Построение модели ИТ-специалиста на основе нечеткой кластеризации для системы поддержки принятия решений в сфере кадрового обеспечения

Р. А. Файзрахманов[1], Д. В. Яруллин[2]
Пермский национальный исследовательский политехнический университет
Пермь, Россия
[1]fayzrakhmanov@gmail.com, [2]d.v.yarullin@ya.ru

***Аннотация*. В работе предлагается подход к построению модели востребованности профессиональных навыков ИТ-специалиста для различных регионов с помощью алгоритмов кластеризации на основе данных, извлеченных из текстов вакансий, размещенных на веб-агрегаторах. Описываются методы извлечения данных о навыках из текстов вакансий на естественном языке. Демонстрируется способ преобразования извлеченных навыков в векторную форму, а также процесс построения матрицы расстояний и матрицы сходства как один из этапов препроцессинга данных. Проводится сравнение работы ряда алгоритмов кластеризации (иерархическая кластеризация по методу Уорда, метод k-средних, спектральная кластеризация, метод распространения близости, DBSCAN), определяются их достоинства и недостатки применительно к задаче. Приводится демонстрация работы выбранных алгоритмов в прототипе системы поддержки принятия решений.**

Ключевые слова: нечеткая кластеризация; вектор; матрица расстояний; интеллектуальный анализ данных; извлечение сущностей

I. Постановка проблемы

В настоящее время значительное число работодателей размещают свои вакансии на популярных сайтах-агрегаторах. Так, только на «HH.ru» представлено более одного миллиона различных компаний [1]. IT-компании используют подобные ресурсы в качестве основного метода поиска новых специалистов [2]. Именно они становятся основным источником информации о том, каких именно специалистов ищут работодатели региона.

Тем не менее, в IT-индустрии сохраняется проблема «кадрового голода». Доступность вакансий не упрощает процесс поиска подходящего кандидата, а сами кандидаты сталкиваются с тем, не все требуемые навыки и компетенции эксплицитно указаны в текстах вакансий. Также вакансии на одну и ту же должность, например, «Веб-программист», могут требовать от соискателя различных навыков.

Кроме того, сотрудники отдела кадров, осуществляющие первичную коммуникацию с соискателем, могут не обладать экспертизой во всех необходимых предметных областях, и на этапе начального взаимодействия выявить несоответствие компетенций соискателя и компетенций, которые ожидает работодатель от специалиста на данной должности, оказывается затруднительно, что влечет за собой ряд издержек и не решает проблему быстрого поиска нужного специалиста.

В связи с этим возникает потребность анализа того, какие именно компетенции и навыки (а также их комбинации) необходимы работодателям в тех или иных компаниях, регионах, странах. Поскольку развитие отрасли информационных технологий остается приоритетным направлением в Российской Федерации [3], в настоящем исследовании мы сосредоточимся на навыках IT-специалистов в различных регионах России.

Данный анализ позволит построить модель, которая структурирует и упорядочит навыки компетенции IT-специалиста по их востребованности в конкретное время и в конкретном регионе.

Такая модель профессионального портрета специалиста способна стать основой для системы поддержки принятия решений в сфере кадрового обеспечения, которая сможет упростить взаимодействие специалистов по кадрам и соискателей.

II. Сбор и подготовка данных

Вслед за существующими исследованиями в данной области [4], минимальную единицу нашей модели мы обозначаем как «навык». В нашей модели под навыком понимается фрагмент информации в рамках предметной области, владение которым позволяет решать определенные профессиональные задачи.

В качестве примеров мы можем привести языки программирования (Python, Java, C++, PHP), системы управления базами данных (MySQL, PostgreSQL) отдельные библиотеки и фреймворки (Spring, Pandas, Angular JS), технологии и протоколы (UDP, контейнерная виртуализация), операционные системы (Linux, macOS), прикладные приложения (Adobe Photoshop). Именно навыки чаще всего указываются работодателем в требованиях к соискателю.

Основным источником данных для нас послужили тексты вакансий в открытом доступе, опубликованные на

агрегаторе «HH.ru». Выбор ресурса был обусловлен как его популярностью, так и наличием открытого API [5].

Мы выгрузили полные тексты вакансий по запросу «Программист» с ограничением по каждому региону. Заметим, что HH.ru имеет ограничение на максимальную величину результатов поиска, поэтому для регионов с большим числом вакансий были собраны только первые 2000. В совокупности по всем регионам было получено 14094 вакансии.

Для каждой вакансии были извлечены данные, содержащиеся в метаполе «Ключевые навыки», если данное метаполе было заполнено. На основе этого был сформирован список навыков, содержащий 3730 позиций.

После этого полный текст каждой вакансии был токенизирован при помощи библиотеки NLTK [6], после чего лемматизирован при помощи модуля pymorphy2 [7]. Из получившихся списков лемм были извлечены совпадающие с леммами навыков первичного списка. Так был построен индекс навыков. Для каждой записи был сформирован список идентификаторов вакансий, в которых данный навык встречается, и указано общее число вхождений навыка в корпус вакансий. Фрагмент индекса приведен на рис. 1:

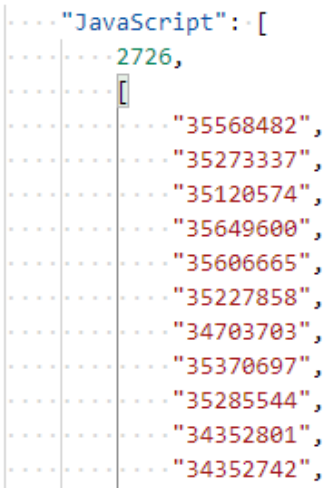


Рис. 1. Фрагмент индекса навыков: навык «JavaScript» и начало списка уникальных идентификаторов вакансий, содержащих этот навык

Для удобства работы с индексом в рамках прототипа системы поддержки принятия решений, был реализован веб-интерфейс, представленный на рис. 2.

Навык	Кол-во вакансий
JavaScript	2726
Git	2580
SQL	1788
CSS	1445
HTML	1409
Java	1400
PHP	1389
ООП	1321
MySQL	1271
1С программирование	1210
Linux	1019
C#	994

Рис. 2. Фрагмент веб-интерфейса для взаимодействия с индексом навыков, отсортированным по частоте встречаемости навыка

Для дальнейшего применения методов кластеризации индекс был переведен в векторную форму (размерность вектора – 14094, по числу вакансий, в качестве признака использовалось наличие или отсутствие навыка в вакансии). При этом были исключены слишком редко встречающиеся навыки, не связанные с какими-либо другими. Минимальное число вакансий, в которых должен был содержаться навык, для общего индекса равняется 100, для региональных индексов – 1% от общего числа вакансий в этом регионе. Также было учтено, что наиболее частотные навыки («JavaScript» и «Git») имеют значительно превосходящую частоту по сравнению с другими, и было принято решение использовать для расчета расстояний в n-мерном пространстве не евклидову метрику, а косинусный коэффициент. Матрица расстояний была визуализирована в виде тепловой карты для подтверждения корректности использования указанной метрики. Данная процедура при анализе проводилась для каждого из регионов (рис. 4).

III. Подходы к кластеризации

Для выявления заранее неизвестных групп взаимосвязанных навыков мы решили обратиться к кластерному анализу. Такой подход, на наш взгляд, позволяет определить, какими компетенциями должен обладать соискатель для той или иной должности в соответствии с реальным запросом работодателей, отраженным в вакансиях региона.

Учитывая специфику нашей задачи, мы решили протестировать ряд методов кластеризации для выявления наиболее подходящих нам алгоритмов.

Первым алгоритмом, к которому мы обратились, был метод k-средних с ожидаемым числом кластеров, равным 10. Результат работы алгоритма оказался неудовлетворительным: из 10 полученных кластеров 5 содержали лишь единственный навык, а большинство навыков были сгруппированы в один кластер. Это указало нам на то, что размеры потенциальных кластеров явно неравны, а их число может превышать 10. Для уточнения этого мы воспользовались иерархической кластеризацией по методу Уорда [8]. Результаты были визуализированы в виде дендрограммы, представленной на рис. 3.

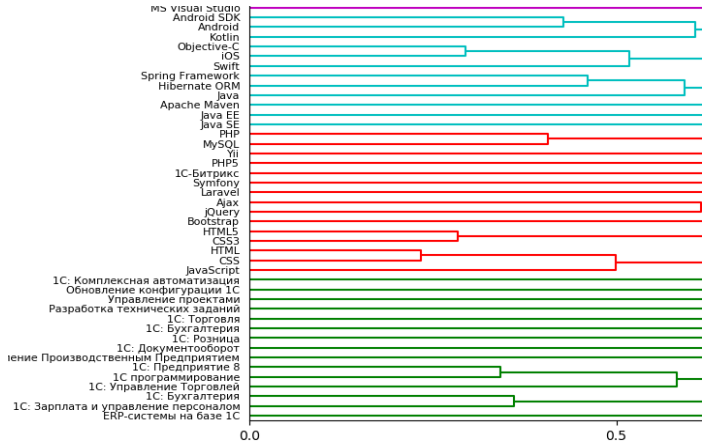


Рис. 3. Фрагмент дендрограммы

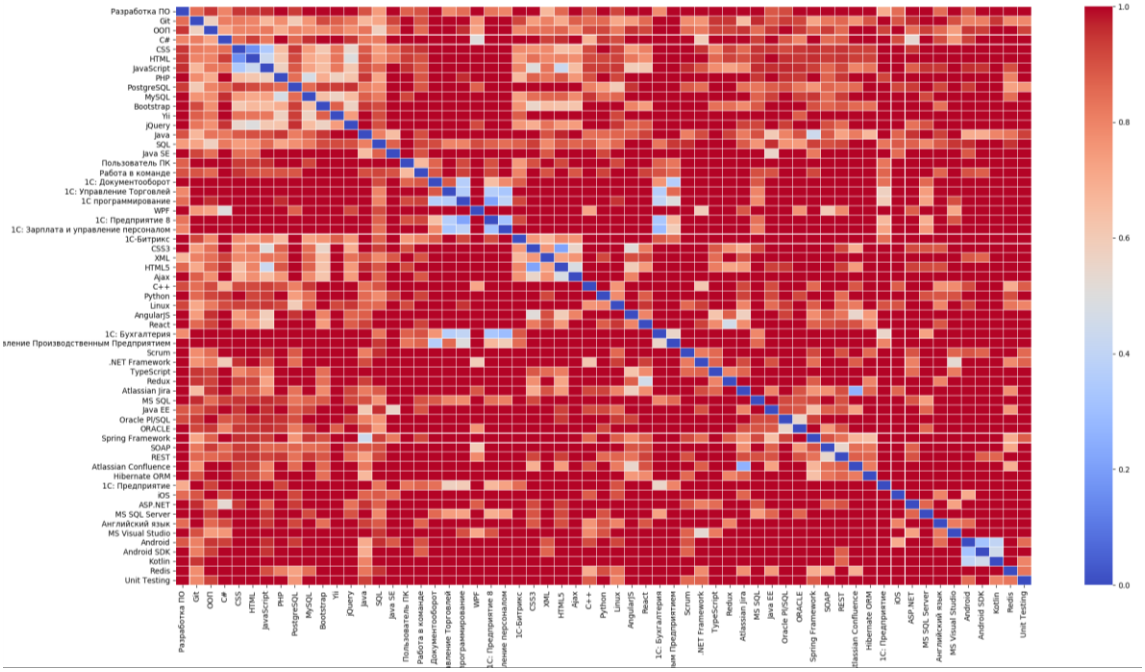


Рис. 4. Тепловая карта навыков в вакансиях Пермского края

Иерархическая кластеризация показала, что между навыками, используемыми в определенных рабочих задачах (в частности, мобильной разработке, веб-разработке, работе с продуктами 1С), действительно присутствует взаимосвязь. Однако число кластеров по-прежнему было неоптимальным: количество реальных рабочих задач явно превышало число выявленных кластеров.

Опираясь на то, что размеры кластеров неравны, мы также попробовали применить алгоритм DBSCAN с ожидаемым числом кластеров, равным 10. DBSCAN показал себя лучше метода k-средних, разбив навыки по ожидаемым, легко интерпретируемым группам. В частности, был выявлен кластер ['JavaScript', 'CSS', 'HTML'], содержащий навыки, необходимые веб-разработчику; кластер ['1С программирование', '1С: Зарплата и управление персоналом', '1С: Управление Торговлей', '1С: Бухгалтерия', '1С: Предприятие 8'] с навыками специалиста 1С; кластер ['Java', 'Hibernate ORM', 'Spring Framework'] с навыками Java-разработчика; кластер ['iOS', 'Objective-C', 'Swift'], легко интерпретируемый как навыки мобильного разработчика под операционную систему iOS.

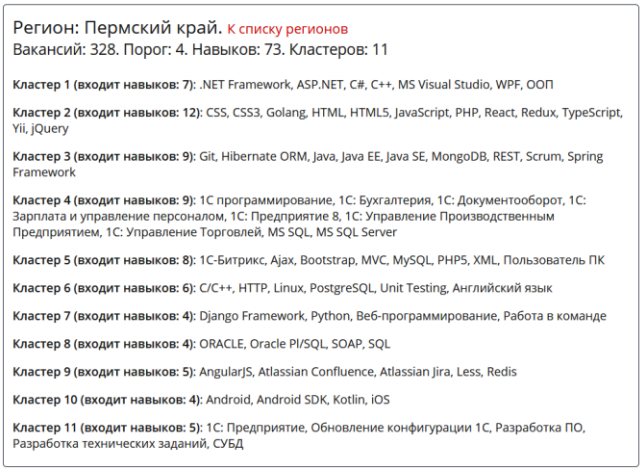
Тем не менее, число кластеров, которое мы предположили заранее, было неверным, данные распределились по 9 кластерам вместо ожидаемых 10. По той же причине нами был отвергнут метод спектральной кластеризации, показавший, тем не менее, даже лучшие результаты. В частности, были выделены два кластера, соответствующие веб-разработке: кластер 1 – ['JavaScript',

'CSS3', 'HTML5', 'Vue.js', 'AngularJS', 'Less', 'Node.js', 'TypeScript', 'React', 'Sass', 'Redux', 'Angular'] – и кластер 2 – ['jQuery', 'Git', 'CSS', 'ООП', 'HTML', 'MySQL', '1С-Битрикс', 'Ajax', 'PHP5', 'PHP', 'Adobe Photoshop', 'Bootstrap', 'CMS Wordpress', 'Веб-программирование', 'Yii', 'Laravel', 'Symfony']. Данные кластеры соответствуют разделению запросов компаний: часть опирается на веб-компоненты и реактивные фреймворки (кластер 1), часть же предпочитает более традиционный подход, связанный с фуллстэк-разработкой на PHP (кластер 2).

В итоге, учтя все особенности нашей задачи, мы остановили свой выбор на алгоритме распространения близости (affinity propagation). Основным плюсом алгоритма для нас стало то, что число кластеров вычисляется в процессе, а сами кластеры ожидаются неравными по размеру. Для более успешной работы алгоритма полученная ранее матрица расстояний была аппроксимирована при помощи Гауссовой функции.

В результате работы были сформированы 13 кластеров, соотносящихся с высказанными ранее предположениями о возможных направлениях работы программистов. В наиболее большой кластер вошли навыки, необходимые для разработки и дизайна клиентской стороны (фронтенда) веб-сервиса: ['JavaScript', 'jQuery', 'Git', 'CSS3', 'HTML5', 'Vue.js', 'CSS', 'HTML', 'Ajax', 'AngularJS', 'Less', 'Node.js', 'Adobe Photoshop', 'Bootstrap', 'TypeScript', 'React', 'Sass', 'Redux', 'Angular']. Выделены небольшие, но выраженные кластеры навыков мобильной разработки: ['Android', 'Kotlin', 'Android SDK'] и ['iOS', 'Objective-C', 'Swift'].

Воспользовавшись методом распространения близости, мы сформировали кластеры навыков по каждому региону, включив инструмент в прототип системы поддержки принятия решений (рис. 5, 6, 7).


Регион: Пермский край. К списку регионов
Вакансий: 328. Порог: 4. Навыков: 73. Кластеров: 11

Кластер 1 (входит навыков: 7): .NET Framework, ASP.NET, C#, C++, MS Visual Studio, WPF, ООП

Кластер 2 (входит навыков: 12): CSS, CSS3, Golang, HTML, HTML5, JavaScript, PHP, React, Redux, TypeScript, Yii, jQuery

Кластер 3 (входит навыков: 9): Git, Hibernate ORM, Java, Java EE, Java SE, MongoDB, REST, Scrum, Spring Framework

Кластер 4 (входит навыков: 9): 1С программирование, 1С: Бухгалтерия, 1С: Документооборот, 1С: Зарплата и управление персоналом, 1С: Предприятие 8, 1С: Управление Производственным Предприятием, 1С: Управление Торговлей, MS SQL, MS SQL Server

Кластер 5 (входит навыков: 8): 1С-Битрикс, Ajax, Bootstrap, MVC, MySQL, PHP5, XML, Пользователь ПК

Кластер 6 (входит навыков: 6): C/C++, HTTP, Linux, PostgreSQL, Unit Testing, Английский язык

Кластер 7 (входит навыков: 4): Django Framework, Python, Веб-программирование, Работа в команде

Кластер 8 (входит навыков: 4): ORACLE, Oracle PI/SQL, SOAP, SQL

Кластер 9 (входит навыков: 5): AngularJS, Atlassian Confluence, Atlassian Jira, Less, Redis

Кластер 10 (входит навыков: 4): Android, Android SDK, Kotlin, iOS

Кластер 11 (входит навыков: 5): 1С: Предприятие, Обновление конфигурации 1С, Разработка ПО, Разработка технических заданий, СУБД

Рис. 5. Кластерный анализа навыков в Пермском крае


Регион: Москва. К списку регионов
Вакансий: 2000. Порог: 20. Навыков: 79. Кластеров: 11

Кластер 1 (входит навыков: 9): 1С-Битрикс, Docker, MySQL, PHP, PHP5, Symfony, Yii, Веб-программирование, ООП

Кластер 2 (входит навыков: 6): C++, C/C++, Linux, Qt, Английский язык, Разработка ПО

Кластер 3 (входит навыков: 8): Django Framework, MongoDB, Nginx, PostgreSQL, Python, Redis, Ruby On Rails, SQL

Кластер 4 (входит навыков: 3): Mac Os, Objective-C, iOS

Кластер 5 (входит навыков: 11): 1С: Бухгалтерия, 1С программирование, 1С: Бухгалтерия, 1С: Документооборот, 1С: Зарплата и управление персоналом, 1С: Предприятие 8, 1С: Торговля, 1С: Управление Производственным Предприятием, 1С: Управление Торговлей, Обновление конфигурации 1С, Работа в команде

Кластер 6 (входит навыков: 10): .NET Framework, ASP.NET, C#, Entity Framework, MS SQL, MS SQL Server, MS Visual Studio, MVC, Transact-SQL, WPF

Кластер 7 (входит навыков: 3): Android, Android SDK, Kotlin

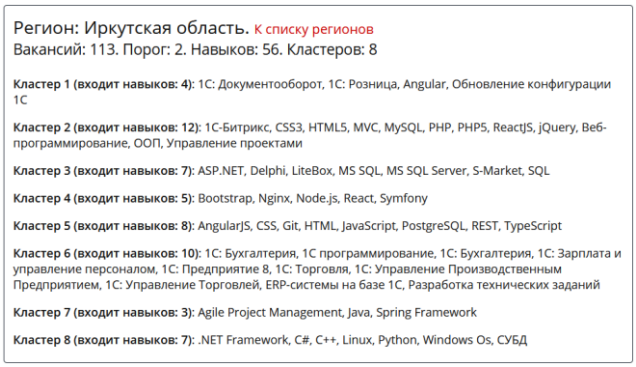
Кластер 8 (входит навыков: 19): API, Adobe Photoshop, Ajax, AngularJS, Bootstrap, CSS, CSS3, Git, HTML, HTML5, JavaScript, Node.js, React, Redux, Sass, TypeScript, Vue, Vue.js, jQuery

Кластер 9 (входит навыков: 3): ORACLE, Oracle PI/SQL, Базы данных

Кластер 10 (входит навыков: 3): JSON API, REST, XML

Кластер 11 (входит навыков: 4): Atlassian Jira, Hibernate ORM, Java, Spring Framework

Рис. 6. Кластерый анализ навыков в Москве


Регион: Иркутская область. К списку регионов
Вакансий: 113. Порог: 2. Навыков: 56. Кластеров: 8

Кластер 1 (входит навыков: 4): 1С: Документооборот, 1С: Розница, Angular, Обновление конфигурации 1С

Кластер 2 (входит навыков: 12): 1С-Битрикс, CSS3, HTML5, MVC, MySQL, PHP, PHP5, ReactJS, jQuery, Веб-программирование, ООП, Управление проектами

Кластер 3 (входит навыков: 7): ASP.NET, Delphi, LiteBox, MS SQL, MS SQL Server, S-Market, SQL

Кластер 4 (входит навыков: 5): Bootstrap, Nginx, Node.js, React, Symfony

Кластер 5 (входит навыков: 8): AngularJS, CSS, Git, HTML, JavaScript, PostgreSQL, REST, TypeScript

Кластер 6 (входит навыков: 10): 1С: Бухгалтерия, 1С программирование, 1С: Бухгалтерия, 1С: Зарплата и управление персоналом, 1С: Предприятие 8, 1С: Торговля, 1С: Управление Производственным Предприятием, 1С: Управление Торговлей, ERP-системы на базе 1С, Разработка технических заданий

Кластер 7 (входит навыков: 3): Agile Project Management, Java, Spring Framework

Кластер 8 (входит навыков: 7): .NET Framework, C#, C++, Linux, Python, Windows Os, СУБД

Рис. 7. Кластерый анализ навыков в Иркутской области

IV. Результаты

В результате исследования была создана обновляемая база данных вакансий с распределением по регионам России.

Были извлечены сведения о навыках IT-специалистов, указываемых в вакансиях, построен автоматически обновляемый индекс навыков как для России в целом, так и индивидуально для каждого региона. Предложен метод векторизации собранных данных для их дальнейшего анализа. Расстояние между векторами навыков визуализировано на тепловых картах для каждого региона.

Также был проведен анализ ряда алгоритмов кластеризации применительно к поставленной задачи. В процессе исследования были выявлены ключевые параметры, необходимые для успешной работы с собранными данными, в результате выбор был сделан в пользу алгоритма распространения близости.

Полученная модель была интегрирована в прототип системы поддержки принятия решений.

Потенциальным следующим шагом работы является анализ вакансий в регионах других стран, размещенных на иных ресурсах-агрегаторах, в частности, федеральных земель Германии.

Список литературы

[1] Поиск персонала и публикация вакансий // Электронный ресурс. URL: https://hh.ru/employer, дата обращения: 30.03.2020

[2] Компании – Хабр Карьера // Электронный ресурс. URL: https://career.habr.com/companies, дата обращения: 30.03.2020

[3] План мероприятий по направлению «Кадры и образование» Национальной программы «Цифровая экономика Российской Федерации» // Электронный ресурс. URL: http://static.government.ru/media/files/k87YsCABuiyuLAjcWDFILEh6itAirUX0.pdf, дата обращения: 30.03.2020

[4] Nikulchev E., Ilin D., Matishuk E. Scalable Service for Professional Skills Analysis Based on the Demand of the Labor Market and Patent Search. Procedia Computer Science. Volume 103, 2017. Pages 44-51. DOI: https://doi.org/10.1016/j.procs.2017.01.008

[5] HeadHunter API: документация и библиотеки // Электронный ресурс. URL: https://github.com/hhru/api, дата обращения: 10.03.2020

[6] Bird S., Loper E., Klein E. Natural Language Processing with Python. O'Reilly Media Inc. 2009.

[7] Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, 2015. Pp 320-332. DOI: https://doi.org/10.1007/978-3-319-26123-2_31

[8] Ward J.H. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association, 1963. 236 p.