Интеллектуальный метод и алгоритм сопоставления учебных дисциплин на основе векторного представления текстов

Д.С. Ботов
Институт информационных технологий
Челябинский государственный университет
Челябинск, Россия
e-mail: dmbotov@gmail.com

Ю.Д. Кленин
Институт информационных технологий
Челябинский государственный университет
Челябинск, Россия
e-mail: jklen@yandex.ru

Аннотация[1]

В статье описывается метод интеллектуального анализа и алгоритм сопоставления программ учебных дисциплин на основе использования подходов векторного представления текстов с помощью модели paragraph2vec, что в отличие от описанных ранее подходов позволяет проводить семантический анализ без необходимости трудоемкого онтологического моделирования предметных областей образовательных программ, описания правил логического вывода или использования экспертных оценок. Приводятся результаты эксперимента по оценке качества предложенного метода для классификации учебных дисциплин на представительном корпусе текстов рабочих программ дисциплин.

1. Введение

В современном образовании в условиях постоянного изменения образовательных стандартов, появления профессиональных стандартов, возрастания объемов учебно-методической документации и изменения требований к ее оформлению все более трудоемким становится процесс разработки образовательных программ и отдельных учебных дисциплин и модулей.

Кроме того взрывной рост объемов образовательного контента в Интернет, стремительного развития массовых онлайн курсов (MOOC) и технологий дистанционного образования приводит к увеличению конкуренции на рынке образовательных услуг и повышает требования к качеству и актуальности содержания образовательных программ.

Все эти факторы приводят к существенному увеличения нагрузки на профессорско-преподавательский состав в задачах разработки и актуализации образовательных программ и курсов.

С учетом применения компетентностного подхода, как основного для формирования образовательных программ, крайне важным становится анализ не только структуры учебной дисциплины, но и целей курса и результатов обучения в их увязке с требованиями образовательных и профессиональных стандартов, а также учет требований рынка труда для профессиональных дисциплин и направленности (профиля) образовательной программы.

И хотя сегодня увеличиваются темпы внедрения средств автоматизации учебного процесса в образовательные организации, постоянно усложняется функциональность информационных систем, позволяющих в том числе разрабатывать учебные планы, рабочие программы дисциплин и прочие виды учебно-методической документации, на сегодняшний день на рынке подобного рода программного обеспечения отсутствуют действительно эффективные средства, позволяющие независимо от предметной области направления подготовки проводить сравнительный анализ содержания, целей и результатов обучения образовательных программ и курсов с целью их дальнейшей актуализации и формирования конкретных рекомендаций по содержанию образования для преподавателей с учетом последних требований образовательных стандартов, развития экономики, науки и техники, изменения требований на рынке труда.

Данное исследование посвящено решению задачи интеллектуальной поддержки процесса сопоставления учебных дисциплин образовательных программ высшего образования с целью снижения трудоемкости разработки новых и актуализации существующих учебных дисциплин и повышения качества образовательного контента.

Труды пятой всероссийской конференции "Информационные технологии интеллектуальной поддержки принятия решений", 16 - 19 мая, Уфа, Россия, 2017

2. Обзор методов интеллектуального анализа образовательных программ

Сравнительный анализ образовательных программ с целью определения сопоставимости образования и принятия решений при планировании образовательных программ и управления образовательными траекториями представлен в работах М.Б. Гузаирова, Н.И. Юсуповой, О.Н. Сметаниной, М.М. Гаяновой, С.В. Тархова, А.С. Пирской, Л.С. Лисицыной и других [1,2,3,4,5,6, 7].

Так в работе О.Н. Сметаниной [5] рассматривается проблема управления образовательным маршрутом в условиях организации и развития академической мобильности. Для решения проблемы предложена модель управления и система поддержки принятия решений с использованием комплекса дискретно-событийных моделей для ситуационного управления и комплекса концептуальных онтологических моделей и методов инженерии знаний для обеспечения информационной поддержки. Степень схожести российских и международных образовательных программ путем сопоставления терминов (ключевых слов), формулировок знаний, умений, навыков, компетенций и наименований дисциплин учебных планов.

В работе Е.А. Черниковой [8] предлагается для сопоставления образовательных курсов использовать онтологические модели, и определяются меры семантической близости курсов на основе анализа ключевых слов содержания дисциплин и результатов обучения с помощью таксономии образовательных целей.

Также онтологические модели применяются в работе А.Ю. Ужва [9] для адаптивного поиска образовательных ресурсов с помощью рассуждений по прецедентам.

Формализованная модель составления учебного плана на основе модульно-компетентностного подхода представлена в работе И.М. Харитонова [10]. Для построения учебного плана используются метод экспертных оценок и метод контент-анализа. А для определения значимости дисциплин (модулей) с учетом потребностей рынка труда используется метод латентно-семантического анализа.

Метод экспертных оценок и когнитивные карты используются в работе И.В. Сибикиной [11] для определения перечня дисциплин, формирующих определенную компетенцию образовательной программы. Для получение перечня значимых дисциплин и перечня компетенций строится лингвистический классификатор на основе синтаксических правил и нечетких множеств.

В работе А.С. Пирской, Л.С. Лисицыной [7] рассматривается модель управления образовательными траекториями студентов на основе результатов освоения компетенций образовательных стандартов с учетом принципа междисплинарности обучения. Для формализации образовательной траектории используется графовая модель в виде план-графа.

Алгоритм автоматизированного составления учебного плана образовательной программы на основе компетентностного подхода предложен в работе С.С. Котова [12]. В работе применяется графовая модель структурно-логических связей дисциплин, рассмотрены различные эвристические алгоритмы формирования учебного плана с использованием экспертных оценок.

Стоит отметить, что основная сложность использования рассмотренных подходов на практике заключается в необходимости привлечения представительного состава экспертов для методов с использованием экспертных оценок, формирования онтологических моделей, правил и/или прецедентов для каждой из предметных областей направлений подготовки образовательных программ.

Прямое сопоставление ключевых слов и названий дисциплин, тем и разделов показывает низкое качество сопоставления образовательных программ и курсов, учитывая имеющиеся различия в подходах разных образовательных организаций к разделению программ на дисциплины и модули, структурирование разделов и тем внутри дисциплины, формулирование результатов обучения по дисциплинам. Все еще более отягощается тем фактом, что в последних изменениях образовательных стандартов третьего поколения (ФГОС ВО) практически отсутствуют (за исключением трех-четырех обязательных дисциплин) требования к наличию и содержанию определенных дисциплин, их структуре и распределению компетенций по дисциплинам.

3. Метод сопоставления учебных дисциплин образовательных программ

В последнее время большую популярность и широкое применение получили модели векторных вложений (word embedding), в частности, созданная сотрудниками компании Google модель word2vec, использующая двухслойные нейронные сети для быстрого и качественного осуществления отражения слов (или любых других токенов, употребляющихся в последовательности) в их векторное представление. Основной идее данной модели было сопоставление слов с контекстами их употребления, ориентирующееся на сопоставление косинусной близости векторов слов количеству их общих контекстов употребления.

Модель показывает высокие результаты как в качестве, так и в скорости работы, имея, при этом, полную языковую независимость. На данный момент, популярность word2vec особенна заметна на соревнованиях по анализу языка, таких как Dialog,

большинство участников которого применяют эту модель в анализе данных.

Позже, в качестве расширения идей word2vec, оригинальными авторами модели была выпущена модель paragraph2vec, фокусирующаяся уже на сопоставление целых документов и отражении их в векторное пространство.

В данном разделе описан метод сопоставления учебных дисциплин различных образовательных программ, который использует современные подходы векторного представления текстов с помощью модели paragraph2vec, что в отличие от описанных ранее подходов позволяет проводить сопоставление без необходимости трудоемкого онтологического моделирования предметных областей образовательных программ и построения баз знаний, без необходимости описания эвристических алгоритмов, правил логического вывода, а также без использования экспертных оценок.

Метод позволяет решать задачи семантического анализа учебных дисциплин, включая информационный поиск с ранжированием семантически близких программ дисциплин, классификацию по различным признакам и кластеризацию учебных дисциплин.

3.1. Модель учебной дисциплины

В результате проведенного ранее анализа описания образовательных курсов на сайтах ведущих российских и зарубежных университетов, а также требований к оформлению рабочих программ дисциплин, установленных Минобрнауки РФ, были определены основные элементы, описывающие учебную дисциплину или модуль [13].

Для формализации описания учебной дисциплины (модуля) была построена концептуальная онтологическая модель, представленная на рисунке 1. На ней отражены наиболее важные с точки зрения формирования образовательных программ концепты, которые присутствуют во всех рабочих программах дисциплин. Концепты можно разделить на три условные части:

1. Образовательные цели и задачи курса, компетенции образовательной программы и результаты обучения по дисциплине (как правило, описываются в виде знаний/умений/навыков).
2. Структура дисциплины: описание разделов, тем курса, содержания лекционных и практических занятий, заданий на самостоятельную работу обучающихся.
3. Описание формы промежуточной аттестации, оценочные средства с примерами заданий и привязкой к проверяемым результатам обучения.

На практике преподаватель при разработке нового курса или актуализации существующего курса в первую очередь анализирует первые две составляющие описания дисциплин.

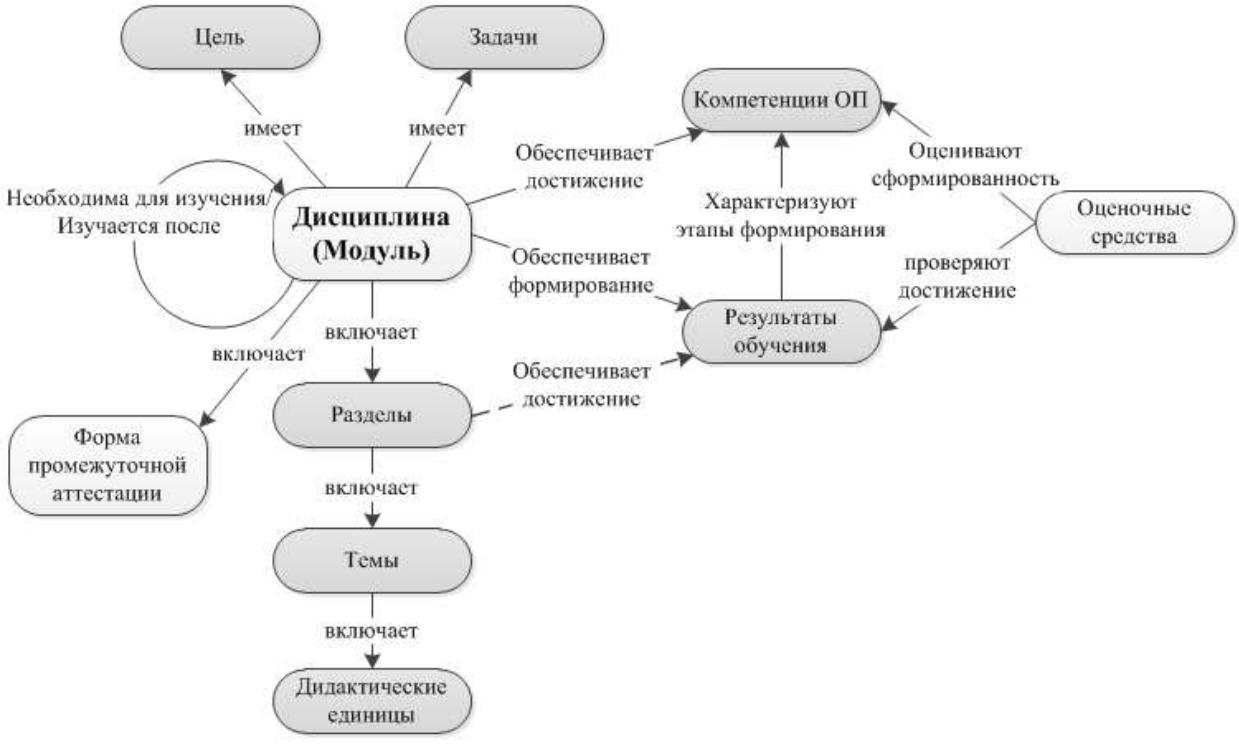

Рис. 1. Онтологическое представление учебной дисциплины (модуля)

3.2. Векторное представление текстов

В последние несколько лет набирают популярность нейросетевые модели языка, которые показывают передовые результаты в задачах определения семантической близости [14,15], в том числе и для русского языка, как это показано в работах по оценке методов определения семантической близости слов RUSSE [16]. Нейросетевые модели такие как word2vec основываются на дистрибутивной гипотезе понимания смысла текста и подходе векторного представления слов (word embedding) с использованием нейронных сетей [15].

В данной работе предлагается использовать нейросетевую модель paragraph2vec [17], которая обучается на текстах входного корпуса, преобразую их в векторные представления в n-мерном пространстве. А затем позволяет определять семантическую близость или связанность для новых текстов путем сравнения векторов обученной модели (например, по косинусной мере близости).

Существует значительное количество публикаций, использующих paragraph2vec как для более базовых вариантов использования векторных моделей – генерации наборов параметров для использования в задачах классификации (например, задачи анализа тональности), так и более специфических случаев использования.

Так в работе [18] авторы представляют вариант использования paragraph2vec в связке с Wikipedia в качестве базы знаний для реализации именованного связывания объектов. При наличии упоминаний именованных сущностей, контекстов употребления этих упоминаний и некоторого набора кандидатов, обученная на базе Wikipedia модель генерирует вектора для контекстов и ищет наиболее близкий к ним вектор среди кандидатов-статей Wikipedia.

Другая группа исследователей, под руководством Hashimoto [19] применила paragraph2vec в задаче тематического моделирования. Авторы обучают модель на базе исследуемого корпуса, а затем осуществляют кластеризацию. Исходя из предположения, что кластеры являются темами, авторы используют модель для генерации слов для каждой темы и сравнивают результаты с латентным размещением Дирихле (LDA). Исходное предположение о том, что кластера векторов текстов соответствуют темам требует дополнительных проверок. В частности, представленные авторами наборы тем и слов не сравнимы по качеству с традиционным LDA.

Еще одно исследование предлагающее сравнение paragraph2vec с LDA представлено в [20]. Здесь, авторы показывают потенциал модели, обученной на Wikipedia и используемой для выявления групп статей энциклопедии, показывая достаточно качественные результаты в задаче поиска наиболее близких концептов статей.

В задачах анализа образовательного контента векторное представление текстов ранее не применялось. Хотя в области Educational Data Mining находят широкое применение методы и алгоритмы машинного обучения. Более подробная информация представлена в обзорах [14,21].

Авторами данного исследования высказывается гипотеза, что методы на основе нейросетевых моделей языка (с помощью word2vec и paragraph2vec) могут показать лучшее качество моделирования и оценки семантики в задачах сопоставления учебных дисциплин при простоте обучения самой модели. При этом стоит учесть один из основных недостатков нейросетевых моделей (в отличие, например, от методов вероятностного тематического моделирования) – отсутствие возможности интерпретировать отдельные признаки полученных на выходе нейронной сети векторов.

3.3. Алгоритм сопоставления содержания учебных дисциплин

На рисунке 2 представлена обобщенная схема алгоритма сопоставления учебных курсов с использованием модели paragraph2vec, обученной на корпусе текстов рабочих программ дисциплин.

На первом шаге алгоритма предусмотрена предварительная обработка документов, включающая выбор из текстов документов наиболее важных фрагментов – целей и задач курса, результатов обучения и компетенций, структуры курса, включая описания тем и разделов курса. Предварительная обработка завершается лемматизацией и отбрасыванием стоп-слов.

После обучения модели paragraph2vec на подготовленном корпусе появляется возможность генерации векторов документов для новых, ранее неизвестных модели paragraph2vec рабочих программ дисциплин.

В дальнейшем можно добавить к полученным векторам новых документов и векторам документов из корпуса дополнительные признаки, которые могут улучшить качество сопоставления рабочих программ и учесть дополнительные факторы для отдельных задач в процессе формирования и актуализации образовательных программ. Данные признаки создаются и отбираются экспертами в ходе традиционного для машинного обучения процесса Feature Engineering. Например, такими признаками могут быть направление подготовки, объем дисциплины в зачетных единицах, распределение часов, период освоения дисциплины (семестр) по учебному плану, форма аттестация и т.д.

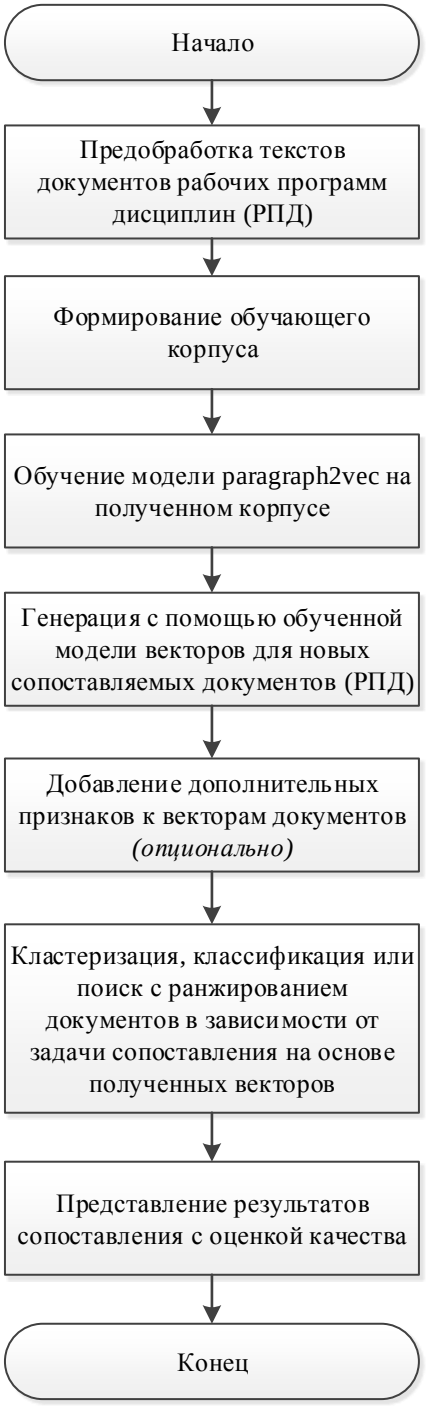

Рис. 2. Алгоритм сопоставления учебных дисциплин с использованием paragraph2vec

Последние этапы обобщенного алгоритма могут варьироваться в зависимости от конкретных задач анализа образовательных программ. Если необходимо подобрать наиболее семантически близкие учебные дисциплины к заданной, то можно применять алгоритмы информационного поиска с ранжированием результатов по степени семантической близости векторов документов (как правило используется косинусная мера близости). Можно провести классификацию учебных дисциплин по определенным областям знаний, направлений подготовки, циклам дисциплин, предварительно обучив классификатор на обучающей выборке.

4. Эксперимент

По причине отсутствия открытых тренировочных корпусов образовательных программ на русском или английском языках в данном исследовании для анализа применимости paragraph2vec в рассматриваемой задаче сопоставления учебных дисциплин, авторы воспользовались текущим вариантом создаваемого собственными силами корпуса рабочих программ дисциплин.

На данный момент, экспериментальный корпус содержит более ста различных учебных дисциплин, которые могут быть разделены на семь групп в соответствии с предметной областью: информационные технологии, экономика, математика и статистика, лингвистика, история, медицина, и право. В каждой группе имеется от 3 до 5 меньших категорий, соответствующих конкретным дисциплинам, каждая из которых содержит 4 документа, представляющих каждый образовательный курс. Пример таких категорий для области математика и статистика: математический анализ, алгебра и геометрия, теория вероятностей и математическая статистика, методы оптимизации.

В рамках эксперимента оценивается качество классификации программ учебных дисциплин в двух задачах: классификации по предметным областям (7 классов) и классификации по категориям в рамках предметных областей (26 классов).

В качестве реализации paragraph2vec используется doc2vec из библиотеки gensim. Обучается модель distributed bag of words, которая, согласно оценкам, приведенным в работе [17], показывает лучшие результаты, чем альтернативная модель distributed memory, на большинстве наборов данных.

Обучающая выборка для модели paragraph2vec и последующего обучения классификаторов составляет 75% от корпуса, а тестовая выборка, по которой производится оценка качества, составляет оставшиеся 25% корпуса. Также при оценке применяется стандартная методика кросс-валидации.

В сравнении участвуют популярные алгоритмы классификации: логистическая регрессия, support vector классификатор, k-ближайших соседей, дерево принятия решений и различные ансамбли лесов деревьев принятия решений.

Результаты оценки основных показателей качества работы классификаторов – точность (precision), полнота (recall) и F-мера, представлены в таблице 1.

Таблица 1. Результаты эксперимента по классификации рабочих программ дисциплин

	Классификация по категориям близких дисциплин в рамках предметной области			Классификация дисциплин по предметным областям		
	Precision	Recall	F-мера	Precision	Recall	F-мера
Logistic Regression	0.83	0.88	0.85	**0.97**	**0.96**	**0.96**
Extra-trees	0.77	0.85	0.79	**0.97**	**0.96**	**0.96**
Random Forest	0.72	0.81	0.75	**0.97**	**0.96**	**0.96**
Decision Tree	0.26	0.35	0.28	0.69	0.65	0.65
k-nearest neighbors	0.77	0.85	0.79	**0.97**	**0.96**	**0.96**
C-Support Vectors	**0.84**	**0.88**	**0.85**	**0.97**	**0.96**	**0.96**

Как и ожидалось, результаты классификации по предметным областям превышают аналогичные показатели для отдельных дисциплин внутри категорий. Однако, как видно из приведенных показателей, все базовые алгоритмы, за исключением дерева принятия решений показали высокие результаты в обеих выбранных задачах классификации. Это означает, что даже в простейшем варианте применения без добавления дополнительных признаков к векторам, paragraph2vec хорошо подходит для семантического анализа программ учебных дисциплин.

5. Заключение

Основные результаты проведенного исследования:

- Предложен метод и алгоритм сопоставления учебных дисциплин на основе использования метода векторного представления текстов путем обучения модели paragraph2vec.
- Проведен эксперимент с реализацией предложенного метода для задач классификации учебных дисциплин на составленном представительном текстовом корпусе рабочих программ дисциплин.
- Результаты эксперимента позволяют сделать вывод о применимости модели paragraph2vec для оценки семантической близости программ учебных дисциплин. Лучшие показатели F-меры (0.85) показал классификатор на основе алгоритма C-Support Vectors.

Направления дальнейшего исследования:

- Оценка применимости предложенного метода в задачах информационного поиска и классификации учебных дисциплин в сравнении с другими подходами к обработке естественного языка (например, взвешенный TF-IDF, LSA, pLSA, LDA, извлечение ключевых слов, word2vec в сочетании с ключевыми словами).
- Извлечение дополнительных количественных и семантических признаков (Feature Engineering) из программ учебных дисциплин и учебных планов для анализа их влияния на качество решения задач интеллектуальной поддержки формирования образовательных программ.
- Определение удобных для конечных пользователей способов визуализации результатов сопоставления учебных дисциплин и образовательных программ.
- Применение векторного представления текстов для сопоставления требований профессиональных стандартов, вакансий рынка труда и результатов освоения образовательных программ в задаче формирования актуальных результатов обучения по профессиональным дисциплинам.

Список используемых источников

1. Гузаиров М.Б., Юсупова, Н.И., Маркелова А.В. Подход к управлению образовательным процессом в университете в рамках Болонского процесса // Матер. 12-й межд. науч. конф. CSIT’2010. Москва – Санкт-Петербург. – Уфа : УГАТУ, 2010. – Т. 2. – С. 187–192.
2. Университетские образовательные программы. Модели и методы для сопоставительного анализа / М. Б. Гузаиров [и др.]. – М.: Изд-во МАИ. 2006. – 117 с.
3. Поддержка принятия решений при управлении академической мобильностью / М.Б. Гузаиров, Н. И. Юсупова, О. Н. Сметанина, В. А. Козырева // Системы управления и информационные технологии. 2011. №3.1. С. 131–136.
4. Гаянова М.М. Информационная поддержка принятия решений при анализе университетских образовательных программ. Дис. канд. техн. наук / Уфимск. гос. авиац. техн. ун-т. – Уфа, 2006. – 143 с.

5. Сметанина О.Н., Методологические основы управления образовательным маршрутом с использованием интеллектуальной информационной поддержки: дис. доктора техн. наук. — УГАТУ, Уфа, 2012.
6. Тархов С.В. Методологические и теоретические основы адаптивного управления электронным обучением на базе агрегативных учебных модулей: дис. д-ра техн. наук: 05.13.10 / Тархов Сергей Владимирович.– Уфа, 2009. – 336 с.
7. Лисицына Л.С., Пирская А.С. Автоматизация управления образовательными траекториями для разработки модульных компетентностно-ориентированных образовательных программ вуза// Тр. Всерос. научно-практ.конференции с международным участием «Информационные технологии в обеспечении нового качества высшего образования (14-15 апреля 2010г., Москва, НИТУ «МИСиС»)». М.: Исследовательский центр проблем качества подготовки специалистов НИТУ «МИСиС», 2010. Кн. 3. С. 75-86.
8. Черникова Е.А., Черников А.С. Формализация и сравнение учебных программ на основе онтологического подхода // Вестник МГТУ им. Н.Э. Баумана. Сер. "Приборостроение". Спецвыпуск "Информационные технологии и компьютерные системы", 2011. - С.101-104.
9. Ужва А.Ю. Автоматизированная разработка онтологической модели предметной области для поиска образовательных ресурсов с использованием анализа текстов рабочих программ.// Современные проблемы науки и образования – 2013 - №1. [Электронный ресурс]. URL: http://www.science-education.ru/107-8324 (дата обращения: 01.03.2017).
10. Харитонов, И. М. Алгоритм формирования учебного плана с применением методики формализованного представления учебной дисциплины (на примере дисциплины «моделирование систем») [Текст] / И. М. Харитонов // Вестник АГТУ. Серия «Управление, вычислительная техника и информатика »: научный журнал - Астрахань, 2011. - № 2. - С. 53.
11. Сибикина, И.В. Оценка значимости дисциплин, формирующих компетенцию на основе лингвистического классификатора / И.В. Сибикина, Н.Ю. Квятковская // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2012. — №2. — С. 182 — 186.
12. Котов С.С., Столбов В.Ю. Управление структурой образовательных программ компетентностного содержания с учетом нечетких социальных предпочтений //Системы управления и информационные технологии.- 2009. -№1.3-С.411-416.
13. Botov D.S. Educational Content Semantic Modelling for Mining of Training Courses According to the Requirements of the Labor Market / D. Botov, J. Klenin // Proceedings of the 1st International Workshop on Technologies of Digital Signal Processing and Storing. Russia, Ufa, 2015. P. 214–218.
14. Melnikov A.V., Botov D.S., Klenin J.D., On usage of machine learning for natural language processing tasks as illustrated by educational content mining // Scientific journal «Ontology of Designing» - v.7, № 1(23)/2017, 2017, P. 34-47.
15. Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed Representations of Words and Phrases and their Compositionality. // Advances in neural information processing systems, 2013, P. 3111-3119.
16. Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N. RUSSE: The First Workshop on Russian Semantic Similarity [Report]. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue”, Moscow, RGGU, 2015, vol. 2, P. 89-105.
17. Q. Le, T. Mikolov. Distributed Represenations of Sentences and Documents. In Proceedings of ICML 2014, P. 1188–1196.
18. Kirsch L. et al. Named Entity Linking using Paragraph Vector // Paper for seminar Knowledge Mining Summer Semester 2016, Hasso Plattner Institute, Potsdam University, 2016.
19. Hashimoto, K., et al., Topic detection using paragraph vectors to support active learning in systematic reviews // Journal of Biomedical Informatics, v.62, 2016. P. 59-65.
20. Dai, A.M., Olah C., Le Q.V., Corrado G.S., Document embedding with paragraph vectors // Proc. of the NIPS Deep Learning Workshop, 2014.
21. Romero, C., Ventura, S. Educational Data Mining: A Review of the State-of- the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2010, 40 (6). P. 601-618.