Определение демографических характеристик пользователей сайта Last.fm на основе анализа их музыкальных интересов

Семёнов А. С., магистрант кафедры КТ университета ИТМО,
semkagtn@gmail.com

Аннотация

В данной работе представлено два подхода к определению пола и возраста пользователей сайта Last.fm[1] на основе прослушиваемых ими музыкальных исполнителей. Достигнутые результаты: точность определения пола — 82,46%, средняя абсолютная ошибка определения возраста — 3,38.

Введение

В современное время миллионы людей по всему миру пользуются социальными сетями и другими онлайн-сервисами. Кроме того, пользователи оставляют о себе огромное количество информации в открытом доступе. Благодаря этому появляется возможность определять «скрытые» характеристики пользователей на основе имеющейся информации. Такими характеристиками, например, являются пол и возраст. Многие сайты предоставляют возможность указывать эти параметры, но часто это не является обязательным требованием. Отсюда возникает задача устранения неполноты в данных. Восстановление недостающих характеристик пользователей позволяет улучшить различные системы рекомендаций [1, 2].

Существует множество исследований, показывающих возможность определять пол и возраст на основе косвенных признаков. В работе [3] исследовалась зависимость между текстом блогеров и их возрастом. Авторы работ [4, 5, 6] предлагают подходы к определению демографических характеристик на основе сообщений пользователей. Текст — не единственная информация, которая может быть использована. Например, в работе [7] учитывалось также поведение пользователей. А в работе [8] для определения демографических характеристик использовались также фотографии и геолокация пользователей.

Использование информации о музыке, которую слушают пользователи, также возможно. В статье [9] была показана корреляция между предпочтением определённых жанров и полом у группы студентов. В работе [10] пол и возраст определялся по истории прослушиваний музыкальных композиций.

[1] http://www.last.fm

В исследовании [11] используется 50 наиболее прослушиваемых композиций каждого пользователя сайта Last.fm.

В рамках настоящей работы предложены подходы к определению демографических характеристик пользователей на данных, которые использовались в последней упомянутой статье. Предполагается, что описанные методы могут помочь улучшить существующие алгоритмы, не использующие информацию о музыке пользователей.

Задача определения пола решалась как задача классификации, а задача определения возраста — как задача восстановления регрессии.

Описание набора данных

Набор данных содержит 96807 пользователей. Каждый пользователь описан 50 исполнителями из списка наиболее прослушиваемых им музыкальных композиций. Эти данные были получены с сайта Last.fm при помощи метода API *User.getTopTracks*[2], который возвращает самые прослушиваемые музыкальные композиции заданного пользователя. Музыкальные композиции отсортированы в порядке убывания по числу прослушиваний. Из этого списка были извлечены исполнители (с сохранением порядка). Таким образом, исполнители, описывающие пользователя, могут повторяться. У каждого пользователя указан пол и число полных лет.

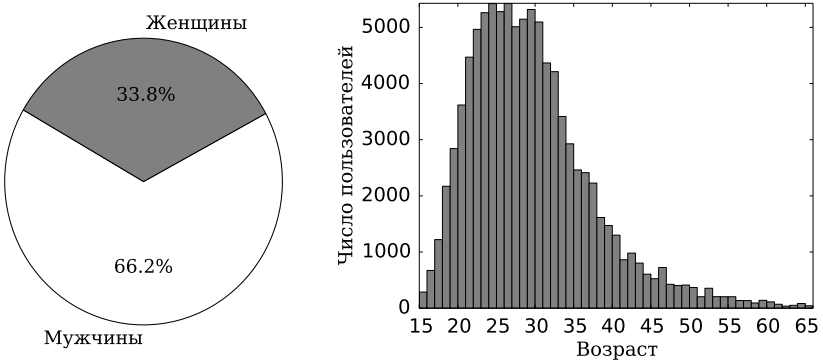
Следует отметить некоторые особенности выборки. Во-первых, мужчин в ней содержится больше, чем женщин (см. Рис. 1a). Во-вторых, возраст пользователь сильно смещён в сторону молодого поколения (см. Рис. 1b).

Выборка разбита на обучающую и контрольную (48404 и 48403 пользователей соответственно) таким образом, чтобы распределение полов в каждой выборке было одинаковым.

Преобразование данных в векторное пространство

В данном разделе описаны подходы, позволяющие преобразовать пользователей в численные векторы, пригодные для использования в качестве признакового описания для решения задачи классификации или восстановления регрессии.

[2] http://www.last.fm/api/show/user.getTopTracks



(a) Пол (b) Возраст

Рис. 1: Распределение демографических характеристик в выборке

Подход на основе матрицы термин–документ

Пользователей можно рассматривать как «документы», в которых терминами являются исполнители. Отсюда на основе подхода ”bag of words” [12] возникает матрица $D = \{d_{ij}\}$, где элемент d_{ij} обозначает степень принадлежности термина i документу j. В качестве функции степени принадлежности часто используется формула TF-IDF:

$$d_{ij} = \mathrm{tf}_{ij} \cdot \log \frac{n}{\mathrm{df}_i}, \tag{1}$$

где tf_{ij} — число встреч термина i в документе j, df_i — число документов, в которых встречается термин i, n — общее число документов. Формула TF-IDF, может быть обобщена следующим образом:

$$d_{ij} = \begin{cases} 0, & \mathrm{tf}_{ij} = 0, \\ l(\mathrm{tf}_{ij}) \cdot g(\frac{n}{\mathrm{df}_i}), & \mathrm{tf}_{ij} \neq 0, \end{cases} \tag{2}$$

где $l(x)$ и $g(x)$ — неубывающие и неотрицательные функции.

Матрица термин–документ описывает каждый документ через вектор терминов. Размерность векторов в нашей задаче оказывается крайне большой — 96891, что обуславливает необходимость применить метод снижения размерности. Ввиду сильной разреженности матрицы для этой цели был выбран метод LSI [13], основанный на сингулярном разложении, а именно его реализация из библиотеки gensim[3]. Произвольным образом была выбра-

[3] https://radimrehurek.com/gensim/

на размерность равная 200.

Таким образом, каждый пользователь описан 200 числовыми признаками. Стоит отметить, что вектор каждого пользователя был нормализован таким образом, чтобы его длина была равной единице.

Подход на основе модели Word2Vec

Модель Word2Vec [14] позволяет преобразовать словарь терминов в численные векторы одинаковой размерности. Таким образом, каждый документ можно представить в виде последовательности численных векторов. Основываясь на предположении о том, что чем выше исполнитель находится в рейтинге пользователя, тем более «значимым» он является, можно построить вектор, представляющий пользователя, следующим образом:

$$d_i = \sum_n f(n) \cdot w_{in}, \tag{3}$$

где w_{in} — термин (исполнитель), который находится у пользователя i на позиции с номером n; $f(n)$ — невозрастающая неотрицательная функция на промежутке $[1; 50]$ (так как каждый документ состоит не более чем из 50 терминов).

Использовалась реализация модели Word2Vec из библиотеки *gensim* с параметрами: *size* = 200 и *window* = 25.

Таким образом, каждый пользователь описан вектором, размерность которого равна 200. Стоит отметить, что векторы пользователей были преобразованы так, чтобы каждая координата имела математическое ожидание равное нулю и среднеквадратическое отклонение равное единице.

Этап обучения и результаты

Для решения задач классификации или регрессии использовался метод опорных векторов (Support Vector Machine) [15], а именно его реализация из библиотеки scikit-learn[4]. В качестве функции ядра использовалось ядро *RBF*, а параметры *gamma* и *C* настраивались на обучающей выборке методом GridSearchCV, который реализован в упомянутой библиотеке. Значения параметра *C* были взяты из множества $\{0{,}1; 0{,}5; 1; 3\}$, значения параметра *gamma* из — $\{0{,}001; 0{,}01; 0{,}1; 1; 8; 16; 32; 64\}$. Выбор параметров основан на рекомендации из статьи [16], где авторы рекомендуют выбирать последовательности, которые растут нелинейно.

[4] http://scikit-learn.org

Качество классификатора или регрессора определялось на контрольной выборке. Метрикой для задачи определения пола была выбрана точность, а для задачи определения возраста — средняя абсолютная ошибка.

Результаты с использованием подхода на основе матрицы термин–документ при различных функциях $l(x)$ и $g(x)$ из уравнения 2 приведены в таблице 1.

$l(x)$	$g(x)$	**Определение пола**	**Определение возраста**
1	1	**82,46%**	**3,54**
$\log x$	1	81,58%	3,64
$\log x$	$\log x$	81,67%	3,57
$\log x$	$\sqrt{x}$	80,04%	3,76
$\sqrt{x}$	1	81,73%	3,63
$\sqrt{x}$	$\log x$	81,76%	3,57
$\sqrt{x}$	$\sqrt{x}$	79,78%	3,77
x	1	79,45%	3,92
x	$\log x$	79,67%	3,85
x	$\sqrt{x}$	78,75%	3,88

Таблица 1: Результаты с использованием подхода термин–документ

Результаты с использованием подхода на основе модели Word2Vec при различных функциях $f(n)$ из уравнения 3 приведены в таблице 2.

$f(n)$	**Опредление пола**	**Определение возраста**
1	78,12%	3,97
$\frac{1}{\log n+1}$	79,61%	3,59
$\frac{1}{n^2}$	78,66%	4,02
$\frac{1}{n}$	80,22%	3,75
$\frac{1}{\sqrt{n}}$	**81,46%**	**3,38**
$51-n$	77,85%	4,03
$\log 51 - \log n$	78,16%	3,95
$\sqrt{51} - \sqrt{n}$	78,12%	3,98

Таблица 2: Результаты с использованием подхода на основе модели Word2Vec

Сравнение лучших результатов, достигнутых в рамках настоящего исследования, (*best*) и полученных ранее в работе [11] (*baseline*) приведено в

таблице 3.

Тип задачи	best	baseline
Определение пола	**82,46%**	78,87%
Определение возраста	**3,38**	3,69

Таблица 3: Сравнение результатов с результатами, полученными ранее

Следует отметить, что размерность пространства признакового описания была выбрана произвольно. Изменение этой величины может сильно влиять на результат.

Заключение

В рамках данной работы были представлены подходы к определению демографических характеристик пользователей сайта Last.fm. Полученные результаты несколько превосходят результаты достигнутые ранее.

Результаты могут быть улучшены путём увеличения размера «сетки» значений, по которой настраивались параметры метода опорных векторов.

Развитием настоящей работы может послужить апробирование предложенных подходов на других наборах данных и в других задачах.

Литература

[1] Swearingen K., Sinha R. Beyond algorithms: An HCI perspective on recommender systems // ACM SIGIR 2001 Workshop on Recommender Systems. – 2001. – Vol. 13(5–6). – P. 1-11.

[2] Adomavicius G., Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions // IEEE Transactions on Knowledge and Data Engineering. – 2005. – Vol. 17(6). – P. 734-749.

[3] Burger J. D., Henderson J. C. An Exploration of Observable Features Related to Blogger Age // AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. – 2006. – P. 15–20.

[4] 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 (SMUC2011). – 2011. – P. 37–44.

[5] Турдаков Д. и др. определение демографических атрибутов пользователей микроблогов // Труды Института системного программирования РАН. – 2013. – Т. 25. – С. 179–192

[6] Schwartz H. A. et al. Personality, gender, and age in the language of social media: The open-vocabulary approach // PloS one. – 2013. – Vol. 8(9). – P. e73791.

[7] 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-Volume 1. – Association for Computational Linguistics, 2011. – P. 763–772.

[8] Farseev A. et al. Harvesting multiple sources for user profile learning: a big data study // Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. – 2015. – P. 235–242.

[9] Christenson P. G., Peterson J. B. Genre and gender in the structure of music preferences // Communication Research. – 1988. – Vol. 15(3). – P. 282–301.

[10] Liu J. Y., Yang Y. H. Inferring personal traits from music listening history // Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. – 2012. – P. 31–36.

[11] Wu M. J., Jang J. S. R., Lu C. H. Gender Identification and Age Estimation of Users Based on Music Metadata // ISMIR. – 2014. – P. 555–560.

[12] Manning C. D. et al. Introduction to information retrieval. – Cambridge : Cambridge university press, 2008. – Vol. 1(1). – P. 412–415.

[13] Deerwester S. et al. Indexing by latent semantic analysis // Journal of the American society for information science. – 1990. – Vol. 41(6). – P. 391.

[14] Mikolov T. et al. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. – 2013.

[15] Suykens J. A. K., Vandewalle J. Least squares support vector machine classifiers // Neural processing letters. – 1999. – Vol. 9(3). – P. 293–300.

[16] Hsu C. W. et al. A practical guide to support vector classification // Tech. rep., Department of Computer Science, National Taiwan University – 2003. – P. 1–16.