О варианте формализации задачи извлечения ключевых навыков и кластерного анализа вакансий при реализации комплексного инструментарий цифровой профориентации

М.Е. Диков, С.Н. Широбокова

Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова, Новочеркасск

Аннотация: В статье предложена общая формализованная модель задачи обработки и извлечения потенциальных ключевых навыков из описаний вакансий для определения востребованности направлений подготовки и возможных сфер трудоустройства выпускников. Формализованная модель использована в программной реализации модуля кластеризации вакансий по полученным множествам ключевых навыков в рамках комплексного инструментария дистанционной профориентации.
Ключевые слова: вакансии, востребованность направлений подготовки, профориентация, цифровизация профориентационной деятельности, формализованная модель, кластеризация, профессии, ключевые навыки.

Взаимоотношения высшего образования и рынка труда в современном российском обществе характеризуются определенными проблемами, в числе которых можно отметить несоответствие спроса абитуриентов на определённые профессии потребностям рынка труда, что в дальнейшем вызывает необходимость перепрофилирования выпускников при выходе в профессиональную среду; несоответствие темпов развития образования и экономики современного общества, консервативность сферы образования и недостаточная чувствительность к изменениям в различных секторах экономики и потребностям заинтересованных сторон [1]. Современным подходом повышения качества профориентации и адаптации направлений подготовки учебных заведений к быстро меняющимся требованиям рынка труда является использование цифровых инструментов.

В рамках разработки комплексного инструментария цифровой профориентации [2] требуется реализовать кластерный анализ вакансий на рынке труда, полученных с помощью поисковых запросов к сервисам, размещающим вакансии, используя ключевые слова и фразы [3-4], характеризующие навыки и компетенции направлений подготовки [5-6].

Кластерный анализ вакансий осуществляется по ключевым словам и фразам, извлеченным из описаний вакансий [7], которые отражают основной контекст. Далее вакансии представляются в векторной форме, которую можно использовать при кластерном анализе.

На рис. 1 приведена концептуальная модель взаимодействия администратора с сервисом, который выполняет предварительную работу по подготовке первичных данных для последующей работы системы.

Для реализации дальнейших этапов анализа данных ранее была проведена формализация предметной области, в рамках которой общая модель вакансий представлена следующим образом:

$$vacancy_i =< Id_i, Srv_i, Url_i, PblAt_i, Nm_i, Dscr_i, Emp_i, Cntr_i, Ar_i, Twn_i,$$
$$SlrFrm_i, SlrT_i, Crr_i, Exp_i, Schd_i, Edc_i, AcTmp_i, AcVp_i, AcKs_i,$$
$$Kwrds_i, KSklls_i >, i = \overline{1, N},$$

где Id_i – уникальный идентификатор, Srv_i – идентификатор сервиса, Url_i – ссылка на сервисе, $PblAt_i$ – дата публикации, Nm_i – наименование, $Dscr_i$ – текстовое описание, Emp_i – наименование работодателя, $Cntr_i$ – страна, Ar_i – регион, Twn_i – населенный пункт, $SlrFrm_i$ – минимальная заработная плата, $SlrT_i$ – максимальная заработная плата, Crr_i – валюта заработной платы, Exp_i – требуемый опыт работы, $Schd_i$ – график работы (вахтовый метод, гибкий график, сменный график и т.д.), Edc_i – уровень образования, $AcTmp_i$ – доступность временного трудоустройства, $AcVp_i$ – доступность для инвалидов и слабозащищенных слоев населения, $AcKs_i$ – доступность для подростков, $Kwrds_i$ – множество ключевых слов (навыков и компетенций направления подготовки), по которому найдена вакансия, $KSklls_i$ – множество ключевых навыков вакансии, N – общее количество вакансий со всех сервисов.

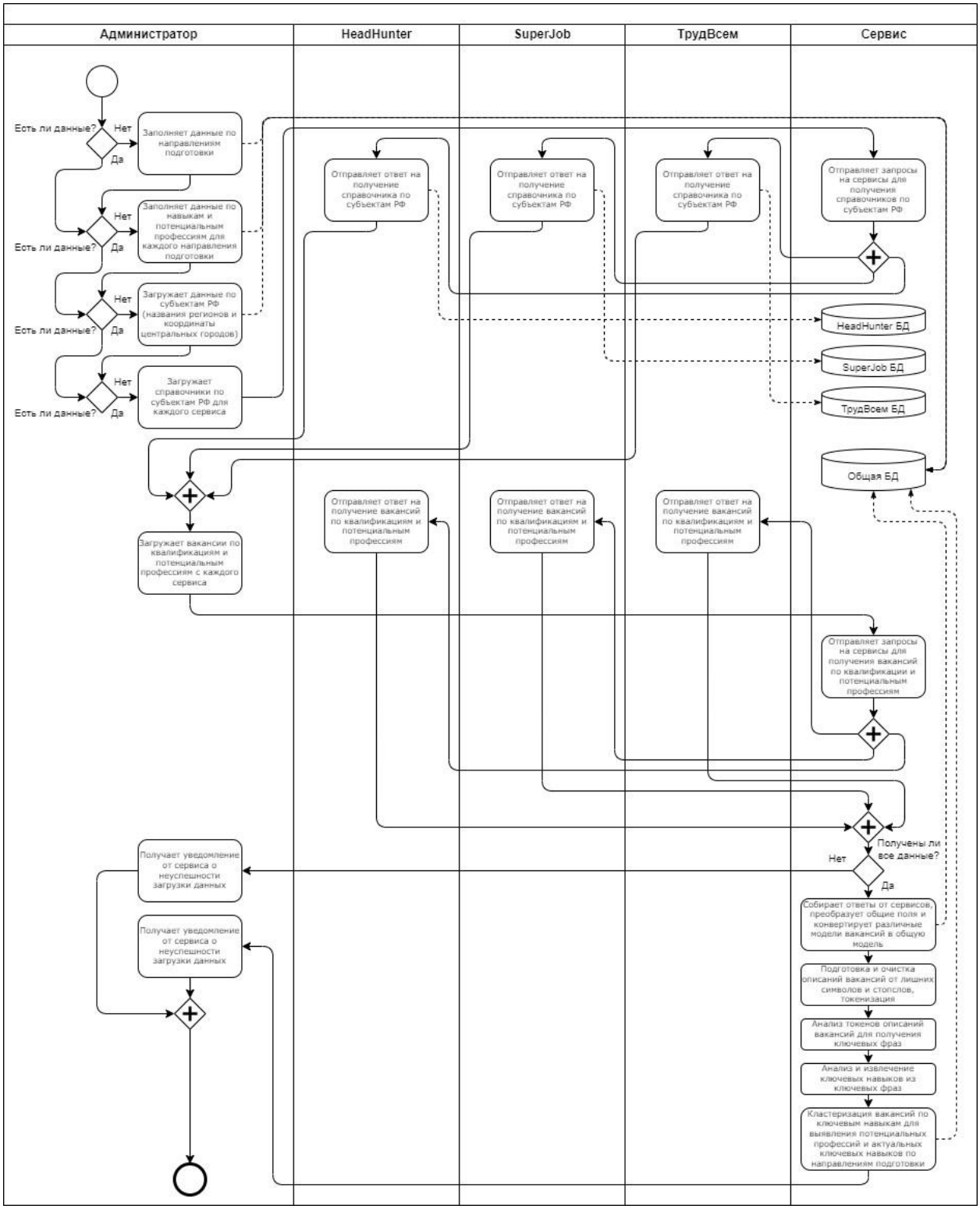

Рис. 1. – Модель процессов предварительной подготовки данных о вакансиях для получения потенциальных профессий и мониторинга востребованности направлений подготовки

Множество ключевых слов, по которому найдена i-я вакансия, можно представить следующим образом:

$$Kwrds_i = \{kwrd_w^i\}, w = \overline{1, W^i},$$

где $kwrd_w^i$ – w-й навык или компетенция направления подготовки, по которому найдена i-я вакансия, W^i – количество навыков и компетенций направления подготовки, по которым найдена i-я вакансия.

Множество ключевых навыков i-й вакансии, полученных от сервиса поиска вакансий:

$$KSklls_i = \{kskll_s^i\}, s = \overline{1, S^i},$$

где $kskll_s^i$ – s-й ключевой навык для i-й вакансии, S^i – количество ключевых навыков i-й вакансии.

Дальнейшим этапом является извлечение ключевых фраз из описаний вакансий [8-9]. Далее опишем общий подход алгоритмизации данного этапа. Введем новые обозначения множеств, а именно:

$$DWrds_i = \{dwrd_j^i\}, j = \overline{1, J^i},$$

где $DWrds_i$ – множество слов в описании i-й вакансии, $dwrd_j^i$ – j-й слово описания i-й вакансии, J^i – общее количество слов описания i-й вакансии.

$$DTkns_i = Tokenizer(DWrds_i) = \{dtkn_k^i\}, k = \overline{1, K^i},$$

где $DTkns_i$ – множество токенов, полученных из текстового описания $Dscr_i$ i-й вакансии с помощью $Tokenizer(DWrds_i)$ – функции преобразования слова в токен, которая включает ряд этапов нормализации (приведение слова к нижнему регистру, очистка от стоп-слов, специальных символов и другие лингвистические операции), $dtkn_k^i$ – k-й токен описания i-й вакансии, K^i – общее количество токенов описания i-й вакансии.

Общее множество токенов для каждого описания вакансий обозначим следующим образом:

$$VDTkns = \bigcup_{i=1}^{N} DTkns^{i}.$$

Далее требуется получить множество ключевых слов и фраз, характеризующих описание вакансии, которое может содержать ключевые навыки, $VDKwrds$ на основе множества $VDTkns$.

$$VDKwrds = KwrdsExtractor(VDTkns) = \{VDKwrds_i\}, i = \overline{1,N},$$

где $KwrdsExtractor(VDTkns)$ – функция извлечения множества ключевых слов и фраз из описаний вакансий, $VDKwrds_i$ – множество ключевых слов и фраз для i-й вакансии, которое представлено в виде:

$$VDKwrds_i = \{vdkwrd_m^i\}, m = \overline{1, M^i},$$

где $vdkwrd_m^i$ – m-е ключевое слово или фраза i-й вакансии, M^i – количество ключевых слов и фраз, полученное по i-й вакансии.

Для дальнейшей реализации кластерного анализа вакансий по словам и фразам требуется объединить множества $Kwrds_i$, $KSklls_i$ и $VDKwrds_i$ для i-й вакансии. Обозначим множество $VKFtrs$ всех потенциально ключевых навыков для i-й вакансии, как:

$$VKFtrs = \bigcup_{i=1}^{N}(Kwrds_i \cup KSklls_i \cup VDKwrds_i) = \{VKFtrs_i\},$$

где $VKFtrs_i$ – множество потенциально ключевых навыков i-й вакансии, представленное следующим образом:

$$VKFtrs_i = \{vkftr_u^i\}, u = \overline{1, U^i},$$

где $vdftr_u^i$ – u-й потенциальный ключевой навык i-й вакансии, U^i – количество потенциальных ключевых навыков i-й вакансии после объединения множеств $Kwrds_i$, $KSklls_i$ и $VDKwrds_i$.

Полученное в результате множество $VKFtrs$ используется для формирования векторного представления каждой вакансии, которое представлено следующим образом:

$$VVctrs = Vectorizer(VKFtrs, H) = \{VVctr_i\}, i = \overline{1, N},$$

где $Vectorizer(VKFtrs, H)$ – функция, которая преобразует множество потенциально ключевых навыков в векторное представление для каждой вакансии, H – размерность векторного представления вакансий, $VVctr_i$ - векторное представление i-й вакансии в виде:

$$VVctr_i = \{vvtm_h^i\}, h = \overline{1, H^i},$$

где $vvtm_h^i$ – h-е элемент вектора i-й вакансии, имеющий численное значение.

Множество $VVctrs$ используется при кластерном анализе для извлечения потенциальных профессий на рынке труда по определенному направлению подготовки, а также для актуализации направления подготовки с помощью нахождения новых ключевых навыков, необходимых для формирования более современной программы подготовки по направлению в учебном заведении.

Для кластерного анализа требуется предварительно вычислить оптимальное количество кластеров, используя различные метрики оценки качества кластеризации (например, интегральная оценка на основе коэффициента силуэта, индекс Калински-Харабаша, индекс Дэвиса-Болдина [10]).

Представим множество результата кластеризации вакансий следующим образом:

$$VClstrs = Clusterizer(VVctrs, C) = \{ClstrIndx_i\}, i = \overline{1, N},$$

где $Clusterizer(VVctrs, C)$ – функция кластеризации, определяющая к какому кластеру относится та или иная вакансия, C – количество кластеров вакансий, $ClstrIndx_i$ – номер кластера, к которому относится i-я вакансия.

В результате работы функции кластеризации множество $VClstrs$ используется для сбора дополнительной статистики по каждому кластеру.

Разработанная формализованная модель использована при реализации *web*-сервиса, направленного на помощь в профориентации потенциальных абитуриентов, а также определения перспективных навыков на рынке труда и поддержки принятия решений при формировании дисциплин учебных планов направлений подготовки.

Литература

1. Кязимов, К. Г. Взаимодействие учреждений образования с субъектами рынка труда // Профессиональное образование в современном мире, 2019, №9(1). С. 2421-2432.

2. Диков М.Е., Широбокова С.Н. О проектных решениях цифрового инструментария профориентации по определению востребованности направлений подготовки на основе анализа описаний вакансий // Инженерный вестник Дона, 2022, № 12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8042.

3. Gugnani A., Misra H. Implicit skills extraction using document embedding and its use in job recommendation // The Thirty-Second Innovative Applications of Artificial Intelligence Conference (IAAI 2020), New York, NY, USA, February 7-12, 2020, AAAI Press. pp. 13286-13293.

4. Zhang M., Jensen K., Sonniks S., Plank B. SkillSpan: Hard and Soft Skill Extraction from English Job Postings // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, Association for Computational Linguistics, 2022. pp. 4962-4984.

5. Широбокова С.Н., Диков М.Е. О варианте формализации задачи анализа востребованности перспективных технологий на рынке труда // Информационные технологии в науке и образовании: материалы Междунар. молодеж. науч.-практ. конф., посвященной 115-летию Южно-Российского

государственного политехнического университета имени М.И. Платова (НПИ), Новочеркасск, 18-19 июня 2022, Новочеркасск, 2022. С. 137-140.9

6. Диков М.Е., Широбокова С.Н. О варианте формализации задачи определения востребованности направлений подготовки и возможных сфер трудоустройства выпускников на основе семантического анализа описаний вакансий // Инженерный вестник Дона, 2022, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7631.

7. Bhola A., Halder K., Prasad A., Kan M.-Y. Retrieving skills from job descriptions: A language model based extreme multi-label classification framework // Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020. pp. 5832-5842.

8. Николаев И.Е. Метод извлечения знаний и навыков/компетенций из текстов требований вакансий // Онтология проектирования, 2023, Т.13, №2(48). С.282-293. DOI:10.18287/2223-9537-2023-13-2-282-293.

9. Фомичев Д.А. Кластеризация вакансий по их описанию с использованием машинного обучения и методов анализа текста // XXVI Международная конференция по мягким вычислениям и измерениям (SCM-2023): Сборник докладов, Санкт-Петербург, 24-26 мая 2023, СПб.: СПбГЭТУ «ЛЭТИ». С. 201-204.

10. Яруллин Д. В. Интеллектуальная система управления подготовкой ИТ-специалистов на основе денотативной аналитики // Прикладная математика и вопросы управления, 2022, №3. С.141-164. DOI 10.15593/2499-9873/2022.3.08.

References

1. Kazimov K.G. Professional`noe obrazovanie v sovremennom mire, 2019, vol. 9, №1. pp. 2421-2432.

2. Dikov M.E., Shirobokova S.N. Inzhenernyj vestnik Dona, 2022, №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8042.

3. Gugnani A., Misra H. The Thirty-Second Innovative Applications of Artificial Intelligence Conference (IAAI 2020), New York, NY, USA, February 7-12, 2020, AAAI Press. pp. 13286-13293.

4. Zhang M., Jensen K., Sonniks S., Plank B. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, Association for Computational Linguistics, 2022. pp. 4962-4984.

5. Shirobokova S.N., Dikov M.E. O variante formalizacii zadachi analiza vostrebovannosti perspektivny`x texnologij na ry`nke truda [On the variant of formalization of the task of analyzing the demand for promising technologies in the labor market]. Informacionnye tekhnologii v nauke i obrazovanii: materialy Mezhdunarodnoj molodyozhnoj nauchno-prakticheskoj konferencii, Novocherkassk, 18-19 june 2022, Novocherkassk: Lik, 2022. pp. 137-140.

6. Dikov M.E., Shirobokova S.N. Inzhenernyj vestnik Dona, 2022, №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7631.

7. Bhola A., Halder K., Prasad A., Kan M.-Y. Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020. pp. 5832-5842.

8. Nikolaev I.E. Ontologiya proektirovaniya, 2023, vol. 13, №2(48). pp. 282-293. DOI:10.18287/2223-9537-2023-13-2-282-293.

9. Fomichev D.A. Klasterizaciya vakansij po ix opisaniyu s ispol`zovaniem mashinnogo obucheniya i metodov analiza teksta [Vacancy clustering by their description using machine learning and text analysis methods]. XXVI Mezhdunarodnaya konferenciya po myagkim vychisleniyam i izmereniyam (SCM-2023): Sbornik dokladov, Saint Petersburg, 24-26 may 2023. SPb.: SPbGETU «LETI». pp. 201-204.

10. Yarullin D.V. Prikladnaya matematika i voprosy` upravleniya, 2022, № 3. pp. 141-164. DOI: 10.15593/2499-9873/2022.3.08.

Дата поступления: 22.01.2024

Дата публикации: 2.03.2024