DOI: 10.46548/CIT-2021-0034-0014

А.И. Мартышкин, А.А. Зоткина

К ВОПРОСУ ПРОФИЛИРОВАНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

Россия, г. Пенза, Пензенский государственный технологический университет

The article discusses the issues of profiling users of social networks. Currently, according to statistics, more than half of the population of our country uses information technologies. Entertainment and interactive services such as social networks, blogs, forums, etc. have become particularly widespread among ordinary users. Such services are characterized by providing a lot of freedom to the user in entering various data about himself: the user can name himself as he likes, specify any place of residence, create a group with any description. The user creates the necessary content himself, searches for the necessary information himself, therefore, the search should be available to anyone, even someone unfamiliar with the skills of searching for data on a computer. Naturally, with the increase in the number of users, the availability of data decreases, it is necessary to implement more and more sophisticated search algorithms that allow people to find the information they need.

As a rule, the search represents to the user a string in which he enters a query to which the program outputs the results found. The user is free to enter any data as a request, intentionally or accidentally distorting them, for example, by making typos. Users who create new content also often make inaccuracies in the names, distortions in words, add icons for decoration, write words in Latin. Thus, there may be errors in the data and in the queries, which must be taken into account when implementing the search.

The article provides an overview of existing methods of fuzzy data retrieval. A review of existing studies has been conducted in which this problem has been analyzed.

Введение. Проблема профилирования пользователей представляет из себя задачу восстановления и определения неизвестных характеристик пользователей по известным. Часто вопрос профилирования решается с использованием данных из социальных сетей, но, кроме этого, существует возможность использовать данные из любых онлайн-сервисов, предоставляющих возможность регистрации и сохранения определенных данных о конкретном пользователе. В качестве неизвестных характеристик пользователей, которые могут быть интересны и которые необходимо восстановить, могут выступать такие демографические характеристики, как пол и возраст. Обычно такая задача решает проблему устранения неполноты в данных: как правило, многие сайты дают возможность вновь регистрируемым пользователям заполнять не все поля в различных анкетах. Устранение неполноты в данных о демографических характеристиках пользователей позволяет улучшить различные системы рекомендаций ввиду того, что подобные признаки порой являются ключевыми в рекомендательных системах [1, 2]. Задача восстановления пола и возраста пользователей обычно сводится к классическим задачам машинного обучения – к задаче классификации или задаче восстановления регрессионной зависимости.

Обзор существующих исследований. Неизвестные характеристики пользователей, которые требуется определить, могут быть очень нетипичными. Например, авторы работы [63] предлагают подход, позволяющий выявлять психотип пользователей, основываясь на данных из их страниц-профилей в социальных сетях. Такие исследования очень востребованы, например, в психологии, т.к. метод определения психологического портрета пользователя даёт возможность выявлять новые

закономерности и строить новые гипотезы относительно поведения людей. Подобные задачи в основном сводятся к задачам классификации, кластеризации или регрессии.

Ещё одним широко распространённым типом задач профилирования является выделение групп пользователей. В работе [4] предлагается подход, позволяющий определять политические взгляды пользователей, путём разбиения их на группы, что может быть использовано в необходимых целях. Авторы работы [5] приводят метод, позволяющий произвести кластеризацию пользователей, основываясь на определенных метриках, что часто оказывается полезным при реализации систем рекомендаций [6].

В качестве информации о пользователях, на основе которой восстанавливаются неизвестные характеристики может быть использована совершенно любая доступная информация. На просторах социальных сетей и онлайн-сервисов доступно огромное множество разноплановых данных, таких как текстовые сообщения пользователей, фотографии, социальные связи, пользовательская активность в течение дня, дата регистрации и др.

Пожалуй, главная сложность, возникающая при решении конкретной задачи профилирования, заключается в сведении ее к задаче машинного обучения (классификации, восстановления регрессии, кластеризации и т.д.).Требуется представить каждого пользователя в виде вектора, пригодного для применения его алгоритмом машинного обучения, причём так, чтобы алгоритм показывал хорошие результаты. Метод построения векторной модели, описывающей пользователей, зависит от типа используемых данных. Случается, что процесс преобразования данных некоторого вида в вектора оказывается весьма нетривиальным шагом. Рассмотрим известные подходы, позволяющие преобразовать пользователей в векторы, основываясь на персональных данных пользователей различного вида. Одним из наиболее часто использующихся видов данных, применяемых для определения характеристик пользователей, является текст, который пользователи пишут в социальных сетях.

В работах [7, 8] применяется подход к построению векторов, описывающих пользователей, на основе подсчёта числа буквенных N-грамм, встречающихся в сообщениях. Похожий подход применен в работе [9], но вместо буквенных N-грамм использовались словесные. Обоснование данного подхода базируется на том, что текст имеет некоторую семантику, которая может быть описана некоторыми частями данного текста. Часто такой подход приводит к большому числу признаков, из-за чего возникает задача выбора наиболее информативных из них.

В работах [3, 10] в качестве известных данных при решении задачи профилирования также применялся текст сообщений, но признаковое описание пользователей формировалось на основе подсчёта количества слов, принадлежащих некоторым непересекающимся семантическим группам. Главным недостатком указанного подхода является то, что составление таких групп слов в основном приходится осуществлять вручную.

В работе [11] предлагается подход к определению пола пользователей Twitter на основе их имён. Авторы работы используют словарь, в котором имеются сведения о женских и мужских именах. Здесь кроме очевидного недостатка, заключающегося в том, что словарь имён необходимо составлять вручную, данный подход, кроме всего, игнорирует тот факт, что имена пользователей социальных сетей зачастую вымышленные (как правило являются псевдонимами). Не забудем еще и то, что существует ряд имён, универсальных для обоих полов.

Авторы работы [12] использовали несколько типов данных, описывающих пользователей. Текст пользовательских сообщений преобразовывался в численные признаки с помощью латентного семантического анализа, плюс ко всему признаки из

текста извлекались программным средством LIWC [13]. Также признаки строились на основе геолокации пользователей и фотографий.

В [14] описан способ, позволяющий определять пол пользователей на основе их фотографий. В качестве численных признаков, описывающих пользователей, используются разности значений яркости всех пар пикселей, из которых состоит изображение пользователей. Данный подход имеет недостатки: во-первых, сложность алгоритма; во-вторых, большая размерность пространства признакового описания; в-третьих, необходимость применения алгоритмов распознавания лиц, точность которых ниже точности распознавания лиц людьми.

Далее следует рассмотреть подходы к решению вопросов, связанных с профилированием пользователей, в которых использовалась информация о музыкальных предпочтениях ввиду того, что большинство пользователей социальных сетей неравнодушны к определенной музыке, которую можно встретить на пользовательских страницах. Первой работой, которую стоит отметить, является [15], причем она вовсе не относится к анализу социальных сетей, но в исследовании изучались музыкальные предпочтения людей в зависимости от их половой принадлежности. В рамках проведенного и описанного в работе исследования проводился опрос среди 239 студентов. Людям предлагалось оценить каждую музыкальную композицию из определенного списка по пятибалльной шкале. После проведенного исследования было показано, что существует корреляция между предпочитаемыми человеком музыкальными жанрами и его полом. Результат исследований указанной статьи даёт все основания полагать, что музыкальные предпочтения конкретного человека содержат в себе определенную «скрытую» информацию о нём.

В исследовании, описанном в работе [16], для определения пола и возраста пользователей использовалась информация с сайта Last.fm. Опытным путем опробовано три способа векторного представления пользователей: на основе даты и времени прослушивания, на основе метаданных о музыке, на основе анализа аудиофайлов музыкальных композиций. Первый способ предполагает вычисление ряда признаков для каждого пользователя, среди которых число прослушиваний в каждый из 24 часов суток, число прослушиваний в каждый из семи дней недели, число прослушиваний в каждый из 12 месяцев в году, доля прослушиваний в рабочее время. Способ, основанный на метаданных о музыке, предполагает подсчёт числа прослушиваний каждой музыкальной композиции, каждого исполнителя, а также подсчёт числа встреч меток исполнителей и композиций. Способ, основанный на анализе аудиофайлов музыкальных композиций, предполагает вычисление признаков с помощью Echo Nest API [17] таких как темп, громкость и т.д. В ходе исследований показано, что наилучшее качество определения пола и возраста пользователей достигается при использовании такой метаинформации о музыке, как название исполнителя и название композиции. Основываясь на результатах исследования, можно сделать вывод, что названия музыкальных композиций и исполнителей являются наиболее информативными данными, поэтому использование этой информации является наиболее целесообразным.

В исследовании [18] для решения задачи определения демографических характеристик пользователей в качестве источника информации также использовался сайт Last.fm. Для пользователей сохранялась информация о 50 наиболее прослушиваемых каждым из них музыкальных композициях. Основываясь только на названиях исполнителей из этих списков, вычислялось векторное представление для каждого пользователя. Применялось два подхода, первый из которых заключается в том, что список исполнителей рассматривается как текстовый документ, далее вычисляется матрица «термин-документ», а затем применяется техника латентного семантического анализа для уменьшения размерности признакового описания. Второй

подход основывается на вычислении gaussian super vector [19]. Авторы исследования решали задачу определения пола как задачу бинарной классификации, а задачу определения возраста – как задачу восстановления регрессии. Наилучшие результаты, достигнутые в исследовании: точность определения пола – 78,87%, средняя абсолютная ошибка определения возраста – 3,69. Результаты данного исследования подтверждают, что в музыкальных предпочтениях пользователей всё-таки содержится «скрытая» информация, описывающая их. Таким образом, каждый пользователь описан последовательностью, каждый элемент последовательности содержит название исполнителя. Стоит отметить, что при использовании такой информации, названия исполнителей у одного и того же пользователя могут повторяться. Следует указать на некоторые особенности задачи в такой постановке. Данная задача очень похожа на задачи профилирования, использующие текстовые данные. С одной стороны, список исполнителей не является полноценным текстом, следовательно семантика, присущая текстовым сообщениям, в данной задаче отсутствует. Поэтому подходы на основе N-грамм или на основе выделения частей речи в данном случае не применимы. С другой стороны, названия исполнителей у каждого пользователя упорядочены по вполне определённому принципу, поэтому здесь имеет место гипотеза о значимости порядка следования исполнителей.

Выводы. В статье рассмотрена проблема профилирования пользователей. Проведен обзор существующих исследований, в рамках которых эта проблема анализировалась.

1. Swearingen K.,Sinha R.Beyond algorithms: An HCI perspective on recommender systems // ACM SIGIR 2001 Workshop on Recommender Systems. Т. 13. Citeseer. 2001. PP. 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 // Knowledge and Data Engineering, IEEE Transactions on. 2005. Т. 17, №6. PP. 734–749.

3. Personality, gender, and age in the language of social media: The open- vocabulary approach / H. A. Schwartz [и др.] // PloS one. 2013. Т. 8, №9. e73791.

4. Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber? / P. Barbera [и др.] // Psychological science. 2015. PP. 1531–1542.

5. Maia M., Almeida J., Almeida V. Identifying user behavior in online social networks // Proceedings of the 1st workshop on Social network systems. ACM. 2008. PP. 1–6.

6. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering / B. M. Sarwar [и др.] // Proceedings of the fifth international conference on computer and information technology. Т. 1. Citeseer. 2002.

7. Miller Z., Dickinson B., Hu W. Gender prediction on Twitterusing stream algorithms with N-gram character features // International Journal of Intelligence Science. 2012. Т. 2, 4A. PP. 143–148.

8. Gender identification on Twitter using the modified balanced winnow / W. Deitrick [и др.] // Communications and Network. 2012. Т. 4, №3. PP. 189–195.

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

10. 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. PP. 763–772.

11. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter / L. Sloan [и др.] // Sociological research online. 2013. Т. 18, №3. P. 7.

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

13. Pennebaker J. W.,Francis M. E.,Booth R. J.Linguistic inquiry and word count: LIWC 2001 // Mahway: Lawrence Erlbaum Associates. 2001. Т. 71. PP. 1–11.

14. Baluja S., Rowley H. A. Boosting sex identification performance // International Journal of computer vision. 2007. Т. 71, №1. PP. 111–119.

15. Christenson P. G., Peterson J. B. Genre and gender in the structure of music preferences // Communication Research. 1988. Т. 15, №3. PP. 282–301.

16. 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. ACM. 2012. PP. 31–36.

17. The Echo Nest · GitHub. – URL: https://github.com/echonest (Дата обращения: 05.11.2021).

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

19. Campbell W. M., Sturim D. E., Reynolds D. A. Support vector machines using GMM supervectors for speaker verification // Signal Processing Letters, IEEE. 2006. Т. 13, №5. PP. 308–311.