Байесовская сеть доверия для представления и обработки данных и знаний о профориентационном типе личности и его цифровых предикторах

В. Ф. Столярова[1], А. О. Хлобыстова[2], М. В. Абрамов[3]

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

[1]vfs@dcsc.pro, [2]aok@dscs.pro, [3]mva@dscs.pro

Аннотация. **Создание автоматизированных систем оценки личностных особенностей является актуальной задачей в рамках профориентационной деятельности. Такие системы служат для поддержки принятия решений в этом направлении, в том числе в условиях ограниченности ресурсов и нехватки экспертов. Одним из ключевых источников данных для таких систем выступают онлайн социальные медиа. В работе для представления и обработки данных и знаний о профориентационном типе личности и его цифровых предикторах предлагается байесовская сеть доверия. Взаимосвязи между узлами сети установлены при помощи анализа статистических данных с привлечением экспертов.**

Ключевые слова: модель Голланда; цифровые следы; байесовская сеть доверия; особенности личности; профориентация

I. Введение

В настоящее время индивидуальный человеческий капитал все чаще выделяют в качестве определяющего фактора для достижения целей организации [2, 6]. При этом удовлетворенность выбранной профессией и условиями труда является одним из его ключевых показателей. Отмечается [8], что удовлетворенность во многом обусловлена соответствием работы интересам и способностям сотрудника, что определяет важность корректного и своевременного карьерного консультирования и профориентации. Кроме того, индивидуальный человеческий капитал может быть одним из факторов успешной реализации социоинженерной атаки [17]. Выявление черт личности, которые связаны с его/ее профессиональными интересами требует проведения опросов и привлечения экспертов, что является ресурсоемким процессом. Таким образом, актуальной является задачи разработки автоматизированных рекомендательных систем в сфере профориентации, а также их математического и алгоритмического обеспечения, которые используют данные из различных источников.

Так, например, подобные системы часто опираются на цифровые следы пользователей информационных систем. Определение профессиональных предпочтений обучающихся с использованием цифровых следов относят к интеллектуальному анализу образовательных данных [14], который возник в рамках цифровой трансформации этой отрасли и общества в целом. Отмечается, что значительная часть студентов выбирает направленность ступеней высшего образования, не имея достаточных знаний о своих возможностях и интересах [12], что приводит к необходимости смены траектории или неудовлетворенности выбранной профессией в будущем.

Существуют и разрабатываются различные автоматизированные рекомендательные системы, которые используют подобные данные. В их основе часто используют данные профилей специализированных информационных систем, которые посвящены выбору профессии [7].

Существует три класса методов, которые используются в качестве математической и алгоритмической основ таких систем.

1. Машинное обучение [4, 7, 9, 12]
2. Глубокое обучение [3, 7]
3. Экспертные знания [7]

Однако существует нехватка подобных систем на основе данных из онлайн социальных сетей русскоязычного сегмента Интернет, таких как VK. Целью исследования является определение структуры вероятностной графической модели данных и знаний о взаимосвязи цифровых следов пользователей онлайн социальной сети и их профессиональных предпочтений. Теоретическая значимость исследования заключается в анализе зависимости между определенными особенностями личности и характеристиками ее/его цифровой самопрезентации, выявлении ключевых признаков. Практическая значимость исследования заключается в предлагаемой структуре сети, которая может служить основой для автоматизированных рекомендательных систем в сфере профориентации.

II. Методы исследования

Существуют несколько способов определения профессиональных предпочтений личности [8], одним из наиболее часто употребляемых является модель Голланда, предложенная в 1997 году [5]. В рамках этой

Работа выполнена при финансовой поддержке РНФ, проект № 23-21-0033

модели выделяют шесть групп профессиональных интересов личности: реалистичный (R), исследовательский (I), артистичный (A), социальный (S), предприимчивый (E), and конвенциональный (C). В рамках профориентационных мероприятий определяется, насколько выражены черты каждой из этих групп, что позволяет сформировать список профессий, которые соответствуют личности. Для оценки таких личностных особенностей существую различные опросные инструменты. В исследовании была использована адаптация и локализация, разработанная Г.В. Резапкиной [15]. В рамках этого тестирования респонденту предлагается в сорока вопросах выбрать одну из двух профессий, которую он считает наиболее близкой себе; при этом каждый из выборов относится к одному из шести типов профессиональных интересов. Согласно полученным баллам, типы могут быть выражены ярко (7–10 баллов) или же слабо (0–3 балла).

Однако в сфере, связанной с анализом личностных особенностей, поступающие на вход системы обработки информации данные и знания часто сопряжены с неопределенностью, поэтому важно учитывать знания экспертов в таких системах. Поэтому в качестве основной модели, позволяющей использовать статистические данные и экспертную информацию, были использованы байесовские сети доверия, которые отражают декомпозицию совместного вероятностного распределения, соответствующую структуре зависимостей переменных, представленной направленным циклическим графом [11, 16].

Для достижения цели исследования при помощи специально разработанного приложения VK Mini Apps – «Психологические тесты»[1] были собраны данные о 1237 прохождения теста Голланда и данные со странички профиля: информация о количестве друзей, подписчиков и подписок, сообществ, медиа (аудио и видео записей), изображений. Кроме того были определены данные о тематиках подписок пользователей, которые являются неотъемлемой характеристикой сообществ в онлайн социальной сети VK. При формировании набора данных рассматривались только открытые профили. В предшествующих работах было приведено описание взаимосвязи между ключевым, наиболее выраженным значением кода Голланда и наиболее часто встречающейся тематикой подписок пользователя [10, 19].

III. Структура знаний о цифровых следах пользователей и цифровых следах

Основу для байесовской сети доверия составили переменные, представленные в табл. 1. Так как в онлайн социальной сети VK сообщества могут быть посвящены более чем 100 тематикам, то для использования при анализе данных, они были укрупнены до 10. Отметим, что не все тематики встречались среди ведущих для пользователей из собранного набора данных.

Для отдельного индивида код Голланда представляет собой набор шести взаимосвязанных упорядоченных значений (от наиболее выраженных к наименее выраженным). При характеризации профессиональных особенностей личности зачастую важен анализ местоположения каждого из шести возможных типов. В исследовании рассматривались наиболее и наименее выраженные типы профессиональной направленности в коде Голланда.

Предварительный анализ показал, что распределения вероятности численных характеристик профиля

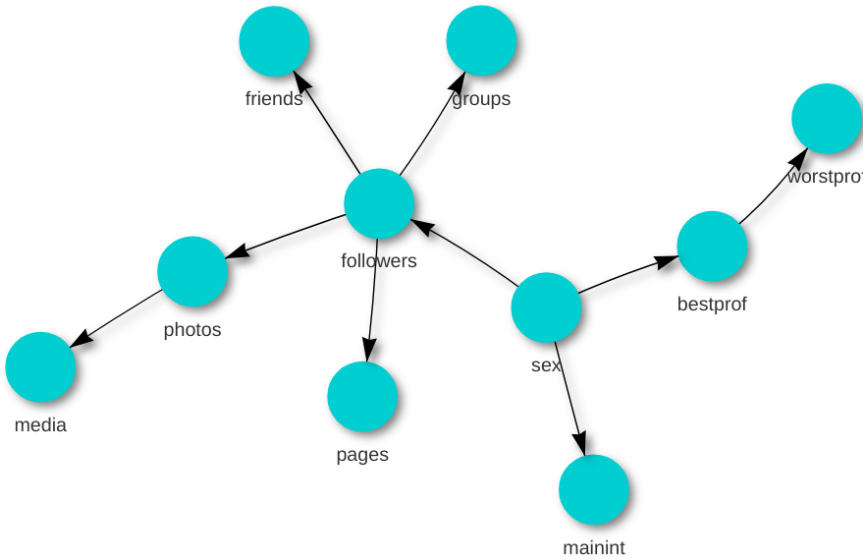

Рис. 1. Обученная при помощи алгоритма Hill Climbing структура байесовской сети

пользователя скошены влево, потому для приведения к симметричному виду использовалось логарифмическое преобразование.

[1] Приложение «Психологические тесты» [Электронный ресурс]. Режим доступа: vk.com/app7794698 (дата обращения: 15.04.2024).

Финальным этапом предобработки данных являлась дискредитация непрерывных переменных, которые были разбиты на шесть интервалов на основе квантилей соответствующего распределения.

ТАБЛИЦА I. ОПИСАНИЕ ПЕРЕМЕННЫХ, ИСПОЛЬЗУЮЩИХСЯ ДЛЯ ОПРЕДЕЛЕНИЯ СТРУКТУРЫ БАЙЕСОВСКОЙ СЕТИ ДОВЕРИЯ

Переменная	**Описание**
sex	Пол, указанный в профиле пользователя 1 —- женский, 2 —- мужской
pages	Логарифм количества интересных страниц для пользователя
followers	Логарифм количества подписчиков
groups	Логарифм количества сообществ
friends	Логарифм количества друзей
photos	Логарифм количества изображений
media	Логарифм количества аудио и видеозаписей на страничке пользователя
mainint	Наиболее часто встречающаяся тематика сообществ пользователя 1. Cities (Города и страны; Туризм) 2. Computer Internet (Компьютер и интернет) 3. Culture (Культура; Музыка) 4. Education (Образование; Финансы) 5. Household (Рецепты; Ремонт; Животные; Услуги) 6. Lifestyle (Красота и здоровье; Отношения; Медицина; Рестораны) 7. Massmedia (СМИ; Персоны; Социальные организации) 8. People Groups (Группы по интересам) 9. Shops (Магазины) 10. Sport (Спорт; Авто)
bestprof	Наиболее выраженный тип профессиональной направленности R, I, A, S, E, C – фактор со значением, которое отвечает коду Голланда
worstprof	Наименее выраженный тип профессиональной направленности (код Голланда) R, I, A, S, E, C – фактор со значением, которое отвечает коду Голланда

Статистическое моделирование проводилось в среде обработки данных R с использованием пакет bnlearn [13]. Рис. 1 был создан с использованием пакета visNetwork [1].

Для получения структуры сети использовался классический алгоритм Hill Climbing (восхождение к вершине).

IV. ОБСУЖДЕНИЕ

Отметим, что переменные в полученной структуре байесовской сети доверия разбиты на блоки. К первому блоку относятся количественные характеристики профиля пользователя онлайн социальной сети, ключевой из которых является количество подписчиков (followers). Эта переменная является связующей для остальных в этом блоке: медиаконтент (media и photo), число друзей, число групп и интересных страниц. При этом эти переменные не обладают сильной корреляционной зависимостью.

Второй смысловой блок описывает личность и предпочтения пользователя онлайн социальной сети (переменные mainint и sex). Третий блок относится к наиболее и наименее выраженным типам в коде Голланда.

Таким образом, можно сделать вывод, что тематики подписок пользователя играют важную роль при определении профессионального типа личности на основе цифровых следов пользователей.

Отметим, что сила взаимосвязи между переменными sex и followers невелика (табл. 2), что свидетельствует о необходимости дальнейших исследований по выявлению отражений личностных особенностей в цифровых следах [18].

ТАБЛИЦА II. СИЛА ДУГ В СТРУКТУРЕ БАЙЕСОВСКОЙ СЕТИ ДОВЕРИЯ, ОТРАЖАЮЩАЯ ИНФОРМАЦИОННЫЕ ПОТЕРИ ПРИ ОТСУТСТВИИ РАССМАТРИВАЕМОЙ ДУГИ

Начало дуги	**Конец дуги**	**Сила**
bestprof	worstprof	-235.7
sex	mainint	-155.9
photos	media	-106.6
followers	friends	-99.8
followers	photos	-95.4
Sex	bestprof	-43.4
followers	pages	-93.1
followers	groups	-39.1
sex	followers	-18.0

Среди ограничений проведенного исследования можно указать нехватку знаний о генеральной совокупности, из которой была взята выборка, в силу особенностей формирования датасета. Приложение, на основе которого собирались данные, было доступно любому пользователю, имеющих аккаунт в онлайн социальной сети VK; для исследования были взяты только открытые аккаунты. Это ограничение не умаляет результатов исследования, и может быть смягчено обращением к экспертным знаниям. За счет этого предложенная в исследовании структура байесовской сети доверия может применяться для различных приложений. Кроме того, примененный алгоритм обучения структуры сети не позволяет определить направления взаимосвязи, поэтому необходимо привлечение дополнительного моделирования и/или экспертных знаний.

V. ЗАКЛЮЧЕНИЕ

Таким образом, в исследовании на основе статистических данных была определена структура байесовской сети доверия, отражающая взаимосвязь между профессиональными предпочтениями индивидов (на основе модели Голланда), индикаторами его личностных особенностей (тематики подписок на сообщества в онлайн социальной сети) и характеристиками активности пользователя. Полученная структура сети позволяет выявить смысловые блоки, которые могут использоваться при разработке автоматизированных рекомендательных систем в сфере профориентации на основе цифровых следов.

Среди дальнейших направлений исследований можно выделить более детальное изучение информации, представленной в сообществах, на которые подписан пользователь. Эта задача опирается на методы обработки естественного языка.

Список литературы

[1] Almende B.V. and contributors, Thieurmel B. _visNetwork: Network visualisation using vis.js Library. [Электронный ресурс] https://cran.r-project.org/web/packages/visNetwork/index.html (дата доступа 15.04.2024)

[2] Bohórquez E., Caiche W., Benavides V., Benavides A. Motivation and job performance: human capital as a key factor for organizational success // Congress in Sustainability, Energy and City. Cham: Springer International Publishing, 2021. С. 123-133.

[3] Ghosh A., Woolf B., Zilberstein S., Lan A. Skill-based career path modeling and recommendation //2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020. С. 1156-1165.

[4] Guleria P., Sood M. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling //Education and Information Technologies. 2023. Т. 28. №. 1. С. 1081-1116.

[5] Holland J.L. Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources, 1997.

[6] Islam M.S., Amin M. A systematic review of human capital and employee well-being: putting human capital back on the track //European Journal of Training and Development. 2021. Т. 46. №. 5/6. С. 504-534.

[7] Herath G., Kumara B.T.G.S., Ishanka U.A.P., Rathnayaka R.M.K.T. Computer-Assisted Career Guidance Tools for Students' Career Path Planning: A Review on Enabling Technologies and Applications //Journal of Information Technology Education: Research. 2024. Т. 23. С. 006.

[8] Hoff K.A., Song Q.C., Wee C.J., Phan W.M.J., Rounds J. Interest fit and job satisfaction: A systematic review and meta-analysis //Journal of Vocational Behavior. 2020. Т. 123. С. 103503.

[9] Kamal A., Naushad B., Rafiq H., Tahzeeb S. Smart career guidance system //2021 4th International Conference on Computing & Information Sciences (ICCIS). IEEE, 2021. С. 1-7.

[10] Khlobystova A., Stoliarova V., Abramov M. Characterization of the Person's Leading Interests in Terms of RIASEC Scores //International Conference on Intelligent Information Technologies for Industry. Cham : Springer Nature Switzerland, 2023. С. 281-290

[11] Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniques. The MIT Press. 1264 с.

[12] Ochirbat A., Shih T. K., Chootong C., Sommool W., Gunarathne W.K.T.M., Wang H.H., Ma Z.H. Hybrid occupation recommendation for adolescents on interest, profile, and behavior //Telematics and Informatics. 2018. Т. 35. №. 3. С. 534-550.

[13] Scutari M., Denis J. Bayesian networs with examples in R. Chapman & Hall. 274 с.

[14] Курзаева Л.В., Савва Л И., Назарова Е.К., Абзалов А.Р., Килиевич Д.А. Анализ и обработка данных цифрового следа обучающихся //Мир науки. Педагогика и психология. 2022. Т. 10. №. 6. С. 72.

[15] Резапкина Г.В. Психология и выбор профессии: программа пред-профильной подготовки. Учебно-методическое пособие для психологов и педагогов. М.: Генезис, 2005. 208 с.

[16] Тулупьев А.Л., Николенко С.И., Сироткин А.В. Основы теории байесовских сетей. СПбГУ. 399 стр.

[17] Тулупьева Т.В., Абрамов М.В., Тулупьев А.Л. Модель социального влияния в анализе социоинженерных атак //Управленческое консультирование. 2021. №. 8 (152). С. 97-107.

[18] Олисеенко В.Д., Хлобыстова А.О., Корепанова А.А., Тулупьева Т.В. Автоматизация оценки темперамента пользователей онлайн социальной сети //Доклады Российской академии наук. Математика, информатика, процессы управления. 2023. Т. 514. №. 2. С. 235-241.

[19] Хлобыстова А.О., Абрамов М.В., Столярова В.Ф. Исследование тенденций взаимосвязи между профориентационными предпочтениями пользователей и их цифровыми следами в социальной сети //Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. №. 3. С. 564-574.