Автоматизированный сбор данных социальных сетей для разработки факторной модели сетевой самопрезентации

Б.А. Низомутдинов, А.С. Тропников, А.Б. Углова

***Аннотация* — На основе проведенного эмпирического исследования информационных образов пользователей были раскрыты ведущие компоненты сетевой самопрезентации: статистический, социально-демографический компонент, визуальный компонент и ценностно-смысловой компонент. Авторы провели анализ скрытых факторов, отвечающих за формирование сетевой самопрезентации посредством информационного образа, изучая прогностические возможности анализа данных социального профиля. Были выделены достоверно значимые различия в содержательном наполнении информационного образа и социально-психологических характеристиках у пользователей с разными стилями сетевой самопрезентации. Представлены алгоритмы сбора и обработки открытой информации из профилей социальных сетей, с последующим факторным анализом, а также методы машинного обучения, для определения тематик сообществ и интересных страниц, на которые подписан пользователи.**

В работе затрагивается этический и правовые вопросы использования сбора массива данных с групп пользователей для создания прогностических моделей на их основе, без уведомления самих пользователей. Также рассматриваются теоретические вопросы междисциплинарного проектирование модели информационного образа, которое становится возможном за счет многостороннего анализа знакового содержания профиля социальной сети: лингво-психологической оценки смыслового содержания контента, социально-психологического обзора коммуникативных практик, реализуемых в сети и разбора особенностей сетевого интерфейса, задающего структуру данному профилю.

***Ключевые слова* — информационные отпечатки, обработка данных, прогностическая модель, информационный образ.**

I. Введение

За последние годы, все большая часть социальных интеракций переносится человечеством на цифровые платформы. Благодаря интенсивному росту аудитории социальных сетей [1] и интернациональной политики предоставления доступного интернета [2], изменилась сама основа социальных наук: теперь вместо проведения экспериментальных исследований в лабораториях или выборочных тестирований на тысяче испытуемых, исследователи могут наблюдать и изучать поведение десятков миллионов людей [3]. Итоги подобных исследований могут предложить беспрецедентные по своим масштабам результаты, способные изменить наше понимание социума и его поведения [4].

Проведение подобных масштабных исследований стало возможным благодаря «цифровым отпечаткам» - любой информации, оставляемой пользователем при использовании сети Интернет. Такие «отпечатки» могут включать в себя множество данных: от IP-адресов и истории поисковых запросов, до статистики нажатых кнопок на сайте и количестве друзей в социальных сетях [5]. Используя такую информацию, исследователи научились создавать различные прогностические модели [6], которые, в свою очередь, используются в персонализированных поисковых системах [7], таргетинговой рекламе [8], системах оценки персональных рисков [9] и т.п. Однако, проведение подобных исследований и применение прогностических моделей ведет к противоречим вопросам в области информационной безопасности и личного пространства.

На данный момент существует ряд крупных исследований, посвященных определению возможных атрибутов пользователей социальных сетей, используя только оставленные ими «цифровые отпечатки» [4,13]. Благодаря этому, мы имеем возможность с высокой точностью спрогнозировать пол или семейный статус человека только на основе его «лайков», даже если он не указывал подобную личную информацию сам [16,17]. Или на основе пользовательских музыкальных предпочтений или семантики комментариев – установить отдельные психологические черты [18,19].

Совокупность всех «цифровых отпечатков» составляет информационный образ пользователя. Структура информационного образа может быть представлена как креолизованный метатекст – комбинацию вербального текста (статистическая информация, комментарии и т.д.) и аудиовизуального текста (фотографии, аватары и т.д.), призванного управлять впечатлением о пользователе и выстраивать

Статья получена 15 декабря 2019.

Б.А. Низомутдинов Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО) (boris-wels@yandex.ru)

А.С. Тропников Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО) (e-mail: flairwow@gmail.com)

А.Б. Углова Российский государственный педагогический университет им. А. И. Герцена (e-mail: anna.uglova@list.ru)

образ его «Я». Использование информационно образа предлагает пользователю новые способы коммуникации, которые встраиваются в структуру представлений человека о себе и задают направление его дальнейшему развитию [14]. Информационный образ является не только отражением повседневных социальных интеракций, но и оказывает большое влияние на развитие личности современного человека. Изучение информационного образа позволит получить новую информацию как о личности отдельных пользователей, так и о групповых социальных процессах в локальных сообществах и обществе в целом [15]. В целом, можно говорить о том, целью нашего исследования является изучение информационного образа пользователей за счет расширения списка «цифровых отпечатков» путем включения в собираемый массив новых данных, полученных машинным анализом изображений, а также поиск значимых взаимосвязей между психологическими особенностями людей и их данных в профиле социальных сетей.

II. Этика исследования

Практика мирового законодательства, в принципе, как и законодательства Российской Федерации ещё не подготовлена для регулируемого процесса сбора массовых открытых данных пользователей. Вступивший в силу 25 мая 2018 года Общий регламент по защите данных (GDPR) был принят Европейским Союзом как один из основных и первых документов, регламентирующих права пользователя в активно развивающиеся среде сбора общедоступных данных.

Благодаря данному постановлению, граждане ЕС (и не только) получили право на полное и всесторонне знание о том, какие данные собираются, предоставляются и обрабатываются как основными Интернет сервисами, так и третьими лицами, выступающими в роле посредника при обработке подобной информации. Кроме того, одним из основных пунктов данного регламента является предоставления пользователям

«Право на забвение» (Right to be Forgotten). Данное право гарантирует человеку возможность запросить полное удаление каких-либо пользовательских данных из систем хранения как первоначального сервиса, так и из систем хранения третьих лиц, которым осуществлялась передача подобного рода информации.

Данное постановление также регулирует множество мелких вопросов, связанных с информационной безопасностью, способами хранения персональных данных, использование данных несовершеннолетних, предоставление информации в конфиденциальной политики сервиса и т.п. Однако, процесс сбора общедоступных данных, которые пользователем самостоятельно и сознательно выкладывает в Инетрнет, остается законным.

В законодательстве РФ есть несколько документов, регламентирующих работу с персональными данными:

- Конвенция о защите физических лиц при автоматизированной обработке персональных данных (Страсбург, 28.01.1981, ратифицирована Федеральным законом от 19.12.2005 года N 160-ФЗ);
- Конституция РФ. Статьи 23 и 24;
- Федеральный закон «О персональных данных» от 27.07.2006 N 152-ФЗ;
- Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 N 149-ФЗ.

Данные нормативные акты регламентируют отношения между оператором персональных данных (лицом, совершающим сбор, выгрузку, обработку, хранение, систематизацию, передачу данных) и человеком, предоставляющим своим персональные данные.

При этом следует отметить, что, согласно статье 22 закона «О персональных данных», исключением является сбор общедоступных данных.

Однако, поскольку данные законодательные документы были составлены ещё до массового применения технологий больших данных при массовом сборе общедоступных данных, то на текущий момент возникают судебные тяжбы между социальными сетями и компаниями, занимающимися выгрузкой пользовательской информации.

Так, в 2017-2018 годах было судебное дело между социальной сетью ВКонтакте и компанией Double Data [10], занимающейся изучением кредитных возможностей людей на основе пользовательских данных (скорингом). Представители социальной сети утверждают что бесконтрольный сбор пользовательских данных является нарушением права пользователей ВКонтакте, в то время как представители скоринг-компании утверждали о законной возможности использования общедоступных данных.

По итогу дела, суд первой инстанции отклонил иск ВКонтакте, обосновав это законностью деятельности компании Double Data [11]. Тем не менее, при повторном судебном разбирательстве арбитражным судом, победителем вышел истец, как обладатель исключительного права на владение и использование базы данных, содержащей информацию о своих пользователях [12]. Однако, представители Double Data снова просят повторного рассмотрения дела, так как считают решения суда не обоснованным и ссылаются на то что сбор информации происходит не из баз данных ВКонтакте, а исключительно из «кэшированной» информации, оставляемой самими пользователями. По итогу, предыдущие постановления судов были аннулированы, дело передано в Арбитражный суд Москвы.

Приведенный пример демонстрирует несостоятельность текущего законодательства при рассмотрении новых проблем, возникающих при обработке общедоступных пользовательских данных для коммерческой реализации. И поскольку рынок систем, занимающихся разработкой прогностических моделей, активно растет – подобных прецедентов, требующих вмешательства со стороны судебных инстанций, будет становиться все больше.

Кроме правовой и этической стороны вопроса при разработке прогностических моделей пользователей,

существуют так же технические и методологические проблемы.

III. Методология исследования

Для создания прогностических моделей требуется кооперация исследователей из разных областей: психологии, социологии, лингвистики, прикладной математики, математической статистики, компьютерного моделирования. Проектирование модели информационного образа становится возможном за счет многостороннего анализа знакового содержания профиля социальной сети: лингво-психологической оценки смыслового содержания выкладываемого контента, социально-психологического обзора коммуникативных практик, реализуемых в сети и разбора особенностей сетевого интерфейса, задающего структуру данному профилю. В связи с этим наиболее удачным с нашей точки зрения является комбинированный междисциплинарный подход, который включает в себя семантические методы, для отбора значимых «цифровых отпечатков», психодиагностический метод для анализ социально-психологических предпосылок, автоматизированных методов выгрузки массива данных из профиля социальных сетей, методов машинного обучения для сбора данных и методов математико-статистического анализа для обработки полученных данных.

На подготовительном этапе был подобран психодиагностический комплекс, который включил в себя ряд тестов: диагностика степени удовлетворенности потребностей Маслоу; ценностный опросник С. Шварца,; методику исследования самоотношения С.Р. Пантелеева; шкалу измерения тактик самопрезентации (С.-Ж. Ли, Б. Куигли и др.).

Следующим важным этапом создания прогностической модели стал отбор пользовательских данных, на основе которых будет проводиться поиск взаимосвязей. Как правило, исследования концентрируются на отдельных аспектах «цифровых отпечатков» пользователя – лайки, сообщения, звукозаписи, статусы и т.д. Помимо подобных данных, также используются данные о социальной структуре в социальных сетях. Исследования в данной области позволяют проанализировать наличие определенных взаимосвязей между членами сообществ, групп или друзей.

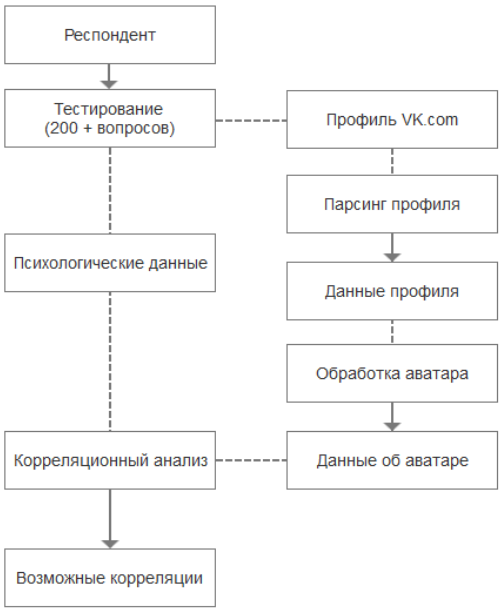

Рис.1 – Методология исследования

Стоит отметить, что анализ пользовательских данных не ограничивается информацией, осознанно оставленной пользователем: многие приложения, сервисы и сайты, открыто предупреждают, что собирают данные о потребителях и передают её третьим лицам «в целях улучшения условий».

Помимо «цифровых отпечатков» оставленных в онлайне, обработке могут быть подвержены и «оффлайновые» данные – данные о геолокации со смартфонов, данные о сердцебиении со спортивных браслетов и т.п. Использование подобных данных позволяет создавать картину пользовательской мобильности в социуме: его перемещениях, предпочтениях, реакциях и пр.

Использование подобных данных может производиться без прямого уведомления и желания пользователя, что вызывает ряд этических проблем. Например, использование персональных данных таким образом может подорвать доверие к цифровым технологиям или вовсе заставить пользователей дистанцироваться от применения таковых.

С другой стороны, применение большого количества персональных данных для разработки прогностических моделей позволяет добиться более точечного моделирования человеческого поведения. Появление все новых автоматизированных средств анализа данных позволяет начинать новые исследования и открывать ранее неподтвержденные корреляции, дающие возможность определить тот или иной аспект человеческой натуры.

IV. Сбор и обработка данных

В исследовании приняли участие 253 человека, большинство из которых являлись студентами университета. В ходе обработки результатов тестирования, были отобраны 157 человек, предоставивших доступ к своим профилям в социальной сети ВКонтакте.

Сбор данных из социальных сетей происходил по страницам, которые испытуемые оставили по завершению психологического тестирования.

На данном этапе исследования использовался метод парсинга контента, без API. Парсер – это программа или скрипт, позволяющая выполнить такой анализ и представить результат в нужном для пользователя виде. API ВКонтакте - это интерфейс, который позволяет получать информацию из базы данных vk.com с помощью http-запросов к специальному серверу. Вам не нужно знать в подробностях, как устроена база, из каких таблиц и полей каких типов она состоит — достаточно того, что API-запрос об этом «знает». Синтаксис запросов и тип возвращаемых ими данных строго определены на стороне самого сервиса.

В данном подходе, специальной программой парсером, загружалась каждая страница в кеш, после

чего, из html кода, по заранее заданным правилам собиралась открытая информация.

Такой подход имеет ограничение в производительности, однако дает быструю возможность для сбора информации.

На следующих этапах исследования запланирован сбор открытых данных через API вконтакте.

Используя средства парсинга, был собран массив данных, содержащих всю общедоступную информацию, размещенную на странице профиля: имя, фамилия, количество подписок на сообщества, количество аудиозаписей и видеозаписей, аватары, статус о работе и семейном положении, возраст, количество друзей и т.д.

На рис. 2 представлена общая схема процесса по сбору данных.

Рис 2. Описание процесса сбора данных

На следующем этапе исследования будет использован метод машинного обучения, для определения тематик сообществ и интересных страниц, на которые подписан пользователи.

Кроме текстовой информации, анализировались фотографии профиля. Используя облачный сервис машинного анализа изображений Azure, собранные фотографии были проанализированы для сбора новых данных. В ходе обработки была получена информация об отображаемых эмоциях на аватаре, определенных предметах, общей композиции, интенсивности цвета, количестве людей и их возрасте. Собранная информация была добавлена в общий массив данных о пользователях.

На следующем шаге исследования, данные, собранные в ходе автоматической выгрузки, были объединены в более крупные категории, описывающие базовые компоненты информационного образа пользователя: информация об имени пользователя (самоименование), информация о возрасте, пространственная локализация, информация о близких отношениях, о профессиональном статусе, ассоциированность в сети, информация об образовании, конфиденциальности, количество друзей, подписчиков, фотографий, видеозаписей, аудиозаписей, групп, подписок, а также эмоции и объекты, выявленные на фотографиях пользователей.

V. Факторный анализ

После автоматизированного анализа статистического компонента информационного образа пользователей Вконтакте, данные были объединены в итоговую матрицу размерностью [157x18], которая затем была подвергнута процедуре факторного анализа по методу «главных компонент» с последующим varimax-вращением. В результате факторного анализа, была проведена редукция и группировка сходных статистических данных в более крупные категории, были выделены обобщенные смысловые размерности, отражающие особенности самопрезентации в социальной сети (таблица 1).

Первый фактор, объясняющий 32 % общей дисперсии, был проинтерпретирован как ось «Стратегия сообщения о своих социально-коммуникативных достижениях». Ведущие смысловые конструкты в составе фактора описывают визуальный компонент информационного образа, выполняющий функцию неявного самораскрытия, а также ассоциативную функции. Визуальный образ, как первое на что мы обращаем внимание, призван сформировать направленность прочтения нейтральной социально-демографической информации о социальных контактах человека.

Таблица 1 – Результаты факторизации статистических данных информационного образа

Наименование факторов	Компоненты информационного образа в составе фактора	Вес компонентов
Фактор 1 - «Стратегия сообщения о своих социально-коммуникативных достижениях» вес фактора: 2,89 % дисперсии –32	Количество фотографий	0,90
	Количество фотоальбомов	0,69
	Количество подписок	0,64
	Количество друзей	0,89
	Информация о других социальных сетях)	0,58
Фактор 2 - «Стратегия профессиональной самопрезентации» вес фактора: 2,29 % дисперсии –25	Информация о пространственной локализации	0,55
	Информация о профессиональном статусе	0,64
	Информация об обучении в высшем учебном заведении	0,89

	Информация об обучении в школе	0,85

Второй фактор (13 % общей дисперсии) был проинтерпретирован как ось «Стратегия профессиональной самопрезентации». Компоненты данного фактора описывают ценность всех этапов профессионального становления от школы, до актуального профессионального статуса и выполняют функцию демонстративную и мотивационную функцию. Презентуемые в информационном образе ценности образования и профессиональных достижений с одной стороны помогают человеку обозначить определенную позицию в социуме, указать на свой профессиональный статут, а с другой - выступают в качестве мотивационного сигнала для активации деловой коммуникации и взаимодействия с профессиональными киберсообществами.

На следующем этапе исследования было построено семантическое пространство, путем размещения оцениваемых информационных образов в пространстве выделенных факторов, представляющих собой совокупность ключевых характеристик самопрезентации в сети. Психологический смысл «размещения» связан с тем, что информационный образ строится с использованием всех статистических элементов социальной сети, однако некоторые из них могут быть более ярко выражены. Семантическое пространство было образовано факторами «Стратегия сообщения о своих социально-коммуникативных достижениях» и «Стратегия профессиональной самопрезентации», что позволило выделить 5 групп информационных образов. В совокупности компонентов теоретически возможны пять крайних варианта:

- В первую группу вошли испытуемые (доля в выборке - 15 человек - 9,6%), информационные образы которых одновременно направлены как на самораскрытие достижений в социальных сетях, так и профессиональную самопрезентацию, что можно условно назвать стратегией презентации своих достижений;

- Во вторую группу вошли испытуемые (доля в выборке - 67 человек – 42,6%), информационные образы которых содержать минимум информации, что является основой для стратегии уклонения от глубокой самопрезентации;

- В третью группу вошли испытуемые (доля в выборке 42 человек – 26,7%), в информационном образе которых много информации социально-коммуникативных достижениях и минимум информации о профессиональной сфере, что указывает на коммуникативную стратегию самопрезентации;

- В четвертую группу вошли испытуемые (доля в выборке 19 человек – 12,2%), в информационные образы которых содержат в основном информацию о профессиональных достижениях, что составляет основу профессиональной самопрезентации

- В пятую группу мы отдельно выделили испытуемых (доля в выборке 14 человек – 8,9%), в информационных образах которых полностью отсутствует информация о профессиональные достижения и очень мало информации в целом, что указывает на использование социальной сети скорее для развлекательно-коммуникативных целей, нежели для самопрезентации.

На третьем этапе для выявления независимого влияния пяти способов самопрезентации на социально-психологические характеристики испытуемых, мы использовали процедуру дисперсионного анализа (непараметрический метод Краскела-Уолисса), которая предполагает анализ компонентов дисперсии изучаемых признаков. Необходимо отметить, что в нашей работе мы придерживаемся конструктивистского подхода, в соответствии с которым создание информационного образа представляет собой результат творческого процесса воссоздания жизненного опыта в соответствие с актуальными потребностями и ценностно-смысловыми установками человека. Данный подход позволяет проверить гипотезу о влиянии способов самопрезентации на зависимые переменные, в качестве которых выступала совокупность социально-психологических характеристик пользователей (таблица 2).

Таблица 2 – Достоверно-значимые различия социально-психологических характеристик пользователей в зависимости от способа самопрезентации

Социально-психологические характеристики	Группы испытуемых с разным типом самопрезентации*					H**	P***
	1 гр	2 гр	3 гр	4 гр	5 гр		
Возраст	28,0	**33,0**	24,6	20,9	**36,4**	2,51	0,05
Оправдание с принятием ответственности	**28,8**	12,8	23,1	18,4	17,4	3,17	0,02
Препятствование самому себе	**27,1**	14,0	22,2	23,1	23,2	2,55	0,05
Извинение	**32,9**	18,4	29,7	30,4	23,6	3,98	0,01
Запугивание	18,4	6,4	**20,7**	14,2	11,6	3,45	0,01
Преувеличение своих достижений	**27,0**	13,6	20,4	17,0	17,4	2,25	0,05
Уклонение.	**81,5**	40,4	67,9	60,3	60,0	2,78	0,03
Силовое влияние.	41,0	17,4	**44,3**	29,6	29,2	2,93	0,03
Тактики защитного типа	**140**	72,0	118	109	106	3,10	0,02
Потребность в безопасности	17,6	11,0	18,1	17,1	**18,8**	4,59	0,01
Саморуководство	5,8	5,6	6,8	4,9	**7,4**	2,63	0,05
Самопривязанность	5,3	4,8	**6,4**	4,4	6,0	3,10	0,02
Внутренняя конфликтность	6,4	4,8	5,1	**6,4**	3,4	2,66	0,04
Социальная власть	4,4	2,7	**4,5**	3,4	3,8	2,61	0,05

*В таблице приведены средние результаты по группе
** H – Критерий Краскела-Уоллиса
*** P – уровень значимости

Из приведенной таблицы видно, что межгрупповая изменчивость выше внутригрупповой, что позволяет сделать заключение о достоверно значимых различиях социально-психологических характеристик пользователей в зависимости от предпочитаемого способа виртуальной самопрезентации:

Испытуемые, предпочитающие первую тактику цифровой самопрезентации (предоставление всей информации), достоверно чаще ориентированы на использование защитных тактик общения такие как оправдание с принятием ответственности (p=0,0226), препятствование самому себе (p=0,0495), извинение (p=0,0076), уклонение (p=0,0383), а также преувеличение своих достижений (p=0,0498);

Вторую и пятую стратегию, заключающуюся в минимизации информации достоверно чаще, используют люди старше 30 лет (p=0,0496);

Третья стратегия цифровой самопрезентации, направленная на демонстрацию визуального контента и социальных связей, присуща людям с высоким стремлением сохранить в неизменном виде свое «Я» (p=0,0247), для которых важен социальный статус и доминирование (p=0,0484), а также тактики силового влияния (p=0,0311) и запугивания (p=0,0154) в коммуникации;

Четвертая стратегия, связанная с предпочтением деловой коммуникации и демонстрацией профессионального статуса, достоверно чаще используется пользователями, которым свойственна высокая требовательность к себе, внутренняя конфликтность и преобладание негативного фона отношения к себе (p=0,0452);

Пятая стратегия цифровой самопрезентации, связанная с полным отсутствием «цифровых отпечатков» о образовании и профессиональном статусе свойственна взрослым людям, старше 30 лет (p=0,0496), с высокой потребностью в безопасности (p=0,0035), выраженной способностью к самоконтролю и саморуководству (p=0,0470).

VI. Заключение

По итогам проведенного исследования, можно сделать вывод, что достоверно значимые различия в дисперсиях позволяют описать индивидуальные социально-психологический профили каждой группы испытуемых. А также теоретически обосновать разработку автоматизированной прогностической модели информационного образа пользователя, позволяющую предсказать на основе анализа информации в социальной сети те или иные психологические характеристики. Изучение взаимосвязей компонентов информационного образа и личностных особенностей показало довольно высокие прогностические возможности анализа данных социального профиля.

На следующих этапах исследования запланировано как улучшение методов сбора информации (использовании API vk.com), и усовершенствование технологии обработки и анализа собранных данных, с использованием машинного обучения.

Библиография

[1] Sandra M., Oded N. Using Big Data as a window into consumers' psychology // Current Opinion in Behavioral Sciences. 2017. № 18. P. 7 – 12.

[2] Michal K. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes // Psychological Methods. 2016. № 21. P. 493 – 506.

[3] Barbakov O.M., Vinogradova M.V., Shatsky A. Social Portrait of Online Mass Media Audience in Russia // Media Watch. 2018. № 9. P. 383 – 396.

[4] Xenos S., Ryan T. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage // Computers in Human Behavior. 2011. Vol. 27, № 5. P. 1658 – 1664.

[5] Settanni M., Azucar D., Marengo D. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis // Personality and Individual Differences. 2018. Vol 124. P. 150 – 159.

[6] Sophie W.F., Susanne B.E., Patti V. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp // SAGE. 2017. № 20. P. 1 – 19.

[7] Kaiqi H., Qiao W., Zhenyang W. Natural color image enhancement and evaluation algorithm based on human visual system // Computer Vision and Image Understanding. 2006. Vol 1, № 103. P. 52 – 63.

[8] Kim Y., Kim J.H. Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study // Information Processing & Management. 2018. Vol. 6, № 54. P. 1101 – 1114.

[9] Gosling S., Gaddis S., Vazire S. Personality Impressions Based on Facebook Profiles // International Conference on Weblogs and Social Media. 2007.

[10] Социальная сеть требует запретить сбор данных для банков // Коммерсантъ [сайт]. URL: https://www.kommersant.ru/doc/3206044

[11] Право на поиск информации в интернете под угрозой. Решение по делу «ВКонтакте» ограничивает работу поисковиков / Сколково [сайт]. URL: http://sk.ru/news/b/pressreleases/archive/2018/01/29/pravo-na- poisk-informacii-v-internete-pod-ugrozoy--reshenie-po-delu-vkontakte-ogranichivaet-rabotu- poiskovikov.aspx

[12] Постановление суда по интеллектуальным правам. / Суд по интеллектуальным правам [сайт]. URL: http://kad.arbitr.ru/PdfDocument/1f33e071-4a16-4bf9-ab17-4df80f6c1556/4c9d2b02-4fbd-4554-82c8- 53282523639c/A40-18827-2017_20180724_Reshenija_i_postanovlenija.pdf (дата обращения: 22.04.2019)

[13] Kosinski M., Matz S., Gosling S. et al. Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines // American Psychologist. 2015. Vol. 70. N 6. P. 543—556.

[14] Korneeva A., Zeremskaya Yu., Loyko O. Virtual space as a sphere of the personal identity's formation // Journal of Economics and Social Sciences. 2016. № 8 (8). C. 31-35.

[15] Wilson R.E., Gosling S.D., Graham L.T. A review of Facebook research in the social sciences // Perspectives on Psychological Science. 2012. 7. 3. 203-220..

[16] Kosinski M., Stilwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // Proc. the National Academy of Science of the United State of America. 2013. Vol. 110. P. 5802-5805. DOI: 10.1073/pnas.1218772110

[17] Ross C., Orr E.S., Sisic M., Arseneault J.M., Simmering M.G., Orr R.R.: Personality and motivations associated with facebook use // Computers in Human Behavior. 2009. Vol. 25. P. 578-586. DOI: 10.1016/j.chb.2008.12.024

[18] Krylova O.S., Vlasov D.A., Shishkov V.V., Alymov A.S., Ishin I.A., Kolesnikov I.E. Petrov A.I. Opisanie informacionnogo obraza pol'zovatelya social'noj seti s uchetom ego psihologicheskoj harakteristiki // International Journal of Open Information Technologies. 2018. Vol. 4. C. 24-37.

[19] Mairesse F., Walker M., Mehl M., Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text // Journal of Artificial Intelligence Research. 2007. Vol. 30. P. 457-500. DOI: 10.1613/jair.2349

Низомутдинов Борис Абдуллохонович, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), ведущий аналитик, ORCID 0000-0002-4090-9564.
Тропников Александр Сергеевич, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), аналитик, ORCID 0000-0002-4179-536X.
Углова Анна Борисовна, к.псих.н, Российский государственный педагогический университет им. А. И. Герцена, ассистент, ORCID 0000-0002-8072-0539.

Automated collection of social network data to develop a factor model of network self-presentation

B.A. Nizomutdinov, A.S. Tropnikov, A.B. Uglova

***Abstract* — On the basis of the conducted empirical research of information images of users, the leading components of network self-presentation were revealed: statistical, socio-demographic component, visual component and value-semantic component. The authors analyzed the hidden factors responsible for the formation of network self-presentation through the information image, studying the prognostic possibilities of social profile data analysis. Significant differences in the content of the information image and socio-psychological characteristics of users with different styles of network self-presentation were identified. Algorithms for collecting and processing open information from social network profiles, followed by factor analysis, as well as machine learning methods to determine the topics of communities and interesting pages to which users subscribe, are presented.**

The paper deals with ethical and legal issues of using data collection from user groups to create predictive models based on them, without notifying the users themselves. Also discusses theoretical issues of interdisciplinary design model information image, which becomes possible through multilateral analysis of the symbolic content of social network profile: linguistic and psychological evaluation of the semantic content of the content of socio-psychological overview of communication practices that are implemented in the network and parsing features of the network interface that specifies the structure of this profile.

***Keywords* — information prints, data processing, prognostic model, information image.**

REFERENCES

[1] Sandra M., Oded N. Using Big Data as a window into consumers' psychology // Current Opinion in Behavioral Sciences. 2017. № 18. P. 7 – 12.

[2] Michal K. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes // Psychological Methods. 2016. № 21. P. 493 – 506.

[3] Barbakov O.M., Vinogradova M.V., Shatsky A. Social Portrait of Online Mass Media Audience in Russia // Media Watch. 2018. № 9. P. 383 – 396.

[4] Xenos S., Ryan T. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage // Computers in Human Behavior. 2011. Vol. 27, № 5. P. 1658 – 1664.

[5] Settanni M., Azucar D., Marengo D. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis // Personality and Individual Differences. 2018. Vol 124. P. 150 – 159.

[6] Sophie W.F., Susanne B.E., Patti V. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp // SAGE. 2017. № 20. P. 1 – 19.

[7] Kaiqi H., Qiao W., Zhenyang W. Natural color image enhancement and evaluation algorithm based on human visual system // Computer Vision and Image Understanding. 2006. Vol 1, № 103. P. 52 – 63.

[8] Kim Y., Kim J.H. Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study // Information Processing & Management. 2018. Vol. 6, № 54. P. 1101 – 1114.

[9] Gosling S., Gaddis S., Vazire S. Personality Impressions Based on Facebook Profiles // International Conference on Weblogs and Social Media. 2007.

[10] The social network requires prohibiting the collection of data for banks // Kommerasnt [site].. URL: https://www.kommersant.ru/doc/3206044

[11] The right to seek information on the Internet is at risk. The decision in the VKontakte case limits the work of search engines / Skolkovo [site]. URL: http://sk.ru/news/b/pressreleases/archive/2018/01/29/pravo-na-poisk-informacii-v-internete-pod-ugrozoy--reshenie-po-delu-vkontakte-ogranichivaet-rabotu- poiskovikov.aspx

[12] Court ruling on intellectual property rights. / Intellectual Property Court [site]. URL: http://kad.arbitr.ru/PdfDocument/1f33e071-4a16-4bf9-ab17-4df80f6c1556/4c9d2b02-4fbd-4554-82c8- 53282523639c/A40-18827-2017_20180724_Reshenija_i_postanovlenija.pdf (дата обращения: 22.04.2019)

[13] Kosinski M., Matz S., Gosling S. et al. Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines // American Psychologist. 2015. Vol. 70. N 6. P. 543—556.

[14] Korneeva A., Zeremskaya Yu., Loyko O. Virtual space as a sphere of the personal identity's formation // Journal of Economics and Social Sciences. 2016. № 8 (8). C. 31-35.

[15] Wilson R.E., Gosling S.D., Graham L.T. A review of Facebook research in the social sciences // Perspectives on Psychological Science. 2012. 7. 3. 203-220..

[16] Kosinski M., Stilwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // Proc. the National Academy of Science of the United State of America. 2013. Vol. 110. P. 5802-5805. DOI: 10.1073/pnas.1218772110

[17] Ross C., Orr E.S., Sisic M., Arseneault J.M., Simmering M.G., Orr R.R.: Personality and motivations associated with facebook use // Computers in Human Behavior. 2009. Vol. 25. P. 578-586. DOI: 10.1016/j.chb.2008.12.024

[18] Krylova O.S., Vlasov D.A., Shishkov V.V., Alymov A.S., Ishin I.A., Kolesnikov I.E. Petrov A.I. Opisanie informacionnogo obraza pol'zovatelya social'noj seti s uchetom ego psihologicheskoj harakteristiki // International Journal of Open Information Technologies. 2018. Vol. 4. C. 24-37.

[19] Mairesse F., Walker M., Mehl M., Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text // Journal of Artificial Intelligence Research. 2007. Vol. 30. P. 457-500. DOI: 10.1613/jair.2349