УДК 004.048
https://doi.org/10.47533/2024.1606-146X.29

Д. Ж. КАЙБАСОВА[1,2]*, Л. С. ЛИСИЦЫНА[3], Н. И. ТОМИЛОВА[2], Э. К. СЕЙПИШЕВА[2]

[1] Astana IT University;
[2]Карагандинский технический университет имени Абылкаса Сагинова;
[3]Университет ИТМО.
*e-mail: *dindgin@mail.ru, lisizina@mail.ifmo.ru, tomilova_kstu@mail.ru, elmira_s89@bk.ru*

ИССЛЕДОВАНИЕ СТРУКТУРНЫХ ЭЛЕМЕНТОВ ОБРАЗОВАТЕЛЬНОГО КОНТЕНТА ДЛЯ ФОРМИРОВАНИЯ КОРПУСА ДОКУМЕНТОВ

Представлены результаты исследования по интеллектуальному сопровождению процесса автоматического извлечения данных из текстовых документов, что позволило сформировать корпус документов для образовательных программ с помощью обработки больших объемов слабоструктурированных текстов без переобработки и адаптации, без трудоёмкой работы по определению соответствующих рабочих планов дисциплин. Предметом исследования является содержание рабочих учебных программ (силлабусов), определяемое как совокупность данных, характеризующих результаты обучения и содержание предмета. В результате работы авторами создан корпус текстов из документов рабочих учебных программ по предметам образовательной программы «Информационные системы». Полученный набор документов позволяет также получить матрицу косинусных расстояний для выявления схожих документов по образовательному контенту рабочих учебных программ.

Ключевые слова: *извлечение данных, корпус документов, обработка естественного языка, неструктурированные данные, образовательный контент.*

Д. Ж. КАЙБАСОВА[1,2]*, Л. С. ЛИСИЦЫНА[3], Н. И. ТОМИЛОВА[2], Э. К. СЕЙПИШЕВА[2]

[1]Astana IT University, Қазақстан;
[2]Әбілқас Сағынов атындағы Қарағанды техникалық университеті, Қазақстан;
[3]АТМО Университеті, Ресей Федерациясы.
*e-mail: *dindgin@mail.ru, lisizina@mail.ifmo.ru, tomilova_kstu@mail.ru, elmira_s89@bk.ru*

ҚҰЖАТТАР КОРПУСЫН ҚАЛЫПТАСТЫРУ ҮШІН БІЛІМ БЕРУ МАЗМҰНЫНЫҢ ҚҰРЫЛЫМДЫҚ ЭЛЕМЕНТТЕРІН ЗЕРТТЕУ

Мәтіндік құжаттардан деректерді автоматты түрде алу процесін зияткерлік қамтамасыз ету бойынша зерттеу нәтижелері ұсынылған, бұл көп жұмысты қажет етпей, қайта өңдеусіз және бейімдеусіз үлкен көлемдегі жартылай құрылымдық мәтіндерден алынған құжаттар корпусын құруға және пәндер бойынша сәйкес жұмыс жоспарларын анықтау мүмкіндік береді. Зерттеу пәні оқу нәтижелері мен оқу пәнінің мазмұнын сипаттайтын деректер жиынтығы ретінде анықталған жұмыс оқу жоспарларының (силлабустардың) мазмұны болып табылады. Жұмыстың нәтижесінде авторлар «Ақпараттық жүйелер» білім беру бағдарламасының пәндері бойынша жұмыс оқу жоспарларының құжаттарынан мәтіндер корпусын құрды. Болашақта жұмыс оқу жоспарларының білім беру мазмұны бойынша ұқсас құжаттарды анықтау үшін косинус қашықтықтарының матрицасын алу жоспарлануда.

Түйін сөздер: *деректерді шығару, құжат корпусы, табиғи тілде өңдеу, құрылымдалмаған деректер, білім беру мазмұны.*

D. ZH. KAIBASSOVA[1,2]*, L. S. LISITSYNA[3], N. I. TOMILOVA[2], E. K. SEIPISHEVA[2]

[1]Astana IT University;
[2]Abylkas Saginov Karaganda Technical University;
[2]ITMO University.
*e-mail: *dindgin@mail.ru, lisizina@mail.ifmo.ru, tomilova_kstu@mail.ru, elmira_s89@bk.ru*

RESEARCH OF STRUCTURAL ELEMENTS OF EDUCATIONAL CONTENT FOR FORMING THE CORPUS OF DOCUMENTS

The results of a study on the intellectual support of the process of automatic data extraction from text documents are presented, which made it possible to form a corpus of documents extracted from large volumes of semi-structured texts without reprocessing and adaptation, without requiring much work to determine the appropriate work plans for subjects. The subject of the study is the content of working curricula (syllabuses), defined as a set of data characterizing the learning outcomes and the content of the subject. As a result of the work, the authors created a corpus of texts from documents of working curricula on the subjects of the educational program "Information Systems". In the future, it is planned to obtain a matrix of cosine distances to identify similar documents on the educational content of working curricula.

Key words: *data mining, corpus of document, natural language processing, unstructured data, educational content.*

Введение. В настоящее время повсеместное применение практически во всех областях деятельности человека цифровых технологий, в том числе в области высшего образования привело к активному росту объема накапливаемой слабоструктурированной информации. Обработка большого объема информации с использованием методов машинного обучения и интеллектуального анализа данных дает возможность трансформировать хранимые данные в полезную информацию для системы интеграции формирования образовательных программ с профессиональными стандартами, что обеспечивает актуальность данного исследования.

Образовательная программа (ОП) в системе высшего образования разрабатывается в соответствии с действующим Государственным классификатором занятий и утвержденным профессиональным стандартом РК. Рейтинг ОП вузов РК, проведенный НПП «Атамекен» в 2020 г. показал, что востребованность выпускников и удовлетворённость работодателей РК качеством и актуальностью ОП составляет 48%. На мировом пространстве большое внимание уделяется качеству разработке ОП путем интеграции их контента с работодателями, интересы, которых представлены в профессиональных стандартах. В РК каждый вуз самостоятельно, «ручным способом» пытается разработать ОП по направлениям подготовки выпускников.

В образовательной программе, разработанной на основе профессионального стандарта, основные трудовые функции проецируются в профессиональные модули. При этом каждый профессиональный модуль предполагает формирование компетенций, связанных с выполнением основных трудовых функций профессии. Соответственно, компетенции и трудовые функции из профессиональных стандартов трансформируются в профессиональные компетенции и результаты обучения.

Используемые методы на сегодняшний день интеллектуальной поддержки формирования образовательных программ на основе онтологических моделей знаний и систем правил [1-3], эвристических алгоритмов создания автоматизированных учебных планов, методов экспертного оценивания и когнитивных карт не позволяют эффективно учитывать и оперативно контролировать изменения на рынке труда и в пространстве образовательного контента. В свою очередь, формирование и актуализация экспертами онтологических моделей, правил и системы прецедентов по всем текущим предметным областям образовательных программ является весьма трудоемким процессом, требующим привлечения представительного штата экспертов по каждой из предметных областей для обеспечения необходимой точности.

Профессиональные компетенции же описывают набор основных типичных черт какой-либо специальности, определяющих конкретную направленность (профиль) образовательной программы, которые проявляются в умении специалиста решать весь объём профессиональных задач в выбранной сфере деятельности с помощью характерных для данной сферы знаний, умений и навыков. Перечень профессиональных компетенций структурируется в соответствии с теми основными видами профессиональной деятельности, к которым должен быть подготовлен выпускник, например:

– научно-исследовательские;
– проектные;
– производственно-технологические;
– организационно-управленческие компетенции.

Профессиональные компетенции проецируются в профессиональные модули. При построении образовательной программы необходимо учитывать междисциплинарность содержания для того, чтобы исключить дублирования дисциплин. Структурные элементы проектирования образовательной программы, в виде диаграмм процессов, на показаны рисунке 1.

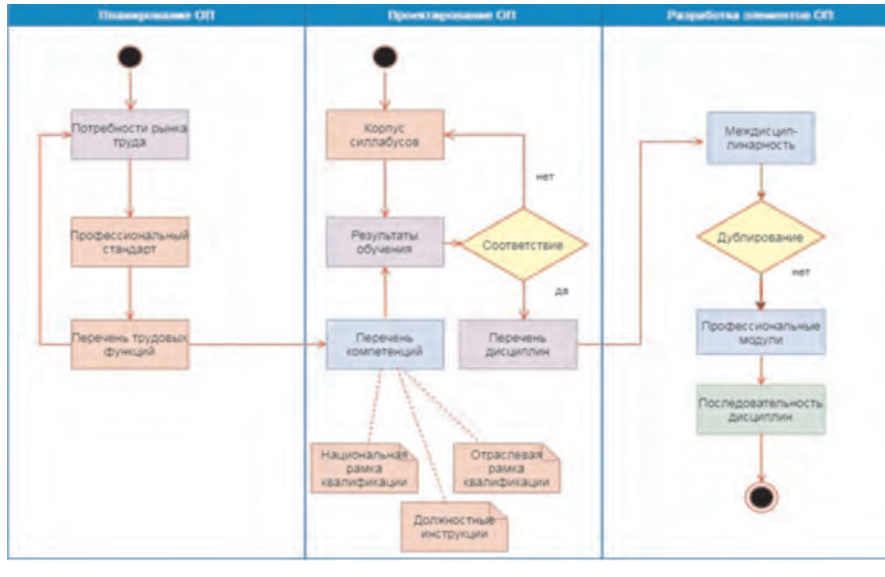

Рисунок 1 – Диаграмма процессов проектирования образовательных программ

На сегодняшний момент проблемой является отсутствие методических указаний у образовательных организаций для получения профессиональных компетенций, сформированных на основе анализа требований рынка труда и профессиональных компетенций в данной сфере деятельности.

Как показано на рисунке 1, структурные элементы образовательной программы взаимосвязаны и формируются поэтапно. Например, для получения элементов профессиональной области на этапе планирования необходимо сформировать базу профессиональных компетенций, полученных на основе анализа профессиональных стандартов и опроса работодателей.

К основным элементам профессионального стандарта относятся [4]:

– наименование направления профессиональной деятельности;

– наименование профессии;

– требования к образованию и опыту работы в рамках ОРК, определяющий квалификационный уровень профессии;

– трудовая функция, определяющая представление совокупности набора трудовых функций, необходимых для достижения определенной задача;

– требования к умениям и знаниям, необходимых для реализации заявленной трудовой функции.

Рабочая учебная программа (силлабус) (согласно ДП КарГТУ 15-2019) включает в себя следующие элементы:

– сведения о преподавателях;

– описание изучаемой дисциплины;

– цели и задачи дисциплины;

– краткое ее содержание, темы и продолжительность их изучения;

– задания самостоятельной работы, требования преподавателя;

– критерии оценки, график выполнения и сдачи заданий по дисциплине;

– критерии оценки знаний обучающихся;

– перечень основной и дополнительной литературы, информационных ресурсов.

Кроме того, рабочая учебная программа дисциплины содержит варианты контрольных работ, курсовых проектов (работ), заданий для студентов заочной формы обучения.

Отметим важные концепты, которые необходимо учитывать при формирования образовательных программ:

1. Цели и задачи изучаемой дисциплины, компетенции и результаты обучения по образовательной программе, ориентированных на последующую подготовку (постреквизиты) по образовательной программе и место дисциплины в структуре образовательной программы, которое предъявляет входные требования к обучающимся с помощью пререквизитов.

2. Описание структуры дисциплины, представленное её разделами, основными темами дисциплины, целями и содержанием лекционных и практических занятий, а также заданиями на самостоятельную работу обучающихся.

Основная часть представленной информации как в образовательных программах, так и в рабочих программах дисциплин образовательной программы состоит из большого объёма слабоструктурированных блоков текста.

Часть из этих блоков представлена в формате таблиц (например, такие как трудоемкость дисциплины, тематический план дисциплины, перечень практических и лабораторных работ) или списков (например, такие как цели и задачи освоение дисциплины и т.д.)

Исследование и выбор метода по отбору текстовых документов, необходимых для ведения обучения, по ключевым компетенциям образовательной программы, невозможен без создания сложных интеллектуальных алгоритмов. Первая часть данных алгоритмов должна предназначена для решения задачи по выделению терминов, характеризующих данную область научного знания, вторая часть для определения сходств этих терминов с онтологией базы знания данной предметной [5, 6]. В настоящее время уже имеется ряд стандартных подходов для извлечения ключевых фраз из отдельных документов, основанных на применении конкретных метрик.

Методы и материалы. Для единообразного понимания вначале определимся с понятием как корпус, корпус текстов. Корпус представляет собой коллекцию взаимосвязанных документов/ текстов на естественном языке [7].

В лингвистике под корпусом понимается подобранный и обработанный по определённым правилам набор текстов, который используется в качестве базы при исследовании языка.

Основное назначение корпусов, их использование в ходе статистического анализа, проверки статистических гипотез, а также при подтверждении лингвистических правил в рассматриваемом языке [8]. Например, авторы работы [9], корпус текстов более широко, как «сформированная по определённым правилам выборка данных из проблемной области», и которая является «видом корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты».

Видим, что основное назначение текстовых корпусов является изображение совокупности коллекции документов, представленных набором слов, которые подчиняются определённым морфологическим, синтаксическим, семантическим и т.п. правилам [10-12]. Для создания крупных текстовых корпусов необходимы исследовательские группы в специализированных институтах [13-14]

По представленным результатам исследований в работах [7-8] можно утверждать, что, несмотря на широкое использование имеющихся универсальных текстовых корпусов, требуется создание оригинальных коллекций, необходимых для решения частных задач.

Постановка задачи для построения специального корпуса текстов представляет собой процесс извлечения данных из содержимого текстовых файлов рабочих учебных программ дисциплин. Решение задачи, связанное с извлечением информации из текстов, относится к компьютерной лингвистике и машинному обучению, связанных с обработкой естественного языка (natural language processing) [15].

Методы, используемые для обработки естественного языка, представляют собой направления по применению компьютерного анализа и синтеза естественного языка. Этапы, применяемые для анализа естественного языка, можно разбить на несколько уровней. Каждый из следующих уровней в любом случае использует информацию, полученную на нижестоящих уровнях. И каждый из следующих уровней ориентирован на следующие виды анализа текста [11]:

– лексический анализ – это уровень, который предназначен для обработки отдельных фрагментов естественного языка – лексем (слов). На этом этапе происходит разделение текста на отдельные слова или предложения. Также этот этап может включать в себя деобусфикацию искаженных слов, очистку от стоп-слов, т.е. шумовых слов, не несущих полезную информационную нагрузку, или преобразование в лексемы, эмоционально значимых объектов, таких как эмодзи;

– морфологический анализ, уровень предобработки определенных характеристик слова, граммем, например, единственное или множественное число, граммемы категории числа. На данном этапе решается две основных прикладных задачи, это лемматизация, т.е. приведение слова к нормальной форме, к лемме и следующая, стемминг, представляющий собой процесс по нахождению основ слова, и не обязательно это корень слова;

– синтаксический анализ, представляет из себя уровень по выделению группы слов и взаимосвязей между этими группами. Результат процесса данного уровня представляет собой древовидную иерархию для каждого предложения из текста;

– семантический анализ, это уровень, который предназначен для смыслового анализа текста. Данный этап анализа предназначен для выделения сарказма или иронии в тексте.

Реализация итеративного процесса, состоящая из двух этапов, стандартным конвейером типичного приложения данных, представленная в работе [16], состоит из сборки и развертывания, изображена на рисунке 2, является отражением конвейера машинного обучения. Вначале на этапе сборки исходные данные необходимо преобразовать в форму, достаточную для передачи этих данных в модели и экспериментов с ними. На этапе развертывания, используемом для оценок и прогнозов пользователям, осуществляется выбор моделей. Первый этап решения задачи служит для автоматической классификации текстов и представляет собой преобразование исходных документов. На этом этапе документы, представляющих собой набор последовательности символов, преобразуются к виду, необходимому и достаточному для алгоритмов машинного обучения в соответствии с целью и задачами классификации. Чаще всего алгоритмам машинного обучения предназначены для работы с векторами в пространстве, которое называют пространством признаков [17]. Второй этап предназначен для построения классифицирующей функции Ф`, с помощью обучения на примерах. Здесь качество классификации зависит в первую очередь от того, как документы смогли быть преобразованы в векторное представление, а также и от алгоритма, который используется на втором этапе. При этом отметим, что методы преобразования текста в вектор специфичны для задачи классификации текстов и сильно зависят как от коллекции документов, так и типа текста, простого, структурированного, а также от языка документа.

В отличии от первого этапа, методы машинного обучения, используемые на втором этапе, не специфичны для задачи классификации текстов и могут быть применимы и в других областях, например, для задачи распознавания образов [18].

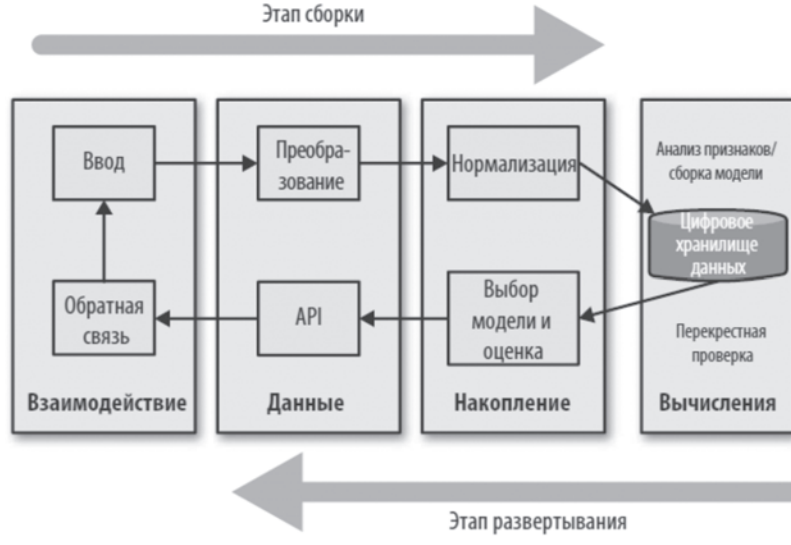

Рисунок 2 – Конвейер приложения данных

Неструктурированная информация должна быть преобразована в хорошо продуманную семантическую модель. Для структурирования предметного содержания квалификационной характеристики рекомендуется создавать ядро системы с использованием онтологической базы знаний [1, 5]. Значительных усилий и времени затрачиваются на разработку базы знаний (БЗ) при разработке интеллектуальных систем, т.е. для накопления знаний, создания модели представления знаний, их структурирования, заполнения БЗ и дальнейшего поддержания ее в актуальном состоянии [19].

База знаний (англ. Knowledge Base, KB), представляющая собой особого рода базу данных, разработана для управления знаниями (метаданными), то есть для сбора, хранения, поиска и выдачи знаний [19]. Анализ работ [9, 11, 15], позволяет определить следующие особенности информационных единиц, представленных в таблице 1, в результате использования которых данные могут быть превращены в знания, т.е. базы данных «перерастают» в базы знаний (БЗ) [5].

Таблица 1 – Виды особенностей информационных единиц

Вид особенности данных*	Свойство
1	2
Внутренняя интерпретируемость	Каждая информационная единица должна обязательно иметь своё уникальное имя, по которому бы информационная система смогла находить ее, а также отвечала бы на запросы, в которых это имя встречается
Структурированность знаний	Наличие у информационных единиц определенной структуры, т.е. эти структуры должны предоставлять возможность для установления отношений типа, таких как часть – целое или род – вид, или элемент – класс

Окончание таблицы 1

1	2
Связность	Между информационными единицами информационной базы должна быть представлена возможность для установления связи различного типа, характеризующих отношения между информационными единицами
Семантическая метрика	Это отношение служит для характеристики ситуационной близости информационных единиц, показывающих силу ассоциативных связей между информационными единицами
Активность	Появление новых данных должны стать источником активности интеллектуальной системы, т.е. выполнение программ в ИС должно инициироваться текущим состоянием информационной базы

*Примечание. Источник [16].

Приведенные особенности предоставляют возможность создать общую модель знаний, называемую «семантическая сеть». Семантическая сеть представляет собой модель знаний, в вершинах которой расположены информационные единицы, а дуги сети отображают виды связей между информационными единицами. В случае иерархичных связей, эти связи определяют отношение структуризации, если связи являются неиерархическими, то они в данном случае определяют отношения иных типов [5, 20].

Результаты и обсуждения. В ходе исследований было принято решение для задач поиска и анализа информации, определения структур и зависимостей в данных, а также для формирования гипотез использовать метод анализа формальных понятий. Так как этот метод является методом анализа данных для визуализации, который с помощью построения решёточной модели особого вида, позволяет сходство группы объектов сохранять в объектно-признаковом описании. Его достоинство состоит в том, что понятия строятся формально, т.е. подмножества слов, связанных друг с другом, связываются с помощью отношения принадлежности «объект – атрибут», в связи с этим данному методу присвоили название анализа формальных понятий.

Для установления же связей между объектами и их свойствами служит формальный контекст. Простой формальный контекст представляет из себя тройку множеств: множество объектов, множество свойств, множество связей между объектами и свойствами.

По результатам работы Ganter B., Stumme G., and Wille R [21], формальным контекстом стали называть тройку вида, представленную в виде (1):

$$K = (G, M, I), \quad (1)$$

где G и M – множества; I – отношение на множестве GxM.

При этом G представляет собой множество объектов, а M есть множество свойств, где gIm означает, что объект g обладает свойством m. Элементы G называются формальными объектами, а элементы M формальными признаками. Эти множества мо-

гут быть частично упорядочены некоторыми отношениями, которые обозначаются $\subseteq$ и $\supseteq$. Отношение $I \subseteq G \times M$ служит для означения какие объекты какими признаками обладают.

Для определения связей между объектами и атрибутами используем следующий порядок. Для подмножеств $A \subseteq G$ и $B \subseteq M$ объектов и атрибутов задаём отображения, т.е. функции, A': $A \rightarrow B$ и B': $B \rightarrow A$, определяемые операторами Галуа в виде (2):

$$A' = \{m \in M \mid \forall\, g \in A\ (g\ I\ m)\},$$
$$B' = \{g \in G \mid \forall\, m \in B\ (g\ I\ m)\} \quad (2)$$

В случае, если пара множеств (A, B) такие, что A'=B, B'=A, то их называют формальным понятием контекста K.

Оператор ", т.е. двукратное применение оператора′, служит оператором замыкания: он идемпотентен, если $A'' = A''$, он монотонен, если $A \subseteq B$ влечет $A'' \subseteq B''$ и он экстенсивен, если $A \subseteq A''$. Множество объектов $A \subseteq G$ называется замкнутым, если $A'' = A$. Аналогично и для замкнутых множеств признаков, подмножеств множества M. Пара множеств (A, B), называется формальным понятием контекста K, если $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$. Множества A и B замкнуты и называются объемом и содержанием формального понятия (A, B) соответственно. Для множества объектов A множество их общих признаков A' служит описанием сходства объектов из множества A, а замкнутое множество A″ является кластером сходных объектов, с множеством общих признаков A′. Отношение ″быть более общим понятием″ задаем следующим образом: $(A, B) \geq (C, D)$ тогда и только тогда, если $A \supseteq C$.

Понятия формального контекста K= (G, M, I), упорядоченные по вложению объемов, образуют решетку B (G, M, I), которую так и стали называть – решеткой понятий.

Таким образом была построена структурная модель для реализации методов интеллектуального анализа текстов, представленная на рисунке 3.

В структурной модели отражены основные процессы работы программного комплекса. Текстовый корпус формируется из коллекции рабочих учебных программ дисциплин образовательных программ, который в свою очередь, является входной информацией. На процессе предобработки данных осуществляется нормализация текстов, реализованная методами удаления различных символов, в том числе, цифр и знаков препинания, а также исключения множества вспомогательных слов, импортированных как шумовые слова (стоп-слова) из библиотеки для работы с естественным языком NLTK. Впоследствии нормализации получаем «мешок слов» для последующего проведения анализа. На следующем этапе формируем векторную модель, основанную на частотном и весовом анализе употребления слов в документах, а также получение матрицы косинусных расстояний для выявления схожих документов по образовательному контенту рабочих учебных программ.

Заключение. Исследования и анализ процесса разработки образовательных программ в соответствии с требованиями современных образовательных и профессиональных стандартов позволили систематизировать жизненный цикл образовательной программы в учебных организациях. Проведённый обзор существующих интеллектуальных методов поддержки принятия решений, используемых при формировании

Рисунок 3 – Структурная модель интеллектуальной системы анализа текстов

образовательных программ и анализе рынка труда, позволил определить ключевые проблемы сопоставления профессиональной области и образовательной программы. Результаты исследованных структурных элементов образовательной программы, послужили одним из элементов стартового этапа для формирования векторной модели, основанной на частотном и весовом анализе употребления слов в документах, а также получения матрицы косинусных расстояний для выявления схожих документов по образовательному контенту рабочих учебных программ.

Благодарность. Исследование финансировалось Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP19677319).

ЛИТЕРАТУРА

1 Bakanova A., Letov N.E., Kaibassova D., Kuzmin K.S., Loginov K.V., Shikov A.N. The use of Ontologies in the Development of a Mobile E-Learning Application in the Process of Staff Adaptation // International Journal of Recent Technology and Engineering. Vol. 8, Issue-2S10, September. 2019. - pp. 780–789.

2 Chung H., Kim J. An Ontological Approach for Ssemantic Modelling of Curriculum and Syllabus in Higher Education // International Journal of Information and Education Technology. Vol. 6, no. 5. 2016. - pp. 365–369.

3 Oprea M. On the Use of Educational Ontologies as Support Tools for Didactical Activities // Proceedings of the International Conference on Virtual Learning (ICVL2012), Nov. 2012. - pp. 67–73.

4 Ботов Д.С. Методы и алгоритмы интеллектуальной поддержки формирования образовательных программ по требованиям рынка труда на основе нейросетевых моделей языка :

диссертация ... кандидата технических наук : 05.13.10 / Ботов Дмитрий Сергеевич; [Место защиты: Уфим. гос. авиац. техн. ун-т]. - Челябинск, 2019. - 160 с.

5 Кайбасова Д.Ж., Ла Л.Л. Методы и модели описания предметной области для разработки интеллектуальной системы в контексте формирования образовательных траекторий // Вестник Семипалатинского государственного университета им. Шакарима. – 2019. – №1(85). – С. 78-83.

6 Кайбасова Д.Ж. Применение онтологического моделирования для формирования образовательных траекторий // Инновационные IT и Smart-технологии: матер. науч. конф., посв. 70-летнему юбилею И.Т. Утепбергенова. – Алматы, 2019. – С. 149-152.

7 Litvinova T., Zagorovskaya O., Litvinova O. Russian text corpora for deception detection studies // International Journal of Open Information Technologies. – 2017. – Vol. 5, Issue 11. – P. 58-63.

8 Zevakhina N., Dzhakupova S. Russian metalinguistic comparatives: a functional perspective: working papers by NRU HSE. – M., 2015. – 30 p.

9 Sojka P., Líška M., Ružicka M. Building Corpora of Technical Texts: Approaches and Tools // Procced. 5th Workshop on Recent Advances in Slavonic Natural Languages Processing. – Brno, 2011. – pp. 71-82.

10 Shang J. et al. Automated phrase mining from massive text corpora // IEEE Transactions on Knowledge and Data Engineering. – 2018. – Vol. 30, Issue 10. – pp. 1825-1837.

11 Большакова Е.И., Воронцов К.В., Ефремова Н.Э. и др. Автоматическая обработка текстов на естественном языке и анализ данных – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.

12 Воронцов К.В. Вероятностное тематическое моделирование: лекции по Машинному обучению // http://www.machinelearning.ru/wiki/images/2/22 /Voron-2013-ptm.pdf. 12.03.2019.

13 Roll U., Correia R. A., Berger Tal O. Using machine learning to disentangle homonyms in large text corpora // Conservation Biology. – 2018. – Vol. 32, №3. – P. 716-724.

14 Campillos L., Deléger L., Grouin C. et al. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT) // Language Resources and Evaluation. – 2018. – Vol. 52(2). – P. 571-601.

15 Глазкова А.В. Формирование текстового корпуса для автоматического извлечения библиографических фактов из русскоязычного текста // International Journal of Open Information Technologies. – 2019. – Vol. 7, №1. – P. 97-103.

16 Бенгфорт Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python: машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. – 363 с.

17 Агеев М.С. Методы автоматической рубрикации текстов, основанные на машинном обучении знаниях экспертов. – дис. канд. физ.-матем. наук: 05.13.11. – М., 2004. – 136 с.

18 Кайбасова Д.Ж. Предварительная обработка коллекции рабочих учебных программ дисциплин для формирования корпуса текстов // Вестник КазНИТУ. – 2019. – №6(136). – С. 541-546.

19 Коровин А.М. Интеллектуальные системы: текст лекций. – Челябинск: Издательский центр ЮУрГУ, 2015. – 60 с.

20 Кайбасова Д.Ж., Ла Л.Л. Ассоциативные правила при поиске признаковых зависимостей для интеллектуального анализа данных // Вестник КазНИТУ. – 2019. – №2(132). – С. 247-251.

21 Ganter B., Stumme G., Wille R. Formal concept analysis: foundations and applications. – Berlin: Springer Science & Business Media, 2005. –349 p.

REFERENCES

1 Bakanova A., Letov N.E., Kaibassova D., Kuzmin K.S., Loginov K.V., Shikov A.N. The use of Ontologies in the Development of a Mobile E-Learning Application in the Process of Staff Adapta-

tion // International Journal of Recent Technology and Engineering. Vol. 8, Issue-2S10, September. 2019. – pp. 780-789.

2 Chung H., Kim J. An Ontological Approach for Ssemantic Modelling of Curriculum and Syllabus in Higher Education // International Journal of Information and Education Technology. Vol. 6, no. 5. 2016.– pp. 365-369.

3 Oprea M. On the Use of Educational Ontologies as Support Tools for Didactical Activities // Proceedings of the International Conference on Virtual Learning (ICVL2012), Nov. 2012. – pp. 67-73.

4 Botov D.S. Metody i algoritmy intellektual'noj podderzhki formirovaniya obrazovatel'nyh programm po trebovaniyam rynka truda na osnove nejrosetevyh modelej yazyka : dissertaciya ... kandidata tekhnicheskih nauk : 05.13.10 / Botov Dmitrij Sergeevich; [Mesto zashchity: Ufim. gos. aviac. tekhn. un-t]. – Chelyabinsk, 2019. – 160 s.

5 Kajbasova D.Zh., La L.L. Metody i modeli opisaniya predmetnoj oblasti dlya razrabotki intellektual'noj sistemy v kontekste formirovaniya obrazovatel'nyh traektorij // Vestnik Semipalatinskogo gosudarstvennogo universiteta im. Shakarima. – 2019. – №1(85). – S. 78-83.

6 Kajbasova D.Zh. Primenenie ontologicheskogo modelirovaniya dlya formirovaniya obrazovatel'nyh traektorij // Innovacionnye IT i Smart-tekhnologii: mater. nauch. konf., posv. 70-letnemu yubileyu I.T. Utepbergenova. – Almaty, 2019. – S. 149-152.

7 Litvinova T., Zagorovskaya O., Litvinova O. Russian text corpora for deception detection studies // International Journal of Open Information Technologies. – 2017. – Vol. 5, Issue 11. – P. 58-63.

8 Zevakhina N., Dzhakupova S. Russian metalinguistic comparatives: a functional perspective: working papers by NRU HSE. – M., 2015. – 30 r.

9 Sojka R., Líška M., Ružicka M. Building Corpora of Technical Texts: Approaches and Tools // Procced. 5th Workshop on Recent Advances in Slavonic Natural Languages Processing. – Brno, 2011. – pp. 71-82.

10 Shang J. et al. Automated phrase mining from massive text corpora // IEEE Transactions on Knowledge and Data Engineering. – 2018. – Vol. 30, Issue 10. – pp. 1825-1837.

11 Bol'shakova E.I., Voroncov K.V., Efremova N.E. i dr. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh – M.: Izd-vo NIU VShE, 2017. – 269 s.

12 Voroncov K.V. Veroyatnostnoe tematicheskoe modelirovanie: lekcii po Mashinnomu obucheniyu // http://www.machinelearning.ru/wiki/images/2/22 /Voron-2013-ptm.pdf. 12.03.2019.

13 Roll U., Correia R. A., Berger Tal O. Using machine learning to disentangle homonyms in large text corpora // Conservation Biology. – 2018. – Vol. 32, №3. – P. 716-724.

14 Campillos L., Deléger L., Grouin C. et al. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT) // Language Resources and Evaluation. – 2018. – Vol. 52(2). – P. 571-601.

15 Glazkova A.V. Formirovanie tekstovogo korpusa dlya avtomaticheskogo izvlecheniya bibliograficheskih faktov iz russkoyazychnogo teksta // International Journal of Open Information Technologies. – 2019. – Vol. 7, №1. – R. 97-103.

16 Bengfort B., Bilbro R., Oheda T. Prikladnoj analiz tekstovyh dannyh na Python: mashinnoe obuchenie i sozdanie prilozhenij obrabotki estestvennogo yazyka. – SPb.: Piter, 2019. – 363 s.

17 Ageev M.S. Metody avtomaticheskoj rubrikacii tekstov, osnovannye na mashinnom obuchenii znaniyah ekspertov. – dis. kand. fiz.-matem. nauk: 05.13.11. – M., 2004. – 136 s.

18 Kajbasova D.Zh. Predvaritel'naya obrabotka kollekcii rabochih uchebnyh programm disciplin dlya formirovaniya korpusa tekstov // Vestnik KazNITU. – 2019. – №6(136). – S. 541-546.

19. Korovin A.M. Intellektual'nye sistemy: tekst lekcij. – Chelyabinsk: Izdatel'skij centr YuUrGU, 2015. – 60 s.

20. Kajbasova D.Zh., La L.L. Associativnye pravila pri poiske priznakovyh zavisimostej dlya intellektual'nogo analiza dannyh // Vestnik KazNITU. – 2019. – №2(132). – S. 247-251.

21. Ganter B., Stumme G., Wille R. Formal concept analysis: foundations and applications. – Berlin: Springer Science & Business Media, 2005. – 349 r.