ПРИКЛАДНЫЕ ОНТОЛОГИИ ПРОЕКТИРОВАНИЯ

УДК 004.85 DOI: 10.18287/2223-9537-2020-10-1-34-49

Модели и методы индивидуализации электронного обучения в контексте онтологического подхода

Д.И. Муромцев

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия

Аннотация

Рассматривается индивидуализация электронного обучения (ЭО) как совокупности процессов создания, развития, использования и утилизации цифрового контента и данных ЭО, описаны соответствующие онтологические модели и методы. Описан технологический стек для построения и реализации индивидуальных траекторий обучения, приведены примеры существующих систем, которые полностью или частично используют указанный стек технологий. Для эффективной обработки образовательных материалов и данных, формируемых системами управления обучением, предложена архитектура, позволяющая осуществить семантическое аннотирование и выделение в данных слоёв концептов различной степени абстракции. Эти слои включают: верхнеуровневые абстракции моделирования, общие концепты учебных материалов и образовательного процесса, специфические концепты для доступа и интеграции данных системы ЭО в терминах предметной области. Впервые в качестве формальной основы для индивидуализированного ЭО предложено использовать семантические модели, включающие аппарат векторных представлений графов знаний, который позволяет эффективно обрабатывать большие и сложные структуры данных, а также обладает гибкостью и выразительностью онтологического подхода. Последовательно рассмотрены основные аспекты, связанные с индивидуализацией в системах ЭО, в том числе: существующие технологии и онтологии для ЭО, моделирование индивидуальной траектории, семантическое аннотирование образовательных материалов, способы оценки знаний в индивидуализированном обучении, а также онтологическое моделирование когнитивного профиля обучаемого.

Ключевые слова: *электронное обучение, индивидуализация, онтологии, графы знаний, векторное представление, семантические технологии.*

Цитирование: *Муромцев, Д.И.* Модели и методы индивидуализации электронного обучения в контексте онтологического подхода / Д.И. Муромцев // Онтология проектирования. – 2020. – Т. 10, №1(35). - С.34-49. – DOI: 10.18287/2223-9537-2020-10-1-34-49.

Введение

Некоторые результаты исследований проблем индивидуализации электронного обучения (ЭО) изложены в работах [1-6]. Важно отметить, что в процессах ЭО или управления знаниями взаимодействие происходит не между человеком и системой управления, а между цифровым следом и цифровыми артефактами обучения и интеллектуальной системой управления знаниями. Эта особенность порождает целый ряд новых проблем, таких как поиск эффективных способов представления данных индивидуализированного обучения, моделирование цифровых артефактов, порождаемых системами ЭО, интеллектуализация анализа образовательных данных, автоматизация процессов адаптации и актуализации образовательного контента и ряд других. Процессы индивидуализации ЭО являются комплексными, затрагивают формирование индивидуальных компетенций у обучающихся и относятся к системам

управления обучением (*Learning Management System, LMS*), добавляя им требования по синтезу индивидуализированного контента и обеспечению автоматического подбора наиболее релевантных (индивидуализированных) средств оценки знаний.

Понятие индивидуализации имеет несколько корней: это и индивидуализированные методики [1-3], и индивидуализированные технологии [4-6]. В контексте данной работы именно второй аспект имеет наибольший интерес, так как напрямую связан с процессами создания, развития, использования и утилизации цифрового контента и данных ЭО.

С точки зрения конечного результата индивидуализированные процессы ЭО играют огромную роль, так как являются основой для функционирования и устойчивого развития экосистемы цифровой экономики и цифрового общества. Тенденции в промышленном производстве по переходу от массового к персонализированному, интеллектуализация всех сфер человеческой деятельности, роботизация рутинного труда формируют новые профессии, предполагающие наличие специалистов с уникальными наборами знаний.

Создание систем индивидуализированного ЭО и управления знаниями предполагает использование специальных технологий, баз знаний (БЗ) и интеллектуальных алгоритмов анализа данных. В работе анализируется применение к решению поставленных задач таких интеллектуальных технологий, как графы знаний, онтологии и машинное обучение (МО).

1 Технологическая основа для построения БЗ

Сегодня основной тенденцией создания электронного образовательного контента является технология МООК (массовых открытых онлайн-курсов, *Massive Open Online Courses, MOOC*). Данная технология ориентирована на массовое обучение. Это порождает несоответствие между потребностью обучающихся и возможностями (содержанием) электронных образовательных курсов (и образовательных программ в целом).

Решение задачи индивидуализации как систем ЭО, в частности, так и систем управления знаниями в различных областях может быть достигнуто за счёт превращения баз данных (БД) и хранилищ медиа-контента в системах *LMS* в полноценные БЗ, предоставляющие соответствующие модели репрезентации знаний и методы логического вывода и интеллектуального поиска. Для этого необходима технологическая база для интеллектуализации задач управления контентом и процессами обучения. Современным подходом при построении БЗ является стек семантических технологий. Его основой является язык *The Resource Description Framework* (*RDF*), представляющий собой семантическую графовую модель и предназначенный для репрезентации полуструктурированных данных о фактах реального мира или абстракциях. Стандартизованный консорциумом *W3C* [7] *RDF* специфицирует архитектуру, синтаксис и семантику, а также базовый словарь *RDF Schema* (*RDFS*) для построения моделей предметных областей (ПрО) [8, 9].

Основным элементом языка *RDF* является тройка вида <субъект, предикат, объект>, где субъекты и объекты могут быть уникальными сущностями или наименованными сущностями для представления более сложных конструкций (вложенных подграфов, множеств и др.). Каждая сущность имеет свой универсальный и уникальный идентификатор ресурса — *URI* (*Uniform Resource Identifier*). *URI* необходимы для того, чтобы была возможность ссылаться на описываемые сущности. Например, идентификатор Университета ИТМО может содержать *http://en.ifmo.ru/ITMO_University*, где префикс *http://en.ifmo.ru/* — адрес в Интернете. Для удобства полные *URI* можно сокращать в префиксы и использовать запись *prefixName:Entity*. Например, термины *RDF* имеют стандартный префикс *rdf:,* что заменяет *<http://www.w3.org/1999/02/22-rdf-syntax-ns#>*.

Неименованные вершины (*blank nodes*) — анонимно заданные сущности без идентификатора или литерала, могут содержать другие отношения и значения. Они используются для описания сложных предикатов и других конструкций в процессе моделирования. Объектами могут быть также простые строковые литералы, представляющие значения атрибутов субъектов. Предикаты обозначают отношения между субъектами или объектами или свойства (атрибуты) субъектов. Формально тройки можно представить как элементы $x_{ijk} = (e_i, r_k, e_j)$, где $E = \{e_1, \ldots, e_{Ne}\}$ — множество всех сущностей (субъектов или объектов), а $R = \{r_1, \ldots, r_{Nr}\}$ — множество всех связей (отношений) на графе.

Множество троек формирует *RDF*-граф, который, в свою очередь, можно определить формально [10]. Пусть U, B, L — непересекающиеся бесконечные множества *URI*, неименованных вершин и литералов соответственно. Тогда *RDF*-граф G можно определить как направленный помеченный мультиграф $G = (E, R, \Sigma, L)$, где:

$E \subset (U \cup B \cup L)$ - конечное множество *RDF*-термов, соответствующих узлам графа;

$R \subseteq E \times E$ - конечное множество дуг, связывающих *RDF*-термы;

$\Sigma \subset$ U - множество уникальных меток, определённых с помощью *URI*;

$L: R \to 2^{\Sigma}$ - отображение дуг на множество меток.

Язык *RDF* используется для описания лишь базовых элементов графа знаний, например, «нечто имеет такой-то тип» или «что-то связано с чем-то», но не позволяет определять классы (группы имеющих сходные атрибуты) или настраивать множества допустимых значений для атрибутов. Расширение *RDFS* для языка *RDF* вводит дополнительные предикаты для построения более сложных моделей, включая иерархии. Это такие предикаты, как *rdfs:Class* для определения классов, *rdfs:Literal* для определения литералов, *rdfs:subClassOf* и *rdfs:subPropertyOf* для определения иерархических отношений. Дополнительно *RDFS* позволяет определять области определения и области значений для отношений между сущностями, а также ряд других возможностей.

Для построения сложных моделей ПрО, использующих в качестве формальной семантики логические выражения, используют язык *Web Ontology Language* (*OWL*) [11], являющийся расширением языка *RDFS*. Существует достаточно большое количество онтологий, разработанных с помощью языков *RDF, RDFS* и *OWL*, которые могут быть полезны при создании индивидуализированных систем ЭО, в том числе следующие.

- *The Academic Institution Internal Structure Ontology* (*AIISO*) [12] является онтологией, описывающей внутреннюю организационную структуру образовательного процесса. *AIISO* предоставляет классы и свойства для описания курсов, модулей, практических и теоретических учебных материалов.
- *The Bibliographic Ontology* (*BIBO*) [13] является онтологией, описывающей библиографические ресурсы. Словарь *BIBO* может использоваться для описания рекомендованной литературы, научных публикаций, методических пособий и монографий.
- *The Ontology for Media Resources* (*MA-ONT*) [14] является онтологией, описывающей медиа ресурсы. С помощью классов и свойств *MA-ONT* производится связывание лекций с видеоматериалами.
- Онтология для описания учебных материалов *TEACH* (*Teaching Core Vocabulary*) [15] является облегчённым словарем, позволяющим преподавателям связывать объекты электронных курсов.
- Онтология *FOAF* (*Friend of a Friend*) [16] определяет некоторые выражения, используемые в высказываниях о ком-либо, например: имя, пол и другие характеристики.

Существуют также системы ЭО, построенные на основе семантических технологий, например, следующие.

- Проект *Metacademy* [17], представляющий собой платформу для открытого персонализированного образования. В основе обучения в данной системе лежат концепты ПрО. Пользователь может составить учебный курс или его дорожную карту на основе концептов, которые он хочет изучить. Все учебные материалы в системе хранятся в онтологиях, что позволяет пользователям осуществлять навигацию по теоретическим материалам. В *Metacademy* весь учебный материал, курсы, лекции, книги связаны друг с другом с помощью концептов ПрО.
- Проект *SlideWiki* [18], в рамках которого создана платформа для создания презентаций для учебных курсов. С помощью семантических технологий платформа позволяет повторно использовать уже опубликованные слайды презентаций, аннотировать дополнительной информацией концепты на слайдах и поддерживать множество языков для одного учебного курса [19].

Однако, несмотря на гибкость и выразительных моделей и широкие возможности по интеллектуализации систем ЭО, перевод существующих курсов в семантический формат является необходимым, но недостаточным условием для достижения целей индивидуализации. Одним из главных препятствий выступает грануляция контента, которая чаще всего отражает структуру курса и виды контента, но недостаточна для построения индивидуальных траекторий обучения (ИТО). Кроме того, в системе отсутствуют модели обучаемого и его знаний, приобретаемых в процессе изучения контента. Можно сказать, что без такой модели построить адекватную систему индивидуализации не предоставляется возможным. Она необходима как для выстраивания персонализированных рекомендаций по прохождению курсов, так и для организации адекватной оценки полученных знаний с учётом индивидуальных потребностей обучаемого. При формировании тестовых и проверочных заданий из всего множества оценочных средств необходимо выбирать только те, которые связаны с ИТО.

Как правило, учебный курс организован линейно и состоит из множества модулей и тем внутри модулей (см. рисунок 1).

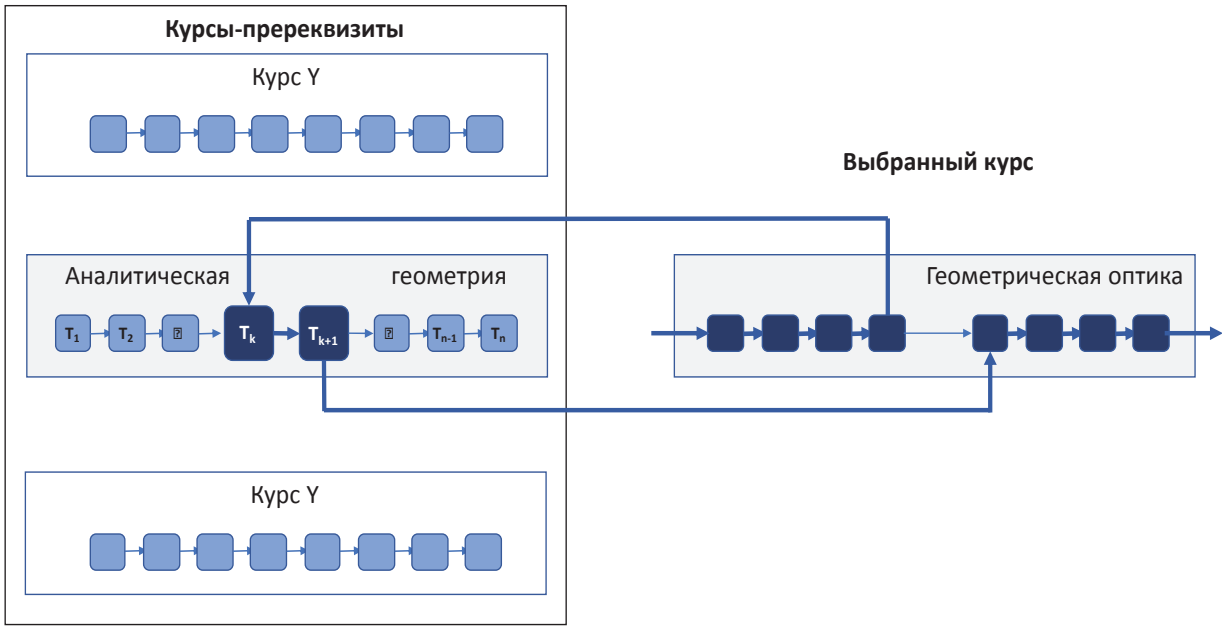

Рисунок 1 - Множество концептов в курсе и позиции точек перехода (изучаемых терминов) от курса к курсу

Выраженные явно или неявно семантические зависимости между модулями задаются в так называемых пререквизитах курса. Т.е. можно выстроить цепочку модулей или курсов, которые необходимо изучить, в зависимости от конкретного выбора обучающегося. Так, ес-

ли обучающийся выбрал вводный курс по геометрической оптике и для понимания, например, явления интерференции ему необходимо также изучить правила сложения векторов, которые вводятся в курсе аналитической геометрии. Такую информацию можно найти в пререквизитах курса по оптике. Но включение в ИТО всего курса по аналитической геометрии или даже модуля по векторной алгебре будет избыточным для данного обучающегося. Достаточно лишь ограничиться необходимыми лекциями из курсов пререквизитов, иначе обучающийся может получить слишком перегруженную ИТО. Аналогичные рассуждения верны и для выбора тестов и заданий из фонда оценочных средств для оценки знаний в процессе прохождения по ИТО. На рисунке 1 показаны гранулы курса и ИТО приведённого примера. Каждая гранула соответствует одному из изучаемых терминов T_i. Из рисунка видно, что точки перехода от курса к курсу могут не совпадать с границами модулей.

Важно отметить, что ИТО не является статической. По ходу освоения элементов курса она может изменяться, дополняясь новыми концептами, если это необходимо для понимания материала и достижения целей обучения. Процесс построения ИТО имеет рекурсивный характер. Так, если в рассмотренном примере при изучении темы «сложение векторов» потребуется изучить понятие «вектор», то включение этой гранулы в ИТО может быть выполнено аналогичным способом, как это было проделано с темой «интерференция».

2 Индивидуальные траектории обучения

Многие *LMS* предлагают средства управления контентом для построения ИТО, позволяя создавать индивидуальную последовательность курсов. Однако есть серьёзное препятствие, связанное с формальным учётом и содержательным анализом данных, необходимых для индивидуализации ЭО.

Управление компетенциями в существующих системах ЭО основано на предположении успешного усвоения курсов, без учёта индивидуальных способностей и интересов обучаемого (и часто без обратной связи). В этом смысле системы являются линейными, а само обучение представляет собой монотонный процесс по достижению целей обучения. Между тем, это грубое приближение к реальным образовательным процессам. Построение ИТО в подавляющем большинстве случаев может быть немонотонным, предполагая возникновение новых точек, через которые должна пройти ИТО. Заранее предположить все возможные точки ИТО сложно по нескольким причинам.

- ИТО возникает в результате проекции друг на друга нескольких онтологических моделей (модели курсов, модели оценки знаний, когнитивная модель обучающегося и др.). Построение полного пространства поиска на основе этих моделей может оказаться вычислительно сложной задачей с множеством противоречий, что требует применения различных эвристик для его оптимизации.
- В процессе обучения производится изменение некоторых из этих моделей. Например, когнитивная модель обучающегося пополняется новыми сущностями по мере продвижения по ИТО, сам обучающийся может вносить коррективы, уточняя свои потребности, возможно изменение моделей курсов, т.к. процесс обновления контента в общем случае может происходить в любое время; модели формирования ИТО также могут меняться по мере накопления данных и изменении образовательного контента, и т.п.
- На процесс обучения могут влиять внешние факторы, например изменяющиеся запросы рынка, различные экономические, социальные и пр. факторы, связанные с образованием.

Индивидуализацию ЭО следует рассматривать не как управление образовательным процессом, а как новую технологию по созданию различных систем интеллектуализации и управления в образовании, которая включала бы такие инструменты и методы как:

- методы онтологического инжиниринга, в том числе автоматическое построение и пополнение онтологий на основе мультимодальных данных;
- методы МО;
- средства семантического анализа и поиска;
- средства выработки рекомендаций и др.

Современные *LMS*, помимо накопленного контента, как правило предоставляют необходимый базовый технологический уровень для построения на их основе индивидуализированных систем обучения, т.к. для хранения данных используются БД, в том числе графовые и *NoSQL* БД, модели репрезентации курсов не являются жёсткими и допускают внесение изменений в их структуру и состав, ведётся достаточно подробное журналирование поведения пользователя в системе, что позволяет выполнять детальный анализ этого поведения. Не представляет труда подключение дополнительных сервисов, что позволяет интегрировать в *LMS* элементы новой технологии и сохранить преемственность в образовательных процессах.

Для удовлетворения описанных требований к системе индивидуализации ЭО необходим стек технологий, обеспечивающий интероперабельность и бесшовную интеграцию различных компонент таких систем. Типовая архитектура, реализующая такой стек технологий, должна включать несколько уровней.

- Интеграционный уровень:
  - провайдеры данных к БД *LMS*;
  - *API* к внешним источникам данных.
- Уровень управления данными:
  - хранилище метаданных;
  - модели МО для семантического анализа логов *LMS* и создания или пополнения онтологий;
  - шаблоны для построения семантических запросов.
- Уровень анализа данных и интеллектуальных сервисов:
  - онтологии курсов, учитывающие индивидуализацию;
  - когнитивные онтологические модели обучающегося;
  - правила порождения ИТО на основе онтологий.
- Уровень приложений и интерфейсов:
  - рекомендательные вопросно-ответные подсистемы для взаимодействия с обучающимся;
  - интерактивная визуализация ИТО.

3 Семантическое аннотирование образовательных материалов и данных результатов обучения

После того, как для подсистемы индивидуализации обеспечен доступ ко всем данным ЭО, необходимо выполнить семантическое аннотирование данных. Этот процесс относится к структурированным данным, например, выгрузкам из БД; полуструктурированным, например, логам системы; неструктурированным, например, текстовому контенту или другой текстовой информации, которая может появиться в системе (эссе и ответы обучающихся на тесты, обсуждения, диалоги и пр.). Для различных видов данных используются различные методы аннотирования. Во всех случаях в качестве результата необходимо получить определённое множество объектов, отражающих ход процесса обучения (пройденные элементы курса, достижения и качества обучающегося и т.п.), а также связи между этими объектами.

Семантические аннотации в общем случае можно определить как некоторые ссылки на метаданные, которые можно выразить через элементы онтологии. В процессе наполнения такой модели экземплярами реальных данных получается граф знаний. Для решения многих сложных задач, в том числе и для задачи индивидуализации ЭО, онтология должна представлять собой развитую концептуальную схему или референтную модель.

Структура референтной модели индивидуализации ЭО представлена на рисунке 2. Такая структура включает слои концептов различной степени абстракции:

- верхнеуровневые абстракции для моделирования ИТО обучающегося;
- общие концепты учебных материалов и образовательного процесса;
- специфические концепты для доступа и интеграции данных системы ЭО в терминах ПрО.

Верхние два уровня практически не зависят от специфики ПрО и конкретных *LMS*. В то же время нижний уровень может быть адаптирован под специфические требования или даже разбит на несколько подуровней при необходимости. В случае систем индивидуализированного ЭО подобные модификации нижнего уровня онтологии могут выполняться достаточно часто, т.к. сам процесс индивидуализации предполагает построение отдельной модели для каждого обучающегося на основе данных, которые им используются или порождаются в процессе обучения.

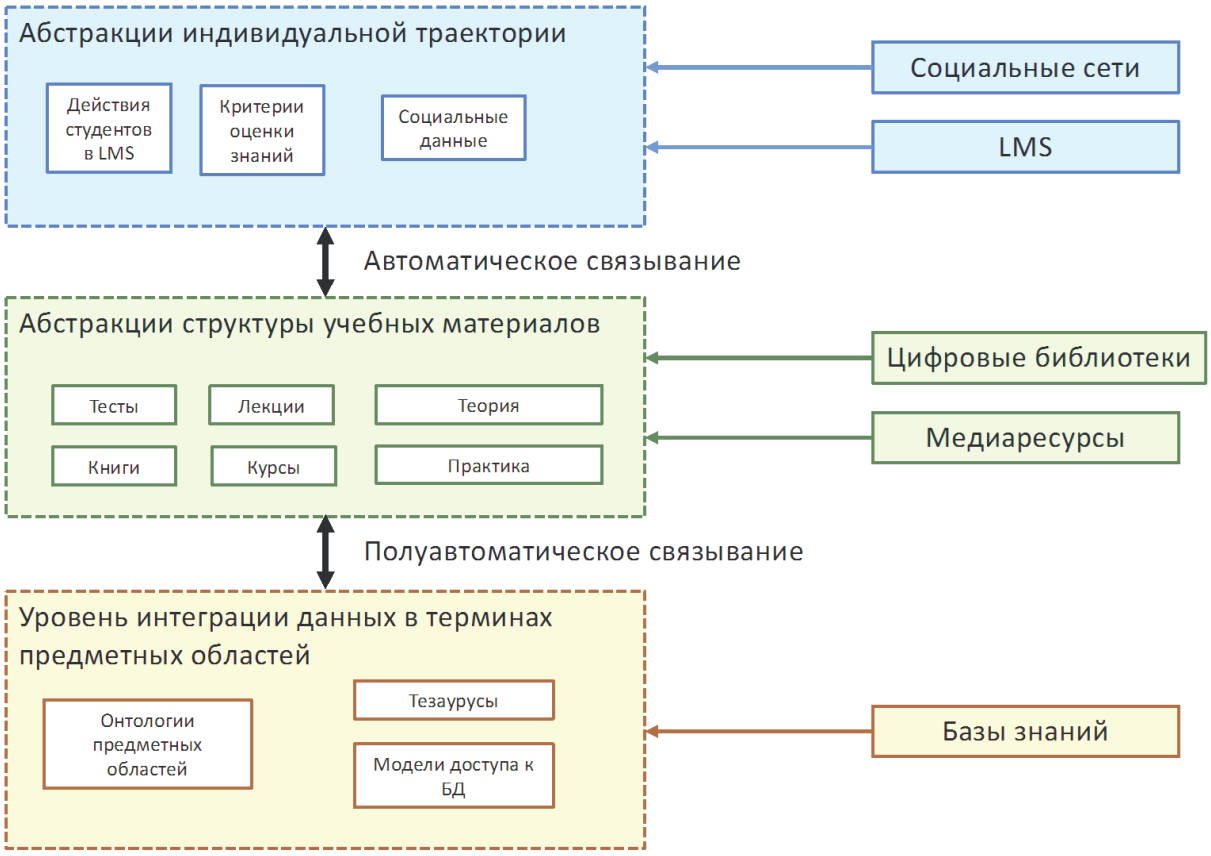

Рисунок 2 - Структура референтной модели индивидуализации электронного обучения

Связывание абстракций индивидуальной траектории и структуры учебных материалов – это часто выполняемая операция над хорошо структурированными данными, т.к. для каждого обучаемого она может выполняться многократно. Также при выполнении этой операции необходимо обеспечить высокий уровень объективности результатов. Эти факторы формируют требования по полной автоматизации операций связывания между этими уровнями,

исключая какое-либо влияние администратора *LMS* или эксперта на результат. Напротив, связывание с уровнем БЗ ПрО предполагает работу со слабоструктурированными и неструктурированными данными и выполняется однократно для каждого курса или его редакции. Точность построения предполагаемы связей будет выше при участии эксперта ПрО в данном процессе, соответственно полная автоматизация представляется нецелесообразной.

Модификация онтологий использует комплекс методов МО и относится к задачам *Information Extraction* [20].

- Распознавание/извлечение именованных сущностей (*Named Entity Recognition/Extraction*) — разграничение позиций упоминаний сущностей во входном тексте. Например, в предложении «Пьер Кюри открыл пьезоэлектричество.» подчёркнутый текст является упоминанием именованных сущностей.
- Связывание/снятие омонимии сущностей или семантическое аннотирование (*Entity Linking/Disambiguation, Semantic Annotation*) - ассоциирование упоминаний сущностей с подходящим и однозначным идентификатором в БЗ. Например, связывание «Пьер Кюри» с сущностью *Q37463* в БЗ *wikidata*.
- Извлечение терминов (*Term Extraction*) — извлечение основных фраз, которые обозначают концепты, релевантные к выбранной ПрО и описанные в корпусе, иногда включая иерархические отношения между концептами. Например, выявление в тексте про МО, что «нейронная сеть» или «k-средних» являются важными концептами в ПрО. Дополнительно можно определить, что оба концепта являются уточнением понятия «искусственный интеллект», а также, что они могут быть связаны с определённым подразделом БЗ.
- Извлечение ключевых слов/фраз (*Keyword/Keyphrase Extraction*) - извлечение основных фраз, которые позволяют категоризировать тематику текста (в отличие от извлечения терминов, задача извлечения ключевых фраз заключается в описании именно текста, а не ПрО). Ключевые фразы также могут быть связаны с БЗ.
- Тематическое моделирование/классификация (*Topic Modeling, Classification*) — кластеризация слов/фраз, которые часто встречаются совместно в сходном контексте. Эти кластеры затем ассоциируются с более абстрактными темами, с которыми связан текст.
- Маркирование/идентификация темы (*Topic Labeling/Identification*) — для кластеров слов, идентифицированных как абстрактные темы, извлечение одиночного термина или фразы, наилучшим образом характеризующей эти темы. Например, определение, что тема, состоящая из {“машинное обучение”, “выборка”, “точность классификации”, “градиентный спуск”} наилучшим образом характеризуется термином «машинное обучение» (которое может быть связано, например, с концептом *Q2539* в *wikidata*).
- Извлечение отношений (*Relation Extraction*) — извлечение потенциальных n-арных отношений из неструктурированных или полуструктурированных (таких как *HTML*-таблицы) источников. Например, из предложения «Пьер Кюри открыл пьезоэлектричество.» можно извлечь открыл (Пьер Кюри, пьезоэлектричество). Бинарные отношения могут быть интерпретированы как *RDF* тройки после связывания предикатов-отношений с соответствующими свойствами в БЗ (таким как *discoverer or inventor* (*P61*)).

4 Семантическая модель индивидуализированного обучения на основе графа знаний

Модель индивидуализированного обучения представляет собой граф знаний изучаемых дисциплин (ГЗД), дополненный определёнными связями между теми концептами, которые входят в множество полученных знаний обучающегося. По совокупности таких связей для различных обучающихся можно судить о том, насколько ГЗД сбалансирован, какие суще-

ствуют предпочтения и тренды при изучении учебных материалов, становится возможной гармонизация учебных материалов и прогнозирование наиболее релевантных вариантов при построении ИТО.

Граф приобретённых знаний обучающегося (ГЗО), формируется из множества концептов описания ПрО и является подмножеством общего графа знаний (ГЗ) всех курсов. Изначально этот ГЗ пустой, и в него помещается некое стартовое множество концептов, подтверждённых входным тестом обучающегося, либо полученным в ходе изучения вводного курса. В этом ГЗ пропущены часть концептов и связей, т.к. объём освоенных знаний по дисциплинам изначально неполный. Он содержит «пробелы в знаниях», которые необходимо выявить в процессе сопоставления с общим ГЗ. Для этого необходимо найти проекцию ГЗО на ГЗД, восстановить пропущенные узлы в ГЗО и, как следствие, добавить связи в ГЗД для фиксации части пройденной ИТО, как это показано на рисунке 3. Поскольку пропуски концептов в ГЗО носят случайный характер, решение задачи поиска подграфа на графе может иметь большой процент ошибок.

Для построения алгоритма решения указанной задачи удобно использовать векторное представление ГЗ. Идея векторного представления [21] основана на дистрибутивном представлении латентных свойств сущностей ГЗ. Латентные свойства для каждой сущности (e_i) задаются с помощью вектора $e_i \in R^{H_e}$, где H_e соответствует числу возможных латентных свойств в модели. Например, возможное объяснение того, что библиотека *NumPy* используется для научных расчётов — это возможность быстрого прототипирования сложных вычислений. В данном примере утверждения, что «*NumPy* является эффективной библиотекой для научных расчётов», а «язык *Python* позволяет быстро прототипировать сложные алгоритмы вычислений» можно промоделировать с помощью вектора-столбца, содержащего эвристические значения для двух латентных свойств, которые также могут быть получены в результате последующего обучения модели:

$$e_{NumPy} = \begin{pmatrix} 0.9 \\ 0.2 \end{pmatrix}, e_{PyFastPrototype} = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix},$$

где вектор-строка (см. пояснения к рисунку 6) e_{i1} соответствует латентному свойству «эффективная библиотека для научных расчётов», а e_{i2} - латентному свойству «быстрое прототипирование сложных вычислений». Следует отметить, что в отличие от данного примера латентные свойства, выделенные в ходе обучения модели на реальных данных, как правило, сложно интерпретировать.

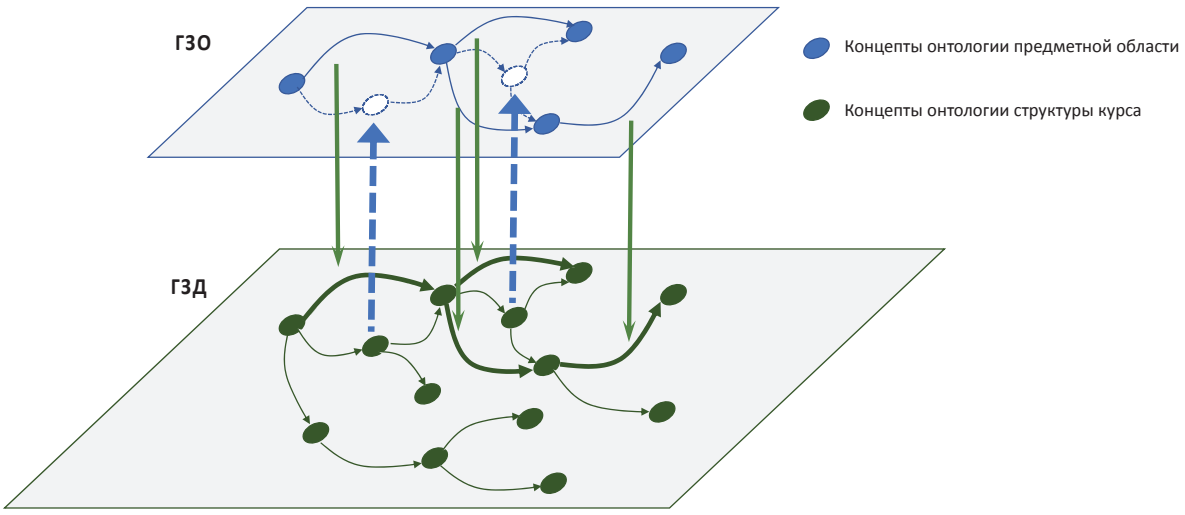

Рисунок 3 - Проекция графа приобретенных знаний обучаемого на граф знаний изучаемых дисциплин

5 Задача обнаружения пропущенных узлов в ГЗО

Для восстановления пропущенных узлов в ГЗО предлагается использовать комбинированный подход, основанный на совместном использовании векторных представлений триплетов из ГЗ и текстового корпуса, основанного на образовательном контенте. Подобные подходы доказали свою эффективность [22] в задаче дополнения ГЗ с использованием предварительно обученных языковых моделей для нейронной сети.

В рамках рассматриваемого подхода узлы и связи ГЗ рассматриваются как текстовые последовательности, состоящие из меток и текстовых описаний соответствующих троек. Например, триплет <*PyTorch, writen In, Python*> может быть представлен как «библиотека *PyTorch* написана на языке *Python*». Эта последовательность слов (токенов) содержит три группы, по одной для каждого элемента в триплете $\left[Tok^{S}_{1....N_s}, Tok^{P}_{1....N_p}, Tok^{O}_{1....N_O}\right]$. В приведённом примере группам токенов соответствуют $Tok^{S}_{1....N_s}$ — «библиотека *PyTorch*», $Tok^{P}_{1....N_p}$— «написана на» и $Tok^{O}_{1....N_O}$— «языке *Python*». Совокупно все эти группы токенов подаются на вход нейронной сети. Для каждой группы рассчитываются отдельные векторные представления. При этом первая и третья группы (субъект и объект) используют общий сегмент векторного представления.

После обнаружения пропущенных концептов ГЗО необходимо установить, какие существуют отношения между этими концептами, что позволит определить последовательность их изучения и перечень необходимых курсов или модулей, которые содержат контент, позволяющий наиболее полно изучить выбранный перечень понятий. Данный аспект особенно важен, так как одни и те же понятия могут излагаться в различных курсах и в различном объёме. Кроме того, содержание курсов периодически обновляется. Эти факторы делают неэффективными любые статические маппинги (проекции) концептов-понятий на концепты-структурные элементы курсов. Более перспективным подходом является метод, основанный на использовании векторного представления для восстановления связей в ГЗ [23]. В различных курсах для связывания терминов может использоваться различный контекст и различное множество связей, а сами связи могут иметь различные области определения и области значений, т.е. одни и те же концепты-термины имеют наборы отличающихся связей.

6 Индивидуализированная оценка результатов ЭО

Тесты и практические задания являются слишком грубыми и случайными для достоверной оценки знаний после прохождения ИТО. А генерация тестов для каждой ИТО потребует большого количества ресурсов и времени преподавателей в условиях массового использования контента. В то же время значительный объём ценной информации о результативности обучения может быть получен за счёт анализа цифрового следа обучающегося путём исследования логов и других цифровых артефактов, которые формируются во время работы в системе ЭО. Похожий подход получил широкое распространение в программной инженерии в задаче автоматизации тестирования программного кода. Применительно к оценке знаний, получаемых в ходе ИТО, цифровые артефакты могут включать следующие виды данных:

- логи поведения пользователя в системе (количество посещений отдельных страниц, время, проведённое на каждой из страниц, действия на страницах и пр.);
- действия пользователя с контентом (динамика просмотра видео, последовательность выполнения заданий, завершённость начатых действий и т.п.);
- активность при взаимодействии с другими пользователями и преподавателем через социальные сервисы (количество вопросов, количество ответов, регулярность отправки сообщений и пр.);

- текстовые данные, которые пишет обучающийся при использовании общего чата или электронной почты;
- данные модулей проверки знаний (закрытые и открытые тесты, результаты выполнения практических заданий и пр.).

Основные методы анализа перечисленных данных включают:

- извлечение именованных сущностей и отношений из текстовых данных;
- статистический анализ логов [24];
- адаптированные методы юнит-тестирования [25].

7 Средства онтологического моделирования для системы индивидуализированного ЭО

Описанные выше подходы к формированию и динамическому изменению ИТО в системе ЭО, когда курсы изучаются не целиком, а отдельными материалами, могут привести к нежелательной ситуации хаотического перемешивания изучаемых тем, так как фактически авторский замысел преподавателя, который был заложен при создании курса, игнорируется системой. Для предотвращения такой ситуации в системе необходим алгоритм, непрерывно отслеживающий семантическое сходство изученного множества тем и содержания дисциплин. Фактически этот алгоритм выполняет аппроксимацию структуры изученных понятий онтологиями курсов, что позволяет сделать формирование ИТО сфокусированным на предмете изучения за счёт присвоения более высоких весов тем материалам, которые содержатся в курсах, наиболее полно излагающих выбранные для ИТО темы.

Алгоритм оценки семантического сходства *GADES* (*a Graph-bAseD Entity Similarity*) [25] учитывает три аспекта, связывающих сравниваемые объекты ГЗ: иерархия, соседство и специфичность.

Анализ иерархического сходства основан на выделении на ГЗ G набора иерархических рёбер, для которых применяются методы вычисления подобия. Иерархические рёбра включают те связи ГЗ, имена свойств которых принадлежат иерархическому отношению, например, *rdf:type* или *rdfs:subClassOf*. Каждое такое отношение определяет связанную сущность через операцию обобщения (таксономические отношения или абстракции более высокого порядка) другого объекта. *GADES* использует иерархические методы подобия, такие метрики измерения таксономического расстояния как d_{tax} и d_{ps} [27] для измерения иерархического сходства между двумя объектами. Обе меры основаны на вычислении наименьшего общего предка (*LCA от англ. lowest common ancestor*): узлы сравниваемых сущностей имеют общего предка, наиболее удалённого от корня дерева иерархии и лежащего на обоих путях от этих вершин до корня.

Вычисление близости соседства сравниваемых объектов. Окружение объекта $e \in E$ определяется как множество пар связь-сущность $N_e = \{(\mathrm{r}, e_i) | (e, r, e_i) \in R\}$, сущности которых находятся на расстоянии одного шага от e. Это определение окружения позволяет рассматривать вместе сущность-соседа и тип отношения ребра графа. *GADES* использует знания, закодированные в иерархиях отношений и классов диаграммы знаний, для сравнения двух пар.

Специфичность сущности e в ГЗ G вычисляется как величина обратно пропорциональная числу её инцидентных рёбер $Incident(e) = \{(e_i, r, e) \in R\}$. *GADES* вычисляет специфичность наименьшего общего предка e_1 и e_2. Суть метода в том, что объекты, общий предок которых содержит более общую информацию, менее похожи, чем сущности, общий предок которых содержит более конкретную информацию.

8 Моделирование когнитивного профиля обучаемого

В процессе изучения формируется ГЗО за счёт пополнения ссылками на концепты изученных ПрО. В совокупности множество этих ссылок формирует когнитивный профиль обучаемого. Для каждой ссылки в процессе обучения вычисляется определённый вес, характеризующий то, насколько хорошо была изучена соответствующая тема.

Онтология обучаемого содержит необходимые концепты и связи, позволяющие моделировать (рисунок 4):

- какие темы и понятия были изучены;
- оценка качества изучения;
- характеристики самого обучаемого, получаемые в процессе анализа его действий.

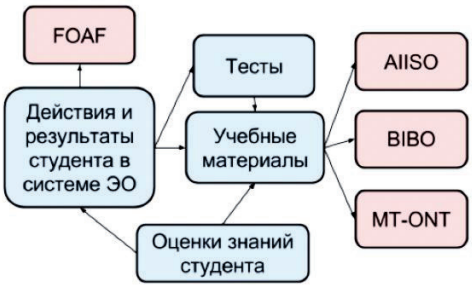

Рисунок 4 - Комплексная модульная онтология в системе ЭО

Важным свойством графовых данных является возможность возникновения различных корреляций между множеством взаимосвязанных узлов. Подобные корреляции могут быть вычислены за счёт включения обработки атрибутов, связей и классов связанных сущностей в алгоритм МО. Для моделирования бинарных отношений на графе удобно использовать трёхсторонний тензор $\underline{\mathbf{Y}}$, в котором две моды образованы идентично на основе конкатенированных сущностей объектов-узлов, а третья мода содержит отношения между ними [28]. Подобный подход получил название тензорная факторизация.

На рисунке 5 приведена иллюстрация процесса моделирования данным методом. Элемент тензора $y_{ijk} = 1$ обозначает факт, что существует отношение (*i-th entity, k-th predicate, j-th entity*). В противном случае, для несуществующих или неизвестных отношений элемент приравнивается нулю. Каждая из возможных реализаций такого тензора может быть интерпретирована как один из возможных миров. Для получения модели всего ГЗ необходимо оценить совместное распределение P($\underline{\mathbf{Y}}$) на множестве $D \in E \times R \times E$ для наблюдаемых троек. Таким образом, строится оценка вероятностного распределения над возможными мирами, которые позволяют предсказать вероятность наличия троек, основываясь на состоянии всего ГЗ.

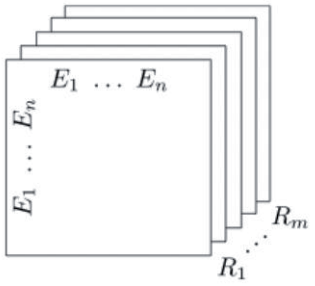

Рисунок 5 - Моделирование отношений с помощью трёхстороннего тензора [27]

Сущности ГЗ могут быть эффективно представлены векторами их латентных свойств. Данные свойства называют латентными, т.к. они напрямую не описаны в данных, но могут быть выведены из имеющихся данных в процессе МО. В работе [29] предложена модель графовых латентных свойств *RESCAL*, представляющая тройки посредством парного взаимодействия этих латентных свойств. Вычисление вероятности существования какой-либо тройки в ГЗ осуществляется с помощью специальной оценочной функции. Тензорное представление графа позволяет эффективным образом вычислять подобные оценки через факторизацию срезов тензора $\boldsymbol{F_k} = \boldsymbol{E}\,\boldsymbol{W_k}\,\boldsymbol{E^T}$, где $F_k \in \mathbb{R}^{N_e \times N_e}$ является матрицей, содержащей все оценки для *k*-й связи (отношения) и *i*-го ряда в матрице $\boldsymbol{E} \in \mathbb{R}^{N_e \times H_e}$. $W^k \in \mathbb{R}^{H_e \times H_e}$ является матрицей весов, элементы которой w_{abk} показывают, насколько латентные свойства a и b взаимосвязаны в *k*-том отношении. Рисунок 6 иллюстрирует описанный принцип вычислений.

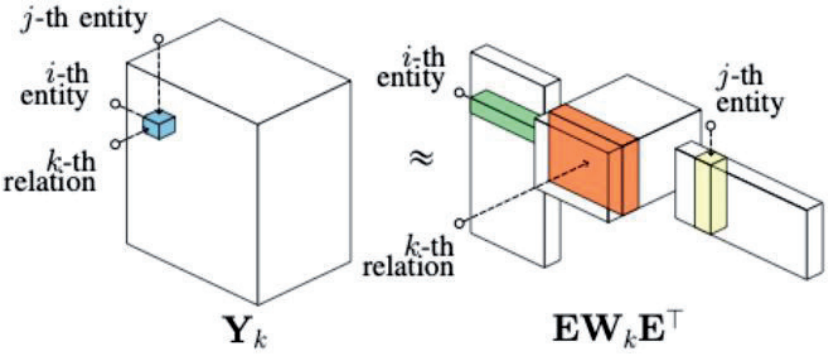

Рисунок 6 - Тензорное представление графа знаний RESCAL [21]

Выводы

Индивидуализация в ЭО является логичным и необходимым этапом эволюции ЭО, которое должно перейти от массовости к персонализации процессов обучения. Этот переход порождает множество методологических, технологических и концептуальных вопросов. Онтологический подход, принятый уже как стандарт де-факто для представления моделей слабоструктурированных данных, отвечает на часть из этих вопросов. Однако вопросы, касающиеся неявных или не декларированных знаний в системах ЭО, не могут быть разрешены непосредственно с помощью онтологий. В статье приведена попытка систематизации подобных вопросов и предложены подходы к их разрешению с помощью аппарата векторных представлений. Проведённый анализ может помочь при создании систем ЭО нового поколения, а также в решении задач обработки и анализа образовательных данных.

Список источников

[1] ***Харабет, Я.К.*** Автоматическое выделение количественных конструкций в русскоязычных научно-популярных текстах / Я.К. Харабет // XVIII Объединённая научная конференция «Интернет и современное общество» (IMS-2015). – Санкт-Петербург. – 2015. – С.23-25.

[2] ***Баранова, Ю.Ю.*** Индивидуализация обучения: возможности и ресурсы в аспекте введения федеральных государственных образовательных стандартов общего образования / Ю.Ю. Баранова // Научное обеспечение системы повышения квалификации кадров. – 2012. – №. 1. – С.123-129.

[3] ***Anderson, J.Q.*** Individualisation of higher education: How technological evolution can revolutionise opportunities for teaching and learning / J.Q. Anderson // *International social science journal*. – 2013. Vol. 64. No. 213-214. – P.305-316.

[4] ***Байдикова, Н.Л.*** Индивидуализация обучения студентов магистратуры в условиях накопительно-балльной системы / Н.Л. Байдикова // Международный научно-исследовательский журнал. – 2016. – № 11(53) Часть 3. – С.9-12.

[5] ***Иванова, Л.А.*** Медиаобразовательное пространство как средство обеспечения индивидуальных учебных траекторий студентов технического вуза / Л.А. Иванова, И.С. Петухова // *Magister Dixit*. – 2011. – №. 4. – С.151-165.

[6] ***Genov, N.*** Introduction to “Challenges of individualisation” / N. Genov // International Social Science Journal. – 2013. – Vol. 64. No. 213-214. – P.193-196.

[7] ***Huang, C.L.*** Generating New Paths for Teacher Professional Development (TPD) through MOOCs / C.L. Huang // Jiao Yu Yan Jiu Yu Fa Zhan Gi Kan. – 2018. – Vol. 14. No. 1. – P.35-71.

[8] ***Resource Description Framework*** (RDF). - https://www.w3.org/RDF/.

[9] ***RDF Schema*** *1.1*. - https://www.w3.org/TR/rdf-schema/.

[10] ***Deibe, M.A.*** Query Processing Over Graph-structured Data on the Web. – IOS Press, 2018. – Vol. 37.

[11] ***Web Ontology Language*** (OWL). - https://www.w3.org/OWL/.

[12] ***Academic Institution Internal Structure Ontology***. - http://vocab.org/aiiso/.

[13] ***Bibliographic Ontology Specification***. - http://bibliontology.com.

[14] ***The Ontology for Media Resources.*** - https://www.w3.org/TR/mediaont-10/.

[15] ***Kauppinen, T.*** Teaching core vocabulary specification / T. Kauppinen, J. Trame, A. Westermann // LinkedScience. org, Tech. Rep. — 2012.
[16] ***FOAF*** Vocabulary Specification. - http://xmlns.com/foaf/spec/.
[17] ***Metacademy***. - https://metacademy.org.
[18] ***SlideWiki****:* elicitation and sharing of corporate knowledge using presentations / Ali Khalili, S. Auer, D. Tarasowa, I. Ermilov // Knowledge Engineering and Knowledge Management. — Springer, 2012. — P.302–316.
[19] ***Crowd-sourcing*** (semantically) Structured Multilingual Educational Content (CoS- MEC) / D. Tarasowa, S. Auer, A. Khalili, J. Unbehauen // Open Praxis. — 2014. — Vol. 6, No.2. — P.159–170.
[20] ***Martinez-Rodriguez, J.L.*** Information extraction meets the semantic web: a survey / J.L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo // Semantic Web. Preprint – 2018. – P.1-81.
[21] ***Nickel, M.*** et al. A review of relational machine learning for knowledge graphs / M. Nickel et al. // Proceedings of the IEEE. – 2015. – Vol.104. No. 1. - P.11-33.
[22] ***Liu, W.*** et al. K-BERT: Enabling Language Representation with Knowledge Graph // arXiv preprint arXiv:1909.07606. – 2019.
[23] ***Lin, Y.*** et al. Learning entity and relation embeddings for knowledge graph completion // Twenty-ninth AAAI conference on artificial intelligence. – 2015.
[24] ***Alspaugh, S.*** et al. Analyzing log analysis: An empirical study of user log mining //28th Large Installation System Administration Conference (LISA14). – 2014. – P.62-77.
[25] ***Peláez, C.*** Unit testing as a teaching tool in higher education //SHS Web of Conferences. – EDP Sciences, 2016. – Vol.26. – P.01107.
[26] ***Traverso, I.*** et al. GADES: a graph-based semantic similarity measure //Proceedings of the 12th International Conference on Semantic Systems. – ACM, 2016. – P.101-104.
[27] ***Paul, C.*** et al. Efficient graph-based document similarity //European Semantic Web Conference. – Springer, Cham, 2016. – P.334-349.
[28] ***Nickel, M.*** A Three-Way Model for Collective Learning on Multi-Relational Data / M. Nickel, V. Tresp, H.P. Kriegel // ICML. – 2011. Vol.11. – P.809-816.
[29] ***Nickel, M.*** Tensor factorization for multi-relational learning / M. Nickel, V. Tresp // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Berlin, Heidelberg, 2013. – P.617-621.

Сведения об авторе

Муромцев Дмитрий Ильич, получил степень бакалавра (1997) и магистра (1999) в области проектирования компьютерных систем в Санкт-Петербургском государственном политехническом университете. Получил степень кандидата технических наук в области компьютерных наук в Университете ИТМО в 2003 г. В настоящее время является зав. кафедрой информатики и прикладной математики Университета ИТМО. Дмитрий Муромцев является членом комитетов по техническим программам и редколлегий ряда международных конференций и журналов. Научные интересы: семантические технологии, Интернет вещей, онтологический инжиниринг, представление знаний и искусственный интеллект. Педагогическая деятельность началась в 2001 году на кафедре компьютерных технологий и управления Университета ИТМО. Дмитрий Муромцев является автором и соавтором более 100 научных и учебно-методических публикаций и 4 книг. AuthorID (РИНЦ): 17726; Author ID (Scopus): 55575780100; ORCID 0000-0002-0644-9242; Researcher ID (WoS): N-6485-2016. *d.muromtsev@gmail.com.*

Поступила в редакцию 02.12.2019, после рецензирования 02.03.2020. Принята к публикации 15.03.2020.

Models and methods of e-learning individualization in the context of ontological approach

D. Mouromtsev

St. Petersburg National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia

Abstract

The article explores the issues of e-learning individualization (EE) as a set of processes for creating, developing, using and utilizing digital content and EE data, describes ontological models and methods for individualizing digital education. The technological stack for building and implementing individual learning paths is considered, as well as examples of existing systems that fully or partially use the specified technology stack. To efficiently process educational materials and data generated by learning management systems, a three-level architecture is proposed that allows semantic annotation and selection of concepts of varying degrees of abstraction in the data layers. These layers include: high-level abstractions of modeling, general concepts of educational materials and the educational process, specific concepts for access and integration of data from the EE system in terms of the subject area. For the first time, it was proposed to use semantic models as the formal basis for an individualized EE, including the vector representations of knowledge graphs, which, on the one hand, allows efficient processing of large and complex data structures, and on the other hand, has the flexibility and expressiveness of the ontological approach. The main aspects related to individualization in EE systems were sequentially considered, including technological aspects and existing ontologies for e-learning, individual trajectory modeling, semantic annotation of educational materials, methods for assessing knowledge in individualized learning, as well as ontological simulation of the cognitive profile of the student.

Key words: *e-learning, individualization, ontologies, knowledge graphs, vector representation, semantic technologies.*

Citation: *Mouromtsev D.* Models and methods of e-learning individualization in the context of ontological approach [In Russian]. *Ontology of designing*. 2020; 10(1): 34-49. DOI: 10.18287/2223-9537-2020-10-1-34-49.

List of figures

References

[1] ***Kharabet YaK.*** Automatic allocation of quantitative constructions in Russian-language popular scientific texts. *XVIII joint scientific conference* "Internet and modern society"(IMS-2015). Saint-Petersburg. 2015. P.23-25.

[2] ***Baranova JJ.*** Individualization of training: opportunities and resources in the aspect of introducing federal state educational standards of general education. *Scientific support of a system for advanced training of personnel*. 2012; 1: 123-129.

[3] ***Anderson JQ.*** Individualisation of higher education: How technological evolution can revolutionise opportunities for teaching and learning. *International social science journal*. 2013; 64(213-214): 305-316.

[4] ***Baydikova NL.*** Individualization of teaching graduate students in a cumulative-point system // International Research Journal. 2016; 11(53) Part 3: 9-12.

[5] ***Ivanova LA, Petukhova IS.*** Media-educational space as a means of providing individual educational trajectories for students of a technical university. *Magister Dixit*. 2011; 4: 151-165.

[6] ***Genov N.*** Introduction to “Challenges of individualisation”. *International Social Science Journal*. 2013; 64(213-214): 193-196.

[7] ***Huang CL.*** Generating New Paths for Teacher Professional Development (TPD) through MOOCs. Jiao Yu Yan Jiu Yu Fa Zhan Gi Kan. 2018; 14(1): 35-71.

[8] ***Resource Description Framework*** (RDF). https://www.w3.org/RDF/.
[9] ***RDF Schema*** 1.1. https://www.w3.org/TR/rdf-schema/.
[10] ***Deibe MA.*** Query Processing Over Graph-structured Data on the Web. IOS Press, 2018. Vol.37.
[11] ***Web Ontology Language*** (OWL). https://www.w3.org/OWL/.
[12] ***Academic Institution Internal Structure Ontology***. http://vocab.org/aiiso/.
[13] ***Bibliographic Ontology Specification***. http://bibliontology.com.
[14] ***The Ontology for Media Resources.*** https://www.w3.org/TR/mediaont-10/.
[15] ***Kauppinen T, Trame J, Westermann A***. Teaching core vocabulary specification. LinkedScience. org, Tech. Rep. 2012.
[16] ***FOAF*** Vocabulary Specification. http://xmlns.com/foaf/spec/.
[17] ***Metacademy***. https://metacademy.org.
[18] ***SlideWiki****:* elicitation and sharing of corporate knowledge using presentations / Ali Khalili, Soeren Auer, Darya Tarasowa, Ivan Ermilov // *Knowledge Engineering and Knowledge Management*. Springer, 2012. P.302–316.
[19] ***Crowd-sourcing*** (semantically) Structured Multilingual Educational Content (CoS- MEC) / Darya Tarasowa, Soeren Auer, Ali Khalili, Jourg Unbehauen // Open Praxis. 2014; 6(2): 159–170.
[20] ***Martinez-Rodriguez JL, Hogan A, Lopez-Arevalo I.*** Information extraction meets the semantic web: a survey. *Semantic Web*. Preprint. 2018. 81 p.
[21] ***Nickel M.*** *et al.* A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*. 2015; 104(1): 11-33.
[22] ***Liu W.*** *et al.* K-BERT: Enabling Language Representation with Knowledge Graph. arXiv preprint arXiv:1909.07606. 2019.
[23] ***Lin Y.*** *et al.* Learning entity and relation embeddings for knowledge graph completion. *Twenty-ninth AAAI conference on artificial intelligence*. 2015.
[24] ***Alspaugh S.*** *et al*. Analyzing log analysis: An empirical study of user log mining. *28th Large Installation System Administration Conference* (LISA14). 2014: 62-77.
[25] ***Peláez C.*** Unit testing as a teaching tool in higher education. *SHS Web of Conferences*. EDP Sciences, 2016; 26: 01107.
[26] ***Traverso I.*** *et al.* GADES: a graph-based semantic similarity measure. *Proceedings of the 12th International Conference on Semantic Systems*. ACM, 2016: 101-104.
[27] ***Paul C. et al.*** Efficient graph-based document similarity. *European Semantic Web Conference*. Springer, Cham, 2016: 334-349.
[28] ***Nickel M, Tresp V, Kriegel HP***. A Three-Way Model for Collective Learning on Multi-Relational Data. *ICML*. 2011; 11: 809-816.
[29] ***Nickel M, Tresp V.*** Tensor factorization for multi-relational learning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, 2013: 617-621.

About the author

Dmitry Mouromtsev received a bachelor's degree (1997) and a master's degree (1999) in computer system design from St. Petersburg state Polytechnic University (Russia). He received the degree of candidate of technical Sciences in computer science in ITMO University in 2003, is currently the head of departments at the Department of Informatics and Applied mathematics of ITMO University. Dmitry Mouromtsev is a member of the technical program committees and editorial boards of a number of international conferences and journals. Research interests include semantic technology, the Internet of things, Ontological engineering, knowledge representation, and artificial intelligence. Pedagogical activity began in 2001 at the Department of Computer technology and management of ITMO University. Since then, he has taught more than 10 training courses on current topics in knowledge-based systems, computer science and more. Dmitry Mouromtsev is the author and co-author of more than 100 scientific and educational publications and 4 books. AuthorID (RCI): 17726; Author ID (Scopus): 55575780100; ORCID 0000-0002-0644-9242; Researcher ID (WoS): N-6485-2016. *d.muromtsev@gmail.com.*

Received December 8, 2020. Revised March 2, 2020. Accepted March 15, 2020.