Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Челябинский государственный университет»

Ботов Дмитрий Сергеевич

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ФОРМИРОВАНИЯ ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ ПО ТРЕБОВАНИЯМ РЫНКА ТРУДА НА ОСНОВЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ЯЗЫКА

Научный руководитель
доктор технических наук,
профессор А.В. Мельников

Челябинск – 2019

Системные проблемы образования

Федеральный проект “Кадры для цифровой экономики”

1. Образование оторвано от работодателя
2. Устаревание федеральных государственных образовательных стандартов в силу отсутствия эффективного инструмента быстрого обновления их содержания и соблюдения баланса «теория-практика»
3. Длительный цикл внедрения профессиональных стандартов (7 лет)
4. Взаимодействие предприятий и учебных заведений без нормативной и единой методической базы
5. Субъективность при выборе, разработке и актуализации профилей образовательных программ и индивидуальных ОТ

Развитие массовых открытых онлайн-курсов

CLASS CENTRAL

By the Numbers: MOOCs in 2018

Основные понятия предметной области

Образовательная программа (ОП) — комплекс основных характеристик образования (объём, ***содержание***, ***планируемые результаты***), который представлен в виде учебного плана, календарного учебного графика, ***рабочих программ учебных предметов, курсов, дисциплин (модулей)***, иных компонентов, а также оценочных и методических материалов.

Компетенция — способность применять знания, умения и опыт в трудовой деятельности.

Результат обучения — это формулировка того, что, как ожидается, будет знать, понимать и / или будет в состоянии продемонстрировать обучающийся после завершения процесса обучения. (ECTS, 2006).

Формирование требований к результатам освоения ОП по ФГОС ВО 3++

Определены только универсальные компетенции (УК) и общепрофессиональные компетенции (ОПК)

Профессиональные компетенции формируются организацией самостоятельно на основе:

- профессиональных стандартов;
- примерных основных образовательных программ;
- анализа требований работодателей, предъявляемых к выпускникам на рынке труда;
- обобщения отечественного и зарубежного опыта.

Жизненный цикл образовательной программы

Анализ предметной области

Выявленные проблемы:

- Существенная разница лексики образовательной и профессиональной областей
- Ограниченность лексики стандартов
- Короткие тексты
- Отсутствие качественной разметки
- Отсутствие полных и обновляемых онтологических моделей и тезаурусов

Обзор работ по интеллектуальной поддержке принятий решений при формировании ОП

Модели	Методы и алгоритмы	Анализ текста
• Онтологические модели • Дискретно-событийные модели • Когнитивные карты • Графовые модели (план-графы)	• Методы экспертных оценок • Логический вывод на правилах • Рассуждения по прецедентам • Эвристические алгоритмы	• Латентно-семантический анализ • Частотный анализ (TF-IDF) • Контент-анализ

Цель исследования

Целью работы является разработка новых эффективных методов и алгоритмов интеллектуальной поддержки формирования образовательных программ, которые обеспечат соответствие содержания образовательных программ актуальным требованиям рынка труда.

Задачи исследования

1. Разработать математическую модель образовательной программы и требований рынка труда в виде графа сущностей документов со структурными и семантическими отношениями.
2. Разработать интеллектуальный метод и алгоритм семантического сопоставления сущностей графов образовательных программ, отдельных учебных дисциплин и онлайн-курсов и сущностей графов требований рынка труда и профессиональных стандартов на основе нейросетевых моделей языка.
3. Разработать методику структурно-семантического анализа и обработки текстовой информации для поддержки принятия решений в процессе формирования образовательных программ под требования рынка труда с использованием методов классификации и кластеризации текстов и предложенного метода семантического сопоставления.
4. Выполнить программную реализацию предложенных методов, моделей и алгоритмов в виде прототипа интеллектуальной системы поддержки принятия решений при формировании образовательных программ.
5. Провести оценку качества и результативности реализованных методов и алгоритмов интеллектуальной поддержки на представительном текстовом корпусе образовательных программ различных вузов и платформ онлайн-курсов.

Векторные представления слов и документов в задачах информационного поиска

- **Частотный анализ:**
  - TF-IDF [*Karen Spärck Jones, 1972*]
  - BM25 [*Stephen E. Robertson et al., 1994*]
- **Тематическое моделирование**:
  - LSA - Latent Semantic Analysis *[Deerwester, Dumais, Furnas, Harshman et al., 1988]*
  - PLSA - Probabilistic Latent Semantic Analysis [*Hofmann, 1999*]
  - LDA - Latent Dirichlet Allocation *[Blei, Ng, Jordan, 2003]*
  - ARTM - Аддитивная регуляризация тематических моделей [*Воронцов, 2014*]
- **Распределенные представления слов (word embedding) и документов**
  - word2vec [Mikolov et al., 2013]
  - GloVe: Global Vectors for Word Representation [Pennington, Socher, Manning, 2014]
  - Paragraph2vec: Distributed Representations of Documents [Le and Mikolov, 2014]
  - Dual Embedding Space Model (DESM) [Nalisnick et al., 2016]
  - Fasttext [P. Bojanowski, E. Grave et al., 2016]

Векторные представления слов и документов в задачах информационного поиска

- Частотный анализ: TF-IDF [*Karen Spärck Jones, 1972*], BM25 [*Stephen E. Robertson et al., 1994*]
- **Тематическое моделирование**:
  - Latent Semantic Indexing *[Deerwester, Dumais, Furnas, Harshman et al., 1988]*
  - Probabilistic Latent Semantic Indexing [*Hofmann, 1999*]
  - Latent Dirichlet Allocation *[Blei, Ng, Jordan, 2003]*
  - Аддитивная регуляризация тематических моделей [*Воронцов, 2014*]
- **Распределенные представления слов (Word embedding)**
  - word2vec [Mikolov et al., 2013]
  - GloVe: Global Vectors for Word Representation [Pennington, Socher, Manning, 2014]
  - Averaging the word representations [Vulic and Moens, 2015, Zamani and Croft, 2016b]
  - Paragraph2vec: Distributed Representations of Documents [Le and Mikolov, 2014]
  - Dual Embedding Space Model (DESM) [Nalisnick et al., 2016]
  - Fasttext [P. Bojanowski, E. Grave et al., 2016]
- Глубокое обучение
  - Deep Structured Semantic Model (DSSM) [Huang et al., 2013]
  - Interaction matrix based approaches with DNN [Hu et al. [2014], Mitra et al. [2017], Pang et al. [2016]
- Обучение ранжированию
  - Learning to rank [Liu , 2009]

Нейросетевая модель естественного языка word2vec [Mikolov et al., 2013]

Дистрибутивная гипотеза: лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения [Harris, 1962]

Семантический вектор слова - это представление слова в качестве точки в многомерном пространстве $\boldsymbol{R}_d$ (вектора в векторном пространстве) с учетом смысла слова в контексте.

ϕ: $\boldsymbol{w} \longmapsto \boldsymbol{R}_d$

Векторное представление текста (документа)

- ***Семантический вектор документа*** - это представление текстового описания сущности документа в качестве точки в многомерном пространстве $\boldsymbol{R}_d$.

  ϕ: $\boldsymbol{d} \longmapsto \boldsymbol{R}_d$

$$d_i = \frac{\sum_{j \in N} w_{i,j}}{|\, d_i \,|}$$

Averaging the word representations
[Vulic and Moens, 2015, Zamani and Croft, 2016b]

$$d_i = \frac{\sum_{j \in N} TFIDF(w_{i,j}) * w_{i,j}}{|\, d_i \,|}$$

Lilleberg J., Zhu Y. and Zhang Y., Support vector machines and Word2vec for text classification with semantic features (2015)

Графовая модель образовательной области

$$G_{edu} = (V_{edu}, E_{edu})$$

$$V_{edu} = \{C, LO\}$$

$$E_{edu} = \{Inc, Req, Match\}$$

$$Inc \subset C \times C \cup LO \times LO$$
$$Req \subset C \times LO$$
$$Match \subset V \times V$$

Графовая модель профессиональной области

$$G_{prof} = (V_{prof}, E_{prof})$$

$$V_{prof} = \{F, S\}$$

$$E_{prof} = \{Inc, Req, Match\}$$

$$Inc \subset F \times F$$
$$Req \subset F \times S$$
$$Match \subset V \times V$$

Сопоставление графов профессиональной области и образовательной программы

Векторное представление вершины графа

Семантический вектор сущности графа - это представление вершины графа в качестве точки в многомерном пространстве $\boldsymbol{R}_{\mathrm{d}}$ (вектора в векторном пространстве) с учетом смысла текста самой сущности и смыслов из контекста связанных сущностей.

$$\phi\colon \boldsymbol{v} \longmapsto \boldsymbol{R}_{\mathrm{d}}$$

Близко лежащие друг к другу точки соответствуют семантически схожим вершинам.

Метод семантического сопоставления сущностей графов на основе семантической близости

Мера семантической близости — это численная мера, показывающая степень схожести двух объектов. Как правило, она выражается в виде скалярной величины в диапазоне [0; 1]

Косинусная мера семантической близости: $$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Семантическое сопоставление — процесс установления отношений на основе меры семантической близости между сущностями ***A*** и ***B*** графов.

Метод семантического сопоставления сущностей графов на основе семантической близости

Этапы поддержки принятия решений при формировании ОП на основе семантического сопоставления сущностей

Этап 0. Построение графов сущностей из документов со структурными отношениями

Этап 1. Определение актуальных требований рынка труда в увязке с требованиями профстандарта.

Этап 2. Определение соответствия элементов образовательной программы (ОП) требованиям рынка труда.

Этап 3. Поиск программ учебных дисциплин и онлайн-курсов под профессиональные компетенции и требования рынка труда.

Этап 4. Сопоставление элементов ОП с отобранными учебными дисциплинами и онлайн-курсами

Построение графов сущностей со структурными отношениями из документов

На текущий момент для сферы подготовки ИТ-специалистов собраны и преобразованы в структуру следующие данные:

- 20 профессиональных стандартов (“Связь, информационные и коммуникационные технологии”)
- 687 тыс. вакансий (headhunter, superjob)
- 12 направлений подготовки бакалавриата и магистратуры
- >3000 учебных курсов (10 вузов, Coursera, Stepic, OpenEdu, Интуит)

Методика структурно-семантического анализа графов сущностей при формировании ОП по требованиям рынка труда (1)

Шаги	Этап 1	Этап 2	Этап 3	Этап 4
Шаг 1. Определение пространства сопоставления	Вакансии ↓ Профессии	Профессии (ур. квалификации) ↓ Образовательная программа	Дисциплины ↓ Тематические кластера дисциплин	Отобранные учебные курсы ↓ Дисциплины ОП
Шаг 2. Семантическое сопоставление сущностей графов с учетом контекста	Требования вакансий ↓ sim() Трудовые функции ПС + Действия/ Требования к умениям/ знаниям	Проф. компетенция ↓ sim() Результаты обучения / Раздел + Тема дисциплины	Содержание и цели курса ↓ sim() Требования рынка труда	Элементы учебных курсов ↓ sim() Элементы дисциплин ОП
Шаг 3. Определение веса (важности) сущностей графа по агрегации структурных и семантических отношений	w(Трудовая функция) ↑ agg(sim) w(Действия/ умения ТФ) ↑ agg(sim) Требования/обязанности вакансий	w(Дисциплины) ↑ agg(sim) w(Результаты обучения) w(Раздел/Темы) ↑ agg(sim) w(Профессиональная компетенция)	w(Дисциплины) ↑ agg(sim) w(Элемент дисциплины) ↑ agg(sim) w(Профессиональная компетенция)	w(Дисциплины ОП) ↑ agg(sim) w(Элемента дисциплины) ↑ agg(sim) w(Элемента учебного курса)

Методика структурно-семантического анализа графов сущностей при формировании ОП по требованиям рынка труда (2)

Шаги	Этап 1	Этап 2	Этап 3	Этап 4
Шаг 4. Вывод результатов	Визуализация графа требований рынка труда путем расширения требований ПС ключевыми словами из требований вакансий	Визуализация графа ОП с оценкой соответствия профессиональным компетенциям	Ранжирование курсов по релевантности профессиональным компетенциям с визуализацией тематических кластеров	Удаление не релевантных элементов ОП Добавление новых дисциплин и отдельных элементов курсов (разделов / тем)
Шаг 5. Принятие решений экспертом	Формирование профессиональных компетенций	Удаление элементов ОП, не соответствующих требованиям рынка труда Дополнение результатов обучения актуальными требованиями рынка труда	Отбор онлайн-курсов и учебных дисциплин вузов	Формирование актуализированной версии ОП / ИОТ

Эксперименты по обучению multi-label классификатора вакансий по профессиям

Multi-labeling Classifier/Strategy	Базовый классификатор	Векторное представление	F1-micro	F1-macro
OneVsRestClassifier	Logistic Regression	TF-IDF	0.8384	0.8349
		Avg. Word2Vec (CBOW)	0.8522	0.8515
LabelPowerset	Logistic Regression	TF-IDF	**0.8767**	**0.8782**
		Avg. Word2Vec (CBOW)	0.8662	0.8682
		Doc2Vec (DBOW)	0.6182	0.6176
Convolutional Neural Networks (CNN)		Char N-grams with punct.	0.8730	0.8764
		Word Emb with lemm	0.8734	0.8796
		Word Emb w/o lemm	**0.8900**	**0.8876**

Эксперименты по обучению классификатора вакансий по профессиям

Метрики оценки качества сопоставления как задачи ранжирования

Average Precision

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \mathrm{rel}(k))}{\text{number of relevant documents}}$$

Mean Average Precision

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Discounted Cumulative Gain

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

Normalized DCG

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p} \qquad \mathrm{IDCG_p} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Этап 1. Определение актуальных требований рынка труда в увязке с профстандартами

Метод сопоставления	**Векторная модель**	**Трудовые функции**		**Требования / Трудовые действия**	
		MAP@1	**MAP@5**	**MAP@1**	**MAP@5**
Сопоставление без предварительной классификации вакансий	Avg. word2vec (CBOW)	0.4504	0.3940	0.3572	0.3450
Сопоставление с предварительной классификацией вакансий	TF-IDF	0.6734	0.5460	0.6134	0.6250
	Avg. word2vec (CBOW)	**0.9967**	**0.9500**	**0.8333**	**0.8167**

Этап 1. Примеры семантического сопоставления

Описание вакансии	Профессиональные стандарты	
Требование вакансии или трудовая обязанность	**Трудовая функция ПС**	**Трудовое действие / требование ПС**
Построение процесса continuous integration/continuous delivery	Разработка процедур интеграции программных модулей	Применять методы и средства сборки модулей и компонент ПО, разработки процедур для развертывания ПО <...>
Работа в проектной команде (разработчики,аналитики,ключевые пользователи).	Разработка требований и проектирование программного обеспечения	Распределение заданий между программистами в соответствии с техническими спецификациями
Общение с заказчиком и командой Бекенд-программистов	Разработка требований и проектирование программного обеспечения	Осуществлять коммуникации с заинтересованными сторонами
Оказание услуг профессионального консалтинга для заказчиков	Анализ требований к программному обеспечению	Проводить оценку и обоснование рекомендуемых решений
Доработка имеющегося функционала в соответствии с запросами пользователей	Проверка работоспособности и рефакторинг кода ПО	Внесение изменений в программный код для устранения выявленных дефектов

Этап 1. Обогащение требований профстандарта ключевыми словами из вакансий

a) Ключевые словосочетания из вакансий для профессии "Администратор БД" (TF)

b) Ключевые слова для профессии “Разработчик web-приложений” (TF-IDF)

Этап 2. Сопоставление результатов обучения и тем/разделов дисциплин ОП с требованиями рынка труда

Векторная модель: average word2vec

Сопоставляемые сущности	MAP@1	MAP@3	MAP@5	MAP@10
Требования ПС ↓ sim() Раздел+тема дисциплины	0,867	0,833	0,816	0,765
Требования ПС ↓ sim() Проф.компетенция + результат обучения	0,905	0,883	0,875	0,813

Этап 3. Семантический поиск курсов по требованиям рынка труда

Модель	MAP@1	MAP@10	nDCG@10
TF-IDF	0,903	0,660	0,757
Avg. Word2Vec	**0,907**	**0,836**	0,746
Avg. Word2Vec + TF-IDF	0,867	0,789	0,717
Avg. FastText	0,857	0,804	**0,811**
ARTM	0,821	0.649	0.729

Этап 3. Применение тематической фильтрации для повышения качества семантического поиска курсов

Визуализация тематического фильтра

Этап 3. Применение тематической фильтрации для повышения качества семантического поиска курсов

Тематическая модель	**Векторная модель**	**F-mera**	**Precision**	**Recall**	**Ср. число тем на запрос**
Модель без регуляризации	fasttext	**0,9437 (+0,1526)**	**0,9289 (+0,2691)**	**0,9673 (-0,0327)**	9,00
	tfidf	0,8717 (+0,1095)	0,8432 (+0,2818)	0,9172 (-0,0828)	10,14
	w2v	0,8956 (+0,052)	0,8828 (+0,1395)	0,9484 (-0,0516)	13,59
Модель с регуляризацией	fasttext	**0,9544 (+0,1634)**	**0,9262 (+0,2735)**	**0,9896 (-0,0104)**	9,17
	tfidf	0,8884 (+0,1261)	0,8619 (+0,3007)	0,9223 (-0,0777)	10,18
	w2v	0,9225 (+0,0789)	0,8835 (+0,1551)	0,972 (-0,028)	13,76

Этап 4. Семантическое сопоставление отдельных учебных дисциплин и онлайн-курсов с ОП

Model	MAP@3	MAP@5	MAP@10
TF-IDF	**0.896**	**0.856**	**0.814**
LSI	0.488	0.451	0.382
LDA	0.292	0.270	0.213
Avr. Word2vec (MOOC)	0.382	0.307	0.229
Weighted Avr. Word2vec (Wiki)	0.459	0.421	0.335
Weighted Avr. Word2vec (MOOC)	0.795	0.736	0.664
Weighted Avr. Word2vec (MOOC) IN-OUT	**0.800**	**0.758**	**0.679**
doc2vec	0.611	0.535	0.406

Архитектура прототипа интеллектуальной системы

Программная реализация прототипа интеллектуальной системы

- Веб-фреймворк Flask
- СУБД PostgreSQL, SQLite
- Предобработка текста: pymorphy2, NLTK
- Извлечение сущностей на основе правил и грамматик: Yargy-парсер
- Нейросетевые фреймворки: Keras + TensorFlow
- Обучение нейросетевых моделей языка: Gensim (word2vec, fasttext)
- Векторизация текста (TF-IDF) и оценка качества: scikit-learn
- Тематическое моделирование: BigARTM
- Визуализация разведочного поиска: Carrot Search FoamTree

Анализ требований к пр | tfidf | **Выдача:** 50

topic_18: приложение, разработка, программист. Курсов в теме:100

Отображать только избранное: No Yes

Начать поиск | Визуализация тем

Введение в программные системы и их разработку

Добавить в избранное

Курс содержит сведения по прикладным вопросам информатики: структуре и классам программного обеспечения, операционным системам, средствам и методам разработки программных продуктов. Особое внимание уделено программному обеспечению, которое используется широким кругом пользователей и специали

- topic_42: программный, проектирование, разработка
- topic_3: система, информационный, корпоративный
- topic_18: приложение, разработка, программист

Архитектурное проектирование программного обеспечения

Добавить в избранное

Рассматриваются вопросы разработки инновационных подходов к созданию и документированию архитектуры программного обеспечения. Цель курса осветить существующие подходы к документированию архитектурного проектирования, выстроить взаимосвязь между функциональными объектами, используемых в качестве еди

- topic_42: программный, проектирование, разработка
- topic_20: процесс, управление, ит
- topic_18: приложение, разработка, программист

2.12 Современные технологии программирования

Аннотация дисциплины

Основная цель, которую необходимо достичь в результате обучения дисциплины – научить создавать программные комплексы с использованием современных средств и технологий. В процессе изучения дисциплины рассматриваются следующие вопросы: объектно-ориентирова

Знания, умения, навыки

По окончанию изучения дисциплины выпускник должен:

0.71 Знать

- 0.71 о конструировании алгоритмов, методы и технологию программирования, методы обработки и способы реализации основных структур данных.

1.41 Уметь

- 0 мыслить категориями объектов реального мира
- 1.41 применять на практике принципы объектно-ориентированного программирования при разработке программного обеспечения, грамотно моделировать работу объектно-ориентированных программ с использованием языка uml

Ограничения разработанного прототипа

1. Рассматриваются только дисциплины профессионального цикла (вариативного компонента ОП)
2. Рассматриваются только профессиональные компетенции, сформированные на основе требований рынка труда и профстандартов
3. Рассматриваются только профессии, которые включены в реестр профессиональных стандартов
4. Рассматриваются только вакансии, которые классификатор определил с одной профессией
5. Рассматривается компетентностный подход и формирование содержания ОП на основе результатов обучения

Апробация результатов исследования

Апробация результатов по формированию образовательных программ на основе предложенных методов и алгоритмов по направлениям подготовки бакалавриата и магистратуры УГС 09.00.00 “Информатика и вычислительная техника” для Челябинского государственного университета и Югорского государственного университета

Заключение и выводы

1. Предложена математическая модель формирования образовательной программы по требованиям рынка труда в виде графа сущностей документов со структурными и семантическими отношениями.
2. Разработаны интеллектуальные методы и алгоритмы семантического сопоставления сущностей графов образовательных программ, отдельных учебных дисциплин и онлайн-курсов, требований рынка труда и профессиональных стандартов на основе нейросетевых моделей языка.
3. Разработана методика интеллектуального анализа и обработки текстовой информации для поддержки принятия решений при создании новых и актуализации существующих образовательных программ на основе предложенных методов и алгоритмов семантического сопоставления, методов глубокого обучения и вероятностного тематического моделирования.
4. Выполнена оценка качества и результативности реализованных методов и алгоритмов на представительном текстовом корпусе образовательных программ и учебных курсов
5. Спроектирована архитектура и разработан прототип интеллектуальной системы поддержки принятия решений при формировании образовательных программ
6. Осуществлена апробация разработанного прототипа интеллектуальной системы при формировании новых и актуализации существующих образовательных программ в вузах.

Спасибо за внимание

dmbotov@gmail.com
+7 952 523 0424

Работа выполнена при финансовой поддержке РФФИ
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”.