ETF Working together Learning for life

European Training Foundation

БОЛЬШИЕ ДАННЫЕ В КОНТЕКСТЕ АНАЛИТИКИ РЫНКА ТРУДА

ОЗНАКОМИТЕЛЬНОЕ ПОСОБИЕ

THE EU AGENCY SUPPORTING COUNTRIES TO DEVELOP THROUGH LEARNING

Доклад подготовлен Марио Мецанцаника и Фабио Меркорио для Европейского фонда образования

Введение

В ознакомительном пособии рассмотрены ключевые концептуальные, методологические и организационные аспекты использования больших данных для аналитики рынка труда (АРТ). Оно предназначено для статистиков, исследователей, аналитиков и разработчиков политики в странах-партнёрах Европейского фонда образования (ЕФО), которые сталкиваются с трудностями в области прогнозирования и распространения результатов анализа спроса на профессии, умения и квалификации.

Большие данные окружают нас повсеместно, но их потенциал и их применение в социальных исследованиях все еще остается новым явлением для многих государственных учреждений и стейкхолдеров в странах-партнёрах ЕФО и за их пределами.

Ознакомительное пособие поясняет, как можно использовать большие данные, чтобы выйти за рамки имеющихся информационных систем рынка труда (ИСРТ) и повысить эффективность существующей статистики. Традиционные методы АРТ, использующие в основном метод опросов, имеют ряд важных ограничений: стоимость, актуальность, точность, применение, интеграция и охват. Эти проблемы можно решить, но это потребует внимания правительств, заинтересованных лиц и их партнёров-доноров.

Источники и анализ больших данных дополняют и обогащают существующую статистику. Аналитику больших данных можно использовать для картирования умений по профессиям, выявления недостающих, или устаревших умений, прогнозирования потребности в новых профессиях или умениях – почти в режиме реального времени. Анализ больших данных позволяет получать более точные (детализированные) сведения в режиме реального времени в разбивке по территории и прогнозировать динамику.

Объём, разнообразие и скорость получения больших данных продолжат расти. Большие объёмы цифровых данных генерируются людьми, организациями, «умными» датчиками, спутниками, камерами видеонаблюдения, в интернете и бесчисленном количеством других устройств. Усилия, направленные на то, чтобы разобраться в них, создают интересные возможности. Создание знаний через данные является главной целью анализа Больших Данных. Другими словами – речь идет о формировании ценности.

Большие данные связаны с немаловажными вызовами и проблемами, в частности, с достоверностью. Это касается качества данных, которое может сильно варьироваться и требует соответствующих подходов, правил и методов. Существуют также вопросы, связанные с защитой данных и конфиденциальностью, требующие мер предосторожности.

Но, прежде чем погружаться в методы анализа Больших Данных, заинтересованная организация или группа заинтересованных лиц должна начать с вопроса: Какова в общих чертах проблема в нашей области? Какое мы видим решение? Кому нужны и кто будет использовать те данные, которые мы получим? Какими будут масштаб, степень детализации и визуализация данных? Кто наполнит смыслом полученные данные?

Области применения анализа Больших Данных очень широки; по счастью, явление и динамику рынков труда и профессиональных навыков можно проверить и проанализировать, используя Большие Данные. Однако ряд важных тем пока не может быть охвачен с помощью анализа Больших Данных, например, особенности и тенденции неформальной занятости, весьма значительной во многих странах.

Большие данные для ИСРТ сочетают в себе некоторые отдельные элементы цифровой трансформации, включая алгоритмы машинного обучения, использование больших объемов интернет-данных и определенную архитектуру вычислительной системы. Эти новые методы и источники данных будут продолжать развиваться. Также должны развиваться наши навыки и понимание в этой области. Данное пособие - это первый шаг.

Европейский фонд образования благодарит команду экспертов, авторов этого ознакомительного пособия - Марио Мецканаика и Фабио Меркорио - за гибкость в процессе разработки документа для, адаптирования информации к потребностям целевых пользователей, и за то, что поделились своим опытом и знаниями, с учетом собственных исследований (CRISP, Университет Милано-Бикокка) и других соответствующих проектов по всему миру, использованные в качестве примеров в данном пособии.

ЕФО выражает благодарность всем организациям, которые предоставили для данного документа примеры и случаи из практики, полезные для иллюстрации ключевых идей. Эксперт ЕФО Эдуарда Кастель-Бранко координировала работу и обсуждения с экспертами и руководила процессом рецензирования, который включал ценные комментарии экспертов ЕФО Майкла Райнера и Мартиньо Рубал Маседы.

СОДЕРЖАНИЕ

КРАТКОЕ ИЗЛОЖЕНИЕ

За последние несколько десятилетий, значимые силы и факторы рынка труда кардинально изменились в своей природе и характеристике, как в развитых, так и в развивающихся странах. С одной стороны технический прогресс, глобализация и реорганизация производственного процесса - с аутсорсингом и офшорингом - радикально изменили спрос на определенные навыки: некоторые рабочие места изживают себя, в то время как создаются новые рабочие места. Некоторые из них являются просто вариантами существующих рабочих мест, в то время как другие - это действительно новые рабочие места, которых не существовало еще несколько лет назад. Примечательно, что старение населения в странах с развитой экономикой усиливает потребность в непрерывном обучении и, вероятно, повлияет на структурный спрос на определенные компетенции: существенно изменились количество и качество спроса на навыки и квалификации, связанные с новым рынком труда. Значительно изменились не только новые навыки необходимые для выполнения новых работ, но и требования к квалификации существующих рабочих мест.

С другой стороны, за последние годы, объем информации о рынке труда (ИРТ) передаваемых через специализированные интернет-порталы и службы, растет в геометрической прогрессии, стимулируя и поддерживая реализацию многих интернет-услуг и инструментов, связанных с рынком труда, таких услуг, как подбор работы, реклама вакансий, услуги по обмену учебными программами и создание сети профессионалов, которые свободно делятся и обмениваются возможностями на рынке труда

В таком динамичном сценарии возникают некоторые вопросы, связанные с надлежащим наблюдением, пониманием и анализом феномена рынка труда и его динамики в Интернете, такие как: Какие профессии и где будут расти в будущем? Какие навыки будут наиболее востребованы компаниями и фирмами в ближайшие несколько лет? Какие навыки человек должен приобрести в процессе обучения в течение всей жизни? Какие рабочие места действительно являются новыми, а какие - просто эволюцией давно существующих рабочих мест, требующих новых или технологических навыков? Какое влияние оказывает распространение цифровых технологий на профессии? Какую роль играют общие навыки в существующих рабочих местах, и какие навыки являются наиболее важными?

Это лишь несколько вопросов, которые находятся на первом плане политических дебатов среди экономистов, политиков и экспертов рынка труда в целом. Сегодня эти вопросы необходимо решать, ориентируясь на парадигмы, основанные на данных, которые позволяют нам своевременно наблюдать и контролировать явления индуктивным способом на очень тонком уровне, т.е. данные используются для построения и подтверждения гипотез, а не наоборот.

Действительно, интернет-аналитика рынка труда предоставляет широкие возможности для мониторинга рынка труда в режиме реального времени, для лучшего понимания динамики рынка труда, фиксируя потребности в навыках и тенденции, сосредоточенные на различных измерениях (например, территория, сектора) на детальном уровне, т.е. большие данные, связанные с анализом рынка труда (Big Data 4 LMI). Неудивительно, что растет интерес к разработке и внедрению реальных ИРТ-приложений, чтобы интернет-данные по рынку труда способствовали деятельности по разработке и оценке политики путем принятия решений, основанных на фактических данных, и это представляет собой цель анализа рынка труда - области, которая становится все более актуальной для разработки и оценки политики рынка

труда Европейского Союза (ЕС).
В 2016 году Европейская комиссия подчеркнула важность направления деятельности в сторону профессионально-технических и образовательных мероприятий, поскольку они ценны для развития профессиональных и универсальных навыков, облегчения перехода к трудоустройству, поддержания и обновления навыков рабочей силы в соответствии с отраслевыми, региональными и местными потребностями. В 2016 году ЕС и Евростат запустили проект "Большие данные Европейской статистической системы", в котором участвуют 22 государства-члена ЕС с целью интеграции больших данных в регулярное производство официальной статистики, посредством пилотных проектов, исследующих потенциал отдельных источников больших данных и создания конкретных приложений. В том же году Европейский центр развития профессионального обучения (СЕДЕФОП) объявил тендер на создание системы анализа онлайн вакансий и разработку системы или инструмента для анализа вакансий и возникающих потребностей в навыках во всех государствах-членах ЕС, реализуя полноценную мультиязычную (32 языка) систему, собирающую вакансии, извлекающую навыки и осуществляющую мониторинг в режиме реального времени во всех 28 государствах-членах ЕС для поддержки принятия решений.

Хотя эти инициативы отличаются друг от друга, общая основа базируется на признании огромной информативной силы, стоящей за информацией о рынке труда в интернете. Эту информативную силу можно использовать, объединив усилия специалистов, статистиков, экономистов и экспертов рынка труда для получения полезных знаний о рынке труда из необработанных данных, чтобы понять динамику и тенденции рынка труда в Интернете и перейти к процессу принятия решений на основе данных с помощью информации о рынке труда.

В данной статье обсуждаются преимущества, потенциал, ограничения, методологические и технические проблемы, вопросы исследования, а также реальные проекты и тематические исследования, связанные с использованием больших данных для получения информации о рынке труда. Мы вводим этот вопрос, через обсуждение роли больших данных в контексте рынка труда, и обзор информации о рынке труда на сегодня. Затем мы обсудим некоторые технические аспекты, необходимые для внедрения аналитики больших данных в информацию о рынке труда. Приводятся примеры недавних приложений и проектов (как в ЕС, так и за его пределами), обсуждаются цели, используемые данные и источники, достигнутые результаты, а также открытые и сложные вопросы по каждому проекту. Наконец, мы обобщаем ряд рекомендаций и шагов для Европейского фонда образования (ЕФО) и его стран-партнеров, а также предлагаем некоторые идеи для проектов, которые возникли на конференции ЕФО «Навыки на будущее: Управление переходным периодом», прошедшей в Турине в ноябре 2018 года.

1. БОЛЬШИЕ ДАННЫЕ И ИНФОРМАЦИЯ О РЫНКЕ ТРУДА: КАК УЛУЧШИТЬ ИРТ В ЦИФРОВУЮ ЭПОХУ - ОБЗОР, СОСТОЯНИЕ ДЕЛ, ПОТЕНЦИАЛ И ОГРАНИЧЕНИЯ

1.1 Общие сведения и определения

Чтобы облегчить чтение данного документа, в этом разделе кратко представлены некоторые термины и базовые понятия, связанные с данными о рынке труда (РТ).

Информация/аналитика о рынке труда

Эти два термина - часто используемые как взаимозаменяемые - относятся к данным, связанным с явлениями и динамикой РТ, которые полезны для поддержки принятия решений, разработки политики и оценки. Однако из использования LMI неясно, что означает I - информация или аналитика.

В частности, «I» как информация описывает все виды данных и информации, используемых для поддержки операционной деятельности, связанной с РТ (без аналитики), а также любую информацию, связанную со спросом и предложением на РТ. Примерами могут служить объявления о вакансиях, навыки, профессии и резюме соискателей.

В сравнении, «I» как аналитика, является развивающейся концепцией во всем сообществе РТ, особенно в Европейском Союзе (ЕС). Несмотря на отсутствие единого определения аналитики РТ, его можно описать как разработку и использование алгоритмов и рамок искусственного интеллекта (ИИ) для анализа данных, связанных с РТ (также известных как информация о рынке труда), в целях поддержки политики и принятия решений (см., например, [1], [2], [3]).

Вопросы и ответы

Когда «I» как информация, становится аналитикой?

Грубо говоря, использование необработанных или агрегированных данных, включая данные, отслеживаемые с течением времени, для поддержки операционной деятельности все еще является Информацией. Информация становится аналитикой, когда автоматизированный алгоритм (сегодня в основном использующий искусственный интеллект) обрабатывает их для получения информации, полезной для целей аналитики (например, прогнозирование для принятия решений, машинное обучение для классификации или извлечение информации для определения навыков в резюме). В частности, способность обрабатывать и анализировать большие объемы данных в режиме реального времени позволяет использовать знания, полученные в процессе анализа, в системах, обычно предназначенных для поддержки оперативной деятельности.

В таком сценарии аналитику РТ следует рассматривать как деятельность, которая - как и ожидалось - дает результат, называемый знаниями о РТ. Здесь применяется общее определение знания, другими словами, озарение и дополнительная информация, извлеченная из опыта (в данном случае информация о РТ), которая может повысить осведомленность и понимание наблюдаемого явления. Эти знания, в свою очередь, позволяют пользователям прогнозировать и анализировать (о чем мы поговорим позже).

Вопросы и ответы

Могут ли информация и аналитика работать вместе в рамках системы (или структуры) для поддержки деятельности по принятию решений?

Да, именно таким образом должны взаимодействовать информация и аналитика о РТ, а именно в рамках системы (бэкэнд), которая собирает информацию о РТ и использует ИИ для создания аналитики РТ. Аналитика о РТ, как результат предоставляется ряду заинтересованных сторон в соответствии с их потребностями и возможностями понимания динамики рынка труда. Этот процесс описывает, как должна работать информационная система рынка труда (ИСРТ).

Информационная система рынка труда

В предыдущем разделе мы пояснили разницу между информацией о РТ (т.е. исходными данными, обычно используемыми для обмена информацией внутри процессов оперативного обслуживания, относящихся к РТ) и аналитикой РТ (инструментами, алгоритмами и процедурами для управления информацией о РТ). Эти две концепции участвуют в реализации ИСРТ, где информационная система обычно определяется как набор взаимосвязанных компонентов (технологических и архитектурных), которые работают совместно для сбора, получения, обработки, хранения и распространения информации в целях облегчения такой деятельности, как планирование, контроль, координация, анализ и принятие решений в бизнес-организациях Таким образом, ценность информации, доступной через информационную систему, двояка: во-первых, она поддерживает операционные процессы, а во-вторых, помогает лицам, принимающим решения, достичь целей анализа.

ИСРТ (восприятие)

ИСРТ можно рассматривать как один из примеров классической информационной системы, которая использует информацию и аналитику РТ для поддержки оперативной деятельности и принятия решений.

По сути, концепция ИСРТ может быть описана как набор инструментов, способных извлекать, анализировать и распространять информацию, связанную с РТ. Тем не менее, нет единого определения того, какой должна быть ИСРТ, и нет универсального практического совета по разработке ИСРТ, поскольку ее архитектура, данные и методы зависят от потребностей анализа, которые зависят от контекста (например, от страны, учреждения, важных вопросов, приоритетов политики и инфраструктуры обработки данных). Практические и различные примеры ИСРТ представлены в пунктах [4], [5], [6], [2], вот лишь некоторые из недавних работ. Некоторые из них будут рассмотрены далее в документе. В этом отношении, доступность данных в Интернете (см. главу 2) освещает значение модернизации и развития ИСРТ, с целью включения интернет-данных и использования алгоритмов ИИ для получения полезных выводов и формулирования прогнозов динамики и тенденций развития РТ (как недавно утверждал Джонсон в [7] и показали Фрей и Осборн [8] для прогнозирования риска роботизации). Эти причины заставили аналитиков и экспертов по РТ включить интернет как дополнительный источник данных и информации по РТ в свою работу, чтобы лучше описать и понять РТ в целом.

Источники данных о РТ

Административные, статистические и интернет-данные - это три основные категории данных, которые могут работать вместе для объяснения какого-либо явления. Этот очень краткий обзор трех основных типов данных подчеркивает их отличительные особенности и сходство.

Административные данные. По сути, надежное определение административных данных - это "наборы данных, собранные правительственными учреждениями или налоговыми службами"[9]. Это означает, что эти данные также относятся к информации, собранной от (или о) физических лиц, которым необходимо принять меры, чтобы стать частью системы, использующей административные данные (например, регистрация фермеров в системе налогообложения и социального обеспечения) или не стать (например, трудовое законодательство Италии гласит, что система должна автоматически отслеживать начало/окончание каждого трудового договора (см. [10]).

Статистические данные. Статистические данные (также известные как данные обследований) собираются в соответствии с конкретной и заранее определенной статистической целью для обеспечения заданного охвата населения, определений, методологии, качества и времени, чтобы удовлетворить аналитические потребности заинтересованных сторон (см., например, [11]). Очевидно, что использование административных данных для статистических целей далеко не простое дело, поскольку оно связано с такими сложными вопросами, как определение совокупности, целевой совокупности и размера выборки, а также с трудностью выбора переменной модели для выборки совокупности.

Хотя статистические и административные данные различаются по целям, они имеют некоторые общие интересные характеристики, как показано в таблице 1.1.

ТАБЛИЦА 1.1 ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ИСТОЧНИКОВ ДАННЫХ О РТ

Тип источника	Категория данных[1]	Параметр генерации	Парадигма модели данных	Качество	Охват	Парадигма анализа	Достоверность	Ценность
Статистический	Структурированный	Периодически	Относительный	Ответственность владельца	Ответственность владельца	Нисходящий и с использованием модели	Ответственность владельца	Объективный
Административный	Структурированный или полуструктурированный	Периодически	Относительный	Ответственность владельца	Ответственность владельца и потребителя	Нисходящий и с использованием модели	Ответственность владельца и потребителя	Объективный
Интернет	Структурированный, полуструктурированный или не структурированный	Почти в реальном времени или в реальном времени	Реляционный или нереляционный (NoSQL)	Ответственность потребителя	Ответственность потребителя	Восходящий и основанный на больших данных	Ответственность потребителя	Не свойственный

Статистические данные часто представляют собой структурированные данные (например, таблицы с числами с четко определенной структурой и типом), в то время как административные данные могут также включать полуструктурированные данные, где

[1] *Структурированные данные* - это четко определенные типы данных, чья структура и повторяющийся шаблон делают их легко доступными для поиска автоматизированной системой. *Неструктурированные данные* - это данные, структура которых не может быть легко определена в виде шаблона или типа, что делает поиск в этих данных сложным (например, свободный текст, аудио, видео и сообщения в социальных сетях). *Полуструктурированные данные* относятся к данным, структура которых частично определена (например, XML документы).

структура частично определена и может появляться свободный текст. Тем не менее, эти данные можно легко хранить, используя классические реляционные парадигмы (например, стандартный язык запросов (SQL)). Гарантия того, что статистические данные являются качественными, это ответственность производителя данных или владельца данных, который также разработал сбор данных/исследование. Это может не сработать для административных данных, качество которых может считаться приемлемым для владельца данных, но низким для потребителя данных. Это неудивительно, поскольку качество данных определяется как "пригодность к использованию", поэтому удовлетворенность качеством может меняться по мере того, как меняется пользователь. Административные данные собираются для мониторинга какого-либо явления, а не для аналитических целей (см., например, [12]). Это также означает, что правдоподобность статистических данных - означающая "степень, в которой данные принимаются или рассматриваются как истинные, реальные и достоверные" [13] - зависит от благонадежности производителя/владельца данных, и это также может быть справедливо для административных данных. Поскольку оба эти вида данных собираются из системы (административные) или предназначены для конкретной цели анализа (статистические), их ценность является неотъемлемой.

Другими словами, данные по своей сути имеют ценность. Очевидно, что эта ценность может быть повышена путем анализа и связывания данных, но она все равно остается.

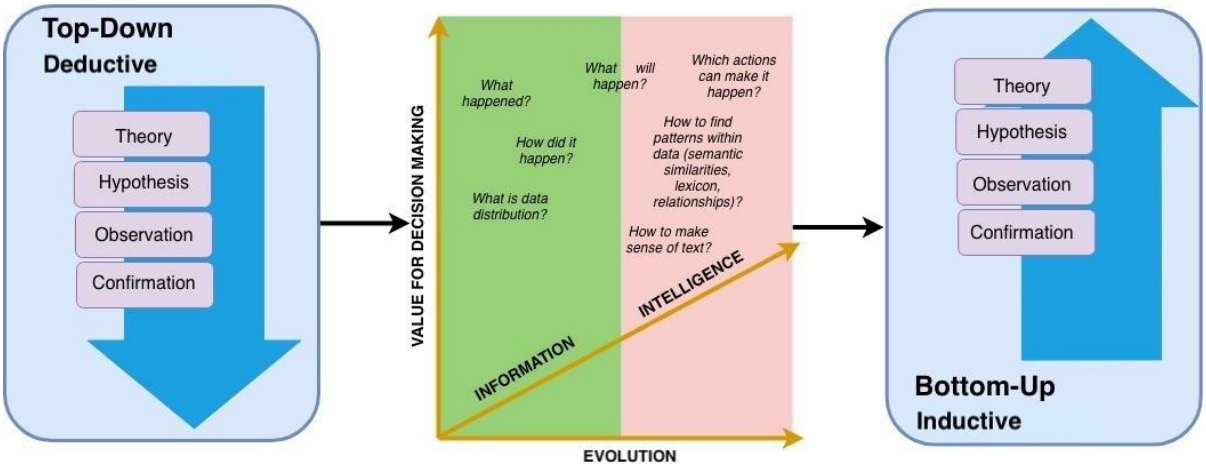
Этот сценарий меняется при работе с интернет-данными, под которыми понимаются все данные, поступающие из интернет-источников. Как можно предположить, эти данные могут иметь любую структуру, поэтому они могут быть структурированными (например, таблицы, собранные из Интернета), полуструктурированными (например, XML[2], такие как твиты) или полностью неструктурированными (все остальное). Эти данные постоянно генерируются из одного или нескольких интернет-источников, над которыми пользователь данных не имеет контроля, и это вынуждает пользователя постоянно отслеживать и собирать данные. Поскольку структура интернет-данных может меняться непредсказуемым образом, объективные парадигмы (которые требуют фиксированной и определенной структуры данных) не могут быть использованы для хранения интернет-данных по мере их поступления из сети. Для решения этой проблемы были разработаны парадигмы NoSQL[3]. Кроме того, качество зависит от способности пользователя выявлять проблемы в данных (дубликаты, недостающие данные, опечатки, синонимы и т.д.), а также от охвата, который должен оценивать и измерять пользователь данных, часто объединяя несколько источников данных в интернете. Следовательно, правдоподобность зависит от надежности пользователя данных, а не их владельца. Наконец, интернет-данные не обладают внутренней ценностью; их ценность зависит от их способности описать и объяснить явление. Другими словами, интернет-данные являются необработанными, и их ценность должна быть раскрыта/исследована пользователем.

Интернет-данные можно сравнить с глыбой гранита, которую должен обработать мастер, который может решить использовать фрезерный станок или резец, чтобы придать ту или иную форму.

Эти различия, в основном между интернет-данными, статистическими и административными данными, также заставляют пользователя изменить свой подход к анализу: от подхода, основанного на модели и использующего процесс "нисходящей", к подходу, основанному на данных и требующему "восходящего" метода, как показано на рисунке 1.1.

[2] Расширяемый язык разметки (XML)-это язык разметки, используемый для определения набора правил для кодирования документов в формате, который легко воспринимается человеком и является машиночитаемым.
[3] NoSQL (не только SQL) относится к растущему движению за поддержку хранения и запроса неструктурированных данных. Роль NoSQL в рамках ИСРТ обсуждается в главе 2.

РИСУНОК 1.1 ИЗМЕНЕНИЕ ПАРАДИГМЫ – ОТ «НИСХОДЯЩЕЙ» К «ВОСХОДЯЩЕЙ»



Вопросы и ответы

Из-за массового присутствия неструктурированных данных/текстов кажется, что использование больших данных делает невозможным выполнение задач качества данных, которые хорошо определены для структурированных данных. Так ли это на самом деле?

Вопрос о применении задач качества (и очистки) данных к интернет-данным все еще остается открытым. Некоторые считают, что интернет-данными следует управлять, как классическими структурированными данными, в то время как другие говорят, что принцип "мусор внутрь, мусор наружу" не применим к большим данным, поскольку объем будет действовать как фактор деноизирования. По нашему опыту, качество больших данных в первую очередь зависит от надежности источников, используемых для сбора данных. Таким образом, ранжирование интернет-источников имеет решающее значение. Можно применять любые методы оценки качества данных: основанные на правилах (если можно определить модель данных) или статистические (для выявления выбросов и деноизирования данных).

Общий Регламент по вопросам защиты данных, связанных с РТ

Общий регламент по защите данных (ОРЗД) вступил в силу в мае 2018 года во всех государствах-членах ЕС. Он представляет собой первый шаг к регулированию использования и обработки персональных данных. Если данные не содержат личной информации, ОРПЗД не применяется. Но в случае, если данные содержат личную информацию, относящуюся к субъекту (например, данные резюме, личные предпочтения, прежний послужной список или личные навыки), то ИСРТ, которая использует эти данные, должна соответствовать ОРЗД.

По сути, ОРЗД направлен на обеспечение основных прав субъекта данных, а также на повышение ответственности компаний, которые контролируют и обрабатывают персональные данные. ОРЗД устанавливает ряд ограничений и запретов на использование персональных данных.

- Во-первых, право доступа субъекта данных к собранной о нем информации, накладывает ограничения на автоматизированное принятие решений компаниями и

организациями, использующие эти данные.

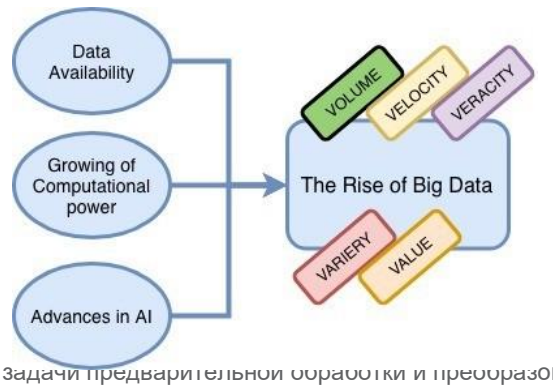
- Во-вторых, организации, назначенные для обработки персональных данных (т.е. обработчики данных), должны уведомить субъектов данных о собранных данных (статьи 13-15).
- В-третьих, прозрачность играет ключевую роль, заставляя обработчика данных работать с данными прозрачным образом (Статья 5, §1a), посредством прозрачной обработки данных (Статья 13, §2 и 14,§2), и уведомлять субъекта данных, если автоматизированный процесс принятия решений применяется к его персональным данным (Статья 22). Кроме того, в статьях 13 и 14 говорится, что при профилировании субъект данных имеет право на "содержательную информацию о задействованной логике".

С технической точки зрения это также относится ко всем процессам и процедурам извлечения, преобразования и загрузки (ИПЗ)[4], которые извлекают данные, преобразуют их из одного формата в другой и, наконец, загружают обработанные и агрегированные данные в хранилища данных для аналитических целей. Таким образом, при расчете ключевого показателя эффективности или общего бизнес-показателя невозможно определить источник информации или то, какие данные, относящиеся к субъекту данных, были использованы. Это также относится к (персональным) данным, связанным с РТ. Грубо говоря, это означает, что обработчик данных среди прочего несет ответственность за обеспечение нижеследующих условий: (i) субъект, к которому относятся данные, не может быть идентифицирован ни прямо, ни косвенно, если идентификаторами субъекта являются его имя, паспортные данные, данные о местоположении, онлайн-идентификатор или один или несколько характерных элементов его физической, физиологической, генетической, психологической, экономической, культурной или социальной идентичности (Статья 4); (ii) данные обрабатываются законно, справедливо и прозрачно по отношению к субъекту данных (Статья 5); и (iii) данные собираются для определенных, явных и законных целей и не обрабатываются далее способом, несовместимым с этими целями (Статья 5).

1.2 Большие данные и ИРТ

Причина растущего интереса к обработке и использованию больших данных заключается в том, что они позволяют менеджерам более эффективно оценивать свой бизнес и, следовательно, создавать знания, которые могут улучшить процесс принятия решений и производительность (см., например, [14]). Это очень общее утверждение, которое применимо и к РТ. Однако, что на самом деле представляют собой большие данные, что делает данные большими, а что нет, а также проблемы и возможности, связанные с работой с большими данными, - все эти вопросы, все еще остаются открытыми для обсуждения.

РИСУНОК 1.2 КЛЮЧЕВЫЕ ЭЛЕМЕНТЫ, ОПРЕДЕЛЯЮЩИЕ РОСТ БОЛЬШИХ ДАННЫХ



4 ИПЗ - это, подход, поддерживающий задачи предварительной обработки и преобразования данных в процессе обнаружения знаний в базах данных (KDD). Данные, извлеченные из исходной системы, подвергаются серии преобразований, которые анализируют, обрабатывают и затем очищают данные перед загрузкой в хранилище данных.

В последние годы сообщество пыталось ответить на эти вопросы, используя различные "модели" Больших Данных, основанные на измерениях (или пяти "V"), которыми должно обладать приложение/подход к Большим Данным. Хотя было предложено несколько моделей, здесь мы предлагаем модель пяти "V", адаптированную для РТ, которая характеризует Большие Данные в отношении пяти фундаментальных измерений: объем, скорость, разнообразие, правдивость и ценность.

Объем. В 2017 году в мире насчитывалось около 4 миллиардов пользователей интернета. Это число растет с большой скоростью: первый миллиард был достигнут в 2005 году, второй миллиард - в 2010 году и третий миллиард - в 2014 году. Около 40% населения имеют доступ к интернету. В 2018 году насчитывается около 2 млрд активных веб-сайтов (без учета "глубокой паутины", то есть веб-страниц, которые не могут быть проиндексированы поисковыми системами), и каждую минуту выполняется более 3,5 млрд поисковых запросов в Google5. Каждую секунду через Интернет проходит больше данных, чем хранилось во всем Интернете всего 20 лет назад. Это дает компаниям уникальную возможность доступа и сбора этих данных для принятия лучших решений и совершенствования своего бизнеса. Например, по оценкам, Walmart может собирать около 2,5 петабайт (т.е. 2,5 квадриллиона байт) данных каждый час от транзакций своих клиентов. Хотя классический подход к Большим Данным измеряет объем в байтах, что хорошо подходит для пользовательских данных, генерируемых системой (например, журналы и транзакции), эта единица измерения не применима к информации о РТ, поскольку масштаб значительно меняется. В сфере РТ может представлять интерес измерение количества собранных записей или элементов, относящихся к спросу или предложению на РТ, или количества рассмотренных источников РТ.

Скорость. Это измерение относится к скорости, с которой генерируются или собираются данные в случае ИРТ. Эти данные собираются из сторонних источников, которые могут позволить собирать данные автономно с помощью (i) интерфейсов прикладного программирования (ИПП)6, (ii) периодически выполняемых пакетных процедур или (iii) практически в режиме реального времени, выполняя краулинг или скрейпинг[7] источника и получая данные через близкие, фиксированные интервалы времени. Очевидно, что чем ниже частота сбора данных, тем выше объем собранных данных и тем больше потребность в больших вычислительных ресурсах и больших ресурсах хранения.

Вариации. Это измерение относится к разнообразию типов данных в источниках Больших Данных, как показано в таблице 1.1. Источник может быть структурированным, полуструктурированным или полностью неструктурированным. Вследствие использования больших объемов неструктурированного контента, лексика, используемая в каждом источнике, различается, а широкомасштабное использование естественного языка означает значительную неоднородность данных.

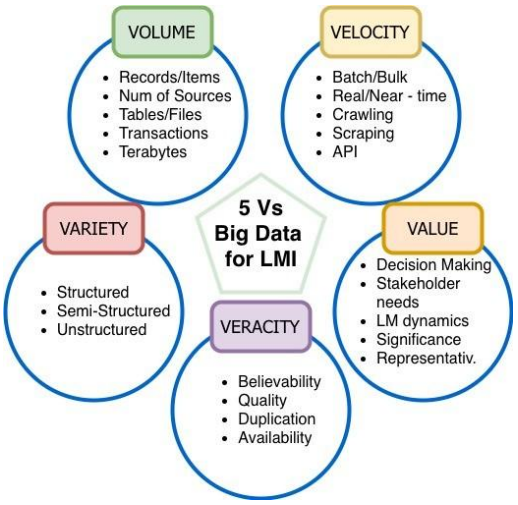
Достоверность. Достоверность данных показывает, насколько точным или правдивым может быть набор данных. Как говорилось выше, качеством интернет-данных нельзя манипулировать в источнике, его приходится оценивать в процессе сбора и хранения данных с помощью

5 Источник: Интернет Live Stats (www.Internetlivestats.com/)
6 В веб-контексте, ИПП относится к набору процедур, которые пользователь использует для связи с сервером (например, для сбора данных). Следовательно, связь между пользователем и сервером регулируется, четко определена и контролируется.
7 Краулинг собирает данные с веб-сайтов в том виде, в котором они есть, в то время как скрейпинг определяет некоторые части веб-сайта, которые необходимо собрать. По сути, скрейпинг анализирует и определяет данные, которые пользователь хочет собрать, в то время как краулинг просто собирает весь веб-контент.

специальных анализов, процедур и инструментов. Предвзятость, аномалии или несоответствия, дублирование и изменчивость - вот некоторые из аспектов, которые необходимо устранить для повышения точности Больших Данных. Как можно предположить, для данного источника данных, чем выше разнообразие, тем выше достоверность. Действительно, использование естественного языка привносит в текст большое количество шума, не содержащего никакой информации (например, предлоги, термины, не относящиеся к интересующей теме, союзы и аббревиатуры, которые необходимо расширить). Все эти вопросы должны быть решены должным образом, чтобы позволить неструктурированным данным производить знания на этапах извлечения информации из данных (KDD).

Ценность. Наконец, данные должны представлять ценность для конкретной области или цели. Другими словами, как говорилось выше, интернет-данные не имеют внутренней стоимости. Их ценность - это знания, которые пользователь извлекает из данных для объяснения какого-либо явления или для поддержки принятия решений посредством анализа и рекомендаций. Стоит отметить важность анализа потребностей заинтересованных сторон, который должен определить, какая часть знаний представляет интерес для данной заинтересованной стороны, а какая нет. В случае с вакансиями, размещенными в интернете, пользователь, ищущий новую работу, будет заинтересован в проведении анализа пробелов в отношении своих навыков, в то время как аналитик РТ может быть заинтересован в наблюдении за РТ в интернете в целом и как на конкретном региональном уровне. Одно и то же знание может иметь разные точки доступа в зависимости от потребностей заинтересованных сторон.

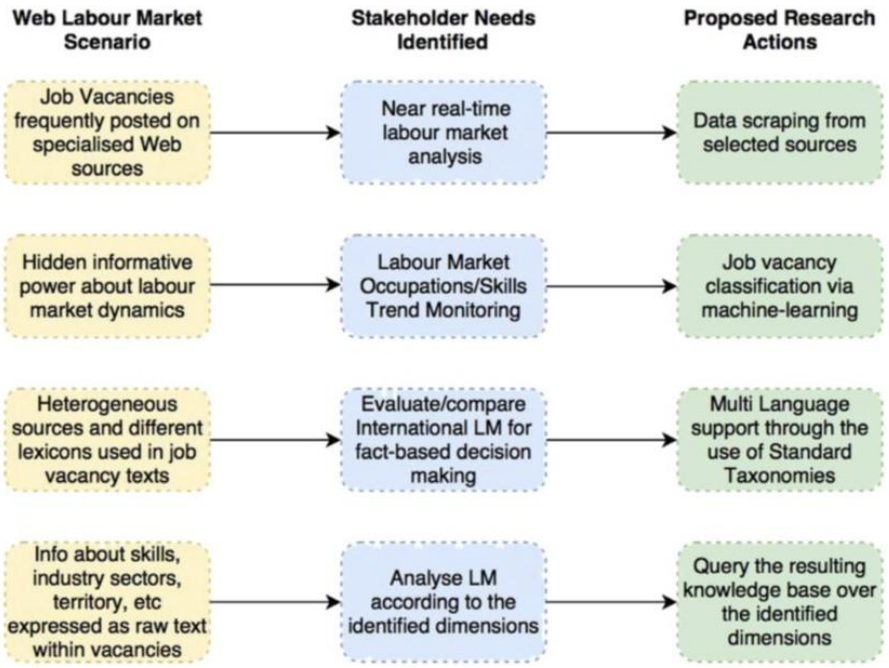
РИСУНОК 1.3 ПЯТЬ "V". МОДЕЛЬ БОЛЬШИХ ДАННЫХ, АДАПТИРОВАННАЯ ДЛЯ ПРИМЕНЕНИЯ В ИРТ



Превращение Больших Данных в информацию о РТ

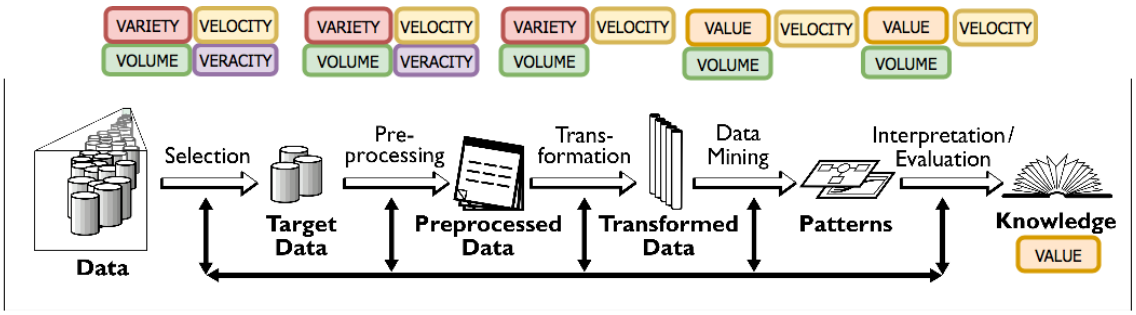
Было рассмотрено извлечение знаний из (больших) данных о РТ. Для этого на рисунке 1.4 приведены основные проблемы, возникающие при работе с ИРТ в Интернете, которые обсуждались в [15].

РИСУНОК 1.4 ОСНОВНЫЕ ЭЛЕМЕНТЫ СЦЕНАРИЯ РТ В ИНТЕРНЕТ, ПОТРЕБНОСТИ ЗАИНТЕРЕСОВАННЫХ СТОРОН И ПРЕДЛАГАЕМЫЕ ДЕЙСТВИЯ



Одним из подходов, позволяющих управлять Большими данными для ИРТ, является процесс KDD. Процесс KDD состоит из пяти основных этапов, как показано в [16] на рисунке 1.5: выбор, предварительная обработка, преобразование, поиск данных и машинное обучение, интерпретация/оценка. Очевидно, что она должна быть адаптирована к интересующей области, усиливая одну задачу или шаг по отношению к другой.

РИСУНОК 1.5 ПРОЦЕСС KDD И "V" БОЛЬШИХ ДАННЫХ, ЗАДЕЙСТВОВАННЫЕ НА КАЖДОМ ЭТАПЕ



Источник: Рисунок основан на иллюстрации из [16].

Выборка. Первым шагом является выборка источников данных. Каждый интернет-источник должен быть оценен и классифицирован точки зрения надежности информации. Например, на этом этапе следует учитывать дату публикации вакансии, частоту обновления сайта, наличие структурированных данных и любые ограничения на загрузку. В конце этого этапа составляется рейтинг надежных веб-источников. На этом этапе задействованы все пять "V" Больших Данных, включая также достоверность, то есть косвенное влияние, искажения и отклонения в данных. Некоторые ключевые вопросы, поставленные на этапе отбора экспертов РТ, включают:

1. [Статистический] Как мы определяем критерии, которые должны быть включены в модель источников, и как мы извлекаем эти критерии (т.е. переменные) из источников? Как мы оцениваем источники?
2. [Технический] Как определить парадигму модели данных (например, реляционная, документная, ключевое значение, графики) для хранения огромных объемов данных в масштабе? Как мы собираем данные автоматически? Нужен ли нам доступ к ИПП, или нам нужно разработать веб-скрепер/сборщик? Как планировать процессы автоматического сбора данных?
3. [Эксперт в области РТ] Как мы выбираем правильные источники? Выбрали ли мы правильные источники?

Предварительная обработка. Этот этап включает очистку данных для удаления искажения из данных или неуместных выбросов (если таковые имеются), принятие решения о том, как обрабатывать отсутствующие данные, а также определение функции для обнаружения и удаления дублирующихся записей (например, дублирующихся вакансий или вакансий с отсутствующими значениями). Качество и очистка данных являются важнейшими задачами в любом подходе к принятию решений на основе данных, чтобы гарантировать достоверность всего процесса, т.е. "степень, в которой данные принимаются или рассматриваются как истинные, реальные и достоверные" (см., например, [12], [13], [17]).

Выявление дублирования вакансий далеко не простое дело. Вакансии обычно размещаются на нескольких сайтах, и это является дублированием, в то время как повторное использование одного и того же текста для объявления аналогичной вакансии таковым не является. Определение подходящих признаков для правильного распознавания дубликатов имеет решающее значение в интернет-сфере РТ. Этап предварительной обработки снижает сложность сценария Больших Данных, смягчая влияние аспекта достоверности за счет качества и очистки данных. Ключевые вопросы, поднимаемые на 2-ом этапе для экспертов РТ:

1. [Статистический] Как мы оцениваем целостность данных? Как мы измеряем точность данных? Как оценить значимость данных?
2. [Технический] Как выявить дубликаты данных? Как определить отсутствующие значения?
3. [Эксперт в области РТ] Как определить синонимы домена РТ, которые помогают повысить точность данных? Как определить критерии, характеризующие отсутствующие значения и дубликаты?

Преобразование. Этот этап включает в себя сокращение и проецирование данных, целью которых является определение единой модели для представления данных, в зависимости от цели задачи. Кроме того, он может включать использование методов сокращения размерности или преобразования для уменьшения эффективного числа переменных или поиска инвариантных представлений данных. Как и шаг 2, шаг преобразования уменьшает сложность набора данных, обращаясь к вариации величин. Обычно это выполняется с помощью методов ИПЗ, которые поддерживают этапы предварительной обработки и преобразования данных в процессе KDD. Грубо говоря, с помощью ИПЗ данные, извлеченные из исходной системы, подвергаются серии процедур преобразования, которые анализируют, обрабатывают и затем очищают данные перед их загрузкой в базу знаний. К концу этого этапа, результатом которого является чистая, четко определенная модель данных, проблема вариации Больших Данных должна быть решена. Ключевые вопросы, поднятые на этапе трансформации для экспертов РТ:

1. [Статистический] Как измерить полноту выявленной целевой модели? Сохраняет ли

целевая модель значимость данных в конце процесса ИПЗ?

2. [Технический] Как разработать процедуры работы с Большими Данными для расширенного преобразования необработанных данных в целевую модель?
3. [Эксперт в области РТ] Как определить формат и классификацию данных для передачи[8]?

Способ поиска данных и компьютерное обучение. Целью данного этапа является определение подходящих алгоритмов ИИ (например, классификация, прогнозирование, регрессия, кластеризация, фильтрация информации), через поиск исследуемых закономерностей в определенной репрезентативной форме, исходя из цели анализа. Более конкретно, в контексте ИРТ, это обычно требует использования алгоритмов классификации текста (например, на основе онтологии или компьютерного обучения) для построения функций классификации для отображения элементов данных в один из нескольких, заранее определенных классов. Этот этап является решающим, поскольку он в основном посвящен извлечению знаний из данных. Ключевые вопросы, возникающие на этапе добычи данных и компьютерного обучения для экспертов РТ:

1. [Статистические и технические] Как выбрать лучший алгоритм? Как настроить их параметры? Как оценить эффективность алгоритма? Как реализовать его на должном уровне?
2. [Эксперт в области РТ] Какие знания следует отобрать, а какие отбросить? Какова значимость знаний, полученных для РТ? Какие новые знания были обнаружены с помощью ИРТ? Как мы можем объяснить результаты процесса поиска информации с точки зрения РТ?

Интерпретация/оценивание. На последнем этапе используются визуальные парадигмы для наглядного представления полученных знаний в зависимости от целей пользователя. В контексте ИРТ это означает учет способности пользователя понимать данные и его основной цели в области ИРТ. Например, государственные учреждения могут быть заинтересованы в определении наиболее востребованных профессий в своем регионе; компании могут сосредоточиться на мониторинге тенденций развития навыков и выявлении новых навыков для определенных профессий, чтобы они могли разработать программы обучения для своих сотрудников. В последние несколько лет много работы было сосредоточено на создании готовых визуальных библиотек, реализующих различные парадигмы повествования и визуального восприятия. Мощным примером является D3.js [18], основанный на данных соответствующей библиотеки для создания динамической, интерактивной визуализации данных даже в контексте Больших Данных (см., например, [19]). Ключевые вопросы, поднятые на этапе интерпретации/оценивания для экспертов РТ:

1. [Статистические и технические] Как выбрать парадигму визуализации? Как выбрать подходящую модель визуализации для знаний, которые мы хотим визуализировать?

2. [Эксперт в области РТ] Как мы предоставляем приемлемые знания в соответствии с потребностями заинтересованных сторон? Как мы определяем визуальные пути навигации для каждой заинтересованной стороны? Как мы получаем обратную связь (если таковая имеется) от пользователей РТ? Как внедрить знания о РТ в бизнес?

Как можно заметить, количество технических и статистических вопросов уменьшается по мере продвижения процесса KDD, в то время как количество вопросов и проблем, с которыми

[8] РТ характеризуется несколькими стандартными классификациями, такими как ISCO/O*NET/SOC для профессий, ESCO для навыков и NACE для классификации экономической деятельности.

сталкивается эксперт РТ, увеличивается. На самом деле, в то время как технические специалисты несут основную ответственность за работу с четырьмя "V" Больших Данных, именно эксперт РТ должен заниматься пятым "V" Больших Данных.

Чтобы прояснить этот момент, в таблице 1.2 мы указываем некоторые пункты РТ для возможного анализа, а также источники информации и основные преимущества/проблемы, связанные с анализом источников.

ТАБЛИЦА 1.2 ВАРИАНТЫ АНАЛИЗА РТ

Элемент РТ	Источники	Преимущества	Проблемы
Профессии и профессиональные навыки: ■ **Требования** ■ **Материальные средства**	■ Объявления о вакансиях ■ Курс обучения	■ Подробная информация ■ Фокус только на представляющую интерес информацию ■ Фокус скорее на терминах, используемых на рынке, чем на классификации ■ Определение схожих профессий на основе требований к квалификации	■ Сбор данных ■ Различение терминов/понятий и искажений ■ Понимание лексики и терминов, в зависимости от области применения ■ Использование стандартных систем классификации для снижения сложности и обеспечения возможности трансграничного сравнения ■ Многоязычное управление
Будущие навыки и новые развивающиеся профессии	■ Объявления о вакансиях ■ Курс обучения	■ Сосредоточенность на ожиданиях РТ ■ Выявление будущих/текущих тенденций и динамики ■ Анализ пробелов между кандидатами и новыми профессиями/навыками для разработки путей обучения ■ Определение схожих профессий на основе требований к квалификации	■ Обучение алгоритма для понимания того, когда термин является навыком ■ Формализация значения новой профессии (Сколько раз и как долго должна размещаться вакансия, чтобы стать новой профессией?)
Уровень цифровых навыков по профессиям	■ Объявления о вакансиях	■ Понимание повсеместного распространения цифровых навыков в профессиях ■ Планирование политики и путей обучения для заполнения разрыва с ожиданиями на РТ	■ Обучение алгоритма пониманию того, что такое цифровой навык ■ Уровень компьютерных навыков (включая мягкие/жесткие нецифровые навыки)
Трансверсальные (также известные как "мягкие") навыки (действительные и необходимые для многих профессий)	■ Объявления о вакансиях ■ Учебная программа	■ Понимание влияния сквозных навыков в рамках РТ ■ Разработка и планирование путей обучения для приобретения мягких навыков	■ Формализация значения сквозных навыков ■ Обучение алгоритмам распознавания сквозных навыков во всех их формах, которые могут значительно отличаться в естественном языке (например, решение проблем, способность решать проблемы, устранение неполадок).
Несоответствие квалификаций	■ Объявления о вакансиях ■ Курс обучения ■ Опрос	■ Способность выявлять переквалификацию/недостаточную квалификацию и устаревание навыков ■ Совершенствование политики РТ в соответствии с тенденциями и динамикой развития РТ ■ Поддержка в разработке образовательных и профессиональных направлений	■ Сбор данных ■ Очистка и интеграция данных ■ Разработка и выбор модели интегрированного анализа ■ Определение показателей несоответствия на различных уровнях детализации

Вопросы и ответы – Прогнозирующие возможности Больших Данных

Как Большие Данные могут способствовать прогнозированию навыков (в краткосрочной и среднесрочной перспективе)? Какие методологические аспекты необходимо рассмотреть, каковы должны быть объем и глубина анализа? Есть ли примеры того, как прогнозирование квалификации было улучшено благодаря использованию Больших Данных?

Большие данные имеют решающее значение для развития навыков по двум причинам. Во-первых, Большие данные, например, полученные из вакансий, размещенных в Интернете, являются единственным источником подробной информации о навыках в качестве альтернативы обследованиям навыков, которые содержат лишь ограниченный набор навыков, которые могут быть оценены. Во-вторых, спрос на навыки может различаться по профессиям, отраслям и регионам. Для того чтобы отследить эти изменения, необходимо иметь очень подробные и детализированные данные, которые можно получить только с помощью Больших Данных.

В методологическом плане существует несколько проблем, которые необходимо решить. Во-первых, данные должны быть привязаны к профессии/сектору/региону. Во-вторых, навыки должны быть классифицированы в значимую классификацию, которая может быть использована для анализа. В-третьих, для прогнозирования изменения навыков во времени необходим последовательный временной ряд.
Существует несколько примеров использования Больших Данных для прогнозирования навыков. В Соединенных Штатах (США) навыки, полученные из вакансий, размещенных в Интернете, используются университетами для прогнозирования тенденций в сфере РТ и адаптации образовательных программ к потребностям РТ.

Вопросы и ответы

Как успешно сотрудничать со статистическими и государственными органами, владеющими данными/реестрами? Как люди могут преодолеть свой скептицизм и опасения по поводу новизны Больших Данных? Есть ли хорошие примеры такого сотрудничества?

Согласно нашему опыту, использование Больших Данных не мешает ни актуальности, ни важности использования официальной статистики и опросов в анализе динамики и тенденций РТ для принятия решений. Для примера, одним из основных критических замечаний, связанных с использованием Больших данных, является то, что они полагаются на статистическую значимость, поскольку Большие данные должны использоваться совместно и сопоставляться с внешними статистическими источниками (например, обследованиями рабочей силы) с целью улучшения возможности наблюдения за динамикой РТ в более широком диапазоне.

Кроме того, следует учитывать, что интернет контингент изменчив и частично наблюдаем по конструкции, и это затрудняет определение стабильной выборки как репрезентативной для всего интернет контингента. Следовательно, можно использовать официальную статистику для оценки достоверности каждого класса собранных интернет-данных, чтобы можно было соответствующим образом взвесить перепредставленные (или недостаточно представленные) классы.

Стоит отметить, что нельзя пренебрегать использованием интернет-данных, поскольку в ближайшем будущем они будут расти. Таким образом, информативность, которую эти данные могут обеспечить при анализе, наблюдении и измерении какого-либо явления, может дать конкурентное преимущество для принятия оперативных и основанных на данных решений. Статья, в которой обсуждаются некоторые инициативы, связанные с использованием Больших Данных во многих реальных проектах, недавно опубликована в Бергамаски и др. (2016).

Вопросы и ответы – Качество данных

Мы знаем, что качество данных - это большая и важная часть науки о данных. Однако каковы основные аспекты и измерения, которые необходимо учитывать и принимать во внимание при реализации проекта Больших Данных? Каковы фундаментальные особенности качества данных в Больших Данных?

Как обсуждалось в Главе 1, оценка и очистка качества данных являются важнейшими и обязательными задачами для обеспечения достоверности любого процесса принятия решений на основе данных. Неудивительно, что критерий достоверности Больших Данных явно относится к вопросам качества, которые остаются значимыми даже при наличии огромного количества данных. Формального и надежного списка вопросов качества данных, которые наиболее актуальны для применения Больших Данных, пока не существует, и это все еще открытая дискуссия в научных кругах. Несомненно, некоторые исследователи полагают, что большого объема данных достаточно, чтобы свести на нет последствия (если таковые имеются) низкого качества данных. Другие считают, что строгий и формальный подход к качеству данных (оценивающий такие параметры, как целостность, полнота, точность и достоверность) должен применяться даже в случае огромных данных. Первый подход игнорирует тот факт, что некачественные данные распространены как в больших базах данных, так и в Интернете; второй подход не учитывает затраты (с точки зрения вычислительного времени и мощности) на оценку качества и очистку данных в огромных массивах данных. Тем не менее, мы считаем, что для целей анализа следует применять оба подхода.

Для примера, сосредоточившись на сборе вакансий из интернет-источников, следует использовать строгий и жесткий подход для ранжирования источников на основе характеристик или переменных источника, оценивая таким образом цельность, точность и полноту (как минимум) для переменных веб-источника. Это может охватывать, но не ограничиваться нижеследующим:

- типология: относится к типологии источника, которым могут быть кадровые агентства, национальные газеты, специализированные сайты, государственные, отраслевые сайты или сайты компаний;
- размер: относится к количеству вакансий, опубликованных на сайте на момент проведения анализа;
- время обновления: относится к частоте, с которой владелец веб-источника предоставляет свежие и обновленные данные, наряду с наличием временных меток, чтобы отметить, когда именно была опубликована вакансия;
- качество описания: определяет, насколько стандартизирована и полна страница с подробным описанием вакансий.

В отличие от этого, идентификация "активных" вакансий также относится к измерению качества данных (в данном случае, точности). Однако, если вакансия опубликована несколько недель или месяцев назад, в ней нет поля, гарантирующего, что вакансия все еще актуальна (т.е. работа еще не занята). В таком случае классический подход не сработает, поскольку доступ к этой информации не может быть автоматизирован и масштабируем (миллионы позиций, которые необходимо обрабатывать одновременно). И наоборот, включение вакансии в анализ независимо от ее обоснованности может иметь непредсказуемые последствия для достоверности анализа. По этим причинам следует построить статистическую модель для вывода даты действия такой вакансии на основе исторических данных, связанных с аналогичными вакансиями (например, с учетом аналогичных источников, аналогичных рекламируемых позиций, компании, которая размещает вакансии). Такая модель гарантирует, что сроки достоверности были оценены в соответствии с характеристиками набора данных.

Мы отсылаем читателя к статье Саха, Барна и Дивеш Шривастава, "Качество данных: Другое лицо больших данных", 2014 30-я Международная конференция IEEE по инженерии данных, IEEE (Институт инженеров электротехники и электроники), 2014; и Садик, Шазия и Паоло Папотти, "Качество больших данных: Чья это проблема?", 2016 IEEE 32-я Международная конференция по инженерии данных (ICDE - Международный совет по открытому и дистанционному образованию), IEEE, 2016 - две последние работы, в которых обсуждается и предлагается руководство по решению проблемы качества данных в приложениях больших данных с методологической точки зрения.

1.3 Литература о Больших Данных для ИРТ

За последние несколько лет некоторые силы и факторы резко изменили характер и характеристики РТ, как в развитых, так и в развивающихся странах. Технический прогресс, глобализация и реорганизация производственных процессов радикально изменили спрос на определенные навыки, усиливая потребность в непрерывном обучении, особенно в тех профессиях, которые являются высокоспециализированными или под значительным влиянием цифровизации. В то же время доступность данных растет благодаря цифровизации, распространению веб-сервисов (включая сервисы, связанные с РТ) и наличию огромного количества технических решений для работы с Большими Данными. В этом сценарии количество обращений к проектам, связанных с РТ, растет быстрыми темпами. Ниже мы приводим примеры некоторых инициатив (европейских и неевропейских), связанных с Большими Данными и информацией/аналитикой по РТ.

Панорама ЕС

Растет интерес к разработке и внедрению реальных приложений ИРТ для интернет-данных по РТ, с целью поддержки разработки и оценки политики через принятие решений на основе фактических данных. В 2010 году Европейская комиссия опубликовала документ *"Новый импульс для европейского сотрудничества в области профессионального образования и обучения"* в поддержку стратегии "Европа 2020" [20], направленной на развитие систем образования в целом и профессионального образования и обучения в частности. В 2016 году, Европейская комиссия подчеркнула важность профессионально-технических и образовательных мероприятий, поскольку они "ценны для развития профессиональных и универсальных навыков, облегчения перехода к трудоустройству, поддержания и обновления навыков рабочей силы в соответствии с отраслевыми, региональными и местными потребностями" [21]. Так же в 2016 году ЕС и Евростат запустили проект ESSnet Big Data [22], в котором участвуют 22 государства-члена ЕС, с целью "интеграции Больших данных в регулярное производство официальной статистики, посредством пилотных проектов, исследующих потенциал отдельных источников Больших данных и создания конкретных приложений". Более того, в 2014 году агентство "Европейский центр развития профессионального образования" (СЕДЕФОП), созданное для поддержки развития европейского профессионального образования и обучения, объявило тендер на проведение технико-экономического обоснования и разработку рабочего прототипа, способного собирать и классифицировать вакансии в интернете из пяти стран ЕС [23]. Смысл проекта заключается в том, чтобы преобразовать данные, извлеченные из вакансий в Интернете, в знания (и, следовательно, ценность) для разработки и оценки политики путем принятия решений на основе фактов. Учитывая успех прототипа, был объявлен дополнительный тендер на создание интернет-монитора РТ для всего ЕС, включая 28 стран-членов ЕС и 24 языка Союза [24].

Также стоит упомянуть проект LMI4All [25], онлайн-портал данных, который объединяет и

стандартизирует существующие источники высококачественных, надежных ИРТ с целью информирования о решениях, касающихся карьеры. Данные предоставляются в свободный доступ через ИПП для использования на веб-сайтах и в сторонних приложениях.

За пределами ЕС: неевропейские проекты

Критической проблемой РТ является возможность потери людьми работы из-за распространения автоматизации и искусственного интеллекта во всех промышленных секторах. Известное исследование [8] использовало машинное обучение на выборке профессий, аннотированных экспертами РТ, для оценки вероятности автоматизации для каждой профессии в США, используя стандартную классификацию профессий (SOC) в качестве системы классификации.

Эта работа заложила основу для дискуссии о риске потери рабочих мест в результате автоматизации. За Фреем и Осборном последовал ряд других исследований, например, работа [26], в которой изучались навыки для оценки риска автоматизации в 21 стране Организации экономического сотрудничества и развития (ОЭСР). Брукфилдский институт инноваций и предпринимательства (BII+E) при поддержке правительства Онтарио, Канада, изучил реальное влияние ИИ на рабочие места (риск потери рабочих мест из-за автоматизации и ИИ) в обрабатывающей промышленности и в сфере финансов/страхования, используя данные РТ, существующую литературу, интервью с более чем 50 заинтересованными сторонами из этих двух секторов и привлечение более 300 жителей Онтарио посредством интервью, общественных консультаций и онлайн-опроса [27]. В исследовании [28] для Всемирного экономического форума оценивалось влияние автоматизации и технологического прогресса на РТ, с использованием в качестве данных как профессии, так и рабочие места.

В то же время, за последние несколько лет в ЕС и за его пределами было разработано все больше коммерческих продуктов по подбору навыков, таких как продукты института Burning Glass, компаний Workday, Pluralsight, платформа EmployInsight, продукты компаний Textkernel и Janzz, для автоматизации деятельности отдела кадров компаний.

Стоит упомянуть Google Job Search ИПП, платную услугу, анонсированную в 2016 году, которая классифицирует вакансии с помощью службы машинного обучения Google в O*NET, стандартной классификации профессий США.

Матрица проектов/функций - Сравнительная модель для уточнения того, какие проекты решают конкретную проблему/задачу, связанную с Большими Данными для ИРТ

ТАБЛИЦА 1.3 обобщает ряд информационных и интеллектуальных проектов РТ и их характристики. В нем также приводится справочная информация по каждому проекту.

TABLE 1.3 МАТРИЦА ПРОЕКТОВ/ФУНКЦИЙ ИРТ

Ссылка	Источники данных	Цель	Вовлеченные страны	Тип	Языки обработки
[23]	Административные и фокус-группы	Оценка вероятности автоматизации	США	Проект	Англ.

[24]	Веб-данные (объявления о вакансиях)	Оценка эффективности ИРТ в интернете, для мониторинга в режиме реального времени	Италия, Германия, Великобритания, Чешская Республика, Греция	Проект и научное исследование	Англ., немецкий, итальянский,чешский, греческий.
[25]	Веб-данные (объявления о вакансиях)	Разработка блока управления РТ в режиме реального времени для ЕС	28 стран ЕС	Система и научное исследование	32 языка
[23]	Данные опросов и интернет-данных (объявлений о вакансиях)	Введение Больших Данных в официальную статистику	ЕС	Проект и научное исследование	Н/П
[27]	Административный (PIAAC[9])	Оценка рисков автоматизации в связи с роботизацией и ИИ	Оценка по США -> применяются к 21 стране ОЭСР	Исследование	Англ.
[29]	Веб-данные (объявления о вакансиях)	Оценка рисков автоматизации в связи с роботизацией и ИИ	Италия	Исследование	Англ.
[30]	Административный (государственный)	Использование наборов государственных данных для достижения более полного понимания потоков РТ	Великобритания	Проект	Англ.

1.4 Большие данные для ИРТ в действии

В этом разделе рассматриваются некоторые рабочие проекты, использующие описанный выше подход KDD для практического применения ИРТ в различных целях. С одной стороны, мы представляем подход WollyBI, который был успешно внедрен в Италии и в Испании (проект Bizkaia).

С другой стороны, мы представляем проект ESSnet Big Data - инициативу ЕС по включению Больших Данных РТ в официальную статистику РТ. Эти два приложения проливают свет на важность использования Больших Данных для анализа динамики и тенденций развития РТ для широкого круга заинтересованных сторон.

Другие проекты и инициативы будут более подробно рассмотрены в Главе 3.

Монитор РТ в режиме реального времени: на примерах Италии и Испании

Проект WollyBI[10] стартовал в начале 2013 года в качестве программного обеспечения как услуги (SaaS)[11] для сбора и классификации вакансий, размещенных в Интернете, по Международной стандартной классификации занятий/Европейской классификации навыков, компетенций, квалификаций и занятий (МСКЗ/ESCO) и извлечения наиболее востребованных навыков из описаний вакансий. Система была разработана таким образом, чтобы

9 PIAAC – Программа международной оценки компетенций взрослых. Данные PIAAC - это уникальный источник данных, который содержит показатели микроуровня по социально-экономическим характеристикам, навыкам, информации, связанной с работой, рабочим задачам и компетенциям. Самое главное, что эти данные сопоставимы между странами-участницами программы. Следовательно, эти данные также позволяют ослабить предположение о том, что структуры задач одинаковы в разных странах.

10 www.wollybi.com, на базе TabulaeX, аккредитованного подразделения Университета Милано-Бикокка, Италия.

11 Будучи веб-сервисом, доступным в любое время любому человеку с действующей учетной записью, SaaS избавляет от необходимости загружать и устанавливать инструменты.

предоставить пользователю пять различных точек входа в зависимости от целей анализа, а именно:

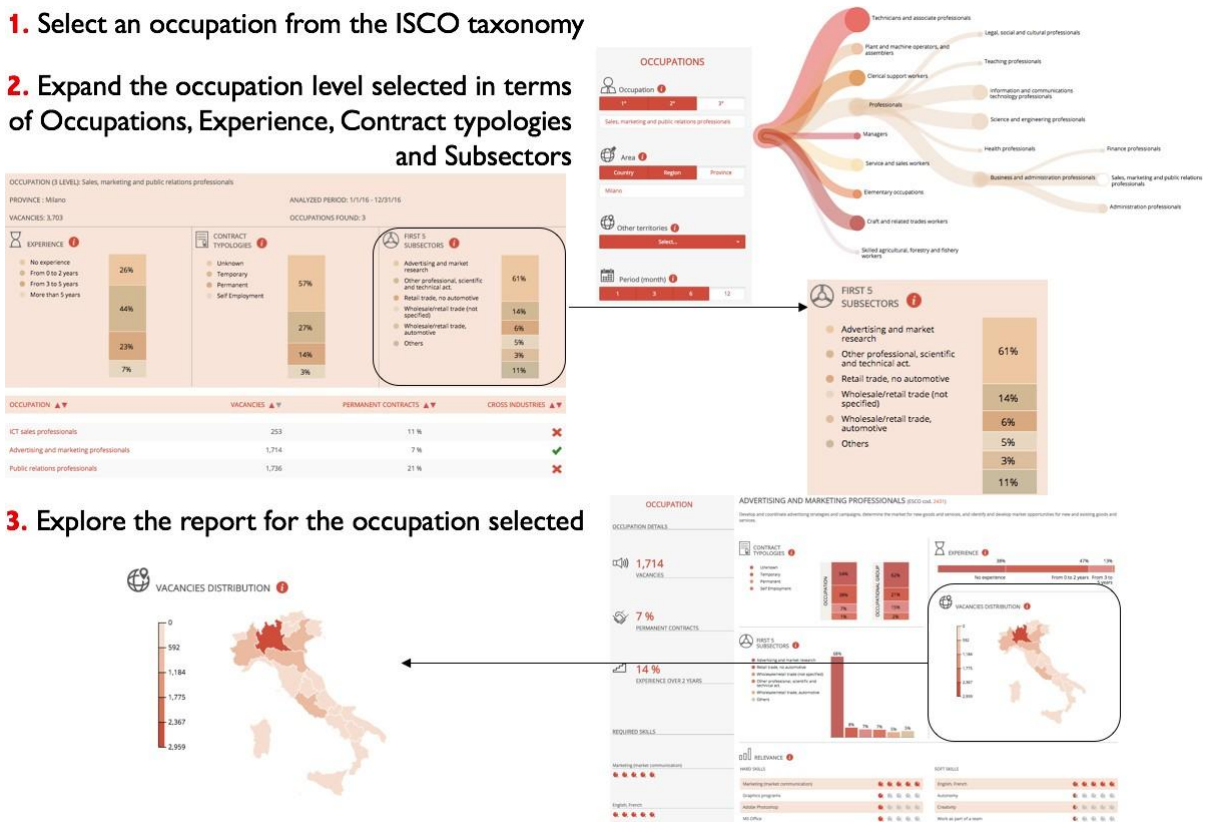
- Географический район - для поиска наиболее часто встречающихся в Интернете профессий и связанных с ними навыков на очень подробном географическом уровне;
- навыки - ввод набора навыков и поиск наиболее часто встречающихся профессий, включающих эти навыки (т.е. анализ пробелов в профиле);
- фирма - для получения рейтинга профессий, определяющих в вакансии конкретный сектор промышленности;
- профессия - для навигации по классификациям МСКЗ/ESCO и использования деталей, связанных с каждой профессией;
- свободные запросы (т.е. настраиваемые) - для свободных классических, нисходящих и восходящих операций над OLAP-кубами[12].

Реализация WollyBI в точности повторяет подход KDD, как показано в [15]. Здесь мы сосредоточимся на навигационных путях. Они отличаются для каждой заинтересованной стороны, но каждая точка входа была разработана на основе правила трех кликов, чтобы результаты были доступны пользователю не более чем через три "следующих" клика.

На рисунке 1.6 показана последовательность снимков из WollyBI, начиная с точки входа в профессию. Во-первых, пользователь должен выбрать группу профессий от первого уровня МСКЗ до третьего, а также некоторые другие параметры, такие как географическая область и временной горизонт анализа. Чем больше размер маркёра списка, тем значимее критерии профессий в конкретной группе. Возвращается первый отчет, показывающий результаты для выбранной группы профессий. Как только пользователь выбирает профессию для глубокого изучения, ему возвращается ряд информации, включая название профессии, ее код МСКЗ, ее определение в соответствии с класификацией МСКЗ, требуемый опыт, типологию контрактов, пять лучших подсекторов и список пяти лучших навыков (как жестких, так и мягких). Демонстрационное видео WollyBI в действии на измерении оккупации доступно на сайте: https://youtu.be/zBNsAS5L04g.

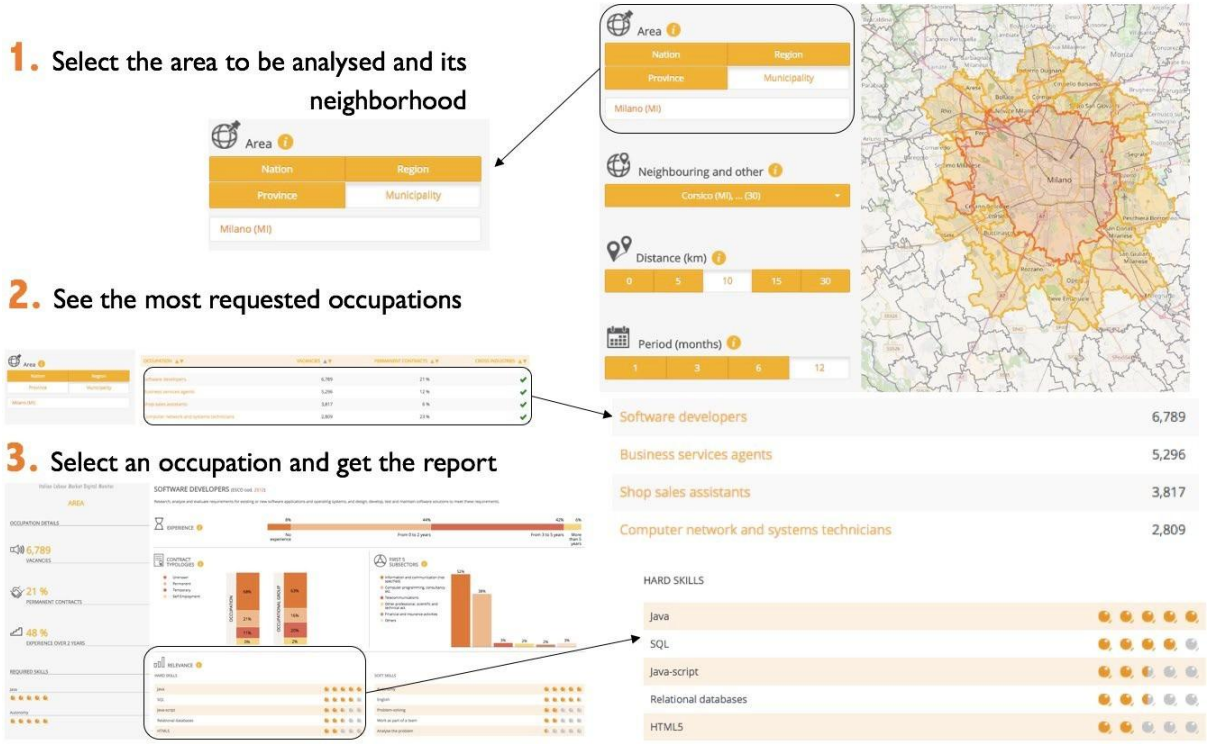
12 OLAP-куб - это многомерная база данных, оптимизированная для приложений хранилища данных и онлайновой аналитической обработки (OLAP).

РИСУНОК 1.6 WOLLYBI - ПОСЛЕДОВАТЕЛЬНОСТЬ СНИМКОВ ИЗ ТРЕХЭТАПНОГО АНАЛИЗА АСПЕКТА ПРОФЕССИИ



На рисунке 1.7 показана последовательность снимков из WollyBI по измерению географического района. Во-первых, пользователь должен выбрать интересующий его географический район, временной горизонт и конкретный уровень МСКЗ (необязательно). Возвращается список профессий четвертого уровня МСКЗ, а также количество вакансий, классифицированных по данному коду профессии, процент от общего количества и галочка, указывающая на наличие межсекторных профессий. Как только пользователь выбирает профессию, ему возвращается ряд информации, включая название профессии, ее код по МСКЗ, ее определение в соответствии с классификацией МСКЗ, требуемый опыт, типологию контрактов, пять лучших подсекторов и список пяти лучших навыков (как жестких, так и мягких). Сюда могут входить навыки, не относящиеся к ESCO навыкам, как указано в предыдущем разделе. Демонстрационное видео WollyBI в действии по аспекту географического района доступно на сайте: https://youtu.be/Xe_OSOHkx20.

РИСУНОК 1.7 WOLLYBI - ПОСЛЕДОВАТЕЛЬНОСТЬ СНИМКОВ ИЗ ТРЕХЭТАПНОГО АНАЛИЗА АСПЕКТА ГЕОГРАФИЧЕСКОГО РАЙОНА



Проект Bizkaia Talent

WollyBI был использован в качестве основы - с позиции как технологических, так и методологических аспектов - для развертывания монитора РТ в Стране Басков, Испания. Баскская обсерватория талантов - первая в мире платформа общественного доступа для высококвалифицированного мониторинга РТ региона, которая была выпущена в 2017 г. Инициатива была разработана Bizkaia Talent совместно с TabulaeX, ответвлением Университета Милано-Бикокка, с целью управления передачей знаний. Она основана на инструменте, который исследует баскский РТ с упором на высококвалифицированных специалистов через анализ Больших Данных из многочисленных правильно отобранных и ранжированных онлайн-источников, на уровне баскского региона.

Цель проекта, поддержанного Министерством экономики и территорий Бискайи, - способствовать повышению конкурентоспособности провинции Бискайи и обеспечить сбор информации о РТ в географической зоне Бискайи, Гипускоа и Алавы в режиме реального времени, используя как международные, так и местные онлайн-источники, такие как университеты или государственное агентство занятости Lanbide. Благодаря анализу Больших Данных, инструмент создает базу знаний о РТ, динамике занятости в любой момент времени или тенденциях во времени, а также о технических и профессиональных требованиях в Стране Басков в отношении высококвалифицированных профилей. Используя ежедневно обновляемые онлайн-данные, инструмент позволяет высококвалифицированным специалистам отслеживать типы профилей, востребованных на РТ Страны Басков, в отношении многочисленных и различных типов и комбинаций критериев, таких как требуемые технические и сквозные навыки, сектор, опыт, географический регион и тип контракта.

База знаний РТ была организована в двух различных точках входа, а именно:

1. для граждан: информационная панель для анализа и просмотра информации в интернете на основе вакансий на РТ, активных в последние 12 месяцев;
2. для аналитиков: информационная панель, визуализирующая данные за последние 12 месяцев, с возможностью выбора интересующего периода: последний месяц, последние три месяца, последние шесть месяцев, последний год. Данные собираются ежедневно и обновляются ежемесячно.

Инструмент находится в открытом доступе и может быть просмотрен любым желающим по следующему веб-адресу: https://basquetalentobservatory.bizkaiatalent.eus/visual/public/index#.

Более подробную информацию об этом проекте можно найти в Главе 3.

Проект Больших Данных ESSnet

С акцентом на статистическую перспективу анализа Больших Данных для ИРТ, Евростат - официальный статистический офис ЕС - запустил проект под названием ESSnet Big Data, направленный на интеграцию Больших Данных об информации РТ в регулярное производство официальной статистики, используя пилотные проекты для изучения потенциала отдельных источников Больших Данных и создания конкретных приложений [23].

Проект, стартовавший в конце 2016 года, состоит из 22 партнеров из ЕС и в точности повторяет подход KDD для сбора данных с ранее ранжированных веб-порталов, очистки и преобразования данных для классификации в соответствии со стандартными классификациями. Здесь стоит отметить важную роль, которую играет репрезентативность Больших Данных, поскольку участники проекта намерены оценить способность Больших Данных быть репрезентативными для всего населения (или стратифицированной выборки) для включения в официальную статистику. Насколько нам известно, это первая государственная инициатива, направленная на включение Больших Данных РТ (например, вакансий) в официальную статистику, поскольку это прольет свет на основную информацию, раскрываемую веб-источниками, которую необходимо правильно извлечь и обработать для получения знаний о РТ, полезных для лиц, принимающих решения, и специалистов РТ.

Вопросы и ответы

Кто является пользователями этих платформ?

Пользователей необходимо правильно определить в начале проектирования системы, чтобы можно было организовать знания в соответствии с потребностями пользователей и их способностью понимать данные. Например, WollyBI использует одни и те же знания для разных типов заинтересованных сторон: агентства занятости, бизнес-ассоциации и профсоюзы, государственные агентства занятости, школы и учебные организации. Bizkaia Talent предназначен как для граждан, так и для аналитиков, в то время как проект ESSnet предназначен для специалистов РТ и аналитиков в целом.

Какова частота обновления этих инструментов РТ?

Решение о частоте обновления должно приниматься с учетом трех элементов: (i) обновление источников, (ii) стоимость с точки зрения вычислительной мощности, необходимой для выполнения обновления, и (iii) потребности заинтересованных сторон. Обычно для целей анализа подходят еженедельные обновления.

В какой степени автоматизирован процесс KDD, развернутый этими платформами, (AI 100%, AI 70%, AI 40% ...)?

Человеческие усилия уменьшаются по мере продвижения процесса KDD. Значительные усилия требуются для идентификации и ранжирования источников, а также для определения потребностей бизнеса и выбора хороших алгоритмов ИИ и соответствующих параметров настройки. После завершения этих действий система работает автономно, с периодическим техническим обслуживанием. Например, очистка данных требует более тщательного обслуживания, поскольку веб-сайт может меняться непредсказуемым образом, в то время как наличие соглашения с владельцем данных предотвращает эту проблему и упрощает скрейпинг. Кроме того, использование облачных вычислений для реализации решения на основе Больших Данных значительно снижает риски и затраты, связанные с поломками, но может оказаться дорогостоящим, если рабочая нагрузка (например, пользователи или обработка в единицу времени) значительно возрастет. В этом отношении реализация ИСРТ Больших Данных-это подход, основанный на участии человека, особенно на ранних стадиях. Опыт-это важнейший навык, который позволяет сократить как затраты, так и человеческие усилия

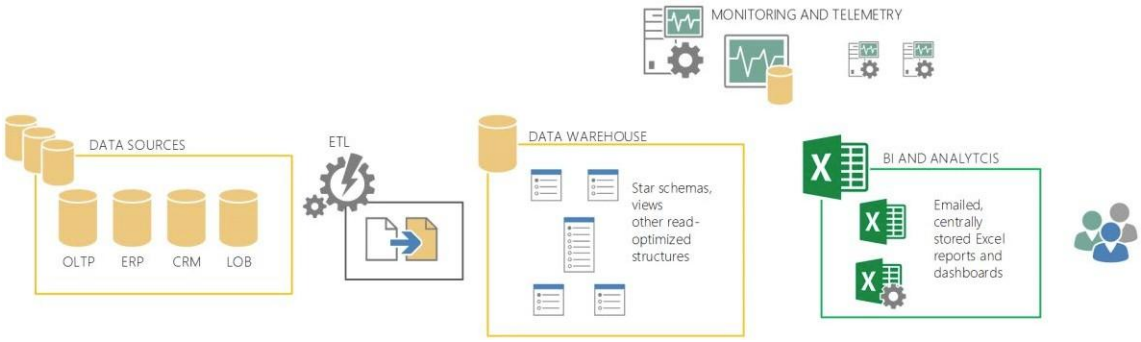
2. ВСТРАИВАНИЕ АНАЛИТИКИ БОЛЬШИХ ДАННЫХ В ИРТ: ПОСЛЕДОВАТЕЛЬНЫЕ ШАГИ

В этой главе мы представим основные строительные блоки, необходимые для разработки архитектуры Больших Данных для ИРТ, и рассмотрим некоторые важные необходимые (предварительные) условия: с чего начать, что рассмотреть, этапы и основные шаги, используя в качестве руководства подход KDD, рассмотренный ранее.

2.1 Компоненты архитектуры Больших Данных

Подход KDD (Рисунок 1.5) представляет собой базовую основу для реализации любой модели управления данными, предназначенной для извлечения знаний из данных, еще до наступления эры Больших Данных, как показано на Рисунке 2.1. Каждый шаг KDD, рассмотренный в главе 1, выполняется для (i) сбора данных из нескольких структурированных источников данных; и (ii) преобразования данных из нескольких и различных моделей и форматов данных в единую унифицированную модель данных с использованием методов и инструментов ИПЗ. Затем данные обычно хранятся в хранилище данных, пригодном для оптимизированных запросов и анализа. Наконец, информационная панель позволяет аналитикам выполнять запросы и использовать полученные знания в виде отчетов и диаграмм для поддержки принятия решений. Это классическая модель Business Intelligence (БА), которая хорошо подходит для структурированных данных.

РИСУНОК 2.1 КЛАССИЧЕСКАЯ АРХИТЕКТУРА БА ДО ПОЯВЛЕНИЯ БОЛЬШИХ ДАННЫХ



Хотя эти типы архитектуры весьма успешны во многих реальных контекстах, они страдают от ограничений, которые препятствуют их использованию в сценариях Больших Данных, как показано в таблице 2.1.

ТАБЛИЦА 2.1 НАИБОЛЕЕ СУЩЕСТВЕННЫЕ ОГРАНИЧЕНИЯ АРХИТЕКТУРЫ БОЛЬШИХ ДАННЫХ

Проблема (наиболее значимая)	Фактор	Концептуальные блоки модели Больших Данных
Данные без схем исключены: манипулировать можно только структурированными источниками данных. Грубо говоря, это означает, что работать можно только с данными, которые подчиняются жесткой, четко определенной модели данных, исключая все "неструктурированные" данные, такие как свободный текст, комментарии и веб-контент в целом.	Изменчивость	Ввод данных; модели NoSQL

Отсутствие адаптивности к изменениям: добавление нового источника требует изменения всего процесса, что затрудняет масштабирование архитектуры на несколько (хотя и структурированных) источников.	Изменчивость, скорость	Озеро данных
Жесткий ИПЗ: процедуры, преобразующие содержимое из исходных форматов в целевые, должны быть точно написаны, чтобы соответствовать желаемой структуре данных (например, хранилищу данных).	Изменчивость	Без схем; подход, основанный на данных ("восходящий", а не "нисходящий")
Времязатратный: чем больше объем данных, подлежащих обработке, тем больше времени требуется для завершения процесса. Процедуры ИПЗ обычно потребляют много времени и памяти, поскольку им необходимо "сканировать" все источники данных в любой момент времени для преобразования исходных данных.	Объем, изменчивость, скорость	Расширение, а не наращивание масштабов

В таблице 2.1 приведены некоторые ключевые проблемы, препятствующие использованию классической архитектуры БА в сценарии Больших Данных, в основном относящиеся к одному или нескольким измерениям Больших Данных ("V" модели Больших Данных, рассмотренные в Главе 1). В последние годы, и научными кругами, и практиками были предприняты большие усилия по определению парадигм, моделей и инструментов, подходящих для сценария Больших Данных.

Здесь мы представим некоторые концептуальные блоки, которые читатель должен знать для лучшего понимания того, как работает модель Больших Данных: ввод данных, модели NoSQL, озеро данных и масштабирование. Затем мы обсудим, как эти блоки могут работать вместе, чтобы реализовать модели Больших Данных для ИРТ. Очевидно, что существует множество решений и инструментов, поскольку платформа Больших Данных продолжает расширяться и развиваться.

Поглощение данных. Этот термин относится к процессу сбора данных из нескольких источников автоматизированным способом. Он может осуществляться в режиме реального времени (каждый элемент данных собирается по мере их поступления из источника) или в рамках пакетного процесса (элементы данных собираются дискретными фрагментами через регулярные промежутки времени). Более того, данные из одного источника могут быть собраны с помощью трех различных подходов: ИПП, краулинг или скрейпинг.

- ИПП – как отмечалось в Главе 1 - это интерфейс прикладного программирования, компонент программного обеспечения, предоставляемый владельцем источника любому программисту для обеспечения сбора данных (например, Twitter, Facebook). Процесс сбора контролируется владельцем данных, который также решает, какие данные могут быть выпущены, а также структуру данных.
- Crawling (краулинг) - это программное обеспечение, которое автоматически индексирует все содержимое веб-страницы и добавляет его в базу данных. Он итеративно следует по всем гиперссылкам, включенным в страницу, а также индексирует эти данные в базу данных (включая изображения, таблицы и стилевые таблицы). Классическим примером краулинга является поисковая деятельность, осуществляемая Google.
- Scraping (скрейпинг) - это двухэтапная деятельность программного обеспечения. Сначала он автоматически запрашивает веб-страницу, затем собирает только ограниченное количество информации со страницы, пропуская остальные данные. Это означает, что скрейпер (частично) знает структуру сайта, поэтому он может определить только тот контент, который представляет интерес для анализа. Например, веб-краулер может загрузить все продукты, перечисленные на сайте электронной коммерции, в то время как скрейпер может собрать только названия продуктов и цены, оставив без внимания ссылки на баннеры, комментарии и метаданные, связанные с оформлением страницы.

Модели NoSQL. В последние годы движение NoSQL выдвинуло на первый план новые парадигмы моделей данных, которые существенно отличаются от классической реляционной модели, лежащей в основе любой архитектуры БА . В конечном итоге возникли четыре парадигмы хранения данных NoSQL (т.е. ключевые значения, базы данных документов, базы данных, ориентированные на столбцы, и графические базы данных). Все эти новые парадигмы имеют ряд интересных особенностей по сравнению с классической реляционной моделью (см., пример, [31]), таких как гибкая схема, которая всегда может меняться в соответствии с данными, возможность горизонтального масштабирования и встроенная поддержка совместного доступа. По этим причинам эти парадигмы стали общей основой любой архитектуры Больших Данных, поскольку они позволяют хранить данные в их естественной форме. Интуитивно реляционную базу данных можно рассматривать как набор таблиц, по которым можно перемещаться с помощью идентификаторов (они же ID). Очевидно, что количество столбцов в данной таблице фиксировано и определено априори. В хранилищах данных NoSQL, напротив, количество столбцов может варьироваться для каждой строки каждой таблицы, что позволяет хранить любые типы данных, независимо от их структуры, поскольку схема может меняться вместе с данными.

Озеро данных. Хранение неструктурированных данных, таких как свободный текст или веб-контент в целом, не позволяет разработать единую схему модели. Точнее говоря, схема неструктурированных данных может свободно изменяться и развиваться с течением времени. Например, представьте себе задачу обработки, в которой миллион объявлений о работе, написанных в виде свободного текста, должны быть упорядочены по столбцам общей электронной таблицы. Задача здесь состоит в том, чтобы определить "модель", которая соответствует всей информации, которую можно извлечь из вакансий.
Очевидно, что схема может меняться по мере того, как меняется схема вакансии. Одна вакансия может содержать несколько контактных адресов, или мест расположения, или навыков, по отношению к другой, и это затрудняет определение "общей схемы" (или модели данных), подходящей для любой вакансии (включая вакансии, которые еще предстоит собрать). Решение этой проблемы, заключается во внедрении "озера данных". Проще говоря, озеро данных - это хранилище данных, которое оставляет все собранные данные в их родном формате, присваивая каждому элементу данных уникальный идентификатор. Это позволяет извлечь элемент, когда он необходим для выполнения определенного действия (например, проанализировать вакансии из источника X или из страны Y).

Масштабирование. В общих чертах, масштабируемость означает, что модель способна продолжать работать, несмотря на рабочую нагрузку. Очевидно, что ни одна модель не может масштабироваться бесконечно, поэтому смысл выражения "несмотря на нагрузку" следует воспринимать как подразумевающий некоторые конкретные требования к масштабируемости, которые считаются важными. Например, архитектура Больших Данных, которая собирает и обрабатывает веб-документы, должна масштабироваться, чтобы гарантировать низкую задержку и высокую пропускную способность. Латентность - это время, необходимое для выполнения действия (или получения результата), измеряемое в единицах времени (минутах для обработки в реальном времени). Пропускная способность - это количество выполненных действий или полученных результатов за единицу времени. Латентность гарантирует, что пользователи не будут бесконечно ждать результатов, в то время как пропускная способность указывает на способность архитектуры обрабатывать данные. Благодаря внедрению хранилищ данных NoSQL, распределенных файловых систем[13] и распределенных вычислений,[14] архитектура Больших Данных может масштабироваться (обеспечивать вертикальную масштабируемость).

13 Распределенные файловые системы можно рассматривать как дисковую структуру, расположенную на нескольких компьютерах.
14 Распределенные вычисления означают способность компьютера распределять свою рабочую нагрузку между несколькими компьютерами, тем самым снижая общую рабочую нагрузку.

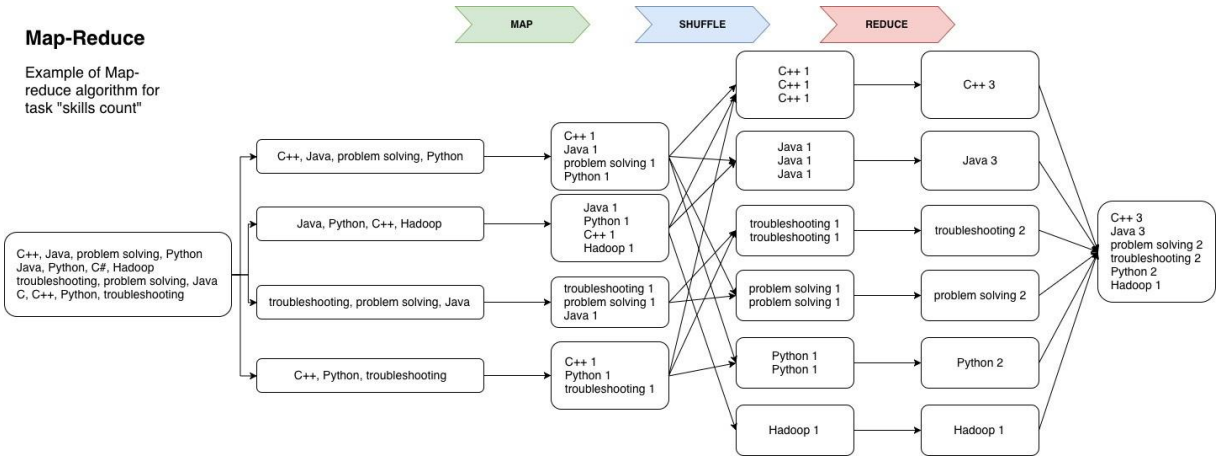
Самый простой способ понять, как система может действительно масштабироваться, уменьшая задержки и увеличивая пропускную способность, - это понять принцип работы MapReduce, основного двигателя любой архитектуры Больших Данных (см. вставку ниже).

MapReduce – демонстрационный пример

Представьте себе миллионы документов, поступающих из различных веб-источников в режиме реального времени. Ваша модель собрала все эти документы в их естественном виде, благодаря использованию хранилищ данных NoSQL. Цель состоит в том, чтобы иметь возможность обрабатывать все эти документы, создавая таблицу подсчета слов, которая суммирует все слова, появляющиеся во всех документах, вместе с их встречаемостью, как показано на рисунке 2.2. Принцип работы MapReduce (введенный [32] во время работы в Google) распределяет рабочую нагрузку на несколько компьютеров таким образом, что каждый компьютер должен выполнить очень простую задачу с помощью функции 'map'. Сначала входной документ разбивается на части (построчно) так, чтобы каждая строка попала к разным картографам (т.е. компьютерам). Каждый компьютер выполняет очень простую функцию, или "карту", которая выдает пару, содержащую слово (ключ) и счетчик (значение), начиная с 1. На следующем этапе каждая пара (или "кортеж") перемешивается (по ключу) в логическом порядке (в данном случае в алфавитном порядке). Понятно, что цель перестановки состоит в том, чтобы все кортежи с одинаковым ключом были распределены одному редуктору, что позволяет сэкономить время. Наконец, каждый редуктор получает для каждого ключа все значения, найденные в кортежах. Теперь задача этого редуктора относительно проста: сложить все значения для своего ключа и выдать через функцию 'reduce' результирующий кортеж, состоящий из ключа и суммарной величины, которая представляет собой количество вхождений этого ключа.

Как видно на рисунке 2.2, благодаря парадигме MapReduce, для выполнения задания задействованы четыре компьютера, что сокращает общее время (латентность) и увеличивает пропускную способность модели при обработке миллионов документов.

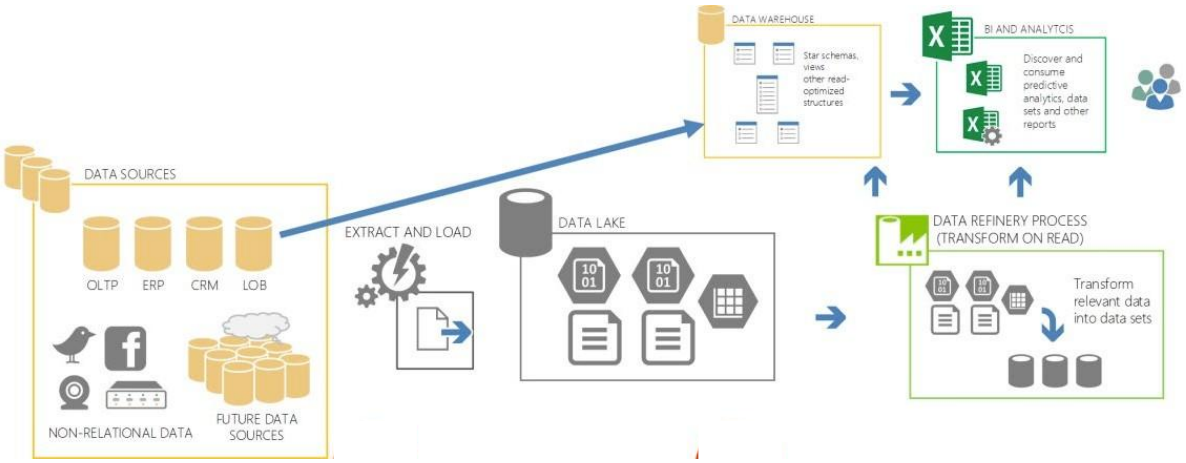
РИСУНОК 2.2 ПРИМЕР АЛГОРИТМА MAPREDUCE, ВЫПОЛНЯЮЩЕГО ЗАДАЧУ "ПОДСЧЕТ НАВЫКОВ"



Благодаря этому концептуальному блоку, введенному Большими Данными, классический подход БА ,

показанный на рисунке 2.1, может перейти к подходу Большие Данные, как показано на рисунке 2.3. Примечательно, что классический рабочий процесс БА , показанный на рисунке 2.1, все еще существует, поскольку реляционные и структурированные данные все еще присутствуют во многих реальных областях, включая веб-данные. Однако рабочий процесс обработки данных теперь идет по двум различным путям: классический подход работает со структурированными данными, а подход Больших Данных, как обсуждалось выше, используется для работы с неструктурированными данными.

РИСУНОК 2.3 ОТ БА К МОДЕЛЯМ БОЛЬШИХ ДАННЫХ



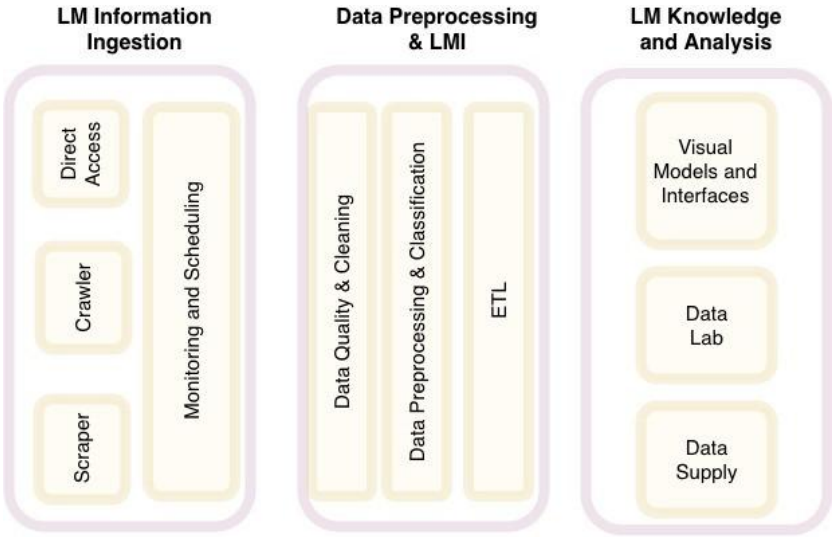
2.2 Современные модели, технологии и инструменты

В этом разделе мы описываем основные элементы классической модели Больших Данных, представляя современные технологии и инструменты. В последние несколько лет платформа Больших Данных стремительно развивается[15]; в связи с этим ниже мы описываем основные строительные блоки, которые могут быть использованы при разработке платформы Больших Данных ИРТ.

Архитектура Больших Данных (восприятие)

С концептуальной точки зрения архитектура Больших Данных для обработки информации РТ должна выглядеть так, как показано на рисунке 2.4. Целью такой модели является сбор информации о РТ в Интернете для извлечения полезных знаний о динамике и тенденциях РТ. Эта информация РТ может быть объявлениями о работе, учебными программами, данными опросов и т.д. Независимо от характера информации РТ, любая модель Больших Данных должна состоять (по крайней мере) из трех макрошагов.

РИСУНОК 2.4 КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ БОЛЬШИХ ДАННЫХ ДЛЯ ИРТ В РЕАЛЬНОМ ВРЕМЕНИ

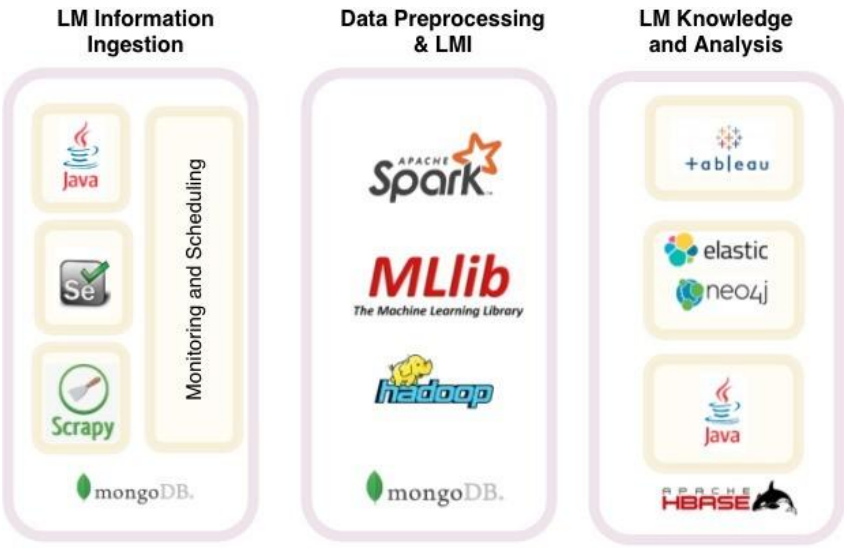


Первый шаг - сбор данных, а второй включает в себя две задачи KDD, а именно предварительную обработку данных и часть деятельности по добыче данных. Последний этап включает в себя три основных модуля, позволяющих использовать полученные знания о РТ.

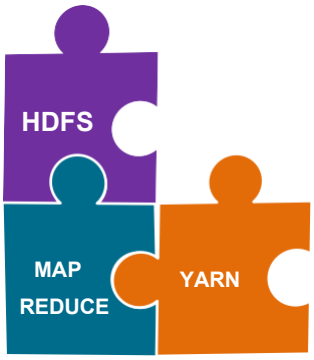
- Визуальные модели и интерфейс позволяет конечным пользователям просматривать базу знаний в интерактивном режиме, используя парадигмы и предопределенные взаимодействия.
- Лаборатория данных - это среда, которую исследователи могут использовать для свободного рассуждения над данными, используя алгоритмы искусственного интеллекта и компьютерного обучения для извлечения из них дополнительных знаний.
- Наконец, модуль обеспечения данных предоставляет полученные знания о РТ третьим лицам. Он действует примерно как служба доставки знаний о РТ.

15 Более подробная информация приведена на сайте https://hadoopecosystemtable.github.io/ (по сост.на март 2019).

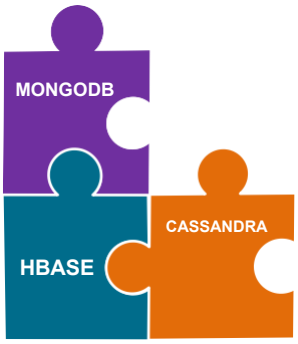
РИСУНОК 2.5 ПРИМЕР МОДЕЛИ БОЛЬШИХ ДАННЫХ ДЛЯ ИРТ В РЕАЛЬНОМ ВРЕМЕНИ - ПЛАТФОРМА HADOOP



Эта модель в точности повторяет подход KDD, рассмотренный в Главе 1. Здесь мы показываем, как такая концептуальная модель может быть развернута с использованием технологий и инструментов Больших Данных. Мы используем модель, показанную на рисунке 2.5, в качестве примера, чтобы представить роль каждого из четырех строительных блоков Больших Данных: Поглощение информации о РТ, MongoDB, предварительная обработка данных и ИРТ, а также знания и анализ РТ.



Поглощение информации РТ. Этот модуль Больших Данных должен быть реализован с учетом способов сбора ИРТ из веб-источников, как обсуждалось выше. Существует (по крайней мере) три различных категории источников интернет ИРТ: (A) агентства занятости и государственные службы занятости; (B) газеты, компании и университетские вакансии; и (C) порталы вакансий. Здесь этап сбора данных может отличаться для каждой категории. Например, (A) может предусматривать прямой доступ, предоставляя ИПП возможность непосредственного сбора данных. В этом случае необходимо определить предпочтительный язык программирования (например, Java, Python, Scala) для подключения через ИПП и получения данных. Как вариант, (B) может не иметь никаких ИПП для использования. Эти источники часто меняют структуру своих веб-страниц, что делает реализацию веб-скрейпера дорогостоящей и трудоемкой. В этом случае для сканирования портала на предмет ИРТ следует использовать краулер. Возможным решением может стать Selenium, автоматизатор веб-браузера, который имитирует просмотр веб-страниц и осуществляет автоматическую навигацию по интернет-контенту. Наконец, данные из (C) можно собрать с помощью скрейпера; структура порталов вакансий меняется нечасто, поэтому скрейпер можно использовать для определения данных, собранных из веб-источников. Существует множество возможных инструментов, таких как Scrapy, основанная на Python система с открытым исходным кодом и совместной работой для извлечения данных с веб-сайтов.



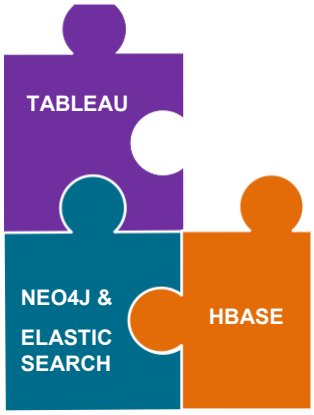
MongoDB. Это документо-ориентированная система баз данных, принадлежащая к семейству систем баз данных NoSQL. По сути, MongoDB хранит структурированные данные в виде JSON[16] -подобных документов, а не в виде таблиц, используемых в классическом реляционном подходе. В результате это позволяет хранить данные в их родном формате, что является преимуществом при работе с неструктурированными данными, схема которых может меняться со временем. Можно использовать и другие хранилища данных NoSQL, такие как HBase и Cassandra, столбцевые распределенные базы данных, способные выполнять массовые операции чтения и записи в реальном времени в очень больших столбцевых таблицах.

Предварительная обработка данных и ИРТ. Этот модуль отвечает за предварительную обработку информации РТ для получения знаний. Важной вехой в обработке Больших Данных является Hadoop (2006), система распределенной обработки с открытым исходным кодом, способная управлять обработкой и хранением данных для приложений Больших Данных, работающих в кластерных системах. Ядро Apache Hadoop состоит из трех основных блоков: (i) Hadoop Distributed File System (HDFS), (ii) MapReduce для распределенных вычислений и (iii) Yet Another Resource Negotiator (YARN) для планирования заданий и управления ресурсами. Вместе эти три элемента реализуют распределенную модель. Если в качестве рабочего примера взять классификацию вакансий в Интернете, то ядро Hadoop позволит:

- хранение миллионов вакансий на нескольких узлах благодаря файловой системе HDFS. HDFS обладает высокой масштабируемостью, надежностью и может использовать любой компьютер как товарное оборудование. Следовательно, информация РТ может храниться на нескольких компьютерах;
- Вычисления можно распределить по нескольким компьютерам благодаря алгоритму MapReduce, описанному ранее. В случае с ИРТ MapReduce можно рассматривать как функцию для поиска текста или для ранжирования профессий после их классификации по стандартной таксономии. MapReduce очень хорошо работает с огромными наборами данных, такими как информация РТ;
- планирование рабочих процессов и управление ресурсами благодаря YARN. YARN отвечает за распределение системных ресурсов между различными приложениями, работающими в кластере Hadoop, и планирование заданий для выполнения на различных узлах кластера. В качестве альтернативы Hadoop, можно использовать проект Apache Spark[17] ориентирован на параллельную обработку данных на кластере, но самое большое отличие заключается в том, что он работает в памяти (RAM). Если Hadoop читает и записывает файлы в HDFS, то Spark обрабатывает данные в оперативной памяти.

[16] *JSON - JavaScript Object Notation - это легкий формат обмена данными*. Его легко читать и писать людям, а компьютерам- разбирать и генерировать.
[17] https://spark.apache.org/ (по сост.на март 2019)

Для этого в Spark была введена концепция, известная как устойчивое распределение Dataset, для обозначения наборов данных, которые хранятся в памяти. В нашем примере модели мы можем объединить Hadoop и Spark, поскольку Spark может работать в автономном режиме, а источником данных служит кластер Hadoop. Наличие обоих решений в нашей модели может оказаться полезным, позволяя выбрать подходящую структуру обработки в зависимости от объема собранных данных. Spark также предоставляет мощную библиотеку компьютерного обучения, включающую самые современные алгоритмы, хотя можно использовать и любую другую библиотеку компьютерного обучения (см., например, scikit-learn[18] или tensorflow[19]).



Знания и анализ РТ. После получения знаний о РТ в Интернете их следует доставить конечным пользователям в соответствии с потребностями заинтересованных сторон. По нашему опыту, знания РТ служат (по крайней мере) трем различным заинтересованным сторонам/целям.

1. Аналитики РТ, с помощью интерактивных информационных панелей, позволяющих анализировать динамику и тенденции развития интернет РТ по заранее заданной схеме. Одним из примеров является WollyBI, рассмотренный в Главе 1, который предлагает четыре различные точки взаимодействия со знаниями РТ в зависимости от потребностей заинтересованных сторон.
2. Исследователи РТ, путем предоставления лаборатории данных типа "песочницы", где исследователи могут применять и тестировать новые алгоритмы и фреймворки для опробования синтезированных знаний о РТ в интернете. Например, база данных графиков (Neo4j) может быть использована для организации данных в виде социальной сети. Если предположить, что узлами могут быть либо профессии, либо навыки, можно провести анализ социальной сети, чтобы найти кластеры похожих профессий или навыков, группы профессий, которые имеют общие навыки, или анализ пробелов, чтобы рекомендовать навыки, которые необходимо приобрести для перехода с одной должности на другую. Аналогичным образом, расширенный текстовый поиск может осуществляться исследователями, создающими поисковую систему с помощью ElasticSearch.
3. Заинтересованные лица, которые могут быть заинтересованы в использовании таких знаний в качестве услуги для реализации собственных продуктов или услуг. Это может быть агентство по трудоустройству, которое использует знания о РТ для рекомендации вакансий или для поддержки профессионального образования и обучения.

С технической точки зрения, для эффективного хранения знаний на узлах кластера можно использовать столбцевое хранилище данных, как в случае с Apache HBase, которое гарантирует линейную масштабируемость и операции чтения/записи в реальном времени в очень больших столбцевых таблицах. Оно является частью платформы Hadoop и может автоматически обрабатывать задания MapReduce для масштабирования на нескольких кластерах.

[18] http://scikit-learn.org/stable/ (по сост.на март 2019)
[19] www.tensorflow.org (по сост.на март 2019)

2.3 Роль ИИ для ИРТ: алгоритмы и структуры для обоснований исходя из ИРТ

2.4 ИИ - это термин, относящийся к моделируемому интеллекту в машинах. Хотя определение ИИ со временем менялось, недавно Европейская комиссия дала хорошее объяснение, определив ИИ как "системы, которые демонстрируют разумное поведение, анализируя свое окружение и предпринимая действия - с определенной степенью автономности - для достижения конкретных целей"[20]. Это определение также применимо к ИРТ, поскольку алгоритмы ИИ (например, классификация, прогнозирование, регрессия и кластеризация) могут быть использованы для поиска интересующих закономерностей в определенной форме представления, в зависимости от цели анализа. Более конкретно, поскольку информация о РТ часто характеризуется текстом, алгоритмы ИИ, работающие с текстами для извлечения знаний, весьма полезны для этой цели.

Контролируемое и неконтролируемое обучение
В контексте компьютерного обучения здесь полезно провести различие между контролируемым и неконтролируемым обучением. Неформально говоря, контролируемое обучение относится к алгоритмам, которые учатся аппроксимировать общую функцию Y=f(x1,...,xn) через процесс обучения таким образом, чтобы при поступлении новых входных данных (x1,...,xn) система могла предсказать соответствующие выходные переменные Y. Очевидно, что на этапе обучения система должна узнать Y для каждого элемента входных данных (x1,...,xn).
Это означает, что контролируемое обучение можно применять только тогда, когда такая целевая функция имеется в данных.

Неконтролируемое обучение относится к алгоритмам, в которых нет информации о соответствующем значении Y для каждого элемента входных данных (x1,…,xn). Следовательно, цель обучения без подкрепления - найти базовую структуру или распределение в данных, чтобы узнать о них больше.
Развернутый обзор алгоритмов ИИ выходит за рамки данной статьи (читатель может обратиться к [33] для ознакомления с моделями компьютерного обучения). По этой причине здесь мы показываем, как специализированные алгоритмы и конвейеры ИИ могут быть использованы к информации о РТ для получения дальнейших знаний на реальном примере из двух этапов. На первом этапе алгоритмы ИИ (контролируемое обучение) используются для классификации вакансий, размещенных в Интернете, в системах SOC (как показано в [29]). На втором этапе алгоритмы искусственного интеллекта (тематическое моделирование) используются в системе "человек в цикле" для извлечения новых появляющихся профессий и составления соответствующего профиля профессий (см. [34]).
Классификация текста с помощью машинного обучения [подконтрольный]. Репрезентативная задача ИИ для ИРТ основана на классификации вакансий по стандартной классификации профессий и навыков. Хотя существует несколько классификаций РТ (например, МСКЗ, O*NET, SOC), в данном примере мы показываем, как классификация вакансий может быть выражена в терминах текстовой классификации. Более конкретно, в контексте ИРТ, это обычно требует использования алгоритмов классификации текста (на основе онтологии, машинного обучения и т.д.) для построения функции классификации, которая отображает элемент данных в один из нескольких предопределенных классов. Примечательно, что элементы представлены вакансиями, размещенными в Интернете, а предопределенные классы взяты из классификации (например, собственной таксономии или публичной классификации, такой как ESCO или O*NET, как показано на рисунке). Таким образом, задача классификации вакансий может быть формально описана в терминах категоризации текста.

[20] Европейская комиссия, *Искусственный интеллект для Европы*, COM(2018) 237. По сост.на март 2019:https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

Классификация ESCO

ESCO - это многоязычная система классификации европейских навыков, компетенций, квалификаций и профессий, разработанная Европейской комиссией. Классификация профессий ESCO соответствует Международной стандартной классификации занятий (МСКЗ-08) до уровня четвертого разряда. Затем он расширяет МСКЗ за счет дополнительного уровня профессий и навыков, организованных в виде графиков, а не дерева (т.е. один навык может относиться к нескольким профессиям).

Категоризация текста направлена на присвоение логического значения каждой паре $(dj,ci) \in D \times C$, где D - набор документов, а C - набор предопределенных категорий. Истинное значение, присвоенное (dj,ci), означает, что документ dj должен быть отнесен к категории ci, а ложное значение означает, что dj не может быть отнесен к категории ci. В варианте моделирования ИРТ набор вакансий J можно рассматривать как набор документов, каждый из которых должен быть отнесен к одному (и только одному) коду профессии в таксономии. Таким образом, классификация вакансии по системе классификации означает присвоение вакансиям одного кода профессий. Эта задача категоризации текста может быть решена с помощью машинного обучения, как указано в [35]. Формально говоря, пусть $J = \{J1,...,Jn\}$ - это набор вакансий, классификация J по таксономии состоит из $|O|$ независимых задач классификации каждой вакансии по заданному таксономией коду профессии oi для $i= 1,...,|O|$. Тогда классификатор - это функция $\psi : J \times O \rightarrow \{0,1\}$, которая аппроксимирует неизвестную целевую функцию $\psi' : J \times O \rightarrow \{0,1\}$. Очевидно, что поскольку в данном случае необходимо иметь дело с классификатором с одной меткой $\forall j \in J$, должно выполняться следующее ограничение: $\sum o \in O \psi(j,o) = 1$.

После того как задача классификации вакансий в интернете смоделирована в терминах категоризации текста, любой алгоритм машинного обучения может быть использован для обучения классификатора, эффективность которого может быть оценена с помощью различных показателей. Очевидно, что классификатор может работать более эффективно на одних классах, но не на других.

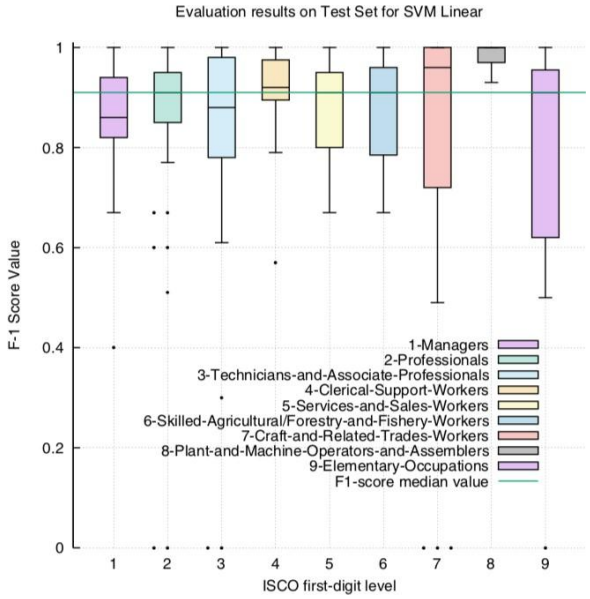
Именно такой случай показан на рисунке 2.6, где классификатор машинного обучения (в данном случае машина опорных векторов) был обучен на наборе из более 60 000 вакансий. F1-score[21] представлен для каждого первого уровня МСКЗ. Блок-схема[22] показывает распределение значения точности для лучшего алгоритма машинного обучения (т.е. линейного SVM) по девяти группам МСКЗ. Таким образом, эффективность каждого алгоритма классификации может быть исследована на конкретной группе профессий. Хотя линейный классификатор SVM достиг общей точности 0,9, его производительность меняется по мере изменения первой цифры МСКЗ. Эти характеристики алгоритмов машинного обучения должны быть тщательно рассмотрены и оценены при оценке любых приложений, которые используют или применяют автоматизированное обучение для поддержки принятия решений.

[21] F1-score (также F-score или F-мера) является одним из наиболее широко используемых показателей для оценки точности классификатора на тестовом наборе данных.

[22] Box-plot - хорошо известный статистический метод, используемый в исследовательском анализе данных для визуального выявления закономерностей, которые иначе могут быть скрыты в наборе данных, путем измерения изменений вариации между различными группами данных. Коробка обозначает положение верхнего и нижнего квартилей соответственно; содержимое коробки обозначает медианное значение, которое является областью между верхним и нижним квартилями и составляет 50% распределения.

Вертикальные линии (также известные как "усы") выходят за крайние точки распределения, указывая на минимальные или максимальные значения в наборе данных. Наконец, точки используются для обозначения верхних и нижних выбросов, а именно элементов данных, которые лежат больше (меньше), чем 3/2 верхнего (нижнего) квартиля соответственно.

РИСУНОК 2.6 ПОДРОБНЫЕ ОТЧЕТЫ О ТОЧНОСТИ МАШИННОГО ОБУЧЕНИЯ



Источник: см. [29].

Новые появляющиеся профессии с использованием тематического моделирования [неконтролируемый]. Ниже мы опишем, как моделирование может быть использовано для определения новых (потенциальных) появляющихся профессий, используя неконтролируемые обучения. Термин "новые (потенциальные) появляющиеся профессии" относится к профессиям, которые еще не были включены в какую-либо систему классификации профессий (т.е. МСКЗ/ESCO в данном случае). Очевидно, что использование нового термина при объявлении вакансии не идентифицирует новую профессию, поскольку этот новый появляющийся термин должен быть подтвержден растущей со временем тенденцией, подтверждающей создание новой (появляющейся) профессии на интернет РТ. С этой целью использование тематического моделирования хорошо подходит для выявления терминов, которые являются статистически значимыми в текстах.

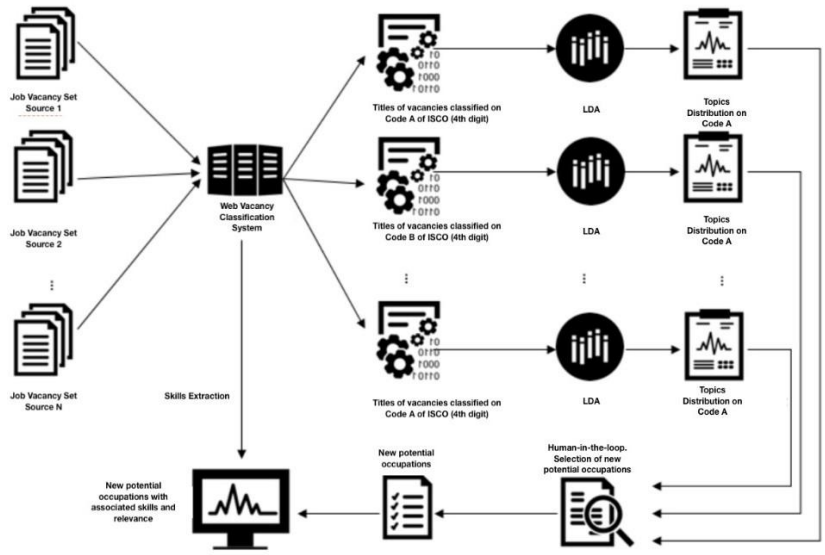
Более конкретно, предположим, что у нас есть коллекция документов - состоящая не более чем из 10 или 15 слов каждый - содержание которых представляет собой смесь тем (например, темы, T_1,T_2,...,Tn), которые характеризуют лексикон каждого подмножества документов. Латентное распределение Дирихле (ЛРД [36]) - это генеративная вероятностная модель, которая рассматривает каждый документ как смесь латентных тем, где каждая тема характеризуется собственным распределением слов. ЛРД позволяет кластеризовать документы по темам на основе частоты слов. В результате каждая тема состоит из слов, которые в основном способствуют ее созданию. Чем выше вероятность того, что тема содержит определенный термин, тем выше релевантность этого термина в соответствующей теме. ЛРД не делает разделения терминов по темам, поскольку термин может принадлежать более чем одной теме, хотя и с разной релевантностью.

Моделирование темы (восприятие)

Идея использования ЛРД для выявления новых потенциальных профессий основана на рассмотрении названия вакансии как документа, содержание которого может быть идеально составлено из ряда (фиксированного, но неизвестного) тем, которые необходимо выявить.

На рисунке 2.7 представлен графический обзор того, как работает этот процесс. После того, как каждая вакансия классифицирована по стандартной классификации (т.е. в нашем случае по четвертому разряду МСКЗ), алгоритм ЛРД применяется к каждому подмножеству вакансий, группируя их по кодам МСКЗ. Этот этап предварительного отбора поможет ЛРД уменьшить пространство признаков и максимизировать производительность ЛРД. Процесс ЛРД возвращает ряд тем вместе с вероятностью распределения слов, составляющих каждую тему[23]. Этот процесс возвращает список лучших терминов для каждого кода МСКЗ, который должен быть проанализирован и уточнен специалистом РТ. Поскольку ЛРД является алгоритмом неконтролируемого обучения, для обеспечения надежности результата обязательно требуется наблюдение человека для проверки конечного результата. Наконец, термины, определяющие новые потенциальные профессии, связывают с вакансиями, в которых эти термины были найдены, а затем связывают с включенными в них навыками. Это позволяет вычислить новые появляющиеся профессии и отфильтровать навыки, востребованные только для них.

РИСУНОК 2.7 ПОДХОД НА ОСНОВЕ ЛРД ДЛЯ ВЫЯВЛЕНИЯ НОВЫХ ПОТЕНЦИАЛЬНЫХ ПРОФЕССИЙ



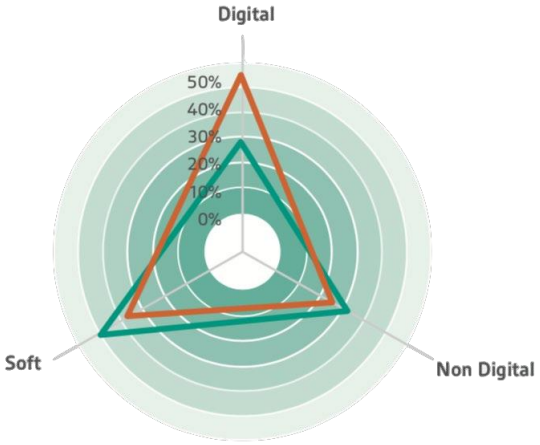
В конце этого процесса можно получить технологическую карточку, в которой указаны:

- показатели навыков новой профессии, определяемые как частота встречаемости каждой категории или группы навыков в рамках профессий, которые могут быть цифровыми, нецифровыми или мягкими;
- распределение навыков по Европейской электронной системе компетенций (e-CF)[24];
- углубленный анализ, позволяющий присвоить соответствующий навык ESCO каждой компетенции e-CF в соответствии с вакансиями.

В качестве примера здесь показан профиль специалиста по большим данным. Данные получены от WollyBI, наблюдающего за итальянскими вакансиями, размещенными в интернете только в 2017 году, и опубликованы в Итальянской обсерватории цифровых компетенций[25]. На цифровые навыки приходится только 29,5%, в то время как 31,5% составляют нецифровые навыки, а 39% - мягкие навыки (Рисунок 2.8). Распределение навыков e-CF показано на рисунке 2.9, где выделены компетенции, требуемые от специалистов по большим данным в целом.

Этот пример показывает, как знаниями о РТ можно манипулировать для ответа на специальные и конкретные исследовательские вопросы, такие как оценка влияния цифровых/мягких/нецифровых навыков в рамках профессий, определение новых развивающихся профессий, которые еще не закодированы в стандартных таксономиях, и понимание ожиданий РТ путем сосредоточения внимания на навыках, которые запрашиваются рынком, а не на навыках, перечисленных в классическом профиле работы.

РИСУНОК 2.8 КОЭФФИЦИЕНТЫ КВАЛИФИКАЦИИ СПЕЦИАЛИСТОВ ПО БОЛЬШИМ ДАННЫМ (ЗЕЛЕНАЯ ЛИНИЯ) ПО ОТНОШЕНИЮ К КОЭФФИЦИЕНТАМ КВАЛИФИКАЦИИ ДРУГИХ ПРОФЕССИЙ ИКТ (КРАСНАЯ ЛИНИЯ)

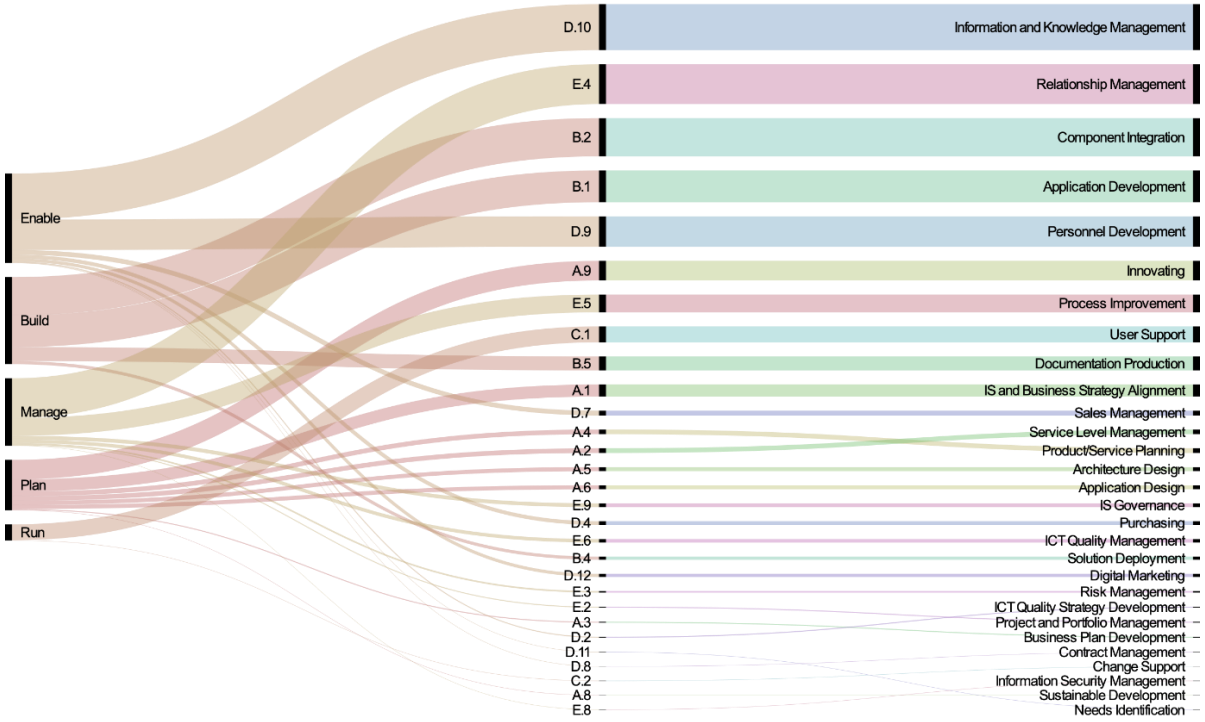


[23] Количество тематик, которые необходимо определить, является входным параметром любого подхода на основе ЛРД и должно быть правильно настроено.

[24] Для получения более общей информации о e- CF: www.ecompetences.eu/. e-CF 3.0 можете найти на сайте:http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework- 3.0_CEN_CWA_16234-1_2014.pdf

[25] Доступно для общественности на сайте: www.assintel.it/assinteldownloads/osservatorio-competenze-digitali-2018-il-volume/[только на итальянском языке].

РИСУНОК 2.9 РАСПРЕДЕЛЕНИЕ НАВЫКОВ E- CF ДЛЯ ФОРМИРУЮЩЕЙСЯ ПРОФЕССИИ СПЕЦИАЛИСТОВ ПО БОЛЬШИМ ДАННЫМ



Вопросы и ответы

Кто пишет алгоритмы ИИ?

На сегодняшний день алгоритмы ИИ по-прежнему пишутся людьми. Однако в основе машинного обучения лежит идея создания систем, которые автоматически обучаются на основе данных. В этом смысле не существует рукописного кода человека, который направлял бы процесс обучения. Человек может только попытаться понять, чему на самом деле научилась система, наблюдая и опрашивая систему как "черный ящик".

Кто гарантирует, что ИИ делает то, что должен делать?

Многие исследователи работают над этим. Одна из актуальных тем касается Объяснимого ИИ, позволяющего пользователю понять, чему учатся машины/алгоритмы ИИ. На сегодняшний день существует мало инструментов для объяснения критериев - и соответствующих характеристик - которыми руководствуется система ИИ при прогнозировании или предложении определенного результата.

Существуют ли законодательство или этические кодексы для поддержки "хорошего" ИИ?

В настоящее время нет, но в мае 2018 года Европейская комиссия опубликовала документ "*Искусственный интеллект для Европы*", в котором также рассматриваются некоторые этические принципы (по сост.на март 2019 года: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe)

3. ИСПОЛЬЗОВАНИЕ АНАЛИТИКИ БОЛЬШИХ ДАННЫХ ДЛЯ ИСРТ: ПОДБОРКА ПРАКТИЧЕСКИХ ПРИМЕРОВ ДЛЯ ИСПОЛЬЗОВАНИЯ В КАЧЕСТВЕ СПРАВОЧНИКА

В этой главе мы рассмотрим некоторые недавние важные проекты с использованием информации о РТ - включая информацию из приложений Большие Данные - для поддержки ИРТ в режиме реального времени, разработанные за пределами ЕС (США и Малави) и внутри ЕС (Великобритания, Нидерланды, Испания и проект ЕС с участием всех стран ЕС).

Информация, изложенная здесь, была согласована и передана главным исследователям проектов, чтобы обеспечить всестороннее описание каждой инициативы. С этой целью каждый проект описывается на основе: (i) цели; (ii) использованных данных/источников; (iii) результатов (или ожидаемых итогов); (iv) достижений; и (v) открытых/проблемных вопросов.

3.1 CyberSeek.org – Соединенные Штаты Америки

CyberSeek.org - это интерактивная карта кликов в области спроса и предложений, а также карьерного пути для сотрудников в сфере кибербезопасности в США. Это совместный проект Burning Glass, CompTIA и Национальной инициативы по образованию в области кибербезопасности (NICE). Он финансируется за счет гранта Национального института стандартов и технологий и был выпущен в 2016 году.

Цель. Цель CyberSeek - предоставить подробные и практические данные о рабочей силе в области кибербезопасности в США, чтобы помочь педагогам, работодателям, политикам и соискателям принять более обоснованные решения и поддержать усилия по устранению разрыва в квалификации в области кибербезопасности.

Использованные данные и источники. CyberSeek предоставляет подробные данные о спросе и предложениях на рабочие места в сфере кибербезопасности в государственном и частном секторах в штатах и районах США. Сюда входят следующие показатели:

- общее количество вакансий;
- общее количество работающих в настоящее время работников;
- соотношение спроса и предложения;
- коэффициент местоположения;
- названия лучших должностей в сфере кибербезопасности;
- вакансии по категориям Структуры рабочей силы для кибербезопасности по NICE;
- спрос и предложение востребованных сертификатов по кибербезопасности;
- общее количество вакансий; средняя заработная плата; требования к навыкам, дипломам и опыту; и возможности перехода между 10 основными рабочими местами в сфере кибербезопасности и пятью рабочими местами, обеспечивающими кибербезопасность.

CyberSeek черпает эти данные из трех первичных источников данных:

- База данных Burning Glass включает сотни миллионов онлайн-объявлений о вакансиях, собранных с 2007 года. Эта база данных использует передовые методы анализа естественного языка, чтобы превратить информацию в каждом объявлении о работе в структурированные, пригодные для использования данные. Это позволяет Burning Glass описать спрос работодателей на конкретные роли или навыки с такой степенью детализации, которая недоступна при традиционных методологиях опроса;
- государственные данные. Бюро трудовой статистики о ныне работающих трудовых ресурсах;

- данные о владельцах сертификатов от пяти различных сертифицирующих организаций: CompTIA, IAPP, ISC^2, ISACA и GIAC.

Результаты (или ожидаемые результаты [26]). CyberSeek добился трех основных результатов:

1. количественная оценка степени дефицита навыков в сфере кибербезопасности по всей стране;
2. определение возможностей перехода в сферу кибербезопасности и внутри нее;
3. определение ключевых навыков и полномочий, необходимых в данной области.

Достижения. С момента выпуска CyberSeek использовался сотнями тысяч пользователей, включая преподавателей, политиков, работодателей, студентов и соискателей. На него ссылались десятки СМИ как на основной источник данных о рынке труда в сфере кибербезопасности. Он также стал финалистом конкурса на лучший некоммерческий сайт года по версии PR Newswire.

Открытые/проблемные вопросы. Текущие проблемы связаны с проблемами включения дополнительных компонентов предложения, помимо существующего контингента (например, студентов или безработных работников); разбивки данных по отраслям; включения данных о провайдерах обучения; и распространения инструмента на новые географические регионы.

3.2 WheretheWorkIs.org – Великобритания

WheretheWorkIs.org - это модель спроса/предложения для рабочих мест средней квалификации, которая определяет рабочие места и регионы, где существуют пробелы в квалификации или избыток работников, чтобы провайдеры обучения могли соответствующим образом адаптировать свои предложения. Это проект Burning Glass Technologies и Института исследований государственной политики, финансируемый JPMorgan Chase & Co в 2016 и 2017 годах в рамках программы New Skills at Work, целью которой является определение стратегий и поддержка решений, способствующих улучшению инфраструктуры РТ и развитию квалифицированной рабочей силы в глобальном масштабе.

Цели. Цель WheretheWorkIs - предоставить уникальный, бесплатный в использовании инструмент, позволяющий провайдерам обучения, исследователям, работодателям и политикам изучить, как предложения и спрос на различные типы рабочих мест в разных районах Великобритании.

Использованные данные и источники. WheretheWorkIs использует два различных вида данных, предоставляемых Burning Glass.

Данные по спросу: с 2012 года в Великобритании в Интернете было размещено 50 миллионов уникальных вакансий, при этом использовалась передовая аналитика естественного языка для преобразования информации, содержащейся в каждом объявлении о вакансии, в структурированные, пригодные для использования данные. Это позволяет Burning Glass описать спрос работодателей на конкретные роли или навыки с такой степенью детализации, которая недоступна при традиционных методологиях опроса.

Спрос на талантливых специалистов начального уровня (с опытом работы менее двух лет) сравнивается с имеющимся предложением новых выпускников или стажеров. Данные по вакансиям Burning Glass нормализуются по данным о занятости, публикуемым Управлением национальной статистики (УНС). Данные дополнительно проверяются по данным Ежегодного исследования часов работы и заработков, публикуемым УНС.

Данные по предложению: Burning Glass использует данные о количестве учащихся, заканчивающих высшее образование и постдипломное обучение (выпускники программ по предметным областям), в качестве косвенного показателя "предложения" талантливых специалистов начального уровня. Данные о предложении поступают от следующих агентств:

- Агентство статистики высшего образования (по Великобритании);
- Агентство по финансированию навыков (Англия);
- Шотландский совет по финансированию;
- Развитие навыков Шотландия;
- Департамент занятости и обучения Северной Ирландии;
- СтатсУэльс.

Результаты. Онлайн-портал показал, что существует огромное несоответствие между средним уровнем квалификации работников, необходимым британским работодателям для начальных должностей, и квалификацией, которой обладают новые соискатели - т.е. спрос на выпускников высших и средних учебных заведений превышает количество кандидатов.

Для выпускников постдипломного обучения, профессиями, которые больше всего превосходят число кандидатов, являются:

- услуги по персональному уходу - 203 758 вакансий;
- продавцы-консультанты и кассиры розничной торговли - 89 898 вакансий;
- супервайзеры по продажам - 20 139 вакансий;
- младшие медицинские работники - 16 108 вакансий;
- профессии текстильной и швейной промышленности - 2 905 вакансий.

Для выпускников высших учебных заведений, профессиями, которые больше всего превосходят число кандидатов, являются:

- специалисты в области преподавания и образования - 150 763 вакансии;
- государственная служба и другие ассоциированные специалисты - 71 873 вакансии;
- уход за детьми и сопутствующие услуги - 29 846 вакансий;
- младшие специалисты по социальному обеспечению и жилью - 20 624 вакансии;
- младшие медицинские работники- 14 390 вакансий.

Этот инструмент позволяет провайдерам обучения и другим лицам углубляться в данные, чтобы понять, какие возможности существуют в конкретных районах.

Достижение. Инструмент был профинансирован для обновления с учетом обновленных данных о "предложениях" в 2017 году. Отзывы работодателей, преподавателей и политиков были положительными. Данные, лежащие в основе инструмента, легли в основу переговоров между Burning Glass и поставщиком информации по карьере для школ по всей Англии.

Открытые/проблемные вопросы. Если данные Burning Glass по "спросу" доступны в режиме реального времени и достаточно детализированы, чтобы рассмотреть отдельные профессии на местном уровне, то данные по "предложению" таковыми не являются: самые свежие данные были опубликованы два года назад, а низкие значения часто не позволяют провести исследование на местном уровне.

[26] Для получения дополнительной информации см.: www.comptia.org/about-us/newsroom/press-releases/2018/06/06/us-cybersecurity-worker-shortage-expanding-new-cyberseek-data-reveals

Обеспечение соответствия спроса и предложения часто является сложной задачей. Некоторые профессии не имеют специальных квалификационных или предметных требований (например, продажи, маркетинг и связанные с ними младшие специалисты). В итоге, несмотря на то, что на эти профессии могут претендовать люди с широким спектром квалификаций, они часто оказываются "недостаточно представленными" в этом инструменте, т.е. когда "возможность трудоустройства" низка.

3.3 Баскская обсерватория талантов провинции Бискайя – Испания

Используя большие данные, Баскская обсерватория талантов анализирует 13 различных онлайн и офлайн государственных и частных платформ для размещения вакансий, ориентированных на людей с высшим образованием, только в Стране Басков, чтобы получить информацию о запрашиваемых профилях с мягкими и жесткими навыками на баскском РТ.

Цели. Обеспечить специалистов, государственные агентства, частный сектор и университеты информацией о видах профилей, уровня образования и навыков, востребуемых на РТ, по отраслям, территориальным районам, и т. д.

Использованные данные и источники. Система проанализировала 126 000 вакансий в Стране Басков за последние 12 месяцев, отфильтровав их для обеспечения 100% соответствия двум критериям: для людей с высшим образованием и для работы в Стране Басков. Были использованы следующие веб-источники: Adecco, Bizkaia Talent, Университет Страны Басков, Indeed, Infoempleo, Jobydoo, Infojobs, Lanbide (баскская государственная служба занятости), Mondragon People, Университет Мондрагона, Monster, Randstad и Studentjob.

Результаты. Как только университеты и государственные структуры поняли, какие профессиональные профили требуются, они смогли адаптировать свое обучение к текущей реальности и тенденциям, используя данные для разработки системы прогнозирования нехватки талантов по каждому виду профиля в Стране Басков и сопоставления предложений о работе (500 в год) и специалистов на платформе (11 000), чтобы предложить нужную вакансию нужному кандидату и наоборот.

Достижение. Платформа представляет собой систему свободного доступа, требующую лишь регистрации на сайтах www.bizkaiatalent.eus или www.bebasquetalentnetwork.eus. Зарегистрированные специалисты и компании также имеют доступ через приложение Bizkaia Talent. Для многих специалистов, живущих за границей (55% из 11 000), знание о спросе на баскский РТ дает им больше возможностей вернуться на работу в регион.

Открытые/проблемные вопросы. Еще предстоит выяснить, как можно использовать данные и различные комбинации времени/навыков/профессий для прогнозирования спроса на кадры в ближайшие годы и работать на опережение, чтобы связаться с нужными специалистами.

3.4 Таксономия навыков на основе данных - Великобритания

Nesta[27] разработала первую в Великобритании таксономию навыков, основанную на данных, которая стала общедоступной. Таксономия была создана на основе исследования, которое Nesta проводила в качестве партнера Центра передового опыта в области экономической статистики (ESCoE[28]). ESCoE финансируется УНС Великобритании, и его целью является анализ возникающих и будущих проблем в измерении современной экономики. В рамках исследования ESCoE изучался потенциал использования естественно возникающих Больших Данных в виде онлайн-рекламы

вакансий для улучшения понимания РТ.

Цели. Нехватка квалифицированных кадров является серьезной проблемой в Великобритании и может существенно препятствовать экономическому росту. Согласно исследованию ОЭСР[29], Великобритания могла бы повысить свою производительность на 5%, если бы снизила уровень несоответствия навыков до уровня лучшей практики ОЭСР.

Несмотря на значимость нехватки навыков, в настоящее время она не оценивается подробно и своевременно. Наилучшие имеющиеся оценки получены в рамках Обзора профессиональных навыков работодателей[30]. Хотя исследование позволяет пролить свет на различные причины нехватки навыков, оно проводится только раз в два года и фокусируется на широких, а не детальных группах навыков.

В перспективе несоответствие профессиональных навыков может усугубиться, поскольку навыки, необходимые для работы, меняются как под влиянием краткосрочных факторов, таких как Brexit, так и под влиянием более долгосрочных тенденций, таких как автоматизация. Исследование Nesta показало, что пятая часть работников занята в профессиях, которые, скорее всего, сократятся в ближайшие 10-15 лет[31].

Первым шагом к измерению дефицита является создание таксономии навыков, которая показывает группы навыков, необходимых работникам в Великобритании сегодня. Затем таксономию можно использовать в качестве основы для измерения спроса на навыки среди работодателей, текущего предложения этих навыков со стороны работников и потенциального предложения на основе курсов, предлагаемых поставщиками образовательных услуг и работодателями.

Использованные данные и источники. Построение таксономии началось со списка из чуть более 10 500 уникальных навыков, упомянутых в описаниях 41 миллиона вакансий, объявленных в Великобритании, собранных в период с 2012 по 2017 год и предоставленных компанией Burning Glass Technologies. Эти навыки включали конкретные задачи (например, андеррайтинг в страховании), знания (биология), программные продукты (Microsoft Excel) и даже личные качества (позитивный настрой). Машинное обучение использовалось для иерархической кластеризации навыков. Чем чаще два навыка встречались в одном и том же объявлении, тем больше вероятность того, что они окажутся в одной и той же ветви таксономии. Таким образом, таксономия отражает кластеры навыков, которые необходимы для разных рабочих мест.

Результаты. Окончательная таксономия имеет древовидную структуру с тремя слоями. Первый слой содержит шесть широких кластеров навыков; они разделяются на 35 групп, а затем еще раз разделяются, чтобы получить 143 кластера конкретных навыков. Каждый из примерно 10 500 навыков входит в одну из этих 143 групп. Та же методология может быть использована для создания следующих слоев.

Таксономия навыков была дополнена для получения оценок спроса на каждый кластер навыков (на основе количества упоминаний в объявлениях о работе), изменения спроса за

[27] www.nesta.org.uk/ и, в частности www.nesta.org.uk/data-visualisation-and-interactive/making-sense-skills/
[28] www.escoe.ac.uk/
[29] www.oecd.org/eco/growth/Labour-Market-Mismatch-and-Labour-Productivity-Evidence-from-PIAAC-Data.pdf
[30] www.gov.uk/government/publications/employer-skills-survey-2017-uk-report
[31] www.nesta.org.uk/report/the-future-of-skills-employment-in-2030/

последние годы и ценности каждого кластера навыков (на основе объявленных зарплат). Оценки спроса дают одну половину картины нехватки навыков. Самое главное, что пользователь может искать в таксономии по названию должности и обнаружить навыки, необходимые для широкого спектра рабочих мест.

Достижение. Таксономия была опубликована только в августе 2018 года. В течение следующего года будет предоставлен ряд примеров с использованием таксономии навыков. Это будет включать оценку нехватки навыков на региональном уровне, автоматическое обнаружение новых и избыточных наборов навыков, а также оценку потенциального предложения навыков на основе доступных курсов и обучения. Сама таксономия также продолжит развиваться, поскольку система добавит четвертый уровень и попытается отразить боковые связи между кластерами.

Открытые/проблемные вопросы. Ни одна таксономия не будет действительно всеобъемлющей, независимо от того, получена ли она от экспертов или создана на основе объявлений о вакансиях. Более того, не существует "правильного способа" группировки навыков. В данном исследовании наиболее важным ограничением было то, что не все вакансии выкладываются в Интернет. В результате спрос на навыки, используемые преимущественно фрилансерами или случайными работниками, может быть недооценен в таксономии. Несмотря на этот риск, подход, основанный на данных, создает наиболее подробную таксономию навыков Великобритании, доступную общественности на сегодняшний день, и ее легче обновить, чем таксономию, составленную экспертами.

3.5 Профессионально-техническое образование и обучение предпринимательству – Малави

Организация Объединенных Наций по вопросам образования, науки и культуры (ЮНЕСКО) и правительство Малави провели в 2018 году обзор профессионально-технического, предпринимательского образования и обучения. В рамках исследования состояния PT был проведен анализ спроса на рабочие места и трудоустройства в Малави с использованием Big Data science и искусственного интеллекта, чтобы продемонстрировать возможности новых медиа и позволить Малави совершить скачок в новую систему, минимизирующую трения на PT благодаря информации о доступных рабочих местах и людях, ищущих работу, практически в режиме реального времени. На крупнейшем в Малави сайте вакансий myJobo.com число пользователей удваивается каждые 10 месяцев и продолжает быстро расширяться.

Использованные данные и источники. Данные о вакансиях в период с 1 января 2016 года по 30 апреля 2018 года были получены с крупнейшего онлайн-портала по поиску работы в Малави: myJobo.com.

Результаты. Агрегирование Больших Данных из 360 000 пунктов данных на myJobo.com показывает разнообразие трендовых рабочих функций, с наибольшим спросом в области администрирования, бухгалтерского учета, управления, инженерного дела, связей с общественностью, образования и здравоохранения. Более подробная информация о названиях вакансий показывает, что в 2016-2018 годах в городских районах Малави наиболее востребованы следующие профессии: бухгалтеры, помощники бухгалтеров, административные помощники, финансовые работники, технические работники, менеджеры проектов и координаторы проектов. Также перечислены 100 лучших вакансий, что позволяет получить детальную картину преобладающего рынка труда для всех соискателей, а также для

государственных органов планирования, образовательных учреждений, тренеров и других. Например, для должности бухгалтера основными навыками являются знание налогов, финансовой отчетности, финансового анализа, корпоративного налогообложения, аудита, бюджетов и прогнозирования. Главные навыки, необходимые для бухгалтера, - это бухгалтерский учет, финансовая отчетность, бюджеты, сверка счетов, Microsoft Excel, главная книга, кредиторская задолженность, внутренний контроль, управление и анализ.

Достижение. Анализ данных, полученных с сайта myJobo.com, позволяет получить важные сведения о тенденциях развития рынка труда Малави, в значительной части городских районов Малави. Эти тенденции дают новую информацию о профессиях, которую можно оценить в количественном выражении, что поможет поставщикам услуг по обучению адаптировать курсы и информировать соискателей о том, какие навыки требуются работодателям. Таким образом, соискатели работы и учащиеся могут персонализировать свой путь обучения, чтобы устранить потенциальные пробелы в навыках, и предложить возможности для повышения квалификации.

Открытые/проблемные вопросы. Возможности, предоставляемые новыми методами с использованием сайтов вакансий, являются полезным дополнением к существующим системам, таким как ИСРТ, и восполняют пробелы в знаниях о РТ за годы между крупномасштабными исследованиями, такими как Обследование рабочей силы Малави, которое в последний раз проводилось в 2013 году. Новые методы позволяют получать информацию практически в режиме реального времени, что поможет соискателям работы, а также специалистам по планированию образования и курсов быть в курсе тенденций в сфере занятости.

3.6 Проекты "Трансферные профессии" (A) и "Индикаторы напряженности" (B) - Нидерланды

Эти проекты были разработаны Департаментом аналитики рынка труда (ДРРТ) (A и B) и Panteia (B) в Нидерландах.

Проект "Трансферные профессии"

Цели. Предоставить ищущим работу лицам, пострадавшим от профессий, где избыток выпускников[32], набор альтернативных профессий и, следовательно, лучшие шансы найти работу (трансферные профессии).

Использованные данные и источники[33]. Проект основан на реальных шагах мобильности соискателей в прошлом. В проекте используются резюме соискателей на сайте werk.nl в период с 2013 по 2017 год. Безработные или частично нетрудоспособные люди, получающие пособие ДРРТ, обязаны зарегистрироваться на сайте werk.nl и разместить на нем свое резюме. Люди, получающие социальную помощь от муниципалитетов, также обязаны зарегистрироваться на сайте werk.nl; соискатели регистрируются на добровольной основе. Система изучает резюме, чтобы определить частоту переходов от одной профессии к другой. Из-за возможного искаженного отбора, база данных рассчитывает частоту переходов только для избыточных профессий (например, административные клерки, секретари, бухгалтеры). Для этих профессий проект выбирает целевые трансферные профессии, когда:

[32] Примеры избыточной занятости: предложение труда превышает спрос или количество краткосрочных безработных более чем в 1,5 раза превышает количество открытых вакансий.

[33] Данные могут быть (полу)структурированными/неструктурированными; источники могут быть административными/статистическими/опросными/веб-сайтами.

- трансферная профессия имеет лучшие возможности трудоустройства, чем избыточная профессия;
- уровень трансферной профессии аналогичен или почти аналогичен уровню избыточной профессии.

Результаты. Для избыточных профессий система периодически публикует несколько трансферных профессий. Эта информация используется наставниками и непосредственно лицами, ищущими работу. Она также используется в презентациях, семинарах и вебинарах.

Проект "Индикаторы напряженности"

Цели. Получение надежного показателя напряженности РТ по регионам и профессиональным группам.

Использованные данные и источники. В проекте используется сочетание Больших Данных, данных опросов и административных данных.

- Большие Данные используют онлайн вакансии, такие как вакансии в базе данных Jobfeed от Textkernel, после устранения дубликатов и проверки надежности.
- Данные обследований включают результаты Национального обследования вакансий Центрального бюро статистики для обеспечения последовательности и учета смещения отбора по секторам и уровню образования.
- Административные данные охватывают людей, получающих пособие по безработице ДРРТ менее шести месяцев.

Результаты. Каждые три месяца система публикует показатель напряженности для 35 регионов в 114 профессиональных группах, и выделяет следующие типы: очень большой дефицит, дефицит, в среднем, избыток, большой избыток.

Система используется для разработчиков политики, прессы, презентаций, семинаров, материалов для нескольких публикаций и дополнительных исследовательских целей.

3.7 Информация о требованиях к навыкам на рынке труда в режиме реального времени - все страны члены ЕС

После успешного пилотного исследования, в 2017 году СЕДЕФОП[34] начал разработку системы для сбора данных с онлайн вакансий во всех странах-членах ЕС, объявив тендер "Информация о рынке труда в реальном времени о требованиях к навыкам: Создание системы ЕС для онлайн-анализа вакансий". Система будет полностью разработана и введена в действие к 2020 году.

В разработке проекта СЕДЕФОП сотрудничает с:

- CRISP - Университет Милано-Бикокка (лидер): CRISP (Межвузовский исследовательский центр общественных услуг) - это исследовательский центр, расположенный в Университете Милано-Бикокка (Италия);
- TabulaeX s.r.l.: аккредитованное ответвление Университета Милано-Бикокка (Италия);
- IWAK (Институт экономики, труда и культуры): институт прикладных исследований, связанный с Университетом Гете во Франкфурте-на-Майне (Германия).

[34] СЕДЕФОП является одним из децентрализованных агентств ЕС. Основанный в 1975 году, СЕДЕФОП поддерживает развитие европейской политики в области профессионального образования и обучения и способствует ее реализации, работая, в частности, с Европейской комиссией, государствами-членами ЕС и социальными партнерами.

Поскольку проект предполагает разработку многоязычной системы классификации, консорциум также сотрудничает с международными экспертами по странам, представляющими все 28 государств-членов ЕС.

Цели. Основной целью данного проекта является разработка полномасштабной системы, которая позволит СЕДЕФОП проводить анализ онлайн вакансий и возникающих потребностей в профессиональных навыках во всех 28 государствах-членах ЕС. Система сбора данных в режиме реального времени будет собирать соответствующую справочную информацию о вакансиях, фирмах и типе требуемого работника (навыки, квалификация и другие атрибуты), что позволит в будущем изучить и проанализировать спрос на профессиональные навыки.

Система сбора данных в реальном времени будет создана с помощью онлайновых инструментов ввода данных, которые систематически посещают сайты и компоненты, участвующие в программном анализе веб-страниц, для сбора определенных форм информации. Конечным результатом будет база знаний, предоставляющая информацию о спросе на рабочую силу, с особым акцентом на требуемые навыки. Эта база знаний может быть использована для проведения различных видов анализа в поддержку заинтересованных сторон и лиц, принимающих решения.

Процесс включает в себя очистку вакансий, размещенных в Интернете, и классификацию извлеченных переменных. Результаты статистического анализа будут доступны с помощью инструментов визуализации.

Использованные данные и источники[35]**.** На первом этапе проекта "Ландшафтная деятельность" были изучены доступные источники данных по всему ЕС. Целью было понять, как работодатели и соискатели используют онлайн-вакансии, а также оценить репрезентативность данных для правильной интерпретации результатов. На этом этапе также был составлен список подходящих веб-порталов для сбора данных.

В списке соответствующих источников по всем 28 государствам-членам ЕС было указано 530 источников, разделенных на:

- Системы поиска работы;
- агентства по трудоустройству;
- сайты по трудоустройству;
- порталы классифицированных объявлений;
- компании;
- веб-сайты государственных служб занятости;
- онлайн-газеты;
- обучение;
- организации по трудоустройству.

Методы получения данных различаются в зависимости от сайтов: скрейпинг (24%), краулинг (18%) и доступ к ИПП (57%), осуществляемый с крупнейшими сайтами в рамках соглашений о лицензировании данных.

После первых шести месяцев работы проект оценил количество уникальных вакансий для всех стран ЕС в год примерно в 60 миллионов.

[35] Данные могут быть (полу)структурированными или неструктурированными; источники могут быть административными, статистическими, опросными или веб-источниками.

В результате сбора информации о вакансиях в Интернете (ИВИ) были извлечены следующие переменные, классифицированные как указано ниже:

- род деятельности--> ESCO v.1 до уровня 4;
- уровень образования --> уровень 1 Международной стандартной классификации образования (МСКО);
- территория --> NUT[36] до уровня 3;
- рабочее время --> пользовательская таксономия ("неполный рабочий день" и "полный рабочий день");
- тип контракта --> пользовательская таксономия ("временный", "постоянный", "самозанятый", "стажировка");
- отрасль деятельности --> уровень NACE[37] до уровня 2;
- зарплата --> пользовательская таксономия, основанная на числовых классах;
- навыки --> ESCO v.1.

Результаты. Результаты проекта имеют три направления:

1. проведение ландшафтных работ в каждом государстве-члене ЕС;
2. реализация полноценной системы онлайн сбора и анализа вакансий во всех 28 государствах-членах ЕС;
3. вклад в подготовку аналитического отчета по результатам проекта и дальнейшее его распространение.

Достижение. Проект все еще продолжается; ранний выпуск результатов, охватывающих семь стран, будет представлен в первом квартале 2019 года.

Открытые/проблемные вопросы. Трудности возникли при попытке поддержать новые модели анализа, полученные из Больших Данных - необходимо ИВИ для поощрения среди политиков и заинтересованных сторон подхода к принятию решений на основе данных. Существует необходимость в разработке панелей визуализации, настраиваемых по типу пользователей (политиков, статистиков и т.д.), и требуется масштабируемая архитектура для поддержки растущих объемов данных. Сохраняются проблемы, связанные с разбивкой данных о навыках по конкретным классификациям (навыки ИКТ, "мягкие" навыки и т.д.), и необходимо сделать больше для поддержки новых исследований и изучения будущего рабочих мест на основе данных ИВИ.

[36] Номенклатура территориальных единиц для статистики, см.: https://ec.europa.eu/eurostat/web/nuts/background

[37] Статистическая классификация экономической деятельности в Европейском сообществе, см: https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE)

4. ВЫВОДЫ И РЕКОМЕНДАЦИИ

4.1 Краткие рекомендации и последующие шаги для ЕФО и стран-партнеров

Возможная разработка системы анализа спроса на рабочую силу в интернете требует изучения ряда макро-тем.

Основная информация. Цели и характеристики системы можно описать следующим образом:

1. Территория, представляющая интерес: Какой РТ мы хотим исследовать? Одна ли это страна, группа стран или континент? Интересует ли нас РТ в целом или мы сосредоточены только на одном сегменте/секторе? Выбор повлияет на таксономии, используемые в проекте, поскольку для сравнения результатов необходимо иметь общую основу.
2. Переменные, представляющие интерес: Какие переменные мы хотим исследовать? Предполагая, что мы заинтересованы в создании базы знаний, ориентированной на спрос на рабочую силу, т.е. используя ИВИ в качестве данных, нам необходимо решить, достаточно ли сосредоточиться на спросе на профессии, хотим ли мы расширить анализ, включив в него навыки (связав запрашиваемые навыки с соответствующими профессиями), или мы хотим использовать данные ИВИ для извлечения из них наибольшей информативности, анализируя уровни образования, территориальные уровни, рабочие часы, тип контракта, отрасль и заработную плату. Это решение, конечно, повлияет на бюджет с точки зрения необходимого пространства для хранения и времени, отведенного на методологический подход.

Процесс отбора источника. При условии использования ИВИ, возникает вопрос: с каких сайтов собирать ИВИ. Этот этап можно разделить следующим образом:

1. Предварительное исследование веб-сайта: Это охватывает многие ключевые вопросы, такие как использование онлайн вакансий работодателями и соискателями (это важно для оценки репрезентативности и достоверности данных) и общий сценарий на местах. На эти вопросы должны ответить эксперты, знающие контекст и наиболее популярные практики.
2. Методология отбора источников: Как уже упоминалось выше, существует множество источников. После предварительного исследования нам необходимо решить, хотим ли мы принять во внимание все виды источников, или есть критерии исключения.
3. Расчет рейтинга и выбор источников на его основе: Расчет рейтинга может быть очень сложной статистической моделью или простой моделью, учитывающей необходимые переменные, так что сайты без этих атрибутов автоматически исключаются. Например, если целью является анализ в режиме реального времени, основным атрибутом источника может быть обновление: сайты, не обновляемые в течение длительного времени (например, более шести месяцев), могут быть исключены. Результатом ранжирования должен стать список наиболее значимых источников: на этом этапе решение о том, рассматривать ли их все или сделать выбор, должно приниматься на основе эксплуатационных возможностей.

Этические и правовые вопросы. Законодательство об использовании данных, если говорить конкретнее, ИВИ, публично размещенных в Интернете, не всегда четко сформулировано. Необходимо провести исследование, чтобы определить границы действий и мероприятий, которые категорически запрещены.

Подход к базам и моделям данных. Прежде чем начать сбор ИВИ с отобранных веб-сайтов, нам необходимо решить, как обрабатывать все данные, собранные в ходе проекта, и создать надежную техническую методологию:

1. Подход, основанный на базе данных: в мире технологий баз данных, существует два основных типа баз данных: SQL и NoSQL (или реляционные и нереляционные базы данных). Разница заключается в том, как они построены, в типе информации, которую они хранят, и в том, как они ее хранят. Реляционные базы данных структурированы, как телефонные книги, в которых хранятся номера телефонов и адреса. Нереляционные базы данных ориентированы на документы и являются распределенными, как файловые папки, содержащие все данные о человеке, от адреса и номера телефона до "лайков" на Facebook и предпочтений в онлайн-покупках. Нам необходимо выбрать тип базы данных, которая будет использоваться для хранения ИВИ.
2. Модель данных: модель данных формируется в зависимости от содержания вакансий и от размеров, которые необходимо учитывать. Модель исходных данных включает в себя структуры, заполняемые процессами ввода, и результаты этапа предварительной обработки. Модель этапных данных включает в себя структуры, используемые для хранения результатов этапа категоризации. Метаданные включают все структуры, используемые для хранения онтологий, таксономий и коллекций услуг. Презентационная модель данных включает все структуры, доступные пользователям для аналитических целей.

Процесс получения данных. Определив источники ИВИ, мы должны решить, как собирать данные. Наиболее распространенными методами получения данных являются скрейпинг, краулинг и прямой доступ к данным, предоставляемый поставщиком, который может предложить ИПП-соединение или напрямую предоставить данные в согласованном формате. ИПП доступ значительно облегчит настройку системы, ускорит и сделает более эффективным весь технический процесс. В любом случае, желательно связаться с веб-мастерами выбранных источников, чтобы проинформировать их о проекте и принять решение о наиболее подходящем подходе для загрузки данных.

Предварительная обработка. Это критически важный этап, после сбора ИВИ. На этом этапе данные очищаются от искажений и подготавливаются к этапу извлечения информации. Он также включает в себя несколько шагов для получения качественных данных. Одним из вопросов, возникающих на этом этапе, является дедупликация, т.е. когда ИВИ должен рассматриваться как дубликат другого ИВИ?

Извлечение информации и ее классификация

1. Таксономии: Для классификации переменных необходимо выбрать таксономии. Существуют различные варианты: для большинства переменных существуют стандартные классификации, принятые в странах и разработанные в течение многих лет (например, ESCO для профессий и навыков, NUT для территориального уровня, NACE для отраслей). Если интересующий нас РТ охватывает более одной страны, важно обеспечить единые таксономии для всех переменных.
2. Методы извлечения информации и присваивание таксономий: При принятии решения о том, какие методы извлечения информации использовать, необходимо изучить состояние дел, чтобы сделать выбор между системами на основе онтологий или машинным обучением (контролируемым или неконтролируемым).

ИПЗ и презентационная зона. Последний этап - принятие решения о том, как информация будет отображаться и визуализироваться. Затем необходимо будет

разработать модель представления данных, пути навигации и панель индикаторов, а также оценить типы использования.

4.2 Идеи для пилотных проектов

В этом разделе мы подводим основные итоги деятельности рабочей группы, обсуждавшей использование Больших данных для ИРТ в рамках семинара " Раскрыты потребности в навыках: Работа с неопределенностью" в рамках семинара "Навыки будущего: Управление переходом", организованного ЕФО в Турине 21 и 22 ноября 2018 г.

Семинар, наряду с деятельностью рабочих групп, предоставил прекрасную возможность для обсуждения того, чем можно поделиться в плане идей, опыта, методов, сложных вопросов и реальных проблем, а также предоставил шанс содействовать взаимному обогащению исследователей и специалистов, работающих в области РТ и Больших Данных, для определения новых направлений, действий и идей для проектов по этой теме. В частности, дискуссии между участниками были сосредоточены в основном на четырех различных темах: (i) преимущества для граждан с точки зрения карьерного роста; (ii) роль Больших Данных для ИРТ; (iii) расширение использования Больших Данных в развивающихся и переходных странах; и (iv) действия для проектов. Ниже мы приводим основные аспекты, связанные с каждой темой, которые возникли в ходе деятельности рабочей группы и которые мы учитывали при создании идей для пилотных схем.

Преимущества для граждан с точки зрения карьерного роста. Использование больших данных для поддержки карьерного роста - одно из самых интересных (и прорывных) применений ИРТ. Это позволит обеспечить такие преимущества для граждан, как:

- способность проводить анализ пробелов между навыками, которыми владеют граждане, и навыками, требуемыми на РТ;
- возможность классификации профессий по стандартной таксономии (такой как ESCO, O*NET или SOC), что позволяет сравнивать одну и ту же работу в разных странах, тем самым поддерживая мобильность между государствами-членами ЕС;
- создание инструмента для профориентационной деятельности;
- повышение осведомленности о том, что РТ быстро меняется, а также о важности обучения на протяжении всей жизни;
- предоставление гражданам возможности связать свой собственный путь обучения с карьерным ростом и формирующимися навыками.

Роль больших данных в ИРТ. Для того чтобы использовать Большие данные для ИРТ, необходимо рассмотреть следующие вопросы:

- необходимость в глубокой и детальной информации о спросе и предложении на РТ в каждой стране;
- необходимость сотрудничества между учреждениями для обмена данными, а также подписания соглашений между владельцами данных для обеспечения надежного и масштабируемого процесса сбора данных;
- необходимость совместного использования источников данных, таких как Интернет для сбора вакансий, исследования, с акцентом на неформальной/серой экономике, и обследования рабочей силы.

Расширение использования Больших Данных в развивающихся странах и странах с переходной экономикой. Использование больших данных для ИРТ было бы особенно полезно для развивающихся стран и стран с переходной экономикой, чтобы облегчить согласование спроса и предложения на специалистов, разработать карьерные траектории, чтобы лучше соответствовать ожиданиям специалистов, и сравнить внутренний РТ с

трансграничными странами, чтобы способствовать мобильности. В этих странах необходимо тщательно рассмотреть следующие вопросы:

- отсутствие доступа к административным данным и статистическим сведениям;
- недостаточная статистика о системе образования в этих странах;
- необходимость обзора веб-источников для оценки проникновения использования интернета для деятельности, связанной с РТ;
- необходимость повышения осведомленности о важности решений, основанных на данных, как для правительства, так и для статистических служб.

Действия по проектам. Учитывая отзывы, собранные во время семинара "Раскрыты потребности в навыках: Работа с неопределенностью" и наш опыт в использовании Больших данных для ИРТ, мы определяем три отдельных действия для проектов, которые могут помочь развивающимся странам и странам с переходной экономикой в использовании Больших данных для исследования РТ, а именно:

1. технико-экономическое обоснование для страны X для выявления, проверки и ранжирования интернет-источников;
2. создать систему сбора информации о РТ в режиме реального времени;
3. определить модели анализа данных для поддержки лиц, принимающих решения, в разработке и оценке политики.

СОКРАЩЕНИЯ

ИИ	Искусственный интеллект
ИПП	Интерфейс прикладного программирования
БА	Бизнес аналитика
СЕДЕФОП	Европейский центр развития профессионального обучения
e-CF	Европейская электронная система компетенции
ESCO	Европейские навыки, компетенции, квалификации и профессии
ESCoE	Центр передового опыта в области экономической статистики
ESS	Европейская статистическая система
ЕФО	Европейский Фонд Образования
ИПЗ	Извлечение, преобразование и загрузка
ЕС	Европейский Союз
HDFS	Распределенная файловая система Hadoop
ИКТ	Информационно-коммуникационные технологии
ИС	Информационная система
МСКЗ	Международная стандартная классификация занятий
KDD	Извлечение информации из данных
ЛРД	Латентное распределение Дирихле
РТ	Рынок труда
ИРТ	Аналитика/информация по рынку труда
ИСРТ	Информационная система рынка труда
NACE	Статистическая классификация экономической деятельности в Европейском сообществе
NICE	Национальная инициатива по образованию в области кибербезопасности
NoSQL	Нереляционные базы данных
NUT	Номенклатура территориальных единиц для статистики
ОЭСР	Организация экономического сотрудничества и развития
ИВИ	Информация о вакансиях в интернете
УНС	Управление национальной статистики
SQL	Стандартный язык запросов
SOC	Стандартная профессиональная классификация

UK	Великобритания
США	Соединенные Штаты Америки
XML	Расширяемый язык разметки текста
YARN	Операционная система, для обработки Больших Данных

ССЫЛКИ

[1] Министерство образования и навыков Великобритании, *LMI Matters*!, 2004.

[2] Комиссия Соединенного Королевства по вопросам занятости и квалификации, *Значимость PT*, 2015. По состоянию на март 2019 г.: https://goo.gl/TtRwvS

[3] Мецканцаника М.,. и Меркорио Ф., «Большие данные для анализа рынка труда: вводное руководство», в *Энциклопедии технологий Больших Данных*, Международная издательская компания, 2018, параграф. 1–11.

[4] ЕФО (Европейский фонд образования), *Информационная система рынка труда*, 2017.

[6] Управление национальной статистики Великобритании, *NOMIS: Национальная онлайн система информации о рабочей силе Великобритании,* 2014.

[7] Джонсон Э., *Могут ли большие данные спасти информационные системы рынка труда?*, концептуальная записка РТИ Пресс №PB-0010-1608, РТИ Пресс, Парк Исследовательский треугольник, СК, рассмотренный, 2017.

[8] Фрей К.Б. и Осборн М.А., 'Будущее занятости: Насколько восприимчивы рабочие места к компьютеризации?', *Технол. Прогноз. Соц. изменения*, том 114, дополнение С, стр. 254–280,2017.

[9] UNECE (Европейская экономическая комиссия Организации Объединенных Наций), *Использование административных и вторичных источников для официальной статистики: Руководство по принципам и практике*, 2015.

[10] Министерство труда и благосостояния Италии, *Годовой отчет о системе СО*, 2012. Последнее посещение в марте 2019 года: http://goo.gl/XdALYd

[11] Пеннек С., и др., 'Использование административных данных в статистических целях, *Экон. Рын.труда.Изд*, том 1, № 10, 2007, стр. 19.

[12] Бозелли Р., Чезарини М., Меркорио Ф. и Мецканцаника М., 'Основанная на моделях оценка деятельности по обеспечению качества данных в KDD', *Инф. Процесс. Менедж.*, том 51, № 2, 2015, стр. 144–166.

[13] Ванг Р.У. и Стронг Д.М., «За пределами точности: Что качество данных значит для потребителей данных», *Ж. Менедж. Инф. Сист.*, том 12, № 4, 1996, стр. 5–33.

[14] МакАфи А., Бринйольфссон Е., Давенпорт, Т.Х., Патил Д. и Бартон Д., 'Большие данные: революция в управлении', *Харв. Биз. Ред.*, том 90, № 10, 2012, стр.60–68.

[15] Бозелли Р., Чезарини М., Меркорио Ф. и Мецканцаника М., 'Классификация онлайн объявлений о работе с помощью машинного обучения', *Будущее покол. Компьют. Сист.*, том 86, 2018,стр.319–328.

[16] Файад Ю., Пятецкий-Шапиро, Г. и Шмид, Р., 'Процесс KDD для извлечения полезных знаний из больших объемов данных', *Коммун. АВМ*, том 39, № 11, 1996, стр.27–34.

[17] Редман Т.С., 'Влияние низкого качества данных на типичное предприятие', *Коммун.*

ABM, том41, № 2, 1998, стр.79–82.

[18] Ботсок М., Огиветский, В. и Хеер Дж., 'D^3 документы, основанные на данных, *IEEE Транз. Виз.Компьют. Граф*, том 17, № 12, 2011, стр.2301–2309.

[19] Бао Ф. и Чен Дж., 'Визуальная основа для больших данных на d3. js', в *Электронике, Компьютере и приложениях, 2014 семинар IEEE*, 2014, стр.47–50.

[20] Европейская комиссия, *Новый импульс для европейского сотрудничества в области профессионального образования и обучения в поддержку стратегии "Европа 2020*, КОМ(2010) 296, Брюсель, 2010. Посл. посещ. март 2019 г.: https://goo.gl/Goluxo

[21] Европейская комиссия, *Новая программа развития навыков для Европы*, КОМ(2016) 381/2, 2016.

[22] ЕвроСтат, *Проект ESSNet Большие Данные*, Европейская комиссия, Страсбург, 2016. Посл. посещ. март 2019 г: https://goo.gl/EF6GtU

[23] СЕДЕФОП, *Информация о требованиях к квалификации на рынке труда в режиме реального времени: технико-экономическое обоснование и рабочий прототип*, СЕДЕФОП регистрационный номер AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14, Уведомление о контракте 2014/S 141-252026 от 15/07/2014, 2014. Посл. посещ. март 2019 г: https://goo.gl/qNjmrn

[24] СЕДЕФОП, *Информация о требованиях к квалификации на рынке труда в режиме реального времени: Настройка системы ЕС для онлайн-анализа вакансий AO/DSL/VKVET-GRUSSO/Real-time LMI 2/009/16. Уведомление о контракте - 2016/S 134-240996 от 14/07/2016*, 2016. Посл. посещ. март 2019 г.: https://goo.gl/5FZS3E

[25] Комиссия по трудоустройству и навыкам Великобритании, *ИРТ для всех*, 2015. Посл. посещ. март 2019 г.:www.lmiforall.org.uk/

[26] Арнц М., Грегори Т. и Зиран У., 'Риск автоматизации для рабочих мест в странах ОЭСР',2016.

[27] Институт инноваций Брукфилда, *Лучше, Быстрее, Сильнее: Максимальное использование преимуществ автоматизации для предприятий и населения Онтарио*, 2018. Посл. посещ. март 2019 г.: https://brookfieldinstitute.ca/report/better-faster-stronger/

[28] Леопольд Т.А., Ратчева В. и Шахири С., 'Будущее рабочих мест: Занятость, навыки и стратегия трудовых ресурсов для четвертой промышленной революции, доклад "Глобальный вызов" на *Всемирном экономическом форуме*, *Женева*, 2016.

[29] Мецканцаника М., Меркорио Ф. и Коломбо Е., 'Цифровизация и автоматизация: выводы из', *Дев. Ски. шан. Концепции мирового труда Измер. Примен. данных. Регистр. Монит местного рынка труда..Евр.*, 2018, стр. 259.

[30] ESCoE (Центр передового опыта в области экономической статистики), *Использование административных и больших данных для улучшения статистики рынка труда*, 2019. Посл. посещ. март 2019 г.: www.escoe.ac.uk/projects/using- administrative-big-data-improve-labour-market-statistics/

[31] Стоунбаркер М., 'База данных SQL и. база данных NoSQL', *Коммун. ABM*, том 53, №4, 2010,стр.10–11.

[32] Дин Дж. и Гемават С., 'MapReduce: упрощенная обработка данных на больших кластерах',

Коммун. ABM, том 51, № 1, 2008, стр.107–113.

[33] Алпайдин, Е., *Введение в машинное обучение*, MIT пресс, 2009 г.

[34] Коломбо, Э., Меркорио, Ф. и Меццанзаника, М., 'Применение инструментов машинного обучения на веб-вакансиях для анализа рынка труда и навыков' в книге *"Терминатор или Джетсоны? Экономические и политические последствия искусственного интеллекта,* 2018**.**

[35] Себастиани, Ф., 'Машинное обучение в автоматизированной категоризации текстов', ABM Комп. обслед. CSUR, том. 34, No 1, 2002, стр.1-47.

[36] Блей, Д.М., Нг, А.Й. и Джордан, М.И., 'Латентное распределение Дирихле', *Инф. сист. аналит. раб.*, том. 3, 2003, стр.993–1022.

[37] Бозелли, Р. и др., "WoLMIS: система анализа рынка труда для классификации вакансий в Интернете", *Инф. сист. аналит. раб.*, том. 51, № 3, 2018, стр.477–502.

www.etf.europa.eu