Информатика и вычислительная техника

УДК 004.855 **DOI: 10.14529/ctcr190201**

СЕМАНТИЧЕСКИЙ ПОИСК УЧЕБНЫХ ДИСЦИПЛИН ПОД ТРЕБОВАНИЯ РЫНКА ТРУДА НА ОСНОВЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ЯЗЫКА

Д.С. Ботов, Ю.В. Дмитрин, Ю.Д. Кленин

Челябинский государственный университет, г. Челябинск, Россия

В условиях роста объема открытого образовательного контента, повышения требований к профессиональному образованию со стороны рынка труда, развития концепции обучения в течение всей жизни (Lifelong Learning) сегодня крайне актуальной является задача регулярного обновления содержания образовательных программ и отдельных дисциплин. В статье рассматривается метод семантического поиска образовательного контента под заданные требования рынка труда, определяемые профессиональными стандартами. В отличие от традиционных подходов сопоставления и анализа содержания образовательных программ, основанных на онтологических моделях и правилах, предлагается использовать подход к распределенному представлению слов (word embeddings) с помощью известных нейросетевых моделей языка word2vec и fastText. В качестве исходных запросов выступают фрагменты профессиональных стандартов – конкретные требования к знаниям, умениям и описания трудовых действий и трудовых функций, а в качестве искомых документов – описания учебных дисциплин и онлайн-курсов, включающие аннотацию, результаты обучения, структуру и содержание основных тем. Приводятся данные экспертной оценки качества ранжирования по метрике NDCG (Normalized Discounted Cumulative Gain) и точности семантического поиска по метрике MAP (Mean Average Precision) на представительном корпусе программ учебных дисциплин вузов по ИТ-направлениям и массовых открытых онлайн-курсов. Лучшие результаты для поиска показывают модели word2vec и fastText, обучаемые без учителя на больших специально подготовленных корпусах текстов программ учебных дисциплин и описаний онлайн-курсов. Для перехода от векторов слов к векторам текстов исследуются разные способы усреднения векторов слов, полученных от нейросетевых моделей, в сочетании с векторной моделью TF-IDF.

Ключевые слова: семантический поиск, семантическая близость, дистрибутивная семантика, word2vec, fastText, учебная дисциплина, массовые открытые онлайн-курсы, рынок труда.

Введение

В современном мире наблюдается стремительный рост количества образовательных программ и массовых открытых онлайн-курсов (Massive Open Online Courses). С одновременным увеличением доступного открытого образовательного контента происходит обновление образовательных и профессиональных стандартов, которое отражает постоянное изменение требований к выпускникам, предъявляемых региональным рынком труда. С учетом новых приоритетов развития цифровой экономики все более значимой становится проблема качества содержательного контента и актуальности образовательных программ, необходимость построения индивидуальных образовательных траекторий обучающихся с использованием информационно-коммуникационных технологий (ИКТ) с целью сокращения разрыва между профессиональным образованием и потребностями динамично развивающихся отраслей цифровой экономики.

Актуальность темы исследования подтверждается и развитием законодательства РФ. В тексте Федерального закона [1] в статье 96 «О профессионально-общественной аккредитации образовательных программ» явно указано требование соответствия качества и уровня подготовки выпускников, освоивших образовательные программы, требованиям профессиональных стандартов и требованиям рынка труда. В Федеральном законе от 3 июля 2016 г. [2] явно прописана проце-

дура подтверждения соответствия квалификации работника положениям профессионального стандарта.

Одновременно с этим завершается переход к компетентностно-ориентированной модели профессионального образования. Крайне важное значение при проектировании образовательных программ имеет грамотное определение профессиональных компетенций и актуальных результатов обучения, которые в новых редакциях Федеральных государственных образовательных стандартов (ФГОС ВО 3++) определяются образовательной организацией самостоятельно, исходя из требований профессиональных стандартов.

Можно сделать вывод об актуальности задачи навигации в пространстве образовательного контента и его сопоставления и актуализации под требования рынка труда, определяемых профессиональными стандартами и работодателями.

1. Обзор подходов к семантическому поиску образовательного контента и требований рынка труда

Методы анализа образовательных программ на основе онтологических подходов и эвристических алгоритмов для сопоставления целей и содержания образования и принятия решений при разработке образовательных программ и управления индивидуальными образовательными траекториями представлены в работах [3–5].

Основная сложность использования данных подходов на практике заключается в необходимости постоянного привлечения представительного состава экспертов для методов с использованием экспертных оценок, решении крайне трудоемких задач формирования и поддержки в актуальном состоянии онтологий, систем правил логического вывода и прецедентов для каждой из предметных областей направлений подготовки образовательных программ. Кроме того, с помощью данных моделей и методов невозможно гибко учесть региональные особенности и постоянно изменяющиеся потребности рынка труда.

Несоответствие между традиционным компетентностным описанием выпускника и конкретными, постоянно изменяющимися требованиями реального рынка труда отмечается авторами проекта по созданию системы мониторинга потребностей рынка труда [6]. В своей работе они предлагают определять соответствие содержания образовательных программ выявленным потребностям рынка труда на основе известной нейросетевой модели word2vec [7], реализующей идею дистрибутивной семантики. Авторами также предлагается идея прогнозирования изменений потребностей рынка труда.

Работа [8] посвящена созданию инструмента поиска образовательного контента – курсов с различных МООС-платформ. В основе лежит преобразование текстов программ курсов в набор составляющих их концептов. Сам поиск работает в два этапа: поиск концептов по ключевому слову; поиск курсов, ближайших к выбранному концепту. Авторами приводятся результаты анкетирования пользователей, показывающих положительную оценку качества работы системы, однако более конкретные количественные показатели качества работы системы не приведены.

Авторы настоящего исследования в работе [9] исследовали применимость различных методов векторного представления текста, тематического моделирования и нейросетевых моделей языка для семантического поиска онлайн-курсов под заданную программу учебной дисциплины. В рамках данной задачи лучшие результаты при экспертной оценке точности поиска показали классическая векторная модель TF-IDF и модель усредненного word2vec со взвешиванием слов по IDF.

В данной работе предлагается определить оптимальный метод поиска и модель языка для задачи семантического сопоставления онлайн-курсов, программ дисциплин университетов с требованиями рынка труда, представленными в профессиональных стандартах. Специфической особенностью поиска является существенное различие лексики описания программ учебных курсов и ограниченной лексики профессиональных стандартов, а также расширенной профессиональной лексики текстов вакансий в системах онлайн-рекрутмента.

2. Метод семантического поиска учебных курсов и требований рынка труда

Для поиска релевантного образовательного контента предлагается использовать следующую информацию, извлекаемую из рабочих программ дисциплин и описаний онлайн-курсов на русском языке:

- название дисциплины;
- цели и задачи дисциплины;
- результаты обучения;
- содержание дисциплины (перечень разделов и тем с описанием их содержания).

Общая схема метода представлена на рис. 1.

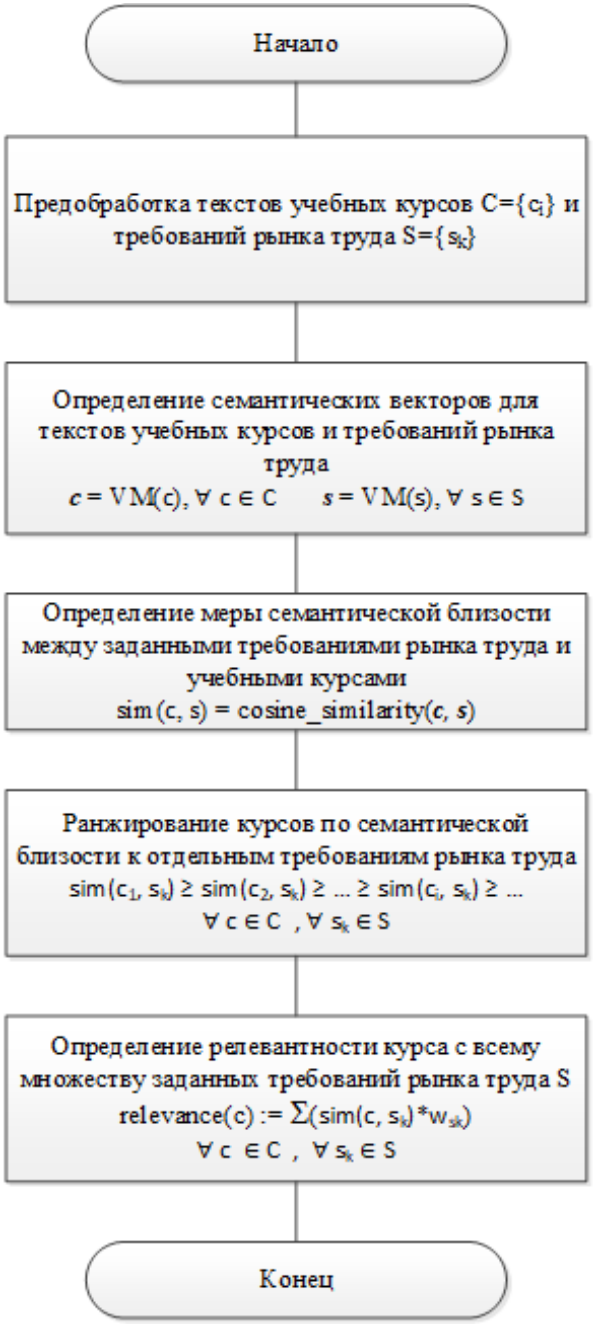

Рис. 1. Схема метода семантического поиска курсов с требованиями рынка труда

В качестве поисковых запросов выступают описания трудовых функций, извлекаемые из текстов профессиональных стандартов, а также отдельные элементы описания функций: трудовые действия, требования к знаниям/умениям.

На первом этапе полученные тексты учебных дисциплин и профессиональных стандартов проходят традиционную для задач обработки естественного языка предобработку по следующим принципам:

1. Токенизация текста.

2. Очистка от всех символов, не являющихся буквами, цифрами, знаками пробела или некоторыми спецсимволами.

3. Лемматизация (используется библиотека pymorphy2).

4. Удаление стоп-слов и служебных частей речи (союзов, предлогов и местоимений).

На втором этапе происходит векторизация текстов – учебных курсов и запросов из требований профессиональных стандартов с использованием одной из моделей векторного представления текстов, их подробное описание приводится в следующем разделе. Затем выполняется сопоставление запросов с каждым из документов из текстовой коллекции учебных дисциплин и онлайн-курсов. В ходе сопоставления определяется косинусная мера между векторами запроса и документа из коллекции – традиционная мера семантической близости в задачах поиска. На основе полученных значений мер семантической близости производится ранжирование и формирование поисковой выдачи под отдельный запрос – конкретное требование профессионального стандарта.

На последнем этапе результаты поиска для отдельных запросов могут быть объединены путем суммирования мер семантической близости для каждого из релевантных документов по отношению ко всем запросам. При этом эксперт или интеллектуальная система поддержки принятия решений может задать различные веса – степень важности для отдельных требований профстандарта. Это позволяет гибко определять итоговый набор учебных дисциплин и онлайн-курсов, подходящих под комплекс заданных требований на уровне всего профстандарта или ограниченных определенными уровнями квалификации.

3. Векторные модели (статистические и нейросетевые модели языка)

В данном разделе описываются основные векторные модели языка, которые могут использоваться в предлагаемом методе семантического поиска.

3.1. TF-IDF

TF-IDF – это классическая схема взвешивания важности слов в коллекции документов. Данный алгоритм состоит из двух компонентов: TF (term frequency, частота термина) – это мера, считающая слова тем важнее, чем чаще они встречаются в коллекции; IDF (inverse document frequency, обратная документная частота) – мера, дающая больший вес тем словам, которые встречаются в меньшем числе документов. Комбинация этих двух схем взвешивания позволяет учитывать важность и релевантность каждого слова по отношению к коллекции документов, при этом снижая влияние общеупотребимой лексики и стоп-слов.

Векторное представление каждого слова в данной модели – это вектор с размерностью в количество уникальных слов в словаре коллекции, с единственным ненулевым измерением, соответствующим уникальному идентификатору этого слова в словаре и равным TF-IDF весу этого слова. Векторы документов же получаются путем сложения векторов всех слов этого документа.

3.2. Word2vec

Word2vec [7] – нейросетевая языковая модель, основывающаяся на неглубокой нейронной сети с одним скрытым слоем. Задачей обучения для этой сети ставится сопоставление слов их контекстам, т. е. определение наиболее вероятного слова по его контексту и наиболее вероятного контекста по заданному слову. В процессе оптимизации по этой задаче нейронная сеть обучается сопоставлять словам в тексте векторные представления – так называемые IN-представления для входящих слов и OUT-представления для выходящих. Как правило, после обучения используются только матрица IN, однако практика показывает [10], что в то время как косинусная близость для векторов IN-IN (оба вектора из IN-представления) выше для функционально близких слов, косинусная близость для IN-OUT векторов выше для слов, которые часто совместно употребляются. В наших экспериментах мы исследуем оба варианта векторного представления.

Для получения векторного представления всего документа, как правило, производится усреднение векторов входящих в документ слов, иногда с учетом индивидуального веса каждого слова (например, по IDF [11]).

Также в текущем исследовании рассматривается и вариант расширения исходного поискового запроса наиболее близкими словами по модели word2vec с дальнейшим формированием вектора расширенного запроса на основе TF-IDF.

3.3. Paragraph2vec

Paragraph2vec [12], также известный как doc2vec, – это развитие модели word2vec, стремящееся к созданию векторов документов таким же образом, как и вектора слов. Для этого в архитектуру нейронной сети вносится изменение: помимо векторов слов, употребляющихся совместно с целевым словом, контекстом также считается и сам документ, представляемый собственным вектором, попадающим в то же пространство, что и отдельные слова. Ранее в работе [9] данная модель показала довольно низкие результаты при оценке качества поиска образовательного контента, поэтому в нашем исследовании ее качество не оценивается.

3.4. FastText

FastText [13] является алгоритмом, схожим с word2vec, однако с одним ключевым отличием: word2vec воспринимает каждый элемент последовательности токенов (слово, число, символ) как отдельную неделимую сущность, не имеющую внутренней структуры, тогда как fastText учитывает тот факт, что слова имеют внутренние взаимосвязанные компоненты. Различные слова, соответственно, могут быть связаны синтаксически или семантически, если они имеют одни и те же компоненты (например, однокоренные слова). Данная модель может более эффективно решать задачи семантической близости в сравнении с word2vec особенно для флективных языков, к которым относится и русский язык.

Таким образом, fastText использует отдельные символьные n-граммы в качестве элементов текста, обучаясь сопоставлять каждый n-грамм с некоторым векторным представлением, генерируя векторные представления целых слов, исходя из векторов входящих в них символьных n-граммов.

4. Описание текстовых корпусов для векторных моделей

В табл. 1 представлены характеристики основных текстовых корпусов, используемых в данном исследовании. Корпуса используются как для обучения нейросетевых моделей word2vec и fastText, так и в качестве коллекции документов, в которой производится поиск.

Таблица 1

Характеристики текстовых корпусов учебных курсов

Корпус	Число источников	Число документов (учебных курсов)	Число токенов	Число уникальных токенов (словарь)
Корпус курсов из МООС-платформ	4 МООС-платформы	2051	249 тыс.	14 тыс.
Корпус рабочих программ дисциплин	10 университетов	976	351 тыс.	16 тыс.
Общее количество	14 источников	3027	600 тыс.	23 тыс.

В табл. 2 представлены детальные данные по МООС-платформам и корпусу рабочих программ дисциплин (РПД), собранные с сайтов 10 крупных российских вузов по направлениям подготовки, связанным с информационными технологиями.

Таблица 2

Детализация корпусов онлайн-курсов и учебных дисциплин

	Stepik	OpenEDU	Coursera	Intuit	РПД
Число документов	554	315	275	907	976
Количество токенов	39 394	93 866	39 669	75 942	351 459
Количество уникальных токенов	6098	6574	6113	6693	15 688
Средняя длина документа	71	297	144	83	360

5. Эксперимент по оценке качества семантического поиска

Для экспериментов в качестве поисковых запросов использовались формулировки трудовых функций, а также отдельных трудовых действий и требований к знаниям и умениям из трех профессиональных стандартов: «Программист», «Системный администратор информационно-коммуникационных систем», «Специалист по информационным системам».

Общее число различных поисковых запросов: 78.

К оценке качества поиска были привлечены эксперты-работодатели, преподающие в вузах профильные дисциплины, связанные с указанными профессиями. Эксперты оценивали релевантность результатов поиска для каждой из моделей по шкале от 1 до 5. После разметки несогласованные оценки экспертов были либо удалены, либо коллегиально пересмотрены.

Для оценки качества использовались традиционные метрики для информационного поиска: Mean average precision (MAP) и Normalized Discounted Cumulative Gain (nDCG).

В табл. 3 приведены параметры обучения для используемых нейросетевых моделей, показавшие лучшие результаты в поиске. Нейросетевые модели были обучены на исходном текстовом корпусе учебных курсов с использованием предобработки текста, описанной выше в методе. Модели, которые были обучены на открытых интернет-корпусах (НКРЯ, Википедиа, Araneum), показывают себя в задаче семантического поиска курсов существенно хуже и в итоговой оценке качества не приводятся.

Таблица 3

Параметры нейросетевых моделей языка (показавших лучшие результаты)

Модель	Архитектура	Размерность	Min частота встречи слова	Эпохи
Word2vec	CBOW	300	3	50
FastText	CBOW	300	3	50

Модель TF-IDF – expansion w2v реализует концепцию расширения исходного короткого запроса из 8–12 слов описания требования профстандарта близкими словами по модели word2vec. Лучшим для этой модели себя показал вариант с расширением каждого слова 3 ближайшими словами.

В табл. 4 представлены результаты оценки точности поиска на глубину выдачи – 1, 5 и 10, а также приведены результаты для двух границ релевантности результатов по оценке экспертов: 2,5 и 3,5. Лучшие результаты показывает модель усредненного word2vec. Стоит отметить, что для более точного поиска по релевантности с 3,5 результаты модели fastText приближаются по качеству к word2vec. Нейросетевые модели с применением сопоставления IN-OUT матриц при расчете семантической близости показывают себя хуже в этой задаче во всех случаях. Модель TF-IDF с расширением запроса по word2vec для поисковой выдачи глубины 5 и 10 показывает лучшие результаты, чем исходная модель TF-IDF без модификации запроса.

Таблица 4

Результаты эксперимента по оценке точности семантического поиска курсов

Модель	Relevance ≥ 2,5			Relevance ≥ 3,5		
	MAP@1	MAP@5	MAP@10	MAP@1	MAP@5	MAP@10
TF_IDF	0,903	0,752	0,660	**0,801**	0,607	0,518
TF-IDF – expansion w2v	0,830	0,768	0,699	0,667	0,618	0,521
Avg. Word2Vec	**0,907**	**0,865**	**0,836**	0,702	**0,647**	**0,616**
Avg Word2vec (IN-OUT)	0,930	0,825	0,749	0,733	0,631	0,542
Avg. Word2Vec – TF-IDF	0,867	0,816	0,789	0,667	0,593	0,562
Avg. Word2Vec – TF-IDF (IN-OUT)	0,833	0,773	0,736	0,733	0,620	0,572
Avg. FastText	0,867	0,811	0,804	0,733	**0,642**	**0,593**
Avg. FastText (IN-OUT)	0,759	0,643	0,604	0,586	0,498	0,448

На рис. 2 показана оценка точности поиска по метрике MAP для глубины выдачи 5 (MAP@5) для разного порога релевантности курсов – от 1,5 до самого точного варианта поиска с оценкой экспертов выше 4,5.

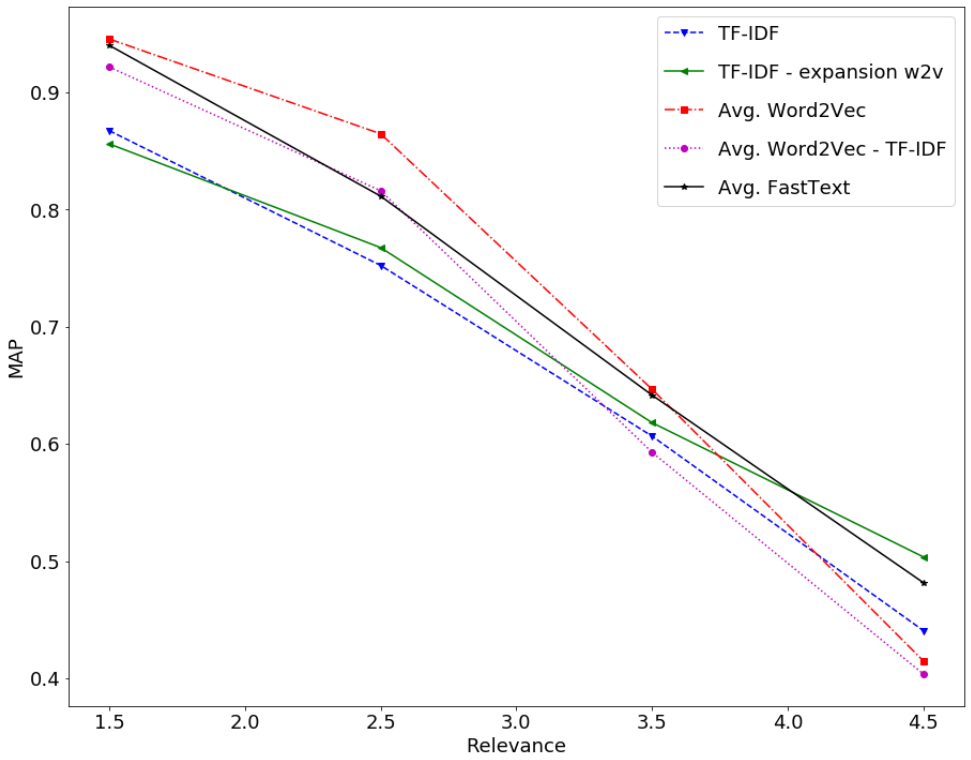

Рис. 2. Оценка качества точности поиска (MAP) векторных моделей при разном значении порога релевантности результатов (от 1,5 до 4,5)

Для самых высоких порогов релевантности лучше всего себя показывает TF-IDF с расширением запроса по word2vec. Кроме этого, можно заметить, что качество модели FastText снижается гораздо медленнее при повышении порога релевантности, чем качество моделей Word2vec. Это делает модель FastText более предпочтительной при высоких требованиях к точности поиска, когда мы ожидаем в качестве результатов более узкие предметные учебные курсы, максимально близкие к исходному запросу – требованию профстандарта, а не более широкие базовые профессиональные курсы, покрывающие требование лишь частично.

В табл. 5 приведены результаты оценки качества ранжирования по метрике nDCG. Лучшие результаты демонстрирует модель FastText, что еще раз свидетельствует об устойчивости модели к поиску более узких предметных курсов, максимально полно отвечающих исходному запросу рынка труда.

Таблица 5

Результаты эксперимента по оценке качества ранжирования (метрика nDCG) семантического поиска курсов

Модель	nDCG@3	nDCG@5	nDCG@10
TF_IDF	0,742	0,729	0,757
TF-IDF – expansion w2v	0,716	0,719	0,733
Avg. Word2Vec	0,685	0,702	0,746
Avg Word2vec (IN-OUT)	0,683	0,709	0,724
Avg. Word2Vec – TF-IDF	0,643	0,677	0,717
Avg. Word2Vec–TF-IDF (IN-OUT)	0,647	0,663	0,726
Avg. FastText	**0,771**	**0,778**	**0,811**
Avg. FastText (IN-OUT)	0,593	0,606	0,689

В табл. 6 приведен пример поисковой выдачи для иллюстрации практического применения подхода объединения запросов с требованиями профстандарта при поиске учебных курсов для одной профессии «Программист» по суммарной семантической близости на основе нейросетевой модели FastText как последний шаг предлагаемого метода семантического поиска программ дисциплин с требованиями рынка труда.

Таблица 6

Пример поисковой выдачи: Топ-10 учебных дисциплин, проранжированных по релевантности по суммарной семантической близости к требованиям профессионального стандарта «Программист»

№	Название учебной дисциплины	Σ(sim)
1	Программная инженерия	3,0026
2	Тестирование и отладка программного обеспечения	2,6837
3	Технологии разработки программного продукта	2,4050
4	Проектирование программных систем	2,3993
5	Технологии разработки программного обеспечения	2,3532
6	Верификация и аттестация программного обеспечения	2,3300
7	Методы тестирования программного обеспечения	2,3129
8	Технологии разработки программного обеспечения для мобильных устройств	1,7825
9	Введение в генерацию программного кода	1,7017
10	Анализ требований к программному обеспечению	1,3147

Заключение

В данной работе предложен метод семантического поиска учебных дисциплин и онлайн-курсов под заданные требования профессиональных стандартов. Проведена экспериментальная оценка качества нейросетевых моделей word2vec и fastText в сравнении со статистической векторной моделью TF-IDF. Модели усредненного word2vec и fastText показали лучшие результаты по точности поиска по метрике MAP, при этом модель fastText также обеспечивает лучшее качество ранжирования выдачи по метрике nDCG.

В качестве дальнейших шагов исследования предлагается применить модели тематического моделирования (PLSA, LDA) с аддитивной регуляризацией для предварительного сужения пространства поиска (коллекции курсов) под заданные требования определенных профессиональных стандартов, что позволит повысить качество поиска на большую глубину выдачи и заранее отсечь курсы, которые нерелевантны для определенной профессиональной области. Также предлагается реализовать оценку важности со стороны работодателей тех или иных требований профстандартов путем семантического анализа текстов вакансий в системах онлайн-рекрутмента на основе нейросетевых моделей языка.

Исследование выполняется при поддержке Российского фонда фундаментальных исследований в рамках проекта № 18-47-860013 р_а «Интеллектуальная система формирования образовательных программ на основе нейросетевых моделей естественного языка с учетом потребностей цифровой экономики Ханты-Мансийского автономного округа – Югры» (договор № 18-47-860013\18).

Литература

1. Федеральный закон от 29 декабря 2012 г. N 273-ФЗ «Об образовании в Российской Федерации. – http://ivo.garant.ru/#/document/70291362/ (дата обращения: 26 декабря 2018).

2. Федеральный закон от 3 июля 2016 г. N 238-ФЗ «О независимой оценке квалификации». – http://ivo.garant.ru/#/document/71433946/ (дата обращения: 26 декабря 2018).

3. Сметанина, О.Н. Методологические основы управления образовательным маршрутом с использованием интеллектуальной информационной поддержки / О.Н. Сметанина. – УГАТУ, 2012. 446 с.

4. Лисицына, Л.С. Автоматизация управления образовательными траекториями для разработки модульных компетентностно-ориентированных образовательных программ вуза / Л.С. Лисицына, А.С. Пирская // Сборник трудов Всероссийской научно-практической конференции с международным участием «Информационные технологии в обеспечении нового качества высшего образования». – М., 2010. – С. 75–86.

5. Черникова, Е.А. Формализация и сравнение учебных программ на основе онтологического подхода / Е.А. Черникова, А.С. Черников // Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». Спецвыпуск «Информационные технологии и компьютерные системы». – 2011. – С. 101–104.

6. Мониторинг соответствия профессионального образования потребностям рынка труда / С.Д. Валентей, П.В. Зрелов, В.В. Кореньков и др. // Общественные науки и современность. – 2018. – № 3. – С. 5–16.

7. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean. – arXiv preprint arXiv:1301.3781, 2013.

8. Architecture of a concept-based information retrieval system for educational resources / R. Pérez-Rodríguez, L. Anido-Rifón, M. Gómez-Carballa, M. Mouriño-García // Science of Computer Programming. – 2016, no. 129. – P. 72–91. DOI: 10.1016/j.scico.2016.05.005

9. Klenin, J. Comparison of Vector Space Representations of Documents for the Task of Information Retrieval of Massive Open Online Courses / J. Klenin, D. Botov, Y. Dmitrin // Proceedings Conference on Artificial Intelligence and Natural Language. – Cham: Springer, 2017. – P. 156–164. DOI: 10.1007/978-3-319-71746-3_14

10. Improving document ranking with dual word embeddings / E. Nalisnick, B. Mitra, N. Craswell, R. Caruana // Proceedings of the 25th International Conference Companion on World Wide Web. – International World Wide Web Conferences Steering Committee, 2016. – P. 83–84. DOI: 10.1145/2872518.2889361

11. Lilleberg, J. Support vector machines and word2vec for text classification with semantic features / J. Lilleberg, Y. Zhu, Y. Zhang // Proceedings of International Conference on Cognitive Informatics & Cognitive Computing (ICCI CC), 2015 IEEE 14th. – 2015. – P. 136–140. DOI: 10.1109/ICCI-CC.2015.7259377*

12. Le, Q. Distributed representations of sentences and documents / Q. Le, T. Mikolov // Proceedings of International Conference on Machine Learning. – 2014. – P. 1188–1196.

13. Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. – arXiv preprint arXiv:1607.04606, 2016. DOI: 10.1162/tacl_a_00051

Ботов Дмитрий Сергеевич, старший преподаватель кафедры информационных технологий и экономической информатики, Челябинский государственный университет, г. Челябинск; dmbotov@gmail.com.

Дмитрин Юрий Владиславович, аспирант кафедры информационных технологий и экономической информатики, Челябинский государственный университет, г. Челябинск; dmitrinyuri@gmail.com.

Кленин Юлий Дмитриевич, аспирант кафедры информационных технологий и экономической информатики, Челябинский государственный университет, г. Челябинск; jklen@yandex.ru.

Поступила в редакцию 20 декабря 2018 г.

DOI: 10.14529/ctcr190201

NEURAL NETWORK-BASED SEMANTIC SEARCH OF EDUCATIONAL PROGRAMMES FITTING LABOR MARKET REQUIREMENTS

D.S. Botov, *dmbotov@gmail.com,*
Yu.V. Dmitrin, *dmitrinyuri@gmail.com,*
Yu.D. Klenin, *jklen@yandex.ru*
Chelyabinsk State University, Chelyabinsk, Russian Federation

With the growth of open educational content, growing demand for professional education from the labor market, and the development of the concept of lifelong learning, the task of updating the content of educational programs today is extremely important. The article discusses the semantic search method to retrieval and ranking of educational content for the specified requirements of the labor market, determined by professional standards. In contrast to traditional approaches of matching and analyzing the content of educational programs based on ontological models and rules, we propose the usage of word embedding and well-known neural network language models word2vec and fastText. The initial requests are specific requirements for knowledge, skills and descriptions of labor activities and labor functions extracted from professional standards. The search results are the descriptions of academic disciplines and online courses, including goals and objectives, learning outcomes, the structure and content of the main topics. We include the results of the expert evaluation of the ranking quality for the semantic search by metrics NDCG (Normalized Discounted Cumulative Gain) and MAP (Mean Average Precision) on the representative corpus of IT disciplines programmes of universities and massive open online courses (MOOC). The best results for the search are shown by the word2vec and fastText models, which are trained without supervision on large specially prepared corpuses of curriculum programs and descriptions of online courses. To move from word vectors to document vectors various combinations of neural network models with the TF-IDF weighting scheme are investigated.

Keywords: semantic search, semantic similarity, distributional semantic, word2vec, fastText, academic discipline, massive open online courses, labor market.

References

1. *Federal'nyy zakon ot 29 dekabrya 2012 g. N 273-FZ "Ob obrazovanii v Rossiyskoy Federatsii"* [Federal Law of December 29, 2012 No. 273-FZ "About Education in the Russian Federation"]. Available at: http://ivo.garant.ru/#/document/70291362/ (accessed 26 December 2018).

2. *Federal'nyy zakon ot 3 iyulya 2016 g. N 238-FZ "O nezavisimoy otsenke kvalifikatsii"* [Federal Law of July 3, 2016 No. 238-FZ "About Independent Assessment of Qualifications"]. Available at: http://ivo.garant.ru/#/document/71433946/ (accessed 26 December 2018).

3. Smetanina O.N. *Metodologicheskie osnovy upravleniya obrazovatel'nym marshrutom s ispol'zovaniem intellektual'noy informatsionnoy podderzhki* [Methodological Basis of Educational Route Management Using Intellectual Information Support]. USATU Publ., 2012. 446 p.

4. Lisitsyna L.S., Pirskaya A.S. [Automation of Management of Educational Trajectories for the Development of Modular Competence-Oriented Educational Programs of the University]. *Sbornik trudov Vserossiyskoy nauchno-prakticheskoy konferentsii s mezhdunarodnym uchastiem "Informacionnye tekhnologii v obespechenii novogo kachestva vysshego obrazovaniya"*. [In Proc. of the All-Russian Scientific-Practical Conference with International Participation "Information Technology in Providing a New Quality of Higher Education"]. Moscow, 2010, pp. 75–86. (in Russ.)

5. Chernikova E.A., Chernikov A.S. [Formalization and Comparison of Curricula Based on the Ontological Approach]. *Herald of the Bauman Moscow State Technical University. Ser. Instrument Engineering. Special Edition "Information Technology and Computer Systems"*, 2011, pp. 101–104. (in Russ.)

6. Valentey S.D., Zrelov P.V., Korenkov V.V., Belov S.D., Kadochnikov I.S. [Monitoring Matching of Professional Education and Labor Market Requirements]. *Social Sciences and Modernity*, 2018, no. 3, pp. 5–16. (in Russ.)

7. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013.

8. Pérez-Rodríguez R., Anido-Rifón L., Gómez-Carballa M., Mouriño-García M. Architecture of a Concept-Based Information Retrieval System for Educational Resources. *Science of Computer Programming*, 2016, no. 129, pp. 72–91. DOI: 10.1016/j.scico.2016.05.005

9. Klenin J., Botov D., Dmitrin Y. Comparison of Vector Space Representations of Documents for the Task of Information Retrieval of Massive Open Online Courses. *Proceedings Conference on Artificial Intelligence and Natural Language*, Springer, Cham, 2017, pp. 156–164. DOI: 10.1007/978-3-319-71746-3_14

10. Nalisnick E., Mitra B., Craswell N., Caruana R. Improving Document Ranking with Dual Word Embeddings. *Proceedings of the 25th International Conference Companion on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 83–84. DOI: 10.1145/2872518.2889361

11. Lilleberg J., Zhu Y., Zhang Y. Support Vector Machines and Word2vec for Text Classification with Semantic Features. *Proceedings of International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th*, 2015. pp. 136–140. DOI: 10.1109/ICCI-CC.2015.7259377

12. Le Q., Mikolov T. Distributed Representations of Sentences and Documents. *Proceedings of International Conference on Machine Learning*, 2014, pp. 1188–1196.

13. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*, 2016. DOI: 10.1162/tacl_a_00051

Received 20 December 2018

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Ботов, Д.С. Семантический поиск учебных дисциплин под требования рынка труда на основе нейросетевых моделей языка / Д.С. Ботов, Ю.В. Дмитрин, Ю.Д. Кленин // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2019. – Т. 19, № 2. – С. 5–15. DOI: 10.14529/ctcr190201

FOR CITATION

Botov D.S., Dmitrin Yu.V., Klenin Yu.D. Neural Network-Based Semantic Search of Educational Programmes Fitting Labor Market Requirements. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2019, vol. 19, no. 2, pp. 5–15. (in Russ.) DOI: 10.14529/ctcr190201