DOI: 10.17323/2587-814X.2020.2.64.83

Спрос на навыки на рынке труда в сфере информационных технологий

А.А. Терников
E-mail: aternikov@hse.ru

Е.А. Александрова
E-mail: ea.aleksandrova@hse.ru

Национальный исследовательский университет «Высшая школа экономики»
Адрес: 194100, г. Санкт-Петербург, ул. Кантемировская, д. 3, корп. 1, лит. А

Аннотация

Сфера информационных технологий является одной из наиболее динамично развивающихся на рынке труда. Предъявляемый спрос на навыки имеет значительную вариацию в зависимости от отрасли и сферы деятельности организаций, специфики рабочего места и организации труда. Для формирования навыков, достаточных для успешного трудоустройства выпускников, со стороны системы образования необходим качественный мониторинг спроса, предъявляемый работодателями. В статье представлен алгоритм, позволяющий определить ключевые комбинации навыков, которые необходимы компаниям в сфере информационных технологий в зависимости от профессиональных групп. Использованные в статье подходы, TF-IDF и *n*-граммы, позволили извлечь и структурировать знания, умения и навыки, полученные из неструктурированной базы данных интернет-вакансий на рынке труда. В результате были найдены ключевые комбинации профессиональных навыков для отдельных профессиональных групп. Предложенный алгоритм позволяет определить и стандартизировать ключевые навыки, которые могут быть использованы для создания системы российских классификаторов по профессиям и навыкам. Кроме того, алгоритм формирует списки ключевых комбинаций навыков, которые высоко востребованы компаниями в каждой конкретной профессиональной группе в сфере информационных технологий.

Ключевые слова: вакансии в сфере информационных технологий; онлайн-вакансии; анализ неструктурированных данных; рынок труда; спрос на навыки; комбинации знаний, умений и навыков работников.

Цитирование: Терников А.А., Александрова Е.А. Спрос на навыки на рынке труда в сфере информационных технологий // Бизнес-информатика. 2020. Т. 14. № 2. С. 64–83. DOI: 10.17323/2587-814X.2020.2.64.83

Введение

Совокупность знаний, умений и навыков (англ. *skills*, далее – ЗУН), предъявляемых работодателями, является одним из наиболее информативных показателей для оценивки спроса на рынке труда, представляя собой описание компетенций, требуемых в различных профессиональных областях (англ. *occupations*). Однако комбинации требуемых ЗУН динамично меняются как с течением времени, так и в зависимости от специфики отрасли, организации, конкретной вакансии. Данные изменения связаны с конъюнктурными колебаниями экономики и реструктуризацией рынка труда. Более того, профессиональные стандарты, формируемые на основе специфики системы образования, теряют свою гибкость к данным изменениям и довольно быстро устаревают. Отдельный интерес представляет собой проблема выявления ключевых наборов ЗУН на рынке труда для различных профессиональных областей в сфере информационных технологий (далее – ИТ), чему и посвящена настоящая статья.

Исследования в области идентификации комбинаций ЗУН, востребованных на рынке труда, интересны именно применительно к области информационных технологий, по нескольким причинам. Во-первых, сфера ИТ крайне динамична, и требования, предъявлемые к ЗУН, интенсивно меняются с течением времени [1–8]. Во-вторых, ЗУН достаточно ясно и просто классифицируются по профессиональным областям вследствие наличия точных формулировок языков программирования, стека технологий, графических интерфейсов пользователя и т.п. [9–11]. В-третьих, внедрение новых технологий сопровождается появлением новых рабочих задач, что требует изменений в части комбинаций ЗУН работников [12–16].

ИТ имеют высокий спрос на рынке труда: технические специалисты с определенными наборами компетенций востребованы в областях, связанных с экономикой, управлением, торговлей и т.д. Таким образом, работодатели формируют спрос на комбинации ЗУН, определяя задачи, предъявляемые в организации к конктретной должности. Не исключено, что уникальные комбинации ЗУН в отдельных сферах деятельности могут формироваться в системе образования не только в ИТ-специальностях, но и будут весьма востребованы среди экономистов, маркетологов, менеджеров, инженеров. Переосмысление образовательной политики в части формирования ЗУН не может происходить без соотнесения со спросом на рынке труда, что требует эффективных инструментов идентификации комбинаций профессиональных ЗУН, востребованных работодателями.

Настоящее иследование посвящено разработке алгоритма, направленного на извлечение и структурирование информации по ключевым ЗУН в разрезе профессиональных областей в сфере ИТ. Целью работы является идентификация тех ЗУН, которые востребованы компаниями-работодателями.

Статья имеет следующую структуру. В первом разделе представлен обзор смежных работ и методов, используемых в целях классификации и кластеризации данных о рынке труда, полученных из открытых интернет-источников. Во втором разделе представлен алгоритм, позволяющий извлечь и классифицировать информацию из неструктурированных объявлений о работе в части ЗУН, а также его применение на данных регионального рынка труда. Третий раздел содержит результаты работы, а именно: списки ключевых ЗУН и их комбинаций для различных профессиональных областей в сфере ИТ. Заключение содержит рекомендации и дискуссию в отношении применения представленного алгоритма.

1. Обзор методов анализа спроса на навыки

Большинство исследований, посвященных анализу спроса на навыки, используют в качестве исходной информации объявления о вакансиях, полученных из открытых интернет-источников [17–25]. Данные источники содержат достаточно обширную информацию о ЗУН и их комбинациях, но зачастую в неструктурированном виде. Основные методы обработки такой информации базируются на техниках обработки естественного языка (англ. *natural language processing*), в частности, на анализе меры TF-IDF (англ. *Term Frequency – Inverse Document Frequency*) и *n*-грамм (последовательности из *n* элементов), а также использовании алгоритмов классификации и кластеризации [17–25].

Онлайн-базы данных вакансий, как правило, имеют неструктурированные текстовые поля,

которые содержат информацию о профессии и необходимых компетенциях. Такие поля заполняются вручную представителями компаний, поэтому требуется применение процедур подготовки данных и алгоритмических методов для извлечения соответствующей информации в стандартизированной форме. Некоторые исследования решают задачу классификации профессиональных должностей по их описанию в онлайн-вакансиях по широко используемыми классификациями профессий и ЗУН, таких как ISCO[1], ESCO[2] и O*NET[3] [17–20]. Другие исследования применяют модели классификации на основе данных, размеченных экспертами [2, 4, 11, 13, 21]. Иначе говоря, выборка анализируется и размечается экспертами в конкретной области, после чего эта информация используется для корректировки алгортмов классификации. Кроме того, исследователи используют подходы кластеризации для определения профессий и навыков при подготовке данных, что позволяет сформировать и разделить профессиональные группы [19, 21–25]. Таким образом, сочетание различных подходов и алгоритмов подготовки и стандартизации данных позволяет заложить основу для анализа спроса на навыки. Краткое описание данных, подходов и алгоритмов, которые используются в смежных работах, представлено в *таблице 1.*

Представленная в таблице информация позволяет обобщить и системтизировать подходы к организации данных, их обработке и методам анализа, выбору критериев для идентификации комбинаций ЗУН.

Все авторы представляют свои алгоритмы извлечения и систематизации информации на основе онлайн-вакансий. Однако способы их реализации различаются в зависимости от поставленной исследовательской задачи. Например, если основная цель исследования связана с процессом сопоставления неструктурированных текстовых полей из объявлений о работе с официальным классификатором по профессиям и навыкам [17, 18, 20, 21], то алгоритмы классификации реализуются на основе нахождения сходных закономерностей в названии должности, их описания и расширенной текстовой информации из официальных классификаторов, включая значительный объем данных ручной разметки экспертов.

Другой подход основывается на ручной корректировке данных после запуска алгоритмов кластеризации [19, 22–25]. Несмотря на различия в целях исследований, в целом применяются общие методы подготовки данных и извлечения стандартизированной информации. Все авторы используют подход TF-IDF и токенизацию (включая удаление стоп-слов и стемминга) для того, чтобы обрабатывать большое количество неструктурированной текстовой информации. Кроме того, *n*-граммы используются для извлечения более чем одного слова. В результате получается набор унифицированных шаблонов (например, профессий и ЗУН). Однако авторы не предоставляют обобщенного алгоритма для сопоставления разных вариантов написания одного и того же шаблона в процессе обработки данных.

Выбор профессиональных областей и ЗУН зависит от информации из официальных классификаций и объема данных. Уровень обобщения таких групп зависит от экспертных оценок, основанных на характеристиках данных. В целом в смежных исследованиях объем данных доступен за однолетний период, и поиск подходящих шаблонов для неструктурированных полей упрощен только для объявлений о работе, опубликованных на одном языке.

Исследователи обращаются к разным метрикам оценки моделей классификации и кластеризации. Авторы используют токенизацию для необработанных текстов и *n*-граммы для построения набора терминов. Анализ сходства различных терминов и описаний с шаблоном реализуется при помощи различных индексов. В случае сохранения порядка слов используется расстояние Левенштайна (англ. *Levenshtein*), но если только пересечение одинаковых терминов является ценным для обнаружения сходства – предпочтительным является индекс Жаккара (англ. *Jaccard*).

[1] International Standard Classification of Occupations, https://www.ilo.org/public/english/bureau/stat/isco/

[2] European Skills/Competences, Qualifications and Occupations, https://ec.europa.eu/esco/portal/home

[3] The Occupational Information Network, https://www.onetonline.org//

Таблица 1.

Смежные работы в области анализа онлайн-вакансий

Основное направление	Данные				Извлеченные группы		Методы обработки данных				Объем ручной разметки данных	Меры схожести	Авторы
	Объем	Период	Источники	Язык	Профессиональные области	ЗУН	TF-IDF	*n*-граммы	Кластеризация	Классификация			
Кластеризация профессиональных областей	1460	4 месяца (апрель – июль 2018)	LinkedIn	Английский	8	96	+	+	Unweighted Pair Group Mean Average method	–	>900	Jaccard	[24]
	12849	7 месяцев (июль –ноябрь 2015 и октябрь – ноябрь 2016)	5 источников	Английский	69	НД*	+	–	Latent Semantic Indexing, Singular Value Decomposition	–	750	Cosine	[19]
Классификация профессиональных областей	75546	27 месяцев (февраль 2013 – апрель 2015)	WollyBI	Итальянский	9	НД	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	57740	Levenshtein, Jaccard, Sørensen–Dice	[17]
	40000	1 месяц (год не указан)	12 источников	Итальянский	62	542	+	–	Weighted Word Pairs (WWP) extraction	LinearSVC & Perceptron classifier.	412	Levenshtein	[21]
Кластеризация профессиональных областей и извлечение ЗУН	2786	3 месяца (осень 2015)	10 источников	Английский	4	180	+	+	Latent Dirichlet Analysis	–	180	Centrality degree	[22]
	2638	3 месяца (май – июль 2018)	Indeed.com	Английский	48	480	+	–	Latent Dirichlet Analysis	–	480	% встречаемости	[23]
	1050	5 месяцев (июль – ноябрь 2017)	6 источников	Английский	2	2335	+	–	Latent Class Analysis, Singular Value Decomposition	–	НД	VARIMAX Rotation	[25]
Классификация профессиональных областей и извлечение ЗУН	6222	4 месяца (июнь – сентябрь 2015)	3 источников	Итальянский	6	НД	+	+	–	SVM (linear & RBF Kernel); Random Forest; NN	1007	Random–Forest importance	[20]
	~2 млн	24 месяца (2016–2017)	WollyBI	Итальянский	22	8	+	+	–	SVM	НД	Levenshtein, Jaccard, Sørensen–Dice	[18]

*НД – «Нет данных»

2. Алгоритм анализа спроса на навыки и описание используемых данных

Предлагаемый алгоритм, позволяющий реализовать анализ спроса на навыки, организован для данных онлайн-вакансий. Эти данные получены из интерфейса прикладного программирования с открытым исходным кодом HeadHunter[4] – крупнейшей российской онлайн-платформы по подбору персонала[5]. Типичная структура онлайн-вакансии представлена в *таблице 2.*

Наряду с представленной структурой данных и методами, использованными в смежных работах, основной интерес данной статьи заключается в организации процесса извлечения ЗУН из неструктурированных данных. Таким образом, определение наиболее востребованных ЗУН в профессиональных группах может быть реализовано досточно точно. Несмотря на использование алгоритмов классификации для агрегирования профессий в смежных работах, текущий набор данных уже кодифицирован производителем данных. Таким образом, на предварительном этапе анализа предположим, что профессиональные группы уже присвоены и существуют. Для того чтобы организовать описание алгоритма, необходимо формализовать и упростить несколько концепций.

Определение 1. Онлайн-вакансия. Пусть I – набор числовых кодов вакансий; H – набор числовых кодов специализаций; S – набор уникальных ЗУН. Пусть $V = \{v_1, \ldots, v_n\} : n \in \mathbb{N}$ – набор вакансий. Тогда онлайн-вакансия v представляет собой 6-элементный кортеж $v = (i, C, d, p, g, K)$, где $i \in \mathrm{I}$, $C \subseteq H : |C| \in \{\overline{1,6}\}$ – подмножество числовых кодов специализаций, d – дата публикации вакансии, p – наименование вакансии, g – описание вакансии, $K \subseteq S : |K| \leq 30$ – подмножество ЗУН для данной вакансии.

Настоящее исследование сосредоточено на секторе ИТ, а предлагаемый подход протестирован на региональном рынке труда (г. Санкт-Петербург). Онлайн-вакансии из сферы ИТ в Санкт-Петербурге были собраны с 2015 по 2019 годы (для этого был использован официальный классификатор HeadHunter для получения вакансий в сфере информационных технологий по числовому коду специализации). Каждое наблюдение для конкретной вакансии (v) содержит числовой код вакансии (i), числовой код специализации по кслассификации HeadHunter (C) и список ЗУН, требуемых в данной вакансии (K). Основные задачи исследования сосредоточены на процессе извлечения и структурирования ЗУН. Таким образом, используется та часть данных, где поле, со-

Таблица 2.

Структура типового объявления портала HeadHunter

Поле	Тип поля		Описание
	Структурированное	Неструктурированное	
Vacancy ID	+		Числовой код вакансии
Specialization ID	+		Набор (от 1 до 6 включительно) числовых кодов специализаций[6]
Published date	+		Длинный формат даты публикации вакансии
Position Name		+	Текст (наименование вакансии)
Job description		+	Текст (описание вакансии)
Key skills		+	Набор текстов (30 – максимум: каждый не более 100 символов), содержащих ЗУН

[4] HeadHunter API, https://dev.hh.ru

[5] SimilarWeb: websites ranking, https://www.similarweb.com/top-websites/russian-federation/category/jobs-and-career/jobs-and-employment

[6] HeadHunter API: Specializations, https://api.hh.ru/specializations

Таблица 3.

Распределение вакансий по кодам специализаций HeadHunter в выборке

Код специализации HeadHunter	Доля, %	Наименование
1.221	20,37	Программирование, Разработка
1.82	7,99	Инженер
1.9	4,90	Web инженер
1.89	4,52	Интернет
1.10	4,18	Web мастер
1.327	3,83	Управление проектами
1.225	3,42	Продажи
1.137	3,21	Маркетинг
1.272	3,11	Системная интеграция
1.295	3,04	Телекоммуникации
1.211	2,99	Поддержка, Helpdesk
1.117	2,90	Тестирование
1.270	2,86	Сетевые технологии
1.273	2,73	Системный администратор
1.25	2,69	Аналитик
1.172	2,62	Начальный уровень, Мало опыта
1.50	2,25	Системы управления предприятием (ERP)
1.400	2,11	Оптимизация сайта (SEO)
1.536	2,10	CRM системы
1.474	2,04	Стартапы
1.359	1,75	Электронная коммерция
1.116	1,58	Контент
1.475	1,51	Игровое ПО
1.113	1,50	Консалтинг, Аутсорсинг
1.246	1,42	Развитие бизнеса
1.420	1,39	Администратор баз данных
1.395	1,04	Банковское ПО
1.203	1,01	Передача данных и доступ в интернет
1.110	0,98	Компьютерная безопасность
1.161	0,86	Мультимедиа
1.296	0,67	Технический писатель
1.274	0,66	Системы автоматизированного проектирования
1.3	0,64	CTO, CIO, Директор по IT
1.277	0,61	Сотовые, Беспроводные технологии
1.30	0,34	Арт–директор
1.232	0,18	Продюсер

держащее ЗУН, является непустым. Полученная выборка состоит из 63869 вакансий, опубликованных с мая 2015 года по сентябрь 2019 года. Каждая вакансия включает от одного до шести профессиональных кодов (специализаций HeadHunter). Распределение по 36 направлениям (внутри группы сферы ИТ) представлено в *таблице 3.*

Несмотря на наличие представленного распределения кодов специализаций HeadHunter, некоторые сферы могут быть удалены или объединены в одну большую подгруппу. В соответствии с классификацией, введенной в работе [20], мы определяем семь групп ИТ-специалистов. После перегруппировки вакансий и удаления некоторых кодов специализаций получено 56000 наблюдений (вакансий). Доля удаленных профессиональных областей составляет 10,2%. Новое распределение среди оставшихся агрегированных профессиональных групп вакансий и их наименования представлены в *таблице 4.*

Таблица 4.

Распределение вакансий в сфере ИТ по укрупненным профессиональным группам

Наименование	Сокращение	Доля, %	Коды специализаций HeadHunter
ИТ–специалисты высокого уровня квалификации	high	13,10	1.327, 1.272, 1.25, 1.113, 1.3
ИТ–специалисты низкого уровня квалификации	low	3,66	1.172, 1.296
Инженеры	engineers	16,18	1.82, 1.295, 1.117, 1.277
Разработчики программного обеспечения	soft	22,67	1.221
Веб и мультимедиа разработчики	web	20,13	1.9, 1.89, 1.10, 1.400, 1.475, 1.161
Системные администраторы и администраторы баз данных	admin	19,30	1.211, 1.270, 1.273, 1.50, 1.536, 1.420, 1.395, 1.203, 1.110
Другие	others	4,96	1.474, 1.359, 1.274

Определение 2. Профессиональная область (группа). Пусть H – набор кодов специализаций HeadHunter, а O – упорядоченный набор укрупненных профессиональных групп. Тогда профессиональная область $o \in O$ представляет собой 2-элементный кортеж $o = (L, a)$, где $L \subseteq H$ – подмножество кодов специализаций, соответствующих укрупненной профессиональной группе с сокращенным наименованием a в текстовом формате.

Для упрощения дальнейшего анализа извлечения ключевых ЗУН для определенных профессиональных групп (O, где $|O| = 7$) процесс обработки данных организован на всей выборке (в течение 5-летнего периода). В качестве допущения для такой агрегации используется относительное распределение вакансий по профессиональным группам (*рисунок 1*). Таким образом, относительное распределение вакансий по укрупненным профессиональным группам меняется незначительно во временной перспективе.

Для целей настоящего исследования ключевые ЗУН и их комбинации должны быть унифицированы и извлечены из набора вакансий (V). Однако, прежде чем вводить алгоритм извлечения ЗУН, каждая онлайн-вакансия, которая может относиться к нескольким профессиональным областям, должна быть сопоставлена с новой структурой данных (онлайн-вакансия ИТ).

Определение 3. Онлайн-вакансия ИТ.

Пусть $J \subseteq V$ – набор онлайн-вакансий ИТ, где $J = \{j_1, ..., j_m\}$: $m \le n$, $m \in \mathbb{N}$. Пусть $\mathbf{c}$ обозначает коды специализаций онлайн-вакансии: $\mathbf{c} = (c_1, ..., c_z) \in C$, где $z \le 6$. Пусть $\mathcal{L}$ –упорядоченный набор меток с отношением $(L, a) \in O \mapsto \mathcal{L}$ в следующей форме: $\mathcal{L} = (\lambda_1, ..., \lambda_q)$, где $q = |O|$. Введем функцию

$$f(\mathbf{c}, \lambda_q) = \begin{cases} 1, \mathbf{c} \subseteq L_q \\ 0 \end{cases},$$

которая ставит в соответствие код специализации o и код укрупненной профессиональной области из $\mathcal{L}$. Введем отношение $H : C \mapsto \mathcal{L}$, которое обеспечивает классификацию по нескольким меткам и отображает набор укрупненных профессиональных областей $\mathcal{L}$ на основе кодов специализаций: $\tilde{O} = H(\mathbf{c}) = \{\lambda \in \mathcal{L} \; f(\mathbf{c}, \lambda) = 1\}$, где $\tilde{O}$ – набор сокращенных наименований укрупненных профессиональных групп. Таким образом, онлан-вакансия ИТ j – это 3-элементный кортеж $j = (i, \tilde{O}, K)$.

В *таблице 5* представлено распределение полученных вакансий по укрупненным группам профессий. Полученное распределение профессиональных областей не является однородным. Иначе говоря, только относительно небольшая доля вакансий (6–30%) относится к одной укрупненной профессиональной группе. Другие вакансии связаны с несколькими такими группами. Таким образом, в последующем анализе необходимо разделение ЗУН, непосредственно связанных с конкретной профессиональной областью.

Для предоставления результатов поиска ключевых ЗУН по профессиональным группам разработан алгоритм извлечения ЗУН из неструктурированных онлайн-вакансий. Здесь и далее алгоритм извлечения ЗУН применен для выборки онлайн-вакансий сектора ИТ.

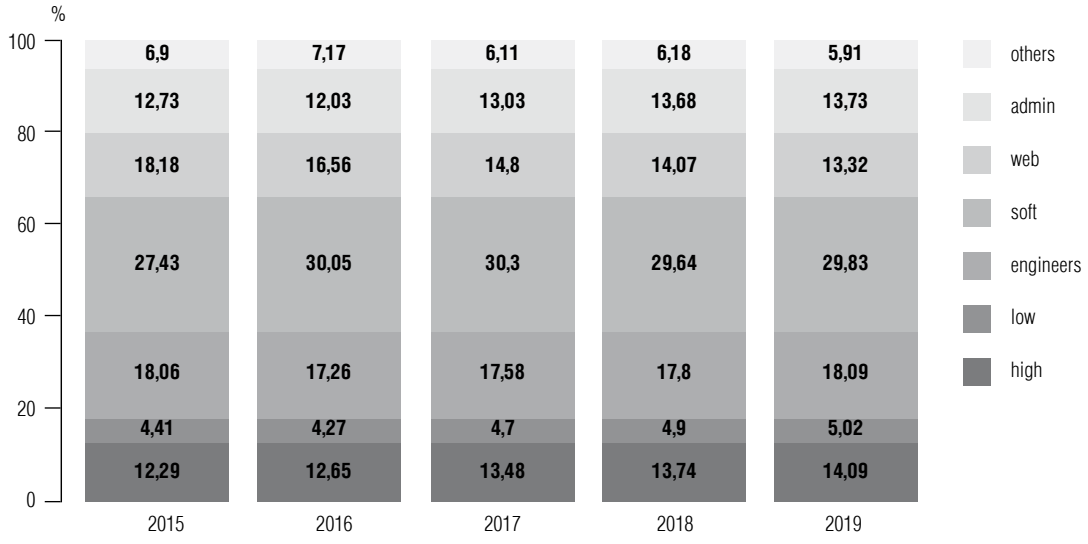

Рис. 1. Распределение количества вакансий по укрупненным профессиональным группам в динамике

Таблица 5.

Распределение укрупненных профессиональных групп в выборке

Сокращение	Количество вакансий	Доля вакансий, соответствующих только одной укрупненной группе профессий, %
high	19266	16,28
low	5383	17,28
engineers	23787	10,15
soft	33333	29,78
web	19312	9,29
admin	20825	6,18
others	7293	15,80

Вход: онлайн-вакансия ИТ (J).

Выход: набор стандартизированных ЗУН ($\tilde{K}$), сопоставленных с J.

1. Пусть $\tilde{S} \subseteq S$ набор уникальных текстовых описаний ЗУН из J
2. Пусть B 2-элементный кортеж: текстовое описание ЗУН и его частота встречаемости в J
3. **foreach** $\tilde{s}_i \in \tilde{S}$ **do**
4. $b_i \leftarrow$ частота встречаемости $\tilde{s}_i$ в J
5. $B[i] = (\tilde{s}_i , b_i)$
6. **end foreach**
7. сортируем B в убывающем порядке по b
8. **procedure** FrequentTerms(h, t)
9. $\tilde{h} \leftarrow$ подмножество h, если $h_i > t, \forall\, h_i \in h$
10. **return** $\tilde{h}$
11. **end procedure**
12. вводим пороговое значение t
13. $\tilde{B} \leftarrow$ FrequentTerms($B(b)$, t)
14. $T \leftarrow$ 3-элементный кортеж ЗУН после ручной разметки $T = (u, x, xs)$, где u идентификатор стандартизированного ЗУН, x – наименование в текстовом формате, xs – набор синонимов в текстовом формате для пары (u, x)
15. **function** Tokenizer(j)
16. нормализация пробельных символов
17. удаление знаков пунктуации
18. строчный регистр
19. стемминг (на английском и русском языках)
20. удаление стоп-слов (на английском и русском языках)
21. **end function**
22. **procedure** NGrams(J, n)
23. **for** j in J **do**
24. $G \leftarrow$ n-граммы размера n для Tokenizer(j)
25. положим G в ngramterms
26. **end for**
27. **return** ngramterms

end procedure
вводим пороговые значения t_1, t_2, t_3
ngram1:= FrequentTerms(NGrams($B(s)$, 1), t_1)
ngram2:= FrequentTerms(NGrams($B(s)$, 2), t_2)
ngram3:= FrequentTerms(NGrams($B(s)$, 3), t_3)
для полученных баз n-грамм проводится ручная разметка (чистка неинформативных терминов)
каждое наблюдение в данных базах n-грамм имеет набор идентификаторов (i, $\tilde{s}$)
procedure MatchTerms(X, Y)
Пусть L набор уникальных комбинаций из X и Y, где $L = \{l_1 | l_1 \in X, l_2 | l_2 \in Y\}$; l_1, l_2 наборы идентификаторов (i, $\tilde{s}$)
for l_1, l_2 in L **do**
$M \leftarrow$ индекс Жаккара: $\frac{|l_1(i) \cap l_2(i)|}{|l_1(i) \cup l_2(i)|}$
if M >0.5 **do**
положим (l_1, l_2) в termsmatched
end if
end for
return termsmatched
end procedure
для каждой комбинации из: ngram1, ngram2, ngram3 проводим MatchTerms(X, Y) → M1, M2, M3
для каждой комбинации из: T, M1, M2, M3 проводим MatchTerms($\tilde{X}$, $\tilde{Y}$), где $\tilde{X} := T$, $\tilde{Y} := T$ {M1,M2,M3}
$\tilde{K} \leftarrow X$ left-join Y
$\tilde{K}$ – полученная база данных со стандартизированными ЗУН, их синонимами и наборами n-грамм

Применение алгоритма к выборке данных онлайн-вакансий ИТ и извлечение ЗУН происходит следующим образом. В выборке представлены 13347 неструктурированных уникальных ЗУН. Описания таких ЗУН не унифицированы. Иначе говоря, каждая компания может ввести свою собственную текстовую строку от 1 до 100 символов. Например, запись одного и того же ЗУН может содержать одно слово/фразу или предложение, содержащее такие слова, разделенные символами пунктуации или пробелами. Для того, чтобы автоматизировать извлечение определенных ЗУН и унифицировать различные формы обозначения одного и того же термина, используются методы анализа и обработки текстовой информации.

В соответствии с [20], на первом этапе предварительной обработки данных можно построить n-граммы слов. Из векторного корпуса (TF-IDF) навыков, представленных в описаниях вакансий, были созданы уни-, би- и триграммы с использованием следующей токенизации: удаление всех знаков препинания и лишних пробелов, замена прописных букв строчными, стемминг слов на английском и на русском языках, удаление стоп-слов. В рамках данных терминов исходная структура и формулировки ЗУН были сохранены. На первом этапе были удалены неинформативные термины. Для базы данных униграмм получено 348 записей из 5234 неуникальных терминов; для биграмм – 577 из 1090; для триграмм – 110 из 303. Следующий этап – создание базы данных синонимов для уже полученных шаблонов. HeadHunter API (рекомендации по ключевым навыкам) позволяет получить часть синонимов для последующей ручной обработки полученных терминов. После получения данных синонимов была сформирована матрица терминов из 707 элементов (1296 записей для ручной проверки). Третий этап заключается в добавлении и корректировке терминов на основе опроса Stack Overflow Developer Survey[7] (108 терминов-наи-

[7] Stack Overflow Annual Developer Survey, https://insights.stackoverflow.com/survey/

менований наиболее популярных ИТ-технологий) и окончательном исправлении соответствующих терминов (база данных с исходными терминами и их синонимами).

В результате было получено 435 стандартизированных терминов (ЗУН) и 420 синонимов для них. Такой набор данных содержит как технические («жесткие»), так и нетехнические («мягкие») навыки для данного сектора. На последнем этапе обработки данных производится пересечение необработанных терминов (кодифицированных при помощи уникальных идентификационных кодов), точно совпадающих с конкретными терминами, полученными из скорректированного вручную списка синонимов HeadHunter и результатов, полученных из матриц TF-IDF (для уни-, би-, триграмм). В результате получаются пары «идентификатор ЗУН – термин» в текстовом формате. Для того, чтобы автоматически определить схожесть нескольких терминов (на основе уникального набора идентификаторов для каждого термина) и сопоставить остальные данные с заданными стандартизированными терминами, используется индекс Жаккара. Например, сходство между двумя наборами слов (терминов) *A* и *B* можно найти при помощи следующей формулы:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Эта мера подходит для сравнения категориальных данных, и ее значение находится в диапазоне от 0 до 1 включительно. Тем не менее, выбор порога отсечения зависит от структуры данных и целей исследования. В качестве порога для идентификации близких терминов используется уровень выше 0,5 (после ручной обработки полученных данных). Таким образом, 53672 вакансий (95,8% от исходной выборки) содержат, по крайней мере, один из стандартизированных ЗУН из ранее полученного набора данных терминов и их синонимов. Процентное распределение ТОП-20 стандартизированных ЗУН (по частоте встречаемости в выборке) среди всего набора данных представлено в *таблице 6.*

Анализ ключевых востребованных (со стороны работодателя) ЗУН (и их комбинаций) основывается на определении специфичных ЗУН для конкретной профессиональной области, которые в то же время достаточно часто встречаются в соответствующих вакансиях.

После подготовки данных набора вакансий получена следующая структура данных: 305217 наблюдений (определенный ЗУН из вакансии), где каждое наблюдение имеет идентификатор стандартизированного ЗУН, идентификатор вакансии и код укрупненной профессиональной группы. Для того, чтобы обеспечить классификацию ЗУН в соответствии с указанными группами вакансий, для каждой из групп вакансий был проведен поиск пар и триплетов ЗУН. Из полученных пар и триплетов ЗУН (отличных от нуля по индексу схожести Жаккара) были извлечены ЗУН с высокой степенью совпадения. Общая схема предложенного алгоритма, реализованного для предложенного набора данных, представлена на *рисунке 2.*

На первом этапе по 435 стандартизованным ЗУН, их парам (C^2_{435} = 94435) и триплетам (C^3_{435} = 1362345) были найдены индексы сходства Жаккара для каждой из семи укрупненных групп вакансий. Далее,

Таблица 6.

ТОП-20 ЗУН по частоте встречаемости в выборке ИТ сектора

ЗУН	Частота присутствия в базе данных ЗУН, %
HTML/CSS	6,73
JavaScript	4,69
1С	3,48
SQL	3,25
PHP	2,63
Git	2,53
Linux	2,32
Java	2,28
MySQL	1,86
Навыки переговоров	1,61
Sales Skills	1,52
Деловая коммуникация	1,51
English	1,45
Testing Framework	1,43
Python	1,40
jQuery	1,36
C/C++	1,30
ООП	1,29
C#	1,28
.net	1,27

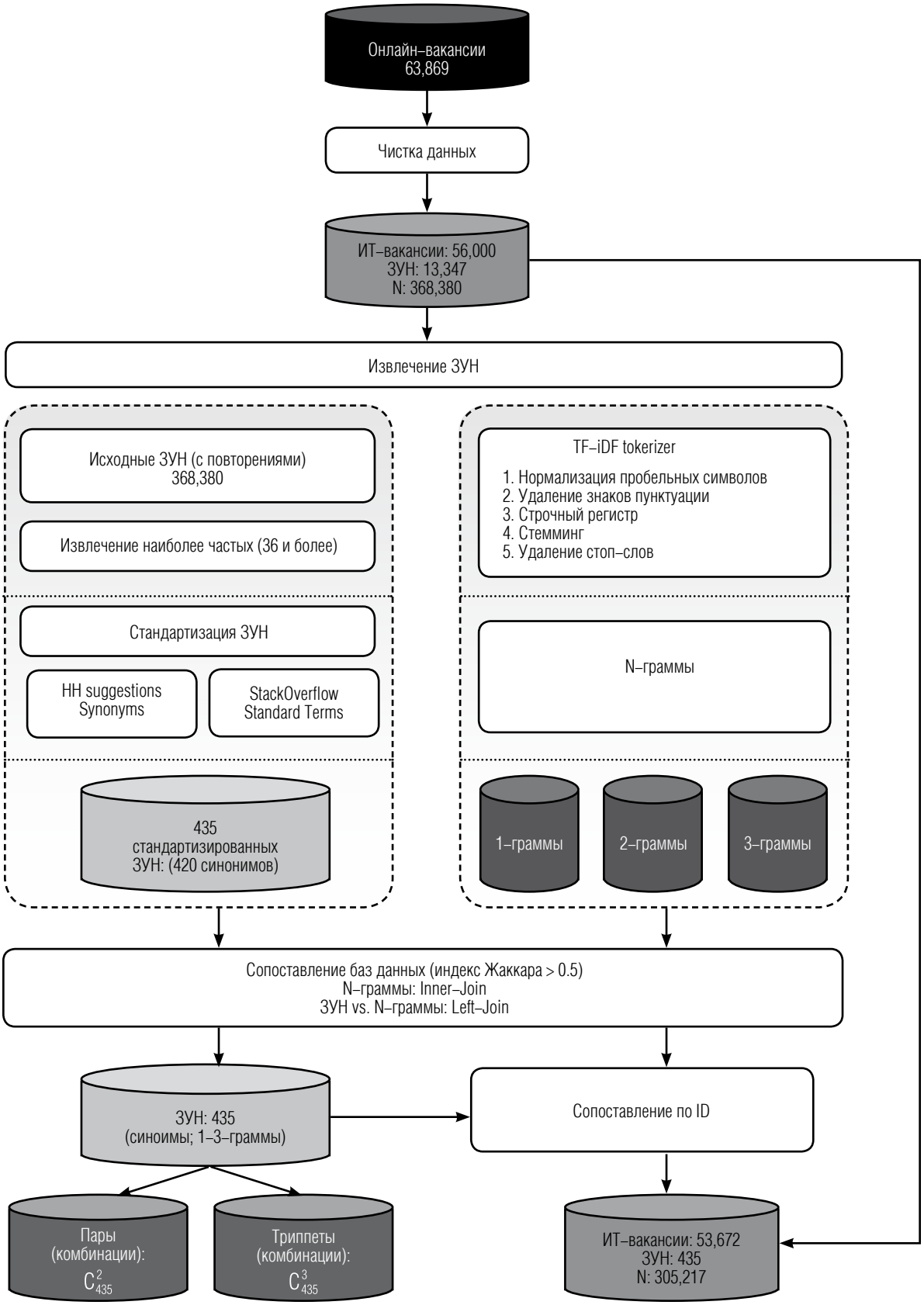

Рис. 2. Схема реализации алгоритма извлечения ЗУН

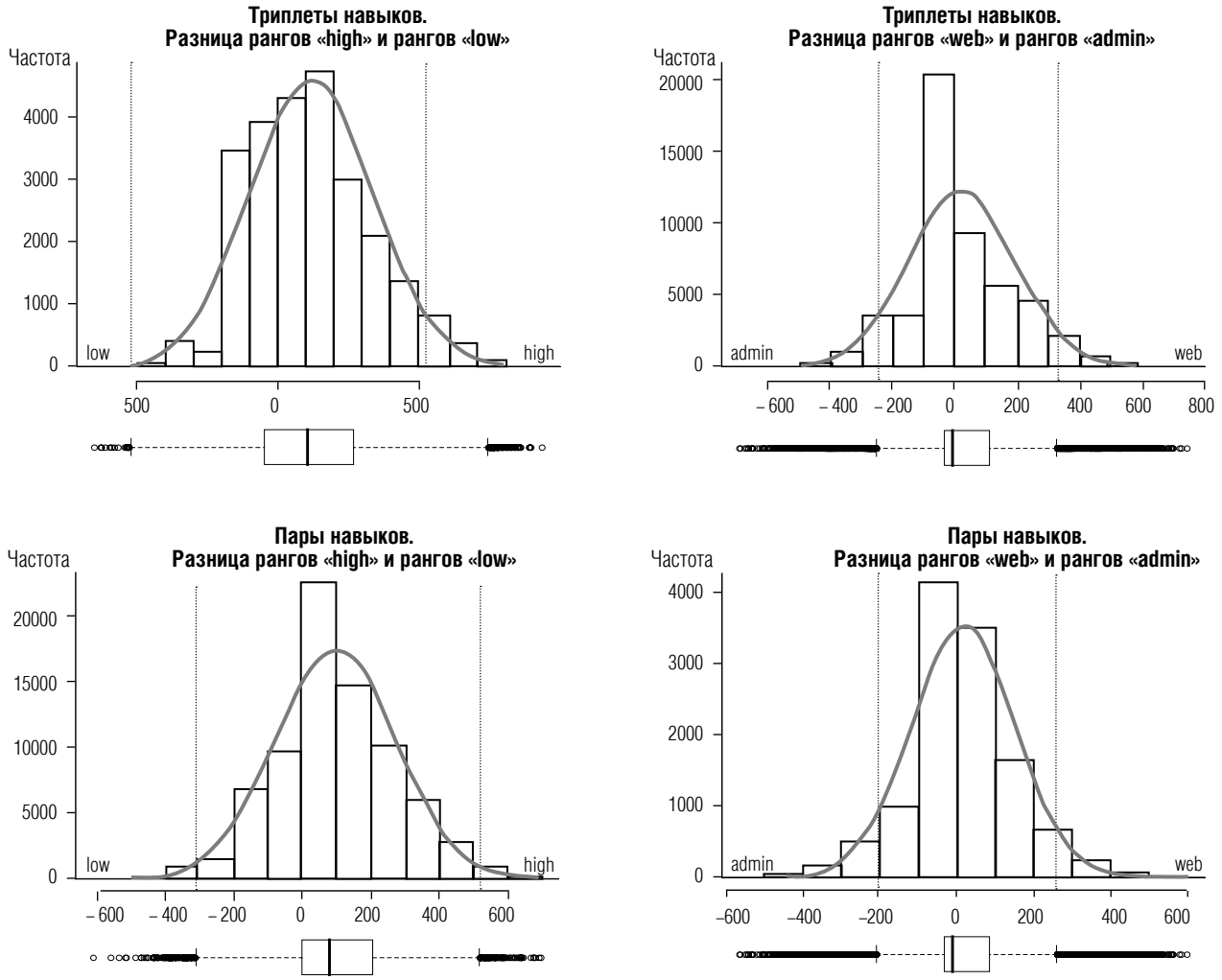

Рис. 3. Распределение разностей рангов индексов схожести Жаккара по укрупненным группам вакансий

используя пары и триплеты ЗУН (уникальные сочетания без повторений), каждый набор ЗУН был проранжирован по индексу Жаккара в пределах их квантильного распределения (с шагом в 0,1% для пар и триплетов).

Для каждой пары и триплета ЗУН такой показатель рассчитывался исходя из количества вакансий, в которые входят те или иные комбинации ЗУН. На следующем этапе, были найдены разности в рангах для каждой пары из укрупненных групп вакансий. Далее, для выявления специфичных наборов ЗУН для каждой профессиональной области были найдены выбросы в предложенном распределении разностей рангов. С точки зрения статистического аппарата, близость распределения разностей рангов к нормальному позволяет обеспечить соответствующий отбор специфичных и при этом ключевых ЗУН, которые определяют различные укрупненные группы вакансий. Например, несколько пар таких групп представлены на *рисунке 3* (в качестве границ для отсечения ЗУН используются границы усов из ящика с усами: 1,5 IQR[8] ниже и выше для подходящей разности в квантильных рангах).

Для извлеченных пар и триплетов применяется процедура добавления уникальных наборов ЗУН для каждой комбинации профессиональных групп. Таким образом, на пересечении ЗУН извле-

[8] IQR – межквартильный размах

каются только значения, составляющие более 95% (по разнице в квантилях) для того, чтобы обнаружить ключевые и специфичные наборы ЗУН. На следующем этапе для каждой пары групп вакансий ($C_7^2 = 21$) были определены ключевые определяющие ЗУН и их комбинации. Таким образом, к ЗУН, которые являются уникальными в определенной группе вакансий в последнем дециле, были добавлены уникальные ЗУН, исходя из их пересечения с другими группами вакансий. В итоге были получены три матрицы 7×7, где на пересечении *i*-й строки и *j*-го столбца ($i \neq j$) находятся уникальные наборы ЗУН (кодированные по-отдельности для уникальных ЗУН, их пар и триплетов) и уникальные ЗУН из *i*-й группы (превышающие порог в 95%) по сравнению с *j*-й группой вакансий. Наконец, данные ЗУН были распределены следующим образом: наличие основных ЗУН, которые определяют, что каждая профессиональная группа была уже установлена, использовалось для пересечения строк (для каждой из заданных матриц), что позволило получить список ключевых и специфиных ЗУН для данной профессиональной группы. Используются следующие пороговые значения: для пар пороговое значение не менее 2/3 отличий от других групп (4 и более отличных групп из оставшихся шести); для триплетов – 100% отличий (6 из 6). Используя приведенную выше последовательность действий, списки таких навыков были получены для каждой конкретной профессиональной группы вакансий, которые, в свою очередь, представляют собой ключевые (наиболее востребованные) навыки присущие конкретной профессиональной группе.

3. Результаты

На этапе определения ключевых и различных ЗУН для разных профессиональных групп в сфере ИТ извлекаются наиболее популярные из них. В соответствии с группами профессий, такие ЗУН могут быть представлены в виде облаков слов по ТОП-50 ЗУН для каждой профессиональной группы по количеству в описании вакансий (*рисунок 4*).

Однако при наличии вакансий, связанных с несколькими профессиональными группами, некоторые ЗУН дублируются. Таким образом, на этапе извлечения пар и триплетов ЗУН, используя пересечение специфичных ЗУН, такое дублирование нивелируется. Наиболее востребованные и в то же время спепцифичные профессиональные пары ЗУН представлены в *таблице 7*, триплеты – в *таблице 8*[9].

Наборы востребованных ЗУН могут быть идентифицированы для разных профессиональных групп. Кроме того, используя пары и триплеты ЗУН, становится возможным получить конкретные их комбинации. Таким образом, предлагаемые методы обработки и извлечения ЗУН могут быть полезны для более широкого понимания и анализа спроса на рынке труда, а также предоставляют дополнительную информацию для системы образования с точки зрения поддержания актуальности и обновления образовательных стандартов в части ЗУН.

Заключение

Применимость представленного в статье алгоритма имеет ряд особенностей.

Используемая база данных имеет предопределенный набор профессиональных групп. С одной стороны, это упрощает реализацию алгоритма. С другой стороны, полученные результаты свидетельствуют о значительном пересечении одинаковых ЗУН для разных групп. Не исключено, что детерминированная структура профессиональных групп не является однозначной, что создает излишнюю вариативность при идентификации уникальных комбинаций ЗУН для укрупненных профессиональных групп. Другими словами, введение классификации или кластеризации для выявления профессиональных групп может улучшить общие результаты. Тем не менее, предоставленный список комбинаций ЗУН построен с точки зрения сохранения специфики наборов ЗУН для конкретной группы профессий.

В одной конкретной вакансии могут быть заявлены совершенно разные профессиональные группы. В таком случае группировка ЗУН (например, на «мягкие» и «жесткие» навыки) может быть использована в качестве дополнительного фактора при классификации. Более того, существуют ЗУН, связанные как с конкретной технологией, так и программными средами для ее реализации, которые не могут быть отделены друг от друга.

Введение большей последовательности слов в *n*-граммах (4 и более) может предоставить больше

[9] Полные списки полученных пар и триплетов могут быть предоставлены авторами по запросу

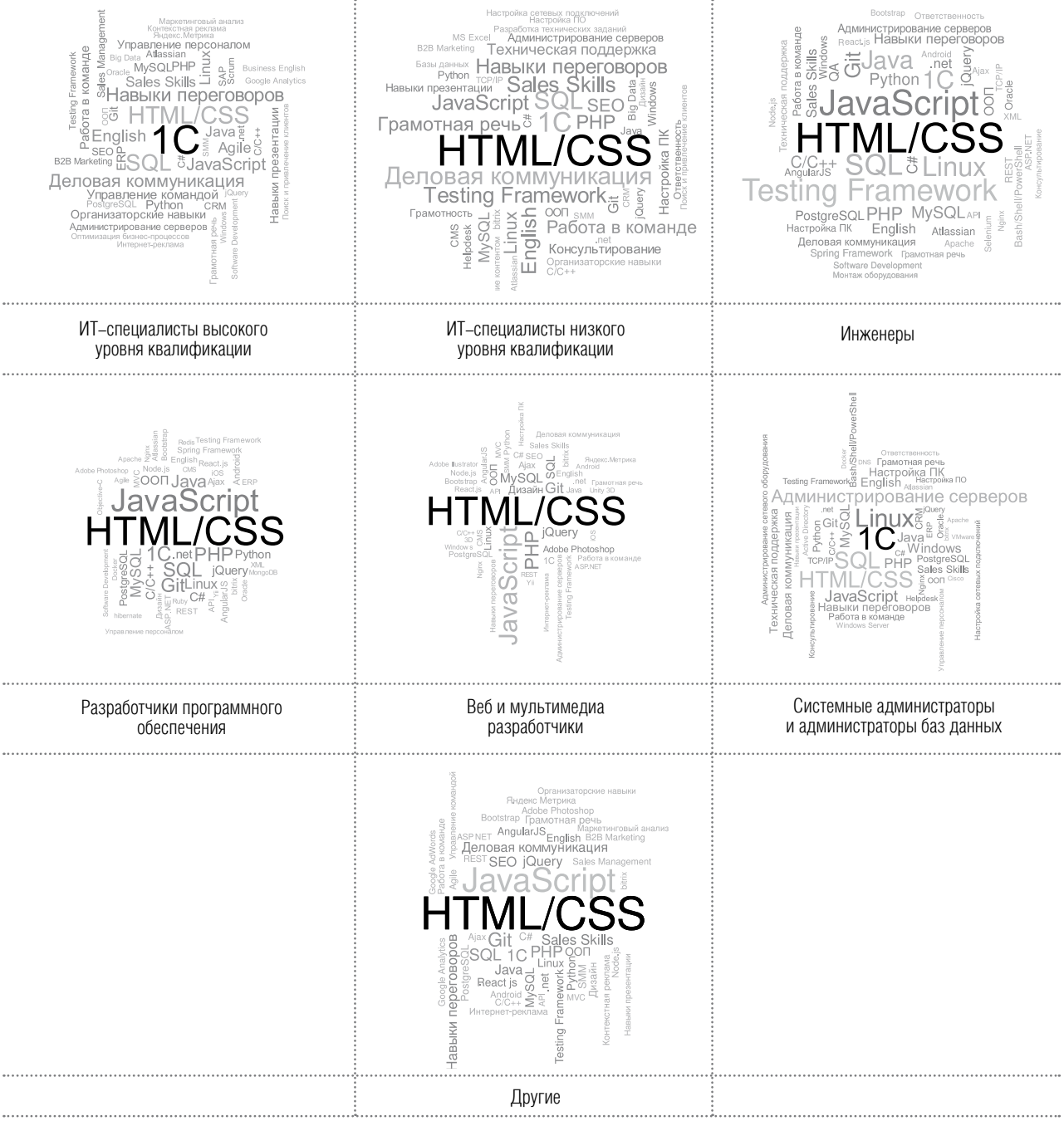

Рис. 4. ТОП–50 ЗУН по профессиональным группам

информации для извлечения ЗУН из неструктурированных баз данных. Однако временная сложность для вычисления таких алгоритмов может быть чрезвычайно высокой, что может потребовать упрощения вычисления метрик подобия (например, с использованием хеш-функций и приближенных формул).

Расширенная реализация алгоритма может быть нацелена на выявление ключевых ЗУН для других секторов рынка труда, учет изменений во времени и организацию межрегиональных сравнений. Результаты могут быть агрегированы с использованием официальной статистики (если цели работы будут направлены на решение общеэкономических вопросов).

Таблица 7.

Ключевые пары ЗУН в профессиональных группах

ЗУН №1	ЗУН №2	Индекс Жаккара	Группа
Dart	Flutter	0,167	high
Billing	Solaris	0,136	high
Arduino	Raspberry Pi	0,125	high
Технические средства информационной защиты	Assembly	0,073	high
Средства криптографической защиты информации	Assembly	0,065	high
Автоматизация производства	САПР	0,125	low
CCNA	OSPF	0,125	low
A/B тесты	Mobile Marketing	0,100	low
Arduino	ARM	0,083	low
Оптимизация бизнес–процессов	Citrix	0,038	low
Системы мониторинга сетей	Google Cloud Platform	0,071	engineers
Cordova	Xamarin	0,065	engineers
Управление персоналом	Яндекс.Метрика	0,014	engineers
Elasticsearch	Node.js	0,011	engineers
MS SharePoint	Windows	0,010	engineers
Firebase	Google Cloud Platform	0,083	soft
Грамотная речь	Составление договоров	0,015	soft
Elasticsearch	Yii	0,011	soft
Scrum	TFS	0,010	soft
Контекстная реклама	Поиск и привлечение клиентов	0,006	soft
Business Intelligence Systems	Olap	0,063	web
3D	Altium Designer	0,022	web
SPA	Unit Testing	0,018	web
Написание статей	Google AdWords	0,017	web
API	Mercurial	0,016	web
Технический аудит сайта	Технический перевод	0,074	admin
Аналитические исследования	Системный анализ	0,033	admin
Apache	Windows Server	0,029	admin
REST	Xsd	0,022	admin
API	Xsd	0,016	admin
Корректура текстов	Adobe Lightroom	0,111	others
Мобильность	Billing	0,111	others
Pandas	Wifi networks	0,100	others
Технический аудит сайта	SMO	0,091	others
A/B тесты	Business Analysis	0,080	others

Таблица 8.

Ключевые триплеты ЗУН в профессиональных группах

ЗУН №1	ЗУН №2	ЗУН №3	Индекс Жаккара	Группа
ARM	GCC	Raspberry Pi	0,019	high
Медиа–планирование	Планирование маркетинговых кампаний	Facebook	0,018	high
CentOS	EJB	NetBeans	0,017	high
Обработка видео	Adobe Premier Pro	SketchUp	0,013	high
Бизнес–планирование	Mobile Marketing	Product Marketing	0,012	high
Автоматизация производства	КИПиА	САПР	0,050	low
Автоматизация технологических процессов	КИПиА	САПР	0,048	low
Автоматизация производства	Автоматизация технологических процессов	САПР	0,043	low
Debian	OSPF	VLAN	0,043	low
Аналитические исследования	Business Analysis	Product Marketing	0,038	low
Обработка видео	Обработка изображений	Adobe Lightroom	0,020	engineers
Ведение переписки на иностранном языке	Написание пресс–релизов	Технический перевод	0,018	engineers
Написание пресс–релизов	Письменный перевод	Технический перевод	0,015	engineers
Обработка изображений	Adobe After Effects	Adobe Lightroom	0,014	engineers
FreeBSD	OSPF	VLAN	0,014	engineers
Поисковая оптимизация сайтов	Работа с биржами	Технический аудит сайта	0,021	soft
Математический анализ	MATLAB	R	0,017	soft
Математическая статистика	MATLAB	R	0,016	soft
Корректура текстов	Написание пресс–релизов	Подготовка презентаций	0,013	soft
Работа с биржами	Технический аудит сайта	SMO	0,013	soft
Microsoft Azure	TensorFlow	Torch/PyTorch	0,020	web
Баннерная реклама	Обработка видео	Adobe Premier Pro	0,016	web
Ведение отчетности	Налоговая отчетность	Billing	0,016	web
Корректура текстов	Написание пресс–релизов	Рерайтинг	0,015	web
Microsoft Azure	Spark	TensorFlow	0,014	web
Математический анализ	Olap	VBA	0,019	admin
Математический анализ	A/B тесты	R	0,018	admin
Chef	LDAP	Wifi networks	0,015	admin
BGP	Chef	LDAP	0,013	admin
Chef	LDAP	OSPF	0,013	admin
Внутренняя оптимизация сайта	Технический аудит сайта	SMO	0,063	others
Внутренняя оптимизация сайта	Российские поисковые системы	SMO	0,048	others
Flask	Pandas	Wifi networks	0,037	others
Мобильность	Электронный документооборот	Billing	0,031	others
Внутренняя оптимизация сайта	Лидогенерация	SMO	0,027	others

В итоге результаты настоящего исследования позволяют выделить несколько значимых моментов. Во-первых, предложенный алгоритм позволяет выявить и стандартизировать ключевые ЗУН, которые могут быть применимы для создания системы русскоязычного классификатора профессий, знаний, умений и навыков. Во-вторых, алгоритм позволяет предоставлять списки наиболее популярных (ключевых) комбинаций ЗУН, которые высоко востребованы компаниями и работодателями для каждой конкретной вакансии. Наконец, гибкость алгоритма позволяет сочетать его с методами классификации и кластеризации данных, которые могут быть полезны для исследований рынка труда. ■

Литература

1. Autor D.H., Levy F., Murnane R.J. The skill content of recent technological change: An empirical exploration // The Quarterly Journal of Economics. 2003. Vol. 118. No 4. P. 1279–1333. DOI: 10.1162/003355303322552801.
2. Bensberg F., Buscher G., Czarnecki C. Digital transformation and IT topics in the consulting industry: A labor market perspective // Advances in consulting research: Recent findings and practical cases / V. Nissen (Ed.). Cham, Switzerland: Springer, 2019. P. 341–357.
3. Christoforaki M., Ipeirotis P.G. A system for scalable and reliable technical-skill testing in online labor markets // Computer Networks. 2015. No 90. P. 110–120. DOI: 10.1016/j.comnet.2015.05.020.
4. Florea R., Stray V. Software tester, we want to hire you! An analysis of the demand for soft skills // 19th International Conference on Agile Processes in Software Engineering and Extreme Programming (XP 2018), Porto, Portugal, 21–25 May 2018. P. 54–67.
5. Goles T., Hawk S., Kaiser K.M. Information technology workforce skills: The software and IT services provider perspective // Information Systems Frontiers. 2008. Vol. 10. No 2. P. 179–194.
6. Johnson K.M. Non-technical skills for IT professionals in the landscape of social media // American Journal of Business and Management. 2016. Vol. 4. No 3. P. 102–122. DOI: 10.11634/216796061504668.
7. Skills for success at different stages of an IT professional's career / L. Kappelman [et al.] // Communications of the ACM. 2016. Vol. 59. No 8. P. 64–70. DOI: 10.1145/2888391.
8. Litecky C.R., Arnett K.P., Prabhakar B. The paradox of soft skills versus technical skills in is hiring // Journal of Computer Information Systems. 2004. Vol. 45. No 1. P. 69–76.
9. Havelka D., Merhout J.W. Toward a theory of information technology professional competence // Journal of Computer Information Systems. 2009. Vol. 50. No 2. P. 106–116.
10. Hussain W., Clear T., MacDonell S. Emerging trends for global DevOps: A New Zealand perspective // IEEE 12th International Conference on Global Software Engineering, Buenos Aires, Argentina, 22–23 May 2017. Vol. 1. P. 21–30. DOI: 10.1109/ICGSE.2017.16.
11. Wowczko I. Skills and vacancy analysis with data mining techniques // Informatics. 2015. Vol. 2. No 4. P. 31–49. DOI: 10.3390/informatics2040031.
12. Bailey J., Mitchell R.B. Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills // Journal of Computer Information Systems. 2006. Vol. 47. No 2. P. 28–33.
13. Brooks N.G., Greer T.H., Morris S.A. Information systems security job advertisement analysis: Skills review and implications for information systems curriculum // Journal of Education for Business. 2018. Vol. 93. No 5. P. 213–221.
14. Casado-Lumbreras C., Colomo-Palacios R., Soto-Acosta P. A vision on the evolution of perceptions of professional practice // International Journal of Human Capital and Information Technology Professionals. 2015. Vol. 6. No 2. P. 65–78. DOI: 10.4018/IJHCITP.2015040105.
15. Föll P., Thiesse F. Aligning is curriculum with industry skill expectations: A text mining approach // 25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, 5–10 June 2017. P. 2949–2959.
16. Stal J., Paliwoda-Pękosz G. Fostering development of soft skills in ICT curricula: A case of a transition economy // Information Technology for Development. 2019. Vol. 25. No 2. P. 250–274. DOI: 10.1080/02681102.2018.1454879.
17. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. Classifying online job advertisements through machine learning // Future Generation Computer Systems. 2018. Vol. 86. P. 319–328.
18. Colombo E., Mercorio F., Mezzanzanica M. AI meets labor market: Exploring the link between automation and skills // Information Economics and Policy. 2019. No 47. P. 27–37. DOI: 10.1016/j.infoecopol.2019.05.003.
19. Data mining approach to monitoring the requirements of the job market: A case study / I. Karakatsanis [et al.] // Information Systems. 2017. No 65. P. 1–6. DOI: 10.1016/j.is.2016.10.009.
20. Lovaglio P.G., Cesarini M., Mercorio F., Mezzanzanica M. Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies // Statistical Analysis and Data Mining. 2018. Vol. 11. No 2. P. 78–91. DOI: doi.org/10.1002/sam.11372.
21. Challenge: Processing web texts for classifying job offers / F. Amato [et al.] // IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, California, USA, 7–9 February 2015. P. 460–463.
22. De Mauro A., Greco M., Grimaldi M., Ritala P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets // Information Processing & Management. 2018. Vol. 54. No 5. P. 807–817. DOI: 10.1016/j.ipm.2017.05.004.

23. Gurcan F., Cagiltay N.E. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling // IEEE Access. 2019. No 7. P. 82541–82552.

24. Pejic-Bach M., Bertoncel T., Meško M., Krstić Ž. Text mining of industry 4.0 job advertisements // International Journal of Information Management. 2020. No 50. P. 416–431.

25. Radovilsky Z., Hegde V., Acharya A., Uma U. Skills requirements of business data analytics and data science jobs: A comparative analysis // Journal of Supply Chain and Operations Management. 2018. Vol. 16. No 1. P. 82–101.

Об авторах

Терников Андрей Александрович

аспирант (аспирантская школа по экономике), преподаватель департамента менеджмента,
Санкт-Петербургская школа экономики и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»,
194100, г. Санкт-Петербург, Кантемировская ул., д. 3, корп., 1, лит. А;

E-mail: aternikov@hse.ru

ORCID: 0000-0003-2354-0109

Александрова Екатерина Александровна

кандидат экономических наук;

директор, Международный центр экономики, управления и политики в области здоровья;
доцент департамента экономики, Санкт-Петербургская школа экономики и менеджмента;
доцент департамента финансов, Санкт-Петербургская школа экономики и менеджмента,
Национальный исследовательский университет «Высшая школа экономики»,
194100, г. Санкт-Петербург, Кантемировская ул., д. 3, корп., 1, лит. А;

E-mail: ea.aleksandrova@hse.ru

ORCID: 0000-0001-7067-5087

Demand for skills on the labor market in the IT sector

Andrei A. Ternikov
E-mail: aternikov@hse.ru

Ekaterina A. Aleksandrova
E-mail: ea.aleksandrova@hse.ru

National Research University Higher School of Economics
Address: 3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia

Abstract

One of the most dynamically changing parts of the labor market relates to information technologies. Skillsets demanded by employers in this sphere vary across different industries, organizations and even certain vacancies. The educational system in the most cases lags behind such changes, so that obsolete skillsets are being taught. This article proposes an algorithm of skillsets identification that allows us to extract skills that are needed by companies from different occupational groups in the information technologies sector. Using the unstructured online-vacancies database for the Russian regional labor market, skills are extracted and unified with the use of TF-IDF and *n*-grams approaches. As a result, key specific skillsets for various occupations are found. The proposed algorithm allows us to identify and standardize key skills which might be applicable to create a system of Russian classification for occupations and skills. In addition, the algorithm allows us to provide lists of the key combinations of skills that are in high demand among companies inside each particular occupation.

Key words: job vacancies in IT sector; online vacancies; unstructured data analysis; labor market; demand on skills of job candidates; combinations of skillsets.

Citation: Ternikov A.A., Aleksandrova E.A. (2020) Demand for skills on the labor market in the IT sector. *Business Informatics*, vol. 14, no 2, pp. 64–83. DOI: 10.17323/2587-814X.2020.2.64.83

References

1. Autor D.H., Levy F., Murnane R.J. (2003) The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, vol. 118, no 4, pp. 1279–1333. DOI: 10.1162/003355303322552801.
2. Bensberg F., Buscher G., Czarnecki C. (2019) Digital transformation and IT topics in the consulting industry: A labor market perspective. *Advances in consulting research: Recent findings and practical cases* (ed. V. Nissen). Cham, Switzerland: Springer, pp. 341–357.
3. Christoforaki M., Ipeirotis P.G. (2015) A system for scalable and reliable technical-skill testing in online labor markets. *Computer Networks*, vol. 90, pp. 110–120. DOI: 10.1016/j.comnet.2015.05.020.
4. Florea R., Stray V. (2018) Software tester, we want to hire you! An analysis of the demand for soft skills. Proceedings of the *19th International Conference on Agile Processes in Software Engineering and Extreme Programming (XP 2018), Porto, Portugal, 21–25 May 2018* (eds. J. Garbajosa, X. Wang, A. Aguiar), pp. 54–67.
5. Goles T., Hawk S., Kaiser K.M. (2008) Information technology workforce skills: The software and IT services provider perspective. *Information Systems Frontiers*, vol. 10, no 2, pp. 179–194.
6. Johnson K.M. (2016) Non-technical skills for IT professionals in the landscape of social media. *American Journal of Business and Management*, vol. 4, no 3, pp. 102–122. DOI: 10.11634/216796061504668.
7. Kappelman L., Jones M.C., Johnson V., McLean E.R., Boonme K. (2016) Skills for success at different stages of an IT professional's career. *Communications of the ACM*, vol. 59, no 8, pp. 64–70. DOI: 10.1145/2888391.
8. Litecky C.R., Arnett K.P., Prabhakar B. (2004) The paradox of soft skills versus technical skills in is hiring. *Journal of Computer Information Systems*, vol. 45, no 1, pp. 69–76.
9. Havelka D., Merhout J.W. (2009) Toward a theory of information technology professional competence. *Journal of Computer Information Systems*, vol. 50, no 2, pp. 106–116.
10. Hussain W., Clear T., MacDonell S. (2017) Emerging trends for global DevOps: A New Zealand perspective. Proceedings of the *IEEE 12th International Conference on Global Software Engineering, Buenos Aires, Argentina, 22–23 May 2017* (ed. R. Bilof), vol. 1, pp. 21–30. DOI: 10.1109/ICGSE.2017.16.
11. Wowczko I. (2015) Skills and vacancy analysis with data mining techniques. *Informatics*, vol. 2, no 4, pp. 31–49. DOI: 10.3390/informatics2040031.
12. Bailey J., Mitchell R.B. (2006) Industry perceptions of the competencies needed by computer programmers: Technical, business, and soft skills. *Journal of Computer Information Systems*, vol. 47, no 2, pp. 28–33.
13. Brooks N.G., Greer T.H., Morris S.A. (2018) Information systems security job advertisement analysis: Skills review and implications for information systems curriculum. *Journal of Education for Business*, vol. 93, no 5, pp. 213–221.
14. Casado-Lumbreras C., Colomo-Palacios R., Soto-Acosta P. (2015) A vision on the evolution of perceptions of professional practice. *International Journal of Human Capital and Information Technology Professionals*, vol. 6, no 2, pp. 65–78. DOI: 10.4018/IJHCITP.2015040105.
15. Föll P., Thiesse F. (2017) Aligning is curriculum with industry skill expectations: A text mining approach. Proceedings of the *25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, 5–10 June 2017* (eds. I. Ramos, V. Tuunainen, H. Krcmar), pp. 2949–2959.
16. Stal J., Paliwoda-Pękosz G. (2019) Fostering development of soft skills in ICT curricula: A case of a transition economy. *Information Technology for Development*, vol. 25, no 2, pp. 250–274. DOI: 10.1080/02681102.2018.1454879.
17. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, vol. 86, pp. 319–328.
18. Colombo E., Mercorio F., Mezzanzanica M. (2019) AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, no 47, pp. 27–37. DOI: 10.1016/j.infoecopol.2019.05.003.
19. Karakatsanis I., AlKhader W., MacCrory F., Alibasic A., Omar M.A., Aung Z., Woon W.L. (2017) Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, no 65, pp. 1–6. DOI: 10.1016/j.is.2016.10.009.
20. Lovaglio P.G., Cesarini M., Mercorio F., Mezzanzanica M. (2018) Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining*, vol. 11, no 2, pp. 78–91. DOI: doi.org/10.1002/sam.11372
21. Amato F., Boselli R., Cesarini M., Mercorio F., Mezzanzanica M., Moscato V., Picariello A. (2015) Challenge: Processing web texts or classifying job offers. Proceedings of the *2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, California, USA, 7–9 February 2015* (eds. M.S. Kankanhalli, T. Li, W. Wang), pp. 460–463.
22. De Mauro A., Greco M., Grimaldi M., Ritala P. (2018) Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management*, vol. 54, no 5, pp. 807–817. DOI: 10.1016/j.ipm.2017.05.004.
23. Gurcan F., Cagiltay N.E. (2019) Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access*, no 7, pp. 82541–82552.
24. Pejic-Bach M., Bertoncel T., Meško M., Krstić Ž. (2020) Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, no 50, pp. 416–431.
25. Radovilsky Z., Hegde V., Acharya A., Uma U. (2018) Skills requirements of business data analytics and data science jobs: A comparative analysis. *Journal of Supply Chain and Operations Management*, vol. 16, no 1, pp. 82–101.

About the authors

Andrei A. Ternikov

Doctoral Student, Doctoral School on Economics;

Lecturer, Department of Management, St. Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: aternikov@hse.ru

ORCID: 0000-0003-2354-0109

Ekaterina A. Aleksandrova

Cand. Sci. (Econ.);

Director, International Centre for Health Economics, Management, and Policy;
Associate Professor, Department of Economics, St.Petersburg School of Economics and Management;
Associate Professor, Department of Finance, St.Petersburg School of Economics and Management,
National Research University Higher School of Economics,
3, Kantemirovskaya Street, Saint-Petersburg 194100, Russia;

E-mail: ea.aleksandrova@hse.ru

ORCID: 0000-0001-7067-5087