

УДК 004.822
МРНТИ 28.23.39

<https://doi.org/10.55452/1998-6688-2026-23-3-312-327>

¹**Серикбаева С.К.,**

PhD, и.о.доцента, ORCID ID: 0000-0002-3627-3321,

*e-mail: inf_8585@mail.ru

²**Рамазанова В.С.,**

магистр, главный инженер, ORCID ID: 0000-0001-6100-3123,

e-mail: ramazanovavs@gmail.com

¹**Самбетбаева М.А.,**

Phd, ассоциированный профессор, ORCID ID: 0000-0001-9358-1614,

e-mail: sambetbayeva_ma_1@enu.kz

^{3,4}**Еримбетова А.С.**

PhD, к.т.н., ассоциированный профессор, профессор-исследователь, в.н.с.,

ORCID ID: 0000-0002-2013-1513,

e-mail: aigerian8888@gmail.com

⁵**Садирмекова Ж.Б.,**

PhD, ассоциированный профессор, ORCID ID: 0000-0002-7514-9315,

e-mail: Zhanna.Sadirmekoval@astanait.edu.kz

¹Евразийский национальный университет им. Л.Н. Гумилева, г. Астана, Казахстан

²Западно-Казахстанский университет им. М. Утемисова, г. Уральск, Казахстан

³META University, г. Алматы, Казахстан

⁴Институт информационных и вычислительных технологий КН МНВО РК,
г. Алматы, Казахстан

⁵Astana IT University, г. Астана, Казахстан

МОДЕЛИ ПОСТРОЕНИЯ ГРАФОВ ЗНАНИЙ ДЛЯ ПОДБОРА МООК НА ОСНОВЕ ИНТЕГРАЦИИ ЗАПРОСОВ РЫНКА ТРУДА И ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ

Аннотация

В статье рассматривается разработка интеграционной модели на основе графов знаний и методов машинного обучения для анализа взаимосвязей между образовательными программами, требованиями рынка труда и онлайн-курсами (МООС). Актуальность исследования обусловлена быстрыми изменениями в сфере информационных технологий, приводящими к несоответствию между подготовкой выпускников и ожиданиями работодателей. Для устранения данного разрыва предложено использование гетерогенного графа знаний, объединяющего данные из трех доменов: университетских программ, вакансий и онлайн-курсов. В модели реализован мета-путь, связывающий навыки, требования вакансий и образовательные курсы, что позволяет формировать рекомендации по обучению и карьерному развитию для различных категорий пользователей. Для обработки текстовых данных применены мультиязычные трансформеры, обеспечивающие анализ русскоязычных и англоязычных описаний. В качестве источников данных использованы карты компетенций образовательных программ, вакансии платформы hh.kz и курсы Coursera, интегрированные в графовую СУБД Neo4j. Полученные результаты демонстрируют возможность выявления недостающих компетенций и автоматического подбора курсов для их формирования. Предложенная модель может применяться при проектировании учебных программ, разработке образовательных траекторий и поддержке принятия решений в образовательной среде.

Ключевые слова: графы знаний, машинное обучение, мультиязычные трансформеры, онлайн-курсы, рынок труда, навыки, рекомендации, образовательные программы, информационные технологии, интеграционные модели.

Введение

В свете быстро меняющихся требований рынка труда и применяемых технологий современное образование сталкивается с необходимостью пересмотра своих методов и подходов к проектированию учебных курсов. Стандартные учебные программы часто не могут адаптироваться к индивидуальным потребностям учащихся и быстро устаревают, не отражая последних достижений в различных областях знания. Существует разрыв между содержанием образовательных программ и реальными потребностями современного сектора информационных технологий, который приводит к недостаточной подготовке выпускников к профессиональной деятельности. Для того чтобы сократить этот разрыв и обеспечить студентов актуальными навыками, необходимы инновационные подходы к рекомендации учебных курсов на основе актуальных требований работодателей.

В образовательном ландшафте графы знаний уже играют важную роль в структурировании и организации учебного контента. Графы, представляющие собой сети взаимосвязанных понятий и идей, обеспечивают глубокое понимание учебного материала, позволяя учащимся визуализировать и исследовать различные связи между знаниями. Благодаря своей способности интегрировать и связывать разнообразные информационные ресурсы, графы знаний способствуют развитию критического мышления и самостоятельного обучения. Машинное обучение революционизирует способы обучения и персонализирует учебные процессы. Эта технология позволяет создавать интеллектуальные системы, способные анализировать данные обучающихся и предлагать индивидуализированные учебные материалы и методики, соответствующие их потребностям и уровню знаний. Благодаря машинному обучению образовательные платформы могут адаптировать учебный материал в реальном времени, учитывая скорость усвоения информации и особенности обучающегося. Кроме того, машинное обучение применяется для анализа эффективности учебных программ, выявления тенденций в образовательных процессах и создания новых методик обучения на основе больших объемов данных.

Интеграция образования, рынка труда и MOOK с помощью графов знаний и машинного обучения еще не до конца изучены, так как отсутствуют некоторые междоменные связи между навыками, полученные современными мультязычными трансформерами для смешанных предложений на русском и английском языках. Графы, представляющие собой знания, позволяют систематизировать и структурировать информацию о необходимых навыках и компетенциях, для того чтобы выявлять связи между ними и требованиями рынков труда. Машинное обучение, в свою очередь, способствует анализу больших объемов данных о востребованных профессиональных навыках и предсказанию их взаимосвязи с образовательными навыками. Путем интеграции этих инструментов в процесс проектирования учебных курсов образовательные учреждения смогут оперативно реагировать на изменения в требованиях рынка труда и обеспечивать выпускников необходимыми знаниями и навыками для успешной карьеры.

Таким образом, в этом исследовании предлагается подход с поддержкой эмбедингов навыков современных мультязычных трансформеров векторизации текстов для решения проблемы междоменных рекомендаций в области высшего образования, рынка труда и курсов MOOK с точки зрения гетерогенного графа знаний. Три домена имеют несколько типов вершин, которые взаимодействуют посредством разных типов связей. Создается мета-путь для рекомендации вакансий и курсов: образовательная программа–навык–навык–вакансия–навык–навык–курс, который может быть использован для различных практических сценариев. Например, для преподавателя, который будет создавать новый курс для каталога элективных дисциплин; для студента или соискателя, который по списку имеющихся навыков сможет выявить неосвоенные навыки интересующей его вакансии и получить рекомендованный список курсов.

Цель данного исследования заключается в разработке методологии использования графов знаний и машинного обучения на основе мультязычных источников данных на русском и

английском языке для рекомендации вакансий и курсов, формирующих отсутствующие навыки и компетенции, которые не были освоены в образовательной программе вуза или на предыдущем образовательном этапе, но востребованы работодателями, на примере рынка труда информационных технологий (ИТ).

Наиболее разнообразное применение в образовании графы знаний получили для студенческой аудитории, которой они помогают в представлении и организации знаний, персонализированном обучении и его аналитике. Графы знаний обеспечивают более структурированный и организованный способ представления знаний, облегчая учащимся связывание и понимание концепций [1, 2, 3, 4, 5]. Их используют для персонализации обучения [6, 7], предоставляя учащимся адаптированные учебные материалы, основанные на их учебных предпочтениях. К графам знаний обращаются для аналитики обучения [8], собирая и анализируя данные об их успеваемости, чтобы определить области, в которых им нужна дополнительная поддержка или где можно улучшить учебные материалы. Для преподавателей они используются как помощь в структуризации учебного материала [9, 10], для менеджеров – при построении образовательной траектории [11], а также как вспомогательный инструмент при принятии решений всеми участниками образовательного процесса (И. Алию [12]).

В образовательных процессах выделяют несколько ключевых тенденций и общих подходов использования графов знаний. Важной тенденцией является стремление к унификации стандартов представления знаний. Авторы, такие как [6, 7, 12, 3], отмечают использование онтологий, в том числе стандартных схем, таких как Schema.org, что способствует более эффективному обмену и взаимопониманию данных в образовательном контексте. Многие исследователи, включая [14, 15, 7, 3, 4, 5, 9], успешно интегрируют методы машинного обучения для повышения эффективности образовательных процессов. Использование моделей, таких как Word2vec, BERT и графовые сверточные сети (GCN), подчеркивает важность применения передовых технологий в адаптации образования к потребностям каждого учащегося. Экспертные знания, такие как представленные в работах [13, 1, 8], играют существенную роль в построении проблемно-ориентированных графов знаний.

Помимо сугубо образовательного контекста, графы знаний нашли свое применение на стыке различных предметных областей, например, интеграция разных уровней образования, интеграция образования и рынка труда. В контексте трудоустройства графы знаний используются для рекомендательных систем [16, 17], которые, исходя из набора навыков пользователей, определяют наиболее подходящие вакансии. Также выделяются варианты применения графов знаний для взаимосвязи образования, науки и трудоустройства. Например, построение или использование графов знаний для взаимосвязи между вакансиями и образованием MOOK методом выравнивания сущностей [18], взаимосвязи между классификаторами профессий и образовательными курсами SAE Institute методами распознавания и классификации сущностей алгоритмом distilbert-base-german-cased [Albert Weichselbraun [19], MOOK с университетскими курсами и вакансиями для рекомендации вакансий и университетских курсов методами определения схожести навыков алгоритмами BM25, Word2vec Bert [20], а также анализа несоответствия между научными публикациями, вакансиями и образованием MOOK [21].

В основе таких работ лежит интеграция посредством извлечения, классификации и выравнивания навыков и применяются алгоритмы FastText, BM25, Word2Vec, Bert, all-mpnet-base-v2 [16, 20, 22], в других обнаружение новых связей между навыками и профессиями с помощью Node2Vec и Preferential Attachment [17]. Для определения сходства узлов используют расчет жаккардового, косинусного сходства, евклидова расстояния [20, 22]. При анализе навыков применяют различные методы расчета корреляции Пирсона, расстояния Кульбака-Лейблера, модели многомерного процесса Хоукса [21]. Для рекомендаций – алгоритмы поиска кратчайшего пути Дейкстры, алгоритмы случайного блуждания, обнаружения сообществ, графовые нейронные сети, рассуждения по пути на основе политики, генетические алгоритмы, оптимизации муравьиных колоний. Для определения наиболее актуальных навыков разных профессий – метод взвешивания терминов TF-IDF [17].

Оценка качества алгоритмов машинного обучения обычно использует меры точности (P), полноты (R) и F1, например, для задач распознавания, классификации и связывания сущностей [19]. Мерами оценки качества рекомендаций выступают средняя средней точности (MAP@K), точность (P@K) и нормализованный дисконтированный совокупный прирост (NDCG@K). Например, [20] оценивает эффективность рейтинга на определенном пороговом уровне, учитывая только лучшие K-курсы, вернувшиеся в качестве кандидатов в эксперимент.

Однако вследствие отсутствия изученных связей по маршруту программа–навык–навык–вакансия–навык–навык–курс в смешанных фразах на русском и английском языках для интеграции высшего образования, рынка труда и MOOK с помощью графов знаний и машинного обучения требуется дополнительное исследование схожести навыков с помощью современных трансформеров векторизации текстов для рекомендации вакансий и курсов MOOK. Кроме того, в отличие от предыдущих работ, данные о карьере, помимо навыков вакансий, дополняются описаниями требований и обязанностей, что обогащает контекст вакансий. Наше исследование предлагает методологию, которая использует гетерогенный граф знаний, объединяющий образовательные и карьерные данные, чтобы на основе содержимого образовательной программы или опыта рекомендовать вакансии и соответствующие им дополнительные курсы для преподавателей или менеджеров образовательных программ, для студентов или сотрудников разного уровня.

Материалы и методы

Набор данных с hh.kz состоит из наименований вакансий, компании, ее отрасли, опыта, зарплаты, даты публикации, ссылки на вакансию, навыков, части требований и обязанностей на русском и английском языке, загруженных с сайта HeadHunter (<https://hh.kz/>) с помощью предоставляемого API (<https://api.hh.ru/vacancies>) и языка программирования Python. Вакансии выбирались по 25 профессиям в сфере информационных технологий из 177 населенных пунктов Казахстана за два месяца с 1.02.2024 по 31.03.2024. Количество вакансий составило 5248.

Набор данных курсов Coursera состоит из бесплатных курсов на английском языке по темам компьютерных наук, информационных технологий и науки о данных по состоянию на 03.05.2024. Данные включают наименования курсов с уровнем их сложности, рейтингом, списком навыков, наименований модулей и уроков с их длительностью изучения и кратким описанием. Данные о курсах были собраны инструментами парсинга BeautifulSoup. Количество курсов составило 953.

Все три набора данных загружались в графовую систему управления базами данных neo4j в виде графов знаний, определенными своими онтологиями. Далее онтологии трех графов знаний интегрировались на уровне навыков из трех источников, посредством расчета косинусной близости эмбедингов навыков. Эмбединги навыков создавались мультязычной предобученной моделью paraphrase-multilingual-mpnet-base-v2, которая поддерживает более 50 языков и прошла тонкую настройку на навыках вакансий. При достижении порога сходства навыков были введены дополнительные связи между тремя видами навыков, а также созданы кластеры схожих навыков вакансий и курсов. Далее навыки вакансий, которые схожи с навыками пользователей, преобразованы в другой тип узлов для удобства дальнейшей обработки.

Моделирование предметных областей

В данном разделе представлен процесс разработки модели графа знаний, который описывает сущности, их взаимосвязи и атрибуты, связанные с навыками и компетенциями в области информационных технологий. Эта модель также учитывает взаимосвязи этих навыков с образовательными программами, вакансиями и MOOK. Процесс моделирования выполнялся с использованием теории множеств, статьи [23, 24], стандарта по графам знаний [25] и инструмента визуализации концептуальных карт Cmap [26]. Согласно стандарту концептуальная модель графа знаний делится на уровень онтологии и уровень экземпляров.

Таблица 1 – Список профессий по информационным технологиям на HeadHunter

№	ID профессии	Название профессии
1.	10	Аналитик
2.	12	Арт-директор, креативный директор
3.	25	Гейм-дизайнер
4.	34	Дизайнер, художник
5.	36	Директор по информационным технологиям (CIO)
6.	73	Менеджер продукта
7.	96	Программист, разработчик
8.	104	Руководитель группы разработки
9.	107	Руководитель проектов
10.	112	Сетевой инженер
11.	113	Системный администратор
12.	114	Системный инженер
13.	116	Специалист по информационной безопасности
14.	121	Специалист технической поддержки
15.	124	Тестировщик
16.	125	Технический директор (CTO)
17.	126	Технический писатель
18.	148	Системный аналитик
19.	150	Бизнес-аналитик
20.	155	Методолог
21.	156	BI-аналитик, аналитик данных
22.	157	Руководитель отдела аналитики
23.	160	DevOps-инженер
24.	164	Продуктовый аналитик
25.	165	Дата-сайентист

В работе представлена онтологическая модель графа знаний предметной области как гетерогенный HAN, помеченный, ориентированный граф $G = \{E, R, A, V, r, f, p\}$, где E – множество классов сущностей; R – множество типов отношений; A – множество атрибутов сущностей или отношений; V – множество значений атрибутов; функция $r: R \rightarrow E \times E$ означает, что каждому типу отношения присваивается пара классов сущностей; функция $f: E \cup R \rightarrow 2^A$ означает, что каждый класс сущности или тип отношения имеет атрибут из множества A ; функция $p: E \cup R \rightarrow 2^{A \times V}$ означает соответствие класса сущности или типа отношения множеству пар атрибут–значение.

Результаты и обсуждение

Проведем анализ предметной области образовательной программы университета. Образовательная программа содержит карту компетенций и каталог элективных дисциплин. Карты компетенций включают компетенции, результаты обучения и дисциплины. Результаты обучения являются единицами компетенций. Результаты обучения формируются дисциплинами, для которых имеется их краткая аннотация в каталоге элективных дисциплин.

Рассмотрим онтологическую модель графа образовательной программы университета. Граф включает множество классов сущностей: $E_{\text{university}} = \{\text{Curriculum}, \text{CUA}\}$. Образовательная программа (ОП) – класс, представляющий собой структурированный план учебных модулей и активностей. CUA состоит из класса фраз компетенций, результатов обучения и аннотаций. Компетенция – отражает специфические способности, которые обучающийся должен развить в рамках образовательной программы. Результаты обучения (РО) – это конкретные знания, умения и навыки, которые студент должен демонстрировать после завершения образовательной программы. Аннотация – предоставляет краткое описание содержания дисциплины, что позволяет студентам и преподавателям быстро оценить ее основные темы и значение в контексте учебной программы.

Определим множество отношений $R_{\text{university}} = \{\text{HAS}\}$. Характеристика отношений и атрибутов приведена в таблице 2.

Таблица 2 – Описание онтологии графа знаний образовательной программы университета.

Relationships and attributes	Description
Curriculum CUA	The educational program has phrases of competencies, units of study and annotations of disciplines.
$f(\text{Curriculum}) = \{\text{code}, \text{title}\}$	The educational program is characterized by attributes such as code and name.
$f(\text{CUA}) = \{\text{code}, \text{description}, \text{embedding}\}$	Each competence has a unique code, competencies are divided into universal competencies (CC), general professional competencies (MIC), professional competencies (PC), the description is divided into phrases, which are also presented in the form of embedding.

Описания компетенций, результатов обучения и аннотаций разделены на фразы по запятым, точкам с запятой и точкам. Эти фразы дополнительно представлены в виде векторного представления (эмбеддинг), генерируемого методами машинного обучения. Визуально модель представлена на рисунке 2.

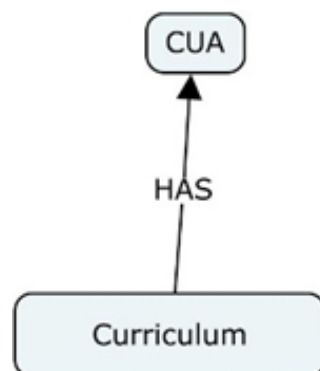


Рисунок 2 – Визуальное отображение онтологической модели графа знаний ОП университета с классом фраз компетенций, результатов обучения и аннотаций

Представим схематично процесс создания графа знаний ОП из файла с описаниями дисциплин, компетенций и результатов обучения на рисунке 3.

В результате построен граф в neo4j, и на рисунке 4 показан подграф с двадцатью пятью узлами.



Рисунок 3 – Процесс построения графа знаний образовательной программы

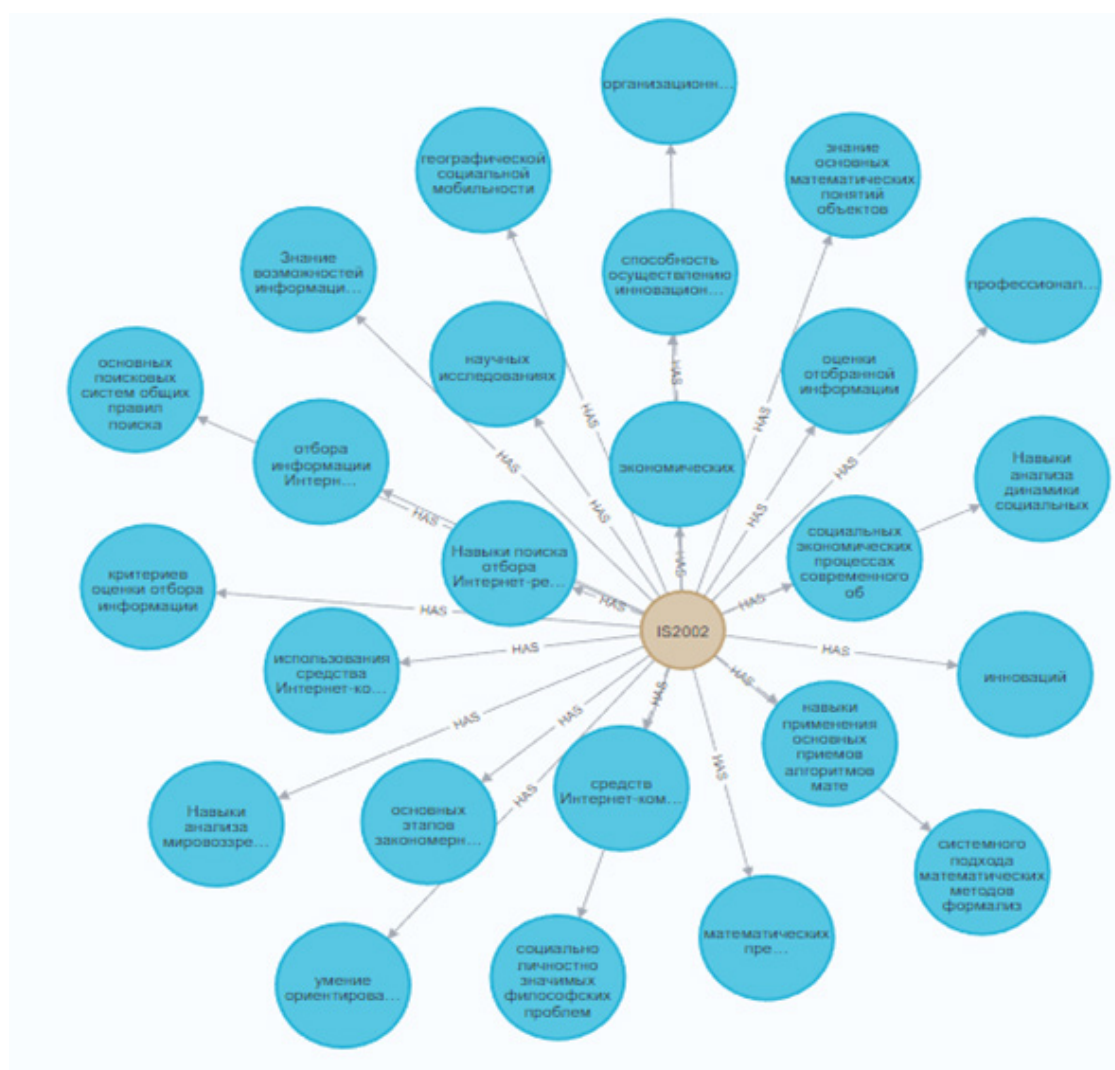


Рисунок 4 – Образец подграфа ОП с узлами CUA и связями HAS

Проведем анализ объявления о вакансии. Ключевыми сущностями являются вакансии, навыки, требования и обязанности. Вакансии – открытые позиции на основе профессий, которые работодатели хотят заполнить. Навыки – знания или умения в вакансии, которые работодатель ожидает от соискателя. Требования и обязанности – это дополнительные навыки, квалификации, рабочие обязанности, необходимые для вакансий.

В третьей таблице показана характеристика отношений атрибутов онтологической модели графа. Граф знаний включает множество классов сущностей $E_{\text{Recruitment}} = \{\text{Vacancy},$

Skill, Requirement, Responsibility} и множество отношений $R_{\text{Recruitment}} = \{\text{REQUIRE, HAS_REQUIREMENT, HAS_RESPONSIBILITY}\}$.

Таблица 3 – Описание онтологии графа знаний объявления о вакансии

Отношения и атрибуты	Описание
$\text{Vacancy} \xrightarrow{\text{REQUIRE}} \text{Skill}$	Вакансия требует определенных навыков.
$\text{Vacancy} \xrightarrow{\text{HAS_REQUIREMENT}} \text{Requirement}$	Вакансия имеет требования.
$\text{Vacancy} \xrightarrow{\text{HAS_RESPONSIBILITY}} \text{Responsibility}$	Вакансия имеет обязанности.
$f(\text{Vacancy}) = \{\text{title, profession, city, company, industry, experience, salary, url, publish_date}\}$	Вакансия характеризуется такими атрибутами, как наименование, название профессии, город, компания, сфера деятельности компании, опыт в годах, уровень зарплаты, ссылка на вакансию, дата публикации.
$f(\text{Skill}) = \{\text{name, embedding}\}$	Навык имеет название и его эмбединг
$f(\text{Requirement}) = \{\text{description, embedding}\}$	Требования содержат фразу текстового описания и его эмбединг.
$f(\text{Responsibility}) = \{\text{description, embedding}\}$	Обязанности содержат фразу текстового описания и его эмбединг.

Названия навыков и описания требований представлены в виде векторного представления (эмбединг), генерируемого методами машинного обучения. Визуально модель представлена на рисунке 5.

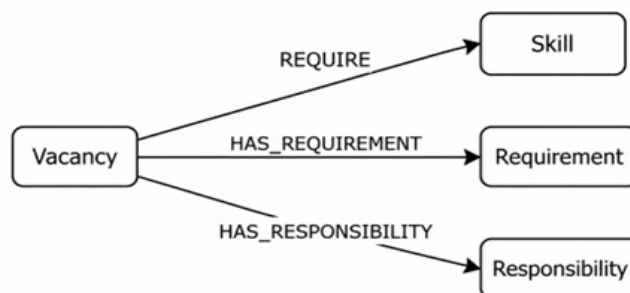


Рисунок 5 – Визуальное отображение онтологической модели графа знаний объявления о вакансии

Процесс создания графа знаний о вакансиях и навыках схематично отображен на рисунке 6.

Основная функция парсинга вакансий инициирует подключение к базе данных Neo4j. Перебирает предоставленный список городов, для каждого города выполняет запросы на получение данных с API. Скрипт отправляет HTTP-запрос на API сайта hh.ru для получения информации о вакансиях. В параметрах запроса указываются регион и количество результатов на странице и номер страницы. После получения ответа функция проверяет статус ответа и возвращает данные в формате JSON. Далее скрипт отправляет запрос к API для получения детальной информации о конкретной вакансии. Извлекает список навыков из ответа и возвращает их в виде строки, разделенной запятыми. Отправляет запрос к API для получения информации о компании по ее идентификатору. Извлекает информацию об отрасли компании из ответа. Добавляет информацию о вакансии в базу данных Neo4j, создавая узел Vacancy с деталями вакансии. Для каждого навыка, требования и обязанности создает соответствующие узлы и связывает их с узлом вакансии. Выполняет несколько запросов для удаления дубликатов узлов вакансий, а также пустых или неполных узлов навыков, требований и обязанностей.

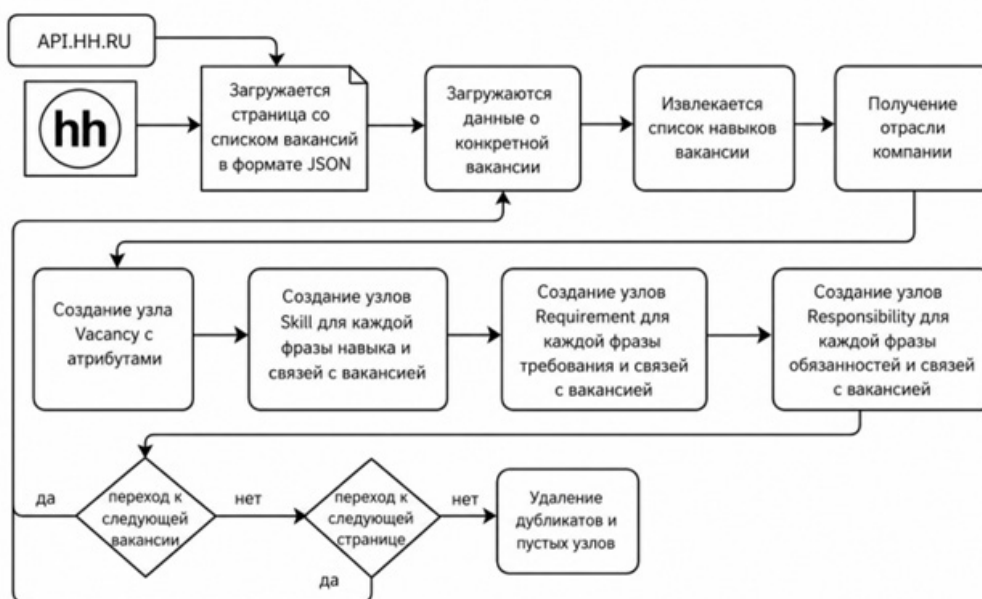


Рисунок 6 – Схема процесса создания узлов и связей графа знаний вакансии

На рисунке 7 отображен подграф вакансии, полученный по онтологии и процессу создания графа знаний вакансий. На изображении представлен подграф, центральным узлом которого является «Senior IT Project Specialist (Aksai)». Этот узел соединен с другими узлами навыков, требований и обязанностей, которые описывают различные аспекты его работы и квалификаций.



Рисунок 7 – Подграф вакансии. Сиреневым цветом обозначен узел с названием вакансии, красным цветом – узлы навыков, синим цветом – узлы требований, желтым цветом – узлы обязанностей

Проведем анализ предметной области MOOK. Ключевыми сущностями выступают курс MOOK, навыки MOOK, модуль MOOK и занятия MOOK. Курс MOOK – курс на сайте MOOK, предназначенный для неограниченного количества участников через интернет. Навыки MOOK – это конкретные знания, умения и навыки, которые студенты достигают в результате прохождения курса. Модуль MOOK – это структурная единица курса, представляющая собой отдельный блок или раздел, посвященный конкретной теме или набору учебных целей в рамках всего курса. Занятие MOOK – составная часть модуля, каждая из которых способствует достижению образовательных целей, определенных для данного модуля.

Рассмотрим онтологическую модель MOOK. Граф включает множество классов сущностей: $E_{MOOK} = \{MCourse, MSkill, Mmodule, MLesson\}$ и множество отношений $R_{MOOK} = \{DEVELOPED_BY, CONSISTS_OF\}$. Характеристика отношений и атрибутов приведена в таблице.

Таблица 4 – Описание онтологии графа знаний курса MOOK

Отношения и атрибуты	Описание
$MSkill \xrightarrow{DEVELOPED_BY} MCourse$	Навыки вырабатываются на курсах MOOK.
$MCourse \xrightarrow{CONSISTS_OF} Mmodule$	MOOK курс состоит из модулей.
$Mmodule \xrightarrow{CONSISTS_OF} MLesson$	Модуль состоит из уроков.
$f(MCourse) = \{title, url, organization, certificate_type, rating, reviews_num, difficulty, students_enrolled\}$	Курс характеризуется такими атрибутами, как наименование, ссылка на курс, организация, тип курса, рейтинг пользователей, количество отзывов, уровень сложности, количество подписанных студентов.
$f(MSkill) = \{name, embedding\}$	Навык имеет название и его эмбединг.
$f(Mmodule) = \{title, description, duration, embedding\}$	Модуль содержит наименование, фразу текстового описания, длительность в часах или минутах, эмбединг фразы описания.
$f(MLesson) = \{title, type, embedding\}$	Занятие имеет наименование, тип занятия, эмбединг наименования.

Названия навыков, названия и описание модулей дополнительно представлены в виде эмбедингов, генерируемых методами машинного обучения. Визуально модель представлена на рисунке 8.

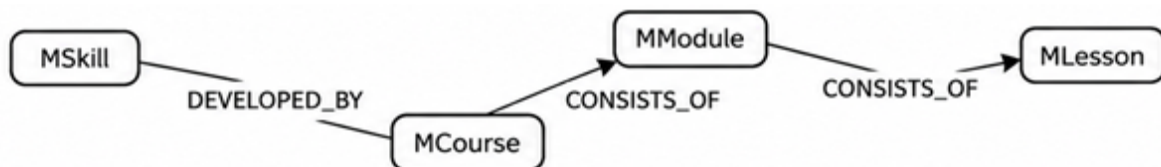


Рисунок 8 – Визуальное отображение онтологической модели графа знаний курса MOOK

Процесс создания графа знаний о вакансиях и навыках схематично отображен на рисунке 9.

Отправляем HTTP-запрос на сервер для получения содержимого страницы, после чего содержимое страницы анализируется с помощью библиотеки BeautifulSoup. Организуется циклический процесс скраппинга страниц, вызывая другие функции для сбора данных. Для каждой страницы и каждого элемента на странице вызывается функция, которая определяет, какие данные необходимо извлечь и как их обрабатывать. После сбора всех данных создается DataFrame. Для каждой строки в DataFrame создается узел курса MCourse. Для каждого на-

выка, модуля и урока в курсе создаются свои узлы MSkill, MModule, MLesson. Между узлами создаются связи, отражающие их взаимосвязи, например, какие навыки развиваются в каждом курсе.



Рисунок 9 – Схема процесса создания узлов и связей графа знаний MOOK

На рисунке 10 показан подграф курса MOOK, который имеет три навыка, связанные с другими курсами, и имеет пять модулей с уроками в каждом модуле.

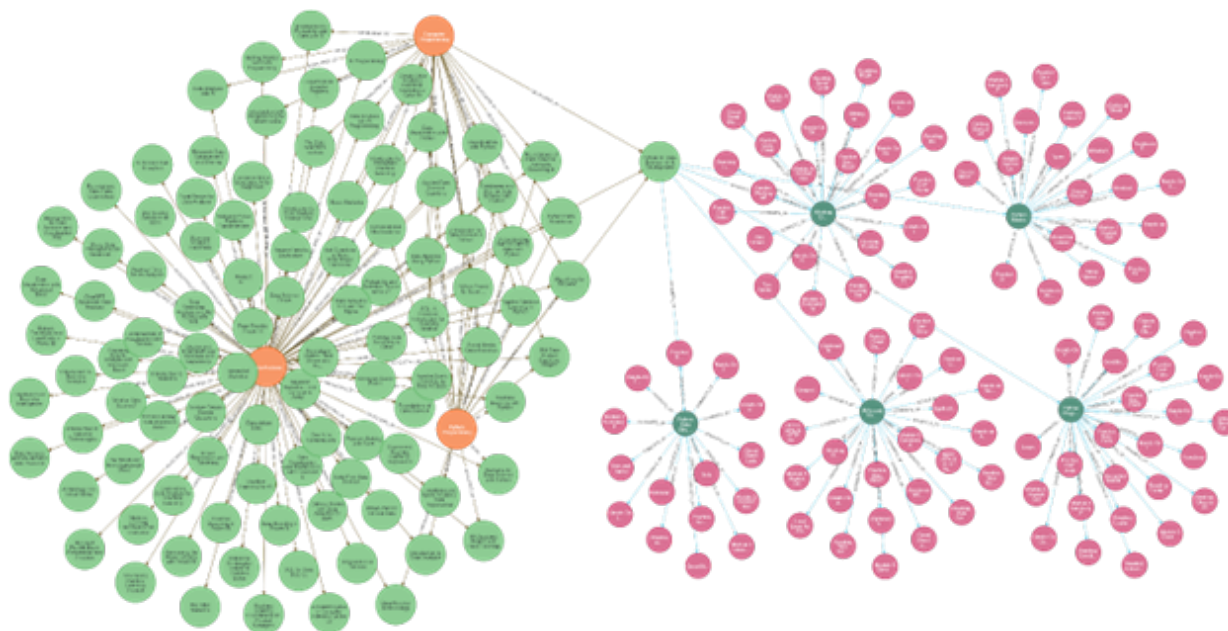


Рисунок 10 – Зеленые узлы – курсы MOOK, оранжевые узлы – навыки курса, темно-зеленые – модули, розовые – уроки модуля

Для экспериментальной оценки эффективности предложенной модели была использована мультязычная предобученная модель paraphrase-multilingual-mpnet-base-v2 (768-мерные эмбединги, контекстное окно 512 токенов, обучена на 50+ языках). Порог косинусного сходства

для связывания навыков между тремя доменами установлен на уровне 0,75. В качестве базовых методов для сравнения применялись BM25 и Word2Vec, широко используемые в аналогичных работах [20, 22]. Оценка качества рекомендаций проводилась на тестовой выборке из 100 вакансий и 200 курсов по метрикам $P@5$, $MAP@10$ и $NDCG@10$. Предложенная модель на основе paraphrase-multilingual-mpnet-base-v2 показала $P@5 = 0,74$, $MAP@10 = 0,68$, $NDCG@10 = 0,71$, тогда как BM25 продемонстрировал $P@5 = 0,52$, $MAP@10 = 0,47$, $NDCG@10 = 0,50$, а Word2Vec — $P@5 = 0,61$, $MAP@10 = 0,55$, $NDCG@10 = 0,59$. Таким образом, мультязычный трансформер превзошел базовые методы по всем метрикам: прирост $P@5$ составил 42% относительно BM25 и 21% относительно Word2Vec, что подтверждает эффективность выбранного подхода для смешанных русско-английских текстов. Качество выравнивания навыков между доменами дополнительно оценивалось по метрике F1, которая составила 0,79, что сопоставимо с результатами [19] для аналогичных задач классификации и связывания сущностей.

Тестовая выборка формировалась методом случайного стратифицированного отбора 100 вакансий и 200 курсов из общего массива данных (5248 вакансий и 953 курса), с сохранением пропорционального распределения по профессиональным категориям (таблица 1) и уровням сложности курсов, что обеспечивало репрезентативность выборки относительно генеральной совокупности. Для каждой вакансии два независимых эксперта в области информационных технологий вручную формировали эталонный список релевантных курсов на основе соответствия требуемых вакансией и формируемых курсом навыков; согласованность экспертной разметки оценивалась коэффициентом каппа Коэна ($\kappa = 0,81$), что соответствует высокой степени согласия. В качестве критериев оценки качества рекомендаций использовались метрики $P@5$, $MAP@10$ и $NDCG@10$, поскольку они дополняют друг друга: $P@5$ отражает долю релевантных курсов среди первых пяти позиций выдачи, $MAP@10$ учитывает точность по первым десяти позициям с учетом ранга каждого релевантного элемента, а $NDCG@10$ – качество ранжирования с учетом убывающей значимости позиции в списке рекомендаций.

Вычислительная сложность предложенного метода определяется преимущественно этапом расчета косинусной близости эмбедингов навыков между тремя доменами и в базовом варианте составляет $O(n \cdot m \cdot d)$, где n и m – количество навыков сравниваемых доменов, а d – размерность эмбединга (768). Применение приближенного индексирования эмбедингов при поиске ближайших соседей позволило снизить практическую сложность до $O((n+m) \cdot \log(n+m))$. На используемой конфигурации (GPU NVIDIA Tesla T4, 16 ГБ видеопамяти) среднее время генерации эмбединга одной фразы составило 1,8 мс, а полное построение графа знаний для всего набора данных (5248 вакансий и 953 курса), включая извлечение эмбедингов, расчет попарного сходства и запись узлов и связей в СУБД Neo4j, заняло около 42 минут. Формирование рекомендаций для одной вакансии в режиме инференса занимало в среднем 120 мс, что свидетельствует о практической применимости предложенного подхода в режиме, близком к реальному времени.

Заключение

Разработанная модель интеграции образовательных программ, требований рынка труда и онлайн-курсов (MOOK) через гетерогенные графы знаний представляет собой инновационный подход к адаптации учебных программ под современные вызовы IT-сектора. Применение машинного обучения и мультязычных трансформеров для анализа данных позволяет строить эффективные рекомендации по формированию навыков, востребованных на рынке. Прототип системы, построенный в Neo4j, демонстрирует возможность использования предложенной модели для создания персонализированных образовательных траекторий и карьерного планирования. Экспериментальная оценка показала, что предложенный подход превосходит базовые методы BM25 и Word2Vec по метрикам $P@5$, $MAP@10$ и $NDCG@10$, достигая $P@5 = 0,74$ и $NDCG@10 = 0,71$, что подтверждает преимущество мультязычных трансформеров при работе со смешанными русско-английскими текстами. Качество выравнивания на-

выков между доменами по метрике F1 составило 0,79. Вместе с тем необходимо отметить ряд ограничений проведенного исследования. Объем тестовой выборки (100 вакансий и 200 курсов) остается относительно небольшим для получения статистически устойчивых оценок, а сами данные ограничены вакансиями ИТ-сферы Казахстана за двухмесячный период и бесплатными англоязычными курсами Coursera по компьютерным наукам и науке о данных, что может снижать обобщающую способность модели на другие отрасли и типы образовательных ресурсов. Порог косинусного сходства (0,75), использованный для связывания навыков между доменами, был подобран эмпирически и не проверялся на устойчивость при изменении предметной области или языковой пары. Масштабирование модели на другие предметные области (например, здравоохранение, инженерные специальности, гуманитарные науки) потребует переобучения либо дополнительной тонкой настройки модели эмбедингов на терминологии соответствующей отрасли, а также пересмотра онтологической схемы графа с учетом специфики новых доменов. Перенос модели на другие образовательные платформы (edX, Udemu, локальные LMS вузов) потребует разработки отдельных модулей извлечения данных с учетом различий в структуре и доступности API, а также проверки совместимости эмбедингов при работе с иными языковыми парами и форматами описания курсов. В перспективе работа может быть расширена за счет добавления данных из других сфер, что повысит универсальность и эффективность предложенного решения.

Информация о финансировании. Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP22783030)».

ЛИТЕРАТУРА

- 1 Zhao, T., et al. Mathgraph: A knowledge graph for automatically solving mathematical exercises. In: Li, G., Yang, J., Gama, J., Natwichai, J., Tong, Y. (Eds.), Database Systems for Advanced Applications. DASFAA 2019, 629–633 (2019). https://doi.org/10.1007/978-3-030-18576-3_45
- 2 Li, W., et al. Constructing low-redundant and high-accuracy knowledge graphs for education. In: González-González, C.S., et al. Learning Technologies and Systems. ICWL SETE 2022, 139–150 (2023). https://doi.org/10.1007/978-3-031-33023-0_13
- 3 Agrawal, G., Deng, Y., Park, J., Liu, H., and Chen, Y.-C. Building knowledge graphs from unstructured texts: Applications and impact analyses in cybersecurity education. Information, 13 (11), 526 (2022). <https://doi.org/10.3390/info13110526>
- 4 Su, Y., and Zhang, Y. Automatic construction of subject knowledge graph based on educational big data. Proceedings of the 2020 3rd International Conference on Big Data and Education (ICBDE '20), 129–135 (2020). <https://doi.org/10.1145/3396452.3396458>
- 5 Qin, Y., Cao, H., and Xue, L. Research and application of knowledge graph in teaching: Take the database course as an example. Journal of Physics: Conference Series, 1607, 012127 (2020). <https://doi.org/10.1088/1742-6596/1607/1/012127>
- 6 Rizun, M. Knowledge graph application in education: A literature review. Acta Universitatis Lodzianis. Folia Oeconomica, 3 (342), 7–19 (2019). <https://doi.org/10.18778/0208-6018.342.01>
- 7 Sun, Y., Liang, J., and Niu, P. Generation of personalized knowledge graphs based on GCN. Journal of Computer and Communications, 9 (9), 108–119 (2021). <https://doi.org/10.4236/jcc.2021.99008>
- 8 Meissner, R., and Thor, A. Creation and utilisation of domain specific knowledge graphs (DSKG) for e-learning. DELFI 2021, 151–162 (2021). <https://dl.gi.de/server/api/core/bitstreams/b99e4bf0-10aa-4e4e-a0f1-96c1497bf3ee/content>
- 9 Chen, P., Lu, Y., Zheng, V.W., Chen, X., and Yang, B. Knowedu: A system to construct knowledge graph for education. IEEE Access, 6, 31553–31563 (2018). <https://doi.org/10.1109/ACCESS.2018.2839607>
- 10 Ke, Q., and Lin, J. Dynamic generation of knowledge graph supporting STEAM learning theme design. Applied Sciences, 12 (21), 11001 (2022). <https://doi.org/10.3390/app122111001>

- 11 Yu, X., Stahr, M., Chen, H., and Yan, R.G. Design and implementation of curriculum system based on knowledge graph. 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 767–770 (2020). <https://arxiv.org/ftp/arxiv/papers/2012/2012.12522.pdf>
- 12 Aliyu, I., Kana, A.F., and Aliyu, S. Development of knowledge graph for university courses management. *International Journal of Education and Management Engineering*, 10 (2), 1–10 (2020). <https://doi.org/10.5815/ijeme.2020.02.01>
- 13 Tel'nov, V.P., and Korovin, Ju.A. Semanticheskij veb i grafy znaniy kak obrazovatel'naja tehnologija podgotovki kadrov dlja jadernoj jenergetiki [Semantic Web and Knowledge Graphs as an Educational Technology for Nuclear Engineering Training]. *Izvestija vysshih uchebnyh zavedenij. Jadernaja jenergetika*, (2), 219–229 (2019). <https://doi.org/10.26583/npe.2019.2.19> (in Russian).
- 14 Nair, L.S., Shivani, M.K., and Cheriyan, S.J. Enabling remote school education using knowledge graphs and deep learning techniques. *Procedia Computer Science*, 215, 618–625 (2022). <https://doi.org/10.1016/j.procs.2022.12.064>
- 15 Chi, Y., Qin, Y., Song, R., and Xu, H. Knowledge graph in smart education: A case study of entrepreneurship scientific publication management. *Sustainability*, 10 (4), 995 (2018). <https://doi.org/10.3390/su10040995>
- 16 Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., and Seveso, A. Skills2Job: A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing*, 101, 107049 (2020). <https://doi.org/10.1016/j.asoc.2020.107049>
- 17 Groot, M., Schutte, J., and Graus, D. Job posting-enriched knowledge graph for skills-based matching. *arXiv preprint, arXiv:2109.02554* (2021). <https://doi.org/10.48550/arXiv.2109.02554>
- 18 Fettach, Y., Ghogho, M., and Benatallah, B. Knowledge graphs in education and employability: A survey on applications and techniques. *IEEE Access*, 10, 85857–85877 (2022). <https://doi.org/10.1109/ACCESS.2022.3194063>
- 19 Weichselbraun, A., Waldvogel, R., Fraefel, A., van Schie, A., and Kuntschik, P. Building knowledge graphs and recommender systems for suggesting reskilling and upskilling options from the web. *Information*, 13 (11), 510 (2022). <https://doi.org/10.3390/info13110510>
- 20 Zhu, G., Chen, Y., and Wang, S. Graph-community-enabled personalized course-job recommendations with cross-domain data integration. *Sustainability*, 14 (12), 7439 (2022). <https://doi.org/10.3390/su14127439>
- 21 Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., and Evans, J.A. Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, 115 (50), 12630–12637 (2018). <https://doi.org/10.1073/pnas.1804247115>
- 22 Chau, H.K. Explainable Course Recommendation: Connecting College Education to Knowledge and Careers Through Skills [Dissertation] (Pittsburgh, PA: University of Pittsburgh, Department of Informatics and Networked Systems, School of Computing and Information, 2024).
- 23 Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J.E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. Knowledge graphs. *ACM Computing Surveys*, 54 (4), Article 71 (2021). <https://doi.org/10.1145/3447772>
- 24 Ramazanova, V.S., Sambetbaeva, M.A., and Zagorul'ko, Ju.A. Obzor i analiz predstavlenij grafov znaniy [Review and Analysis of Knowledge Graph Representations]. *Trudy universiteta*, (1), 414–422 (2024). https://doi.org/10.52209/1609-1825_2024_1_414 (in Russian).
- 25 IEEE Standard for Framework of Knowledge Graphs. *IEEE Std 2807-2022*, 1–52 (2023). <https://doi.org/10.1109/IEEESTD.2022.10017167>
- 26 Zagorulko, G.B. Model, Methods, and Tools for Comprehensive Support of DSS Development in Poorly Formalized Subject Areas [Dissertation] (Novosibirsk: A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, 2020). (in Russian).

^{1*}**Серикбаева С.К.,**

PhD, доцент м.а., ORCID ID: 0000-0002-3627-3321,

*e-mail: inf_8585@mail.ru

²**Рамазанова В.С.,**

магистр, бас инженер, ORCID ID: 0000-0001-6100-3123,

e-mail: ramazanovavs@gmail.com

¹**Самбетбаева М.А.,**

PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9358-1614,

e-mail: sambetbayeva_ma_1@enu.kz

^{3,4}**Еримбетова А.С.,**

PhD, т.ғ.к., қауымдастырылған профессор, профессор-зерттеуші, ж.ғ.к.,

ORCID ID: 0000-0002-2013-1513,

e-mail: aigerian8888@gmail.com

⁵**Садирмекова Ж.Б.,**

PhD, қауымдастырылған профессор, ORCID ID: 0000-0002-7514-9315,

e-mail: Zhanna.Sadirmekova1@astanait.edu.kz

¹Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан

²М. Өтемісов атындағы Батыс Қазақстан университеті, Орал қ., Қазақстан

³META University, Алматы қ., Қазақстан

⁴Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан

⁵Astana IT University, Астана қ., Қазақстан

ЕҢБЕК НАРЫҒЫНЫҢ СҰРАНЫСЫ МЕН БІЛІМ БЕРУ БАҒДАРЛАМАЛАРЫН БІРІКТІРУ НЕГІЗІНДЕ МООС ТАҢДАУ ҮШІН БІЛІМ ГРАФТАРЫН ҚҰРУ МОДЕЛЬДЕРІ

Аңдатпа

Мақалада білім беру бағдарламалары, еңбек нарығы талаптары және онлайн-курстар (МООС) арасындағы байланыстарды талдауға арналған білім графтары мен машиналық оқыту әдістеріне негізделген интеграциялық модель қарастырылады. Зерттеудің өзектілігі ақпараттық технологиялар саласындағы жылдам өзгерістермен байланысты, бұл түлектер дайындығы мен жұмыс берушілер талаптарының сәйкес келмеуіне әкеледі. Осы алшақтықты азайту үшін университет бағдарламалары, бос жұмыс орындары және онлайн-курстар деректерін біріктіретін гетерогенді білім графын пайдалану ұсынылады. Модельде дағдыларды, вакансия талаптарын және оқу курстарын байланыстыратын мета-жол жүзеге асырылған, бұл пайдаланушылардың әртүрлі санаттары үшін білім алу және мансаптық даму бойынша ұсыныстар қалыптастыруға мүмкіндік береді. Мәтіндік деректерді өңдеу үшін орыс және ағылшын тілдеріндегі сипаттамаларды талдайтын көптілді трансформаторлар қолданылды. Деректер көздері ретінде білім беру бағдарламаларының құзыреттер карталары, hh.kz платформасының вакансиялары және Coursera курстары пайдаланылып, Neo4j графтық деректер қорына біріктірілді. Алынған нәтижелер жетіспейтін құзыреттерді анықтауға және оларды қалыптастыруға арналған курстарды автоматты түрде таңдауға болатынын көрсетті. Ұсынылған модель білім беру бағдарламаларын жобалау және білім беру ортасында шешім қабылдауды қолдау үшін қолданылуы мүмкін.

Түйін сөздер: білім графиктері, машиналық оқыту, көп тілді трансформаторлар, онлайн курстар, еңбек нарығы, дағдылар, ұсыныстар, білім беру бағдарламалары, ақпараттық технологиялар, интеграциялық модельдер.

¹**Serikbayeva S.,**

PhD, Acting Associate Professor, ORCID ID: 0000-0002-3627-3321,

*e-mail: inf_8585@mail.ru

²**Ramazanov V.,**

Master's degree, Chief Engineer, ORCID ID: 0000-0001-6100-3123

e-mail: ramazanovavs@gmail.com

¹**Sambetbayeva M.,**

PhD, associate professor, ORCID ID: 0000-0001-9358-1614,

e-mail: sambetbayeva_ma_1@enu.kz

^{3,4}**Yerimbetova A.,**

Cand. Eng. Sci., Associate Professor, professor-researcher, Leading Researcher,

ORCID ID: 0000-0002-2013-1513,

e-mail: aigerian8888@gmail.com

⁵**Sadirmekova Zh.,**

PhD, Associate Professor, ORCID ID: 0000-0002-7514-9315,

e-mail: Zhanna.Sadirmekova1@astanait.edu.kz

¹L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

²M. Utemisov West Kazakhstan University, Ural, Kazakhstan

³META University, Almaty, Kazakhstan

⁴Institute of Information and Computational Technologies CS MSHE RK, Almaty, Kazakhstan

⁵Astana IT University, Astana, Kazakhstan

MODELS FOR BUILDING KNOWLEDGE GRAPHS FOR MOOC SELECTION BASED ON THE INTEGRATION OF LABOR MARKET DEMANDS AND EDUCATIONAL PROGRAMS

Abstract

The article presents the development of an integration model based on knowledge graphs and machine learning methods for analyzing relationships between educational programs, labor market requirements, and online courses (MOOCs). The relevance of the study is driven by rapid changes in the information technology sector, which lead to a mismatch between graduate competencies and employer expectations. To reduce this gap, a heterogeneous knowledge graph integrating data from university curricula, job vacancies, and online courses is proposed. The model implements a meta-path connecting skills, job requirements, and learning courses, enabling recommendations for educational and career development for different categories of users. Multilingual transformers were applied to process textual descriptions in Russian and English. The data sources include competency maps of educational programs, job postings from the hh.kz platform, and courses from Coursera, all integrated into the Neo4j graph database. The obtained results demonstrate the capability to identify missing competencies and automatically recommend relevant courses for their acquisition. The proposed model can be applied in curriculum design, development of learning trajectories, and decision-making support within educational environments.

Keywords: knowledge graphs, machine learning, multilingual transformers, online courses, labor market, skills, recommendations, educational programs, information technology, integration models.

Received April 2, 2026; revised April 2, June 6, August 24, 2026; accepted August 31, 2026.