:

1) Получение признаков.
2) Опциональное вычленение признаков.
3) Обучение модели.
4) Качественная оценка моделирования.

Получение значений атрибутов

Рассмотрим признаки, применяемые при обучении классификаторов юзеров по демографическим характеристикам. Исходная информация, применяемая для установления значений признаков, включает в себя сведения профиля и переписку пользователя. Однако принцип машинного обучения базируется на признаковых характеристиках объекта исследования. При этом объекты выступают в качестве комплекса векторов в пространстве

характеристик. В том случае, если используется машинное обучение с учителем, объект также включает в себя значение целевого признака.

В первую очередь, следует установить какие признаки можно получить из собранной информации. Поскольку текст построен из слов, за признак принимаются конкретные слова и их последовательность (то есть n-граммы, где n – длинна последовательности). За основу принимается бинарное значение: 1 – искомое слово встречалось в тексте, 0 – нет.

В качестве признака при установлении пола или возраста пользователя могут также использоваться части речи в определенной последовательности (POS n-граммы). Текст также может рассматриваться как последовательность некоторых символов – за признак принимают символьные n-граммы. Популярен также подход, рассматривающий в качестве признака последовательности фонем (единица звука). Значение при этом играет именно произношение слов.

В отдельный класс признаков можно выделить статистические, а именно их числовые значения, вычлененные из текста. На пример, в качестве статистического признака может использоваться размер одного сообщения юзера, частотность знаков препинания и прочее. Зачастую анализируется профиль пользователя целиком. Так для определения пола юзера используется его имя, и даже цвет страницы социальной сети.

Полезным для предсказания атрибутов является и социальный граф. К примеру, с его помощью через соответствующие алгоритмы могут устанавливаться определенные сообщества (их совокупность), в которых состоит пользователь. Эта информация в дальнейшем и принимается за признак [16].

Из исходных данных можно получить любое количество комбинаций признаков, однако, применение большого числа атрибутов не всегда может быть эффективным. Качество классификаторов может меняться обратнопропорционально. В таком случае происходит переобучение.

Одного действенного метода при определении демографических признаков не существует – выбор тех или иных атрибутов главным образом зависит от поставленной задачи и имеющихся в распоряжении данных. Рассмотрим несколько примеров. Так в работе [5] устанавливается пол посетителей канала YouTube на основании их комментариев и взаимосвязи пользователя с видео – анализируется факт просмотра дынным юзером конкретного видео. При этом используются также статистические характеристики (средний размер комментария), возраст и пол, словесные n-граммы.

В работах [12,13 и 14] авторы анализируют пол пользователей Twitter. В первом и во втором случае признаки извлекаются из анализа текстов сообщений на базе словесных и символьных n-грамм, а в третьем случае – атрибут устанавливается исходя из информации профиля, в том числе цвета страницы пользователя (текста, ссылок), для решения задачи используются также имена, преобразованные в некоторые последовательности фонем.

В исследовании [17] авторы работали над определением возраста пользователей, которые используют для общения голландский язык. Их возраст разбивался на несколько промежутков, а за признак принимали словесные и символьные 1,2 и 3-граммы.

Возраст и пол юзеров – признаки, которые наиболее часто применяются при исследовании. Однако есть и такие труды, в которых для анализа применяются другие атрибуты, к примеру, место проживания или политические предпочтения. В работе [9] проводится анализ политических взглядов юзеров Твиттера. К рассмотрению принимались такие классы, как республиканцы, демократы и неопределенная категория. В этом случае признаками служили хэштеги, сообщества юзеров и прочее.

В данной статье были разобраны лишь некоторые труды, посвященные выявлению демократических характеристик. Основная масса работ базируется на применении n-граммы для определения признаков, полученных из текстов сообщений юзеров. При этом существует огромное количество характеристик, которые могут быть использованы. Для проведения наиболее эффективного исследования возникает задача отбора из имеющегося числа атрибутов наиболее информативных.

Уменьшение количества наиболее информативных признаков

При формировании комплекса признаков, которые извлекаются из текста, каждый отдельных юзер рассматривается через совокупность присущих ему атрибутов. Если подсчитать количество возможных признаков, то их число будет значительно превышать количество самих пользователей. К примеру, в исследовании [13] подборка насчитывать порядка 180 тыс. юзеров, что касается признаков, то в выборке содержится около 15 млн. атрибутов. В таких условиях может возникнуть переобучение. При этом классификатор может неверно предсказывать результат для новых сведений. Выйти из сложившейся ситуации можно с помощью сокращения количества характеристик. Не следует допускать включения большого числа атрибутов в выборку еще на этапе подбора. В свою очередь, когда дело касается анализа текста сообщений, обойтись без этапа эффективного подбора информативных признаков невозможно.

Существуют такие методы отбора информативных атрибутов, которые не применяют характеристики целевых признаков. К подобным способам относят фильтрацию атрибутов по частоте. Данный метод заключается в том, что для каждого признака рассчитывается количество случаев, в которых данный атрибут присутствует, а в дальнейшем отбираются те, которые характеризуются наибольшей частотой (в свою очередь, редкие показатели не используются). При способе с высокой дисперсией удаляются те атрибуты, характеристики которых незначительно меняются у объектов. Данные способы не берут во внимание значения целевых признаков. Информативность же атрибута оценивается в плане его целевой характеристики. К примеру, такой признак, как «окончание имени» напрямую связан с полом юзера – зачастую женские имена оканчиваются на гласную букву.

Кое-какие алгоритмы машинного обучения характеризуются встроенной возможностью выбора информативных атрибутов из общего числа. Количество признаков в модели при отборе определяет регуляризация (например, LASSO). Ее смысл заключается в том, что сложность модели сводится на нет вместе с вероятностью ошибок при тестировании.

Приведенные выше способы позволяют уменьшить величину исходной информации посредством устранения малоинформативных атрибутов. Существуют также и другие подходы к сокращению размерности сведений, они предполагают изменение признакового пространства в полной мере.

Что касается текстовой информации, то для нее применимы способы тематического моделирования [19] – каждый текст распределяется на темы. При этом исходные сведения можно представить в виде матрицы (объект на признаки).

Проблеме выбора информативных признаков посвящены исследования [18, 20, 21].

Применяемые алгоритмы машинного обучения

Демографические атрибуты зачастую состоят из нескольких основных компонентов. Так признак пола может принимать лишь два значения – мужской или женский, а возраст имеет свои числовые пределы. Следовательно, основная задача заключается в выборе подхода – будь то классификация или регрессия.

Наивный байесовский классификатор – наиболее простой из всех алгоритмов, который предполагает, что все атрибуты независимы друг от друга. Данный способ классификации основывается на теореме Байеса. Его применение для выявления пола можно рассмотреть в исследованиях [12-14]. Польза этого классификатора заключается также в том, что его можно использовать при онлайн-обучении (при добавлении информации модель обновляется, а не пересчитывается).

Если необходимо распределить объекты на две категории – применяется линейный классификатор, наиболее популярным алгоритмом которого является метод опорных

векторов (не онлайновый). Идея способа заключается в поиске разделяющей гиперплоскости с наиболее значительным зазором до объектов групп. В работах [6,10,13,18] был использован данный метод.

Качественная оценка

Эффективность того или иного алгоритма можно определить при помощи оценки его качества. С этой целью измеряют такие его качества, как достоверность, точность, полнота и F-мера.

Точность позволяет установить, для каких объектов классификатор предсказал верное решение. Полнота и достоверность анализируются в пределах единой группы, именуемой как положительная. F-мера – объединение показателя достоверности и полноты.

Зачастую качество алгоритма определяется на основе кросс-валидации – совокупность размеченных сведений разбивается на категории, для каждой из которых осуществляется обучение на оставшихся. Проверка реализуется на определенной группе сведений. Вариация характеристик усредняется по всей совокупности данных.

Заключение

В статье были рассмотрены основные методы формирования социально-демографического профиля юзеров Интернет-ресурсов. Анализ работ, посвященных данной теме, позволяет сделать вывод относительно того, что основная масса исследований посвящена изучению сообщений пользователей ресурсов. Twitter представляет наибольший интерес для авторов, поскольку профиль его пользователей не отражает демографические признаки в полной мере, а сообщения имеют свою специфику.

Множество трудов посвящены определению пола пользователей, есть также работы, в которых устанавливается их возраст, место проживания и политические предпочтения.

Для достижения поставленных задач зачастую применяют метод машинного обучения с учителем, этапы которого подробно рассмотрены в данной статье.

Для оптимизации алгоритмов можно обозначить следующие направления: установление всех признаков в совокупности (с учетом их взаимосвязи) и изучение возможности разработки классификаторов, которые не зависели бы от источника исходных сведений.

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.
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 usergenerated 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.

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.