УДК 004.04
DOI 10.17513/snt.40461

АЛГОРИТМ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЯ ПРИ ФОРМИРОВАНИИ ВХОДНОГО НАБОРА ДАННЫХ ДЛЯ ПОСЛЕДУЮЩЕГО УСТАНОВЛЕНИЯ ВЗАИМОСВЯЗЕЙ «ИСПОЛНИТЕЛЬ – ЗАДАЧА»

Пучкова М.А.

ФГБОУ ВО «МИРЭА – Российский технологический университет», Москва, e-mail: puchkova@mirea.ru

Целью исследования является разработка алгоритма формирования входного набора данных с учетом их разнородности для последующей генерации разноуровневых вариативных связей «исполнитель – задача», как инструмента интеллектуальной поддержки принятия управленческих решений в организационных системах. В качестве новизны разработанного алгоритма выступает установление синтеза между экспертами и методами интеллектуального анализа данных при формировании слабо формализованного входного признакового пространства. Основой для реализации указанного подхода идентификации исполнителя выступают ключевые слова, формирующиеся из разнородной совокупности текстовых данных с применением методов векторизации. Подробно рассмотрена последовательность этапов представленного алгоритма. Также некоторые этапы детализированы на примере формирования входного набора данных для учебной дисциплины в целях решения задачи распределения дисциплин между преподавателями. Представлена возможность применения трехуровневой взвешенной экспертной оценки, необходимой для реализации этапа выбора семантически значимой части или частей исходного документа, а также при проведении очистки данных и формировании словаря исключений. Рассмотрены на примере образовательной организационной структуры результаты формирования облака ключевых слов учебной дисциплины, полученные с применением предложенного алгоритма интеллектуальной поддержки формирования входного набора данных. В отличие от существующих решений представленный алгоритм содержит применение технологий искусственного интеллекта с учетом разнородности входного набора данных и позволяет улучшить качество входного набора данных в целях увеличения точности при последующем установлении разноуровневых связей между исполнителем и задачами в рамках подхода, реализующего интеллектуальную поддержку принятия решения.

Ключевые слова: алгоритм, входные данные, обработка данных, интеллектуальная поддержка, формирование данных, оптимизация, генерация связей

AN ALGORITHM FOR INTELLIGENT DECISION SUPPORT IN THE FORMATION OF AN INPUT DATA SET FOR THE SUBSEQUENT ESTABLISHMENT OF PERFORMER-TASK RELATIONSHIPS

Puchkova M.A.

MIREA – Russian Technological University, Moscow, e-mail: puchkova@mirea.ru

As a tool for intellectual support of managerial decision-making in organizational systems, the article presents an algorithm for generating an input data set, taking into account their heterogeneity, for the subsequent generation of multi-level variable performer-task relationships. The novelty of the developed algorithm is the establishment of a synthesis between experts and data mining methods in the formation of a weakly formalized input feature space. The basis for the implementation of this approach to the identification of the performer are keywords formed from a heterogeneous set of text data using vectorization methods. The sequence of stages of the presented algorithm is considered in detail. Also, some stages are detailed using the example of forming an input data set for an academic discipline in order to solve the problem of distributing disciplines among teachers. The possibility of applying a three-level weighted expert assessment is presented, which is necessary to implement the stage of selecting a semantically significant part or parts of the source document, as well as during data purification and the formation of a dictionary of exceptions. Using the example of an educational organizational structure, the results of the formation of a keyword cloud of an academic discipline obtained using the proposed algorithm for intellectual support of the formation of an input data set are considered. Unlike existing solutions, the presented algorithm contains the use of artificial intelligence technologies, taking into account the heterogeneity of the input data set and allows improving the quality of the input data set in order to increase accuracy in the subsequent establishment of multi-level relationships between the performer and the tasks within the framework of an approach that implements intelligent decision support.

Keywords: algorithm, input data, data processing, intelligent support, data generation, optimization, link generation

Введение

Проблема идентификации связей между исполнителем и задачей актуальна для организаций вне зависимости от типа организационных структур, видов экономической деятельности, организационно-правовой формы, целей деятельности и т.д. Особенно актуальна указанная проблема при рассмотрении потоковых задач, характеризующихся наличием множества возможных исполнителей со схожим квалификационным уровнем. Таким ярким примером является

распределение дисциплин между преподавателями в организациях высшего образования [1; 2]. Процесс распределения задач трудозатратен по времени, а также усложняется при наличии множества сотрудников со схожими должностными обязанностями. При этом индивидуальные особенности сотрудника способствуют определению его предрасположенности к некоторым из имеющихся задач.

Для решения данной проблемы в работе [3] предложен подход, основанный на генерации разноуровневых рекомендаций возможных связей между исполнителем и задачей, а также обозначена потребность в решении задачи обработки входных данных для последующего формирования совокупности ключевых слов, являющихся основой дальнейшей генерации вариативных связей.

Источником входных данных выступают документы, содержащие в себе совокупность ключевых слов с различными весовыми коэффициентами значимости. Однако разнотипность и разнородность входного набора данных обуславливает потребность в разработке соответствующего алгоритма интеллектуальной поддержки его формирования.

В современных исследованиях рассматриваются различные методы и алгоритмы обработки текстовых данных. Работа [4] посвящена исследованию влияния предварительной обработки текстовых данных на эффективность нейросетевой модели для концептуальной разметки токенов. Кроме того, рассматривается предобработка данных с учетом лемматизации, удаления стоп-слов и разделителей предложений. В работах [5; 6] представлены особенности кластерного анализа слабоструктурированных текстовых данных, а также интеллектуальный метод и алгоритм обработки для классификации текстовых данных. Автор работы [7] рассматривает задачи и методы оптимизации состава исполнителей программ и проектов в системе стратегического планирования в условиях цифровизации государственного управления.

Однако указанные методы не ориентированы на обработку текстовых данных для последующей генерации ключевых слов и дальнейшего формирования разноуровневых связей между исполнителями и задачами, что обуславливает актуальность рассматриваемого алгоритма.

Цель исследования – разработка алгоритма, позволяющего сформировать входной набор данных с учетом их разнородности, для последующей генерации разноуровневых вариативных связей «исполнитель – задача». В рамках проводимого исследования понятие разнородности данных используется автором в контексте их различий по типу, структуре, формату, семантической значимости т.д. В качестве новизны разработанного алгоритма выступает установление синтеза между экспертами и методами интеллектуального анализа данных при формировании слабо формализованного входного признакового пространства.

Материалы и методы исследования

Источниками информации о задаче и исполнителе выступают различные как по типу (pdf, doc, почтовый формат, текстовые сообщения из используемых корпоративных систем), так и по содержанию (техническое задание, внутренний документ, резюме исполнителя и т.д.) документы. Ключевыми характеристиками для документа, на основании которого реализуется формирование входного набора данных, является формат и объем данных, принадлежность к характеризующемуся объекту, семантическая значимость, а также источник поступления, способствующий определению уровня документационной значимости. В исследовании формирование входного набора данных и последующее тестирование работы представленного алгоритма рассмотрено на примере источников информации в отношении кафедры прикладной математики Института информационных технологий РТУ МИРЭА.

Поскольку документы, характеризующие как исполнителя, так и задачу, представляют собой формализованные совокупности текстовых данных, основой формирования связи «исполнитель – задача» выступают ключевые слова. При этом в зависимости от семантической значимости ключевые слова обладают различными весовыми коэффициентами. Формирование ключевых слов целесообразно осуществлять с применением методов векторизации. При обработке текстовой информации для последующего решения задач в различных областях широко используется метод TF-IDF. Исследования [8-10], направленные на обнаружение фейковых новостей, разработку системы рекомендаций по вакансиям, а также поиск документов в медицинской карте, рассматриваются с применением данного метода.

Выявление семантически значимых ключевых слов из совокупности документационной базы требует многоэтапной аналитической обработки и включает в себя в том числе применение методов векторизации, позволяющих перевести текстовую информацию в числовое пространство признаков,

использование метрик оценки для определения точности и полноты результатов, а также для верификации и корректировки результатов предусматривается этап экспертной оценки с привлечением специалистов соответствующей предметной области. Проведение экспертного оценивания целесообразно осуществлять с применением многоуровневой взвешенной оценки. Корректировка количества экспертных групп и их весовых коэффициентов реализуется с учетом особенностей рассматриваемой задачи.

Результаты исследования и их обсуждение

Разработанный алгоритм формирования входного набора данных, как инструмент интеллектуальной поддержки принятия управленческих решений в организационных системах, включает в себя определенную последовательность этапов.

1 этап. Осуществление выбора основного документа D в качестве входного признакового пространства.

2 этап. Реализация разделения документа на составные части: $D = \{d_1 \ldots d_n\}$. Данный этап направлен на осуществление дальнейшего выбора наилучшего элемента для обработки и анализа для получения содержательного результата. На рисунке 1 приведен пример разделения рабочей программы дисциплины на несколько содержательных блоков, характеризующихся различной степенью концентрации информации.

На рисунке 2 представлен разработанный алгоритм интеллектуальной поддержки формирования входного набора данных для последующей генерации связей «исполнитель – задача» с описанием этапов.

3-5 этапы. Проведение экспертной оценки каждой из составных частей документа, по результатам которой определяется необходимая для дальнейшей загрузки и обработки часть или же части основного документа. В целях улучшения достоверности и точности возможна обработка нескольких частей в рамках одного исходного документа. В случае проведения некорректной оценки осуществляется повторение цикла экспертной оценки до итогового определения оптимальной части/частей документа. Некорректной считается оценка, при которой сумма количества частей документа с равной максимальной экспертной оценкой составляет не более 50% от общего числа рассматриваемых частей документа.

На примере рабочей программы дисциплины, разделенной на четыре части, рассмотрим проведение их оценки с привлечением экспертов из трех групп, имеющих различный вес с учетом их компетенций. Оценка проводится по пятибалльной шкале, где 5 – высокая семантическая значимость, 1 – семантическая значимость отсутствует (табл. 1 и 2).

Учитывая представленные корректирующие коэффициенты значимости экспертной группы, экспертная оценка каждой из частей документа будет иметь следующий вид:

$$K = 0{,}5 \cdot \bar{x}_1 + 0{,}3 \cdot \bar{x}_2 + 0{,}2 \cdot \bar{x}_3,$$

где $\bar{x}_1, \bar{x}_2$ и $\bar{x}_3$ – средняя оценка каждой из трех представленных групп соответственно.

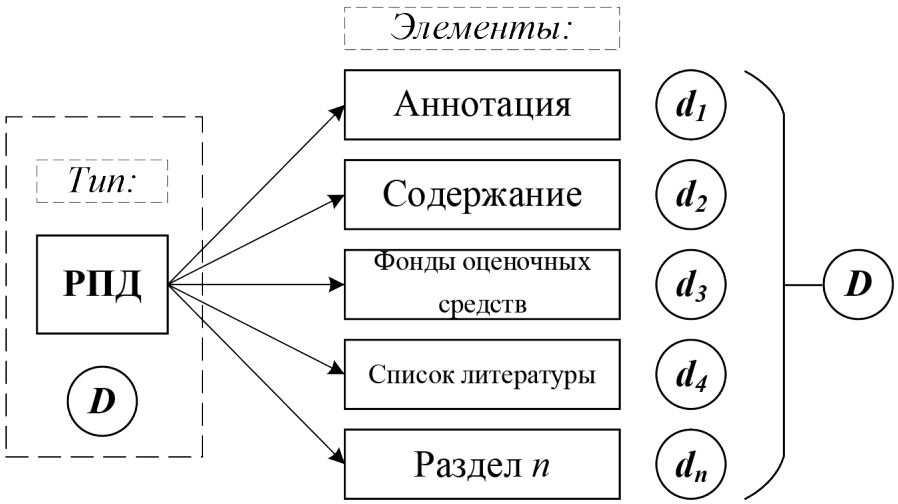

Рис. 1. Схема разделения документа рабочей программы дисциплины на составные части
Источник: составлено автором

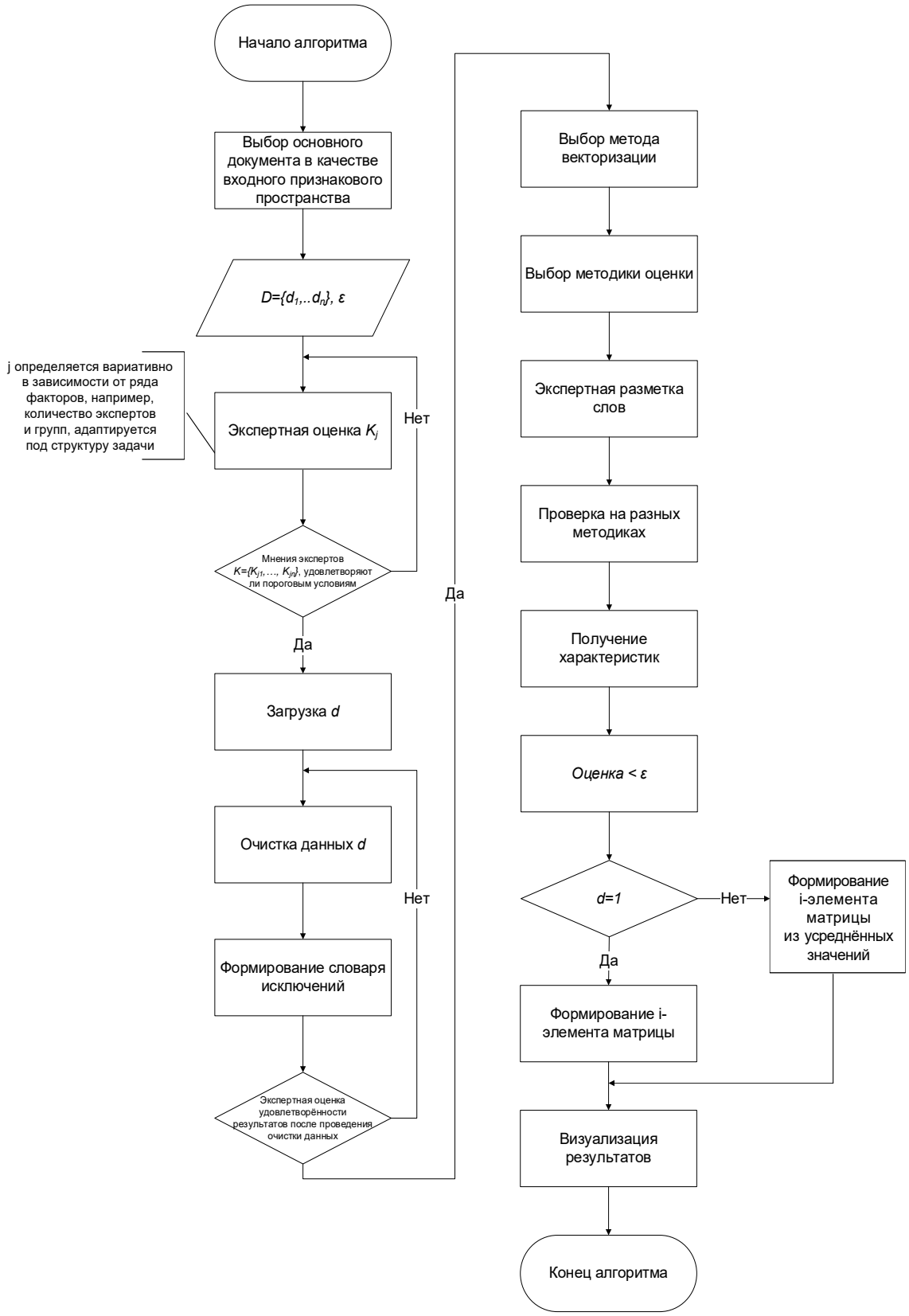

Рис. 2. Блок-схема алгоритма интеллектуальной поддержки формирования входного набора данных
Источник: составлено автором

Таблица 1

Пример результатов экспертной оценки по пятибалльной шкале

Группа	Коэффициент	Эксперт	Экспертная оценка от 1 до 5			
			Часть 1	Часть 2	Часть 3	Часть 4
Старшая	0,5	Эксперт 1	1	2	5	4
		Эксперт 2	2	3	4	4
		Эксперт 3	2	1	5	4
Средняя	0,3	Эксперт 1	4	2	3	5
		Эксперт 2	2	5	4	3
Младшая	0,2	Эксперт 1	4	5	3	3
		Эксперт 2	5	3	2	4
		Эксперт 3	1	5	3	4
		Эксперт 4	2	3	4	5

Примечание: составлено автором по данным проведенного исследования.

Таблица 2

Пример групповых результатов экспертной оценки

Фрагмент документа	Средняя оценка группы			Оценка с учетом коэффициента		
	Старшая	Средняя	Младшая	Старшая	Средняя	Младшая
Часть 1	1,7	3	3	0,85	0,9	0,6
Часть 2	2	3,5	4	1	1,05	0,8
Часть 3	4,7	3,5	3	2,35	1,05	0,6
Часть 4	4	4	4	2	1,2	0,8

Примечание: составлено автором по данным проведенного исследования

Таким образом, наивысшую экспертную оценку семантической значимости имеют сразу две части – третья и четвертая (4 балла). Итоговая экспертная оценка первой и второй частей составляет 2,35 балла и 2,85 балла соответственно. Проведенная оценка считается корректной, так как сумма количества частей документа с равной максимальной экспертной оценкой составляет не более 50% от общего числа рассматриваемых частей документа.

6-8 этапы. Выполнение очистки данных, сопровождаемой экспертной оценкой, для формирования словаря исключений. На примере рабочей программы дисциплины рассмотрим выполнение данных этапов. Предварительная очистка текста включает в себя три шага: очистить строки от кодов компетенций; удалить слова, обозначающие вид занятий; удалить отдельно стоящие цифры, обозначающие номер семестра и объем часов.

В ходе эксперимента выявлена необходимость дополнительной очистки содержания дисциплины для улучшения получаемого результата. При проведении первых нескольких попыток выделения ключевых слов обнаружены слова, не отражающие содержательную часть дисциплины (рис. 3). По результатам экспертной оценки формируется список исключенных слов, который может быть дополнен при анализе содержания других дисциплин. Отрицательный результат при формировании ключевых слов без использования разработанного алгоритма интеллектуальной поддержки формирования входного набора данных подтверждает его практическую значимость.

9 этап. Выбор метода векторизации. Преимуществом TF-IDF является быстрота и простота применения, однако данный метод не учитывает семантических, то есть смысловых отношений слов [11; 12]. Word2Vec предсказывает вероятность слова по его контексту и наоборот, данная модель ограничена локальным контекстным окном, но эффективна для семантических задач [13; 14]. Выбор метода осуществляется, исходя из специфики задачи и входного набора данных.

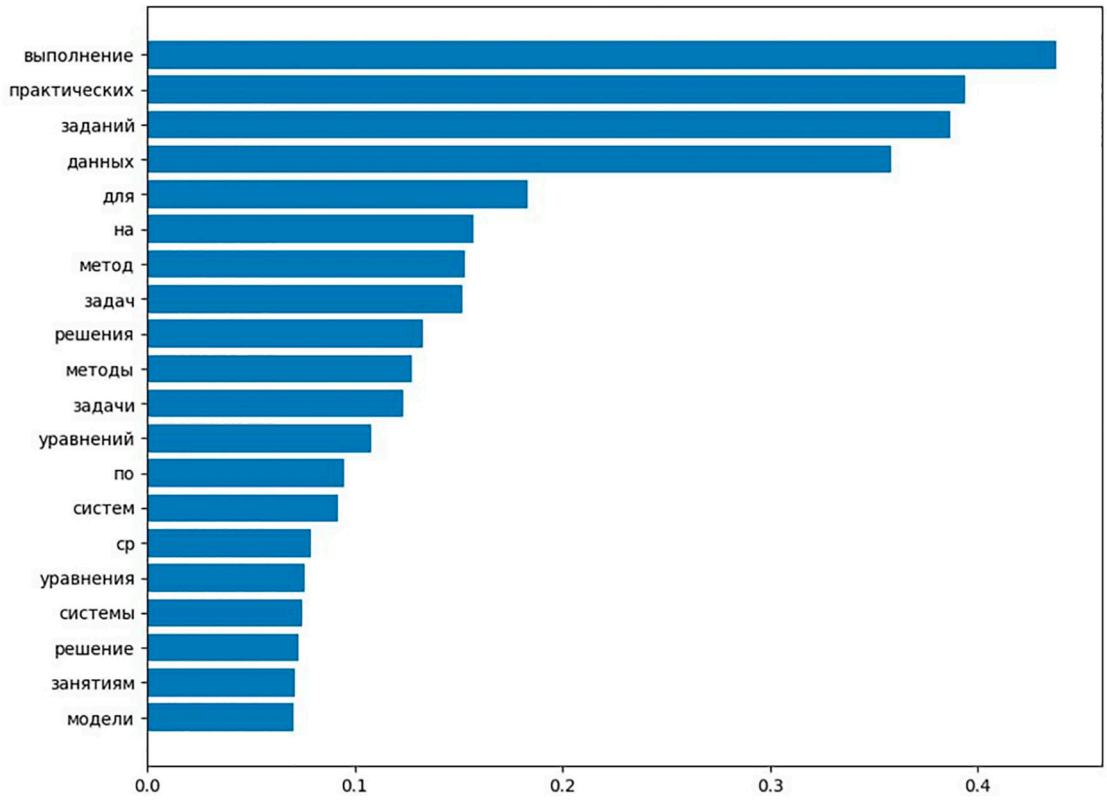

Рис. 3. Результаты выявления ключевых слов после предварительной очистки текста содержания дисциплин
Источник: составлено автором

10 этап. Осуществление выбора методики оценки. Для проверки точности полученных результатов работы алгоритма целесообразно провести оценку, используя различные метрики: Accuracy, которая отражает долю верно классифицированных объектов относительно общего количества; Precision, характеризующая долю истинно положительных случаев среди всех объектов, классифицированных как положительные; Recall, демонстрирующая способность модели идентифицировать все релевантные положительные примеры; и Fβ-мера, которая представляет собой гармоническое среднее между Precision и Recall с возможностью регулировки весов [15].

11-14 этапы. Выполнение экспертной разметки слов с предварительным сокращением количества результатов для уменьшения времени экспертного анализа и улучшения эффективности за счет концентрации на наиболее релевантных терминах. Затем выполнение проверки, последующее получение характеристик и переход к этапу оценки с учетом допустимого диапазона значений.

15 этап. Проверка количества обработанных частей документа. С учетом широкой вариативности задач и особенностей исполнителей источником формирования входного набора могут выступать несколько семантически важных частей одного исходного документа, в связи с чем в алгоритме предусмотрено формирование двух видов элементов матрицы.

16 этап. Визуализация результатов является заключительным этапом работы алгоритма. На рисунке 4 представлено облако слов, сформированное с применением метода векторизации TF-IDF для дисциплины «Большие данные».

Такое сочетание методов интеллектуальной поддержки и экспертного анализа способствует обеспечению высокой достоверности получаемых результатов. Результаты, полученные с применением алгоритма интеллектуальной поддержки формирования входного набора данных, в дальнейшем будут использованы для генерации разноуровневых связей в системе рекомендаций «исполнитель – задача», включающей оценку их достоверности, проведение которой выходит за рамки текущего исследования.

Рис. 4. Визуализация результатов работы алгоритма для дисциплины «Большие данные»
Источник: составлено автором

Такая система рекомендаций реализует двунаправленную функциональность: с одной стороны, она обеспечивает генерацию набора возможных исполнителей для конкретной задачи, с другой – формирует перечень релевантных задач для конкретного исполнителя.

Ключевыми этапами формирования входного набора данных для последующего установления взаимосвязей «исполнитель – задача» являются: формирование многоуровневой экспертной группы; верификация входного набора данных; предобработка данных с учетом экспертной оценки; интеллектуализация формирования ключевых слов.

Основным условием применения представленного алгоритма является наличие семантически значимого документа, выступающего в качестве входного признакового пространства. В качестве незначительного ограничивающего фактора может выступать формат исходного документа, в частности его представление в виде изображения. В подобных случаях необходимо дополнение этапов работы алгоритма в части извлечения текстовых данных из графического представления для реализации дальнейшей обработки.

Заключение

Представленный в исследовании алгоритм выступает составной частью подхода, реализующего интеллектуальную поддержку принятия решения при формировании связи «исполнитель – задача» в целях повышения эффективности функционирования организационных систем, направленного на генерацию разноуровневых рекомендаций возможных связей. Стоит отметить, что проблема идентификации связей между исполнителем и задачей является актуальной для организаций вне зависимости от типа организационных структур. Основой для реализации указанного подхода выступают ключевые слова с различными весовыми коэффициентами значимости, источником формирования которых являются разнотипные документы. При этом разнородность входного признакового пространства обуславливает потребность в разработке алгоритма интеллектуальной поддержки формирования входного набора данных. В отличие от существующих решений разработанный алгоритм отличается установлением синтеза между экспертами и методами интеллектуального анализа данных при формировании слабо формализованного входного признакового пространства. Применение алгоритма позволяет улучшить качество входного набора данных в целях увеличения точности при последующем установлении разноуровневых связей между исполнителем и задачами в рамках подхода, реализующего интеллектуальную поддержку принятия решения.

Список литературы

1. Смоленцева Т.Е. Технология непрерывной оценки остаточных знаний на примере потоковых дисциплин высших учебных заведений // Современные наукоемкие технологии. 2025. № 1. С. 158-165. URL: https://top-technologies.

ru/ru/article/view?id=40292 (дата обращения: 26.06.2025). DOI: 10.17513/snt.40292.

2. Смоленцева Т.Е. Алгоритм модели классификации потоковых дисциплин для разработки рекомендаций по оценке остаточных знаний // Современные проблемы науки и образования. 2025. № 2. С. 46. URL: https://science-education.ru/article/view?id=33972 (дата обращения: 22.07.2025). DOI: 10.17513/spno.33972.

3. Пучкова М.А., Смоленцева Т.Е., Калач Е.В. Концепция формирования инструментария генерации связей «преподаватель–дисциплина» в структуре организаций высшего образования // Вестник Воронежского института ФСИН России. 2024. № 3. С. 93-99. URL: https://vi.fsin.gov.ru/upload/territory/Vi/nauchnaja_dejatelnost/%D0%92%D0%B5%D1%81%D1%82%D0%BD%D0%B8%D0%BA/v_fsin_2024_3.pdf (дата обращения: 03.07.2025).

4. Babina O.I., Zinoveva A.Yu., Nerucheva E.D. Dataset preprocessing effects on Bi-LSTM-based concept tagging of text tokens // Terra Linguistica. 2024. Vol. 15, Is. 3. P. 109-123. URL: https://human.spbstu.ru/userfiles/files/articles/2024/3/109-123.pdf (дата обращения: 30.06.2025). DOI: 10.18721/JHSS.15310.

5. Бова В.В., Кравченко Ю.А., Родзин С.И. Методы и алгоритмы кластеризации текстовых данных (обзор) // Известия ЮФУ. Технические науки. 2022. № 4(228). С. 122-143. URL: https://izn.dev.rdcenter.ru/index.php/izv_tn/article/view/682 (дата обращения: 03.07.2025). DOI: 10.18522/2311-3103-2022-4-122-143.

6. Ефанов С.В., Иванова Е.Н., Чернецкая И.Е. Метод и алгоритм интеллектуальной обработки текстовой информации // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2024. Т. 14. № 3. С. 22-35. URL: https://uprinmatus.elpub.ru/jour/article/view/202 (дата обращения: 04.07.2025). DOI: 10.21869/2223-1536-2024-14-3-22-35.

7. Писарева О.М. Задачи и методы оптимизации состава исполнителей программ и проектов в системе стратегического планирования // МИР (Модернизация. Инновации. Развитие). 2022. Т. 13. № 3. С. 385-401. URL: https://www.mir-nayka.com/jour/article/view/1314 (дата обращения: 28.06.2025). DOI: 10.18184/2079-4665.2022.13.3.385-401.

8. Hersianty Meida V., Amalia Larasati E., Puspitasari D., Wibowo Wahyu D. Penerapan algoritma TF-IDF dan Cosine Similarity dalam sistem rekomendasi lowongan pekerjaan // JATI (Jurnal Mahasiswa Teknik Informatika). 2024. Vol. 9. Is. 1. P. 1619-1625. URL: https://www.researchgate.net/publication/388349464_PENERAPAN_ALGORITMA_TF-IDF_DAN_COSINE_SIMILARITY_DALAM_SISTEM_REKOMENDASI_LOWONGAN_PEKERJAAN (дата обращения: 01.07.2025). DOI: 10.36040/jati.v9i1.12406.

9. Vijay R., Singhal D. Predictive Modeling for Fake News Detection Using TF-IDF & Count Vectorizers // International Journal of Electronic Security and Digital Forensics. 2024. Vol. 16. Is. 4. P. 503-519. URL: https://www.inderscience.com/offers.php?id=139672 (дата обращения: 01.07.2025). DOI: 10.1504/IJESDF.2024.139672.

10. Heryawan L., Novitaningrum D., Nastiti K.R., Mahmudah S.N. Medical Record Document Search with TF-IDF and Vector Space Model (VSM) // International Journal on Advanced Science, Engineering and Information Technology. 2024. Vol. 14. Is. 3. P. 847-852. URL: https://ijaseit.insightsociety.org/index.php/ijaseit/article/view/19606 (дата обращения: 01.07.2025). DOI: 10.18517/ijaseit.14.3.19606.

11. Tawil A.Al., Almazaydeh L., Qawasmeh D. [et al.] Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT // Computers, Materials and Continua. 2024. Vol. 81. Is. 2. P. 3395-3412. URL: https://www.techscience.com/cmc/v81n2/58675 (дата обращения: 22.07.2025). DOI: 10.32604/cmc.2024.057279.

12. Lan F. Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method // Advances in Multimedia. 2022. P. 7923262. URL: https://onlinelibrary.wiley.com/doi/10.1155/2022/7923262 (дата обращения: 01.07.2025). DOI: 10.1155/2022/7923262.

13. Sharma A., Kumar S. Ontology-based semantic retrieval of documents using Word2vec model // Data & Knowledge Engineering. 2023. Vol. 144. P. 102110. URL: https://www.sciencedirect.com/science/article/pii/S0169023X2200101X?via%3Dihub (дата обращения: 22.07.2025). DOI: 10.1016/j.datak.2022.102110.

14. Ярушкина Н.Г., Мошкин В.С., Константинов А.А. Применение языковых моделей word2vec и bert в задаче сентимент-анализа текстовых сообщений социальных сетей // Автоматизация процессов управления. 2020. № 3(61). С. 60-69. URL: https://www.elibrary.ru/item.asp?id=44130587 (дата обращения: 02.07.2025). DOI: 10.35752/1991-2927-2020-3-61-60-69.

15. Rainio O., Teuho Ja., Klén R. Evaluation metrics and statistical tests for machine learning // Scientific Reports. 2024. Vol. 14. Is. 1. P. 6086. URL: https://www.nature.com/articles/s41598-024-56706-x (дата обращения: 03.07.2025). DOI: 10.1038/s41598-024-56706-x.