Раздел 2
МОДЕЛИ, СИСТЕМЫ, СЕТИ В ТЕХНИКЕ

Section 2
MODELS, SYSTEMS, NETWORKS IN THE TECHNIQUE

УДК 004.912
doi:10.21685/2227-8486-2022-3-7

МОДЕЛЬ ОБРАБОТКИ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ НА РУССКОМ ЯЗЫКЕ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ПОДДЕРЖКИ ИНФОРМАЦИОННОГО УПРАВЛЕНИЯ В ДИНАМИЧЕСКИХ ОРГАНИЗАЦИОННЫХ СИСТЕМАХ

Е. А. Макарова[1], Д. Г. Лагерев[2]

[1, 2] Брянский государственный технический университет, Брянск, Россия
[1] m4karova.e@yandex.ru, [2] lagerevdg@yandex.ru

Аннотация. *Актуальность и цели.* Применение моделей интеллектуального анализа данных повышает эффективность и ускоряет процесс принятия решений в динамических организационных системах, особенно при необходимости обработки большого количества значимых слабоструктурированных текстовых данных. Однако для эффективного использования таких данных требуется применение специальных методов их предварительной обработки. Рассмотрен процесс сбора и обработки слабоструктурированных текстовых данных на русском языке с целью дальнейшего использования в моделях интеллектуального анализа данных и последующей передачи результатов обработки лицу, принимающему решения (ЛПР). *Материалы и методы*. Описана модель обработки слабоструктурированных данных на основе использования интеллектуальных технологий с минимальным привлечением эксперта – специалиста в предметной области принятия решений. Рассмотрен процесс обработки текстовых данных с включением дополнительных этапов, позволяющих добиться лучшего качества обработанных данных. Предложен подход разметки данных, сочетающий сильные стороны машинной и человеческой обработки. *Результаты*. Для проверки эффективности модели приведены примеры построения процесса обработки с включением дополнительных этапов, позволяющих оперативно исправлять, выполнять очистку от дублей и предоставить ЛПР информацию в удобной для восприятия форме. Выполнена апробация предложенной модели для обработки текстовых данных в динамических организационных системах, таких как система управления персоналом и система управления здравоохранением на уровне региона.

По итогам апробации показана эффективность по снижению трудозатрат лица, принимающего решения, и, соответственно, увеличение скорости и обоснованности принятия решений, особенно для повторных решений. *Выводы*. Результаты исследования могут быть использованы для информационного управления в различных динамических организационных системах. Использование предложенной модели позволяет сократить время работы эксперта с сохранением качества обработки слабоструктурированных данных.

Ключевые слова: информационное управление, динамические организационные системы, интеллектуальный анализ данных, обработка естественного языка, слабоструктурированные данные, управление кадрами, управление здравоохранением

Для цитирования: Макарова Е. А., Лагерев Д. Г. Модель обработки слабоструктурированных текстовых данных на русском языке для интеллектуальной поддержки информационного управления в динамических организационных системах // Модели, системы, сети в экономике, технике, природе и обществе. 2022. № 3. С. 104–125. doi:10.21685/2227-8486-2022-3-7

MODEL OF PROCESSING SEMI-STRUCTURED TEXT DATA IN RUSSIAN FOR INTELLECTUAL SUPPORT OF INFORMATION MANAGEMENT IN DYNAMIC ORGANIZATIONAL SYSTEMS

E.A. Makarova[1], D.G. Lagerev[2]

[1, 2] Bryansk State Technical University, Bryansk, Russia
[1] m4karova.e@yandex.ru, [2] lagerevdg@yandex.ru

Abstract. *Background.* The use of data mining models increases the efficiency and speeds up the decision-making process in dynamic organizational systems, especially when it is necessary to process a large amount of meaningful semi-structured text data. But, for the effective use of such data, the use of special methods for their preliminary processing is required. This article discusses the process of collecting and processing semi-structured text data in Russian for further use in data mining models, and subsequent transfer of processing results to the decision maker. *Materials and methods.* A model for processing semi-structured data based on the use of intelligent technologies with minimal involvement of an expert – a specialist in the subject area of decision making is described. The process of processing text data with the inclusion of additional stages, allowing achieving the best quality of the processed data, is considered. A data markup approach is proposed that combines the strengths of machine and human processing. *Results.* To test the effectiveness of the model, examples are given of constructing a processing process with the inclusion of additional stages that allow you to quickly correct, clean up duplicates and provide decision makers with information in a form that is easy to understand. The proposed model was tested for processing text data in dynamic organizational systems, such as a personnel management system and a healthcare management system at the regional level. Based on the results of the approbation, the effectiveness of reducing the labor costs of the decision maker and, accordingly, increasing the speed and validity of decision-making, especially for repeated decisions, has been shown. *Conclusions.* The results of the study can be used for information management in various dynamic organizational systems. The use of the proposed model makes it possible to reduce the expert's work time while maintaining the quality of semi-structured data processing.

Keywords: information management, dynamic organizational systems, data mining, natural language processing, semi-structured data, personnel management, healthcare management

For citation: Makarova E.A., Lagerev D.G. Model of processing semi-structured text data in russian for intellectual support of information management in dynamic organizational systems. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve = Models, systems, networks in economics, technology, nature and society*. 2022;(3):104–125. (In Russ.). doi:10.21685/2227-8486-2022-3-7

Введение

Высокие темпы изменений в экономике и социальной жизни ставят новые вызовы перед управленцами в разных сферах – решения нужно принимать все быстрее, при этом необходимо учитывать все большее количество информации о ситуации, чтобы поддержать обоснованность принятых решений на должном уровне. Экономические изменения в 2022 г. влекут за собой необходимость смены контрагентов и логистических цепочек, что диктует необходимость быстрого заключения новых контрактов. Увеличение количества сделок свойственно разным сферам социально-экономической жизни. Растет рынок поглощения и слияния, в том числе в таких социально важных сферах, как медицина [1], где каждой сделке предшествует сбор подробной информации о юридических лицах и ситуации в регионе или сфере действия.

Для получения полной картины к структурированной информации, которой обычно оперируют лица, принимающие решения (ЛПР) – финансовые отчеты и т.д., необходимо добавлять **слабоструктурированные** источники текстовых данных. Существует множество источников таких данных – от открытых данных в интернет-СМИ и социальных сетях до информации из полей свободного ввода в профессиональных закрытых базах данных (например, медицинских). В публичных источниках наблюдается рост объема информации, только социальными сетями на момент 2021 г. пользуется более 67 % населения России [2].

В настоящее время для поддержки принятия решений часто используются модели и методы интеллектуального анализа данных (ИАД), что позволяет повысить оперативность и обоснованность принимаемых управленческих решений. Обычно добавление текстовых данных из слабоструктурированных источников в модель анализа может значительно увеличить точность модели [3]. Несмотря на развитие исследований в области анализа и генерации текстов [4] и то, что современные системы поддержки принятия решений (СППР) часто поддерживают возможность добавления текстовой информации в модели интеллектуального анализа данных, вопросы выбора методов сбора и обработки данных все еще остаются в зоне ответственности ЛПР. По некоторым оценкам, сбор и обработка может занимать до 70 % трудозатрат [5] в процессе анализа данных. Особенно эти трудозатраты заметны в динамических организационных системах [6], поскольку решения требуется принимать оперативно, с учетом регулярно обновляемой информации. К подобным системам относятся системы управления предприятием, здравоохранением, процессом образования и т.д. Различные финансовые системы также чувствительны к обновлению информации: например, исследовано влияние пресс-релизов ЦБ РФ на ожидания и поведение участников денежного рынка [7].

Но в то же время исследования показывают, что наличие положительного эффекта от использования текстовых данных в моделях ИАД во многом зависит от методов их предобработки. Например, исследователи вопроса эффективности добавления текстовых данных в модели прогнозирования фи-

нансовых событий пришли к выводу, что прирост точности модели достигается только при условии правильного сбора и предобработки текстовых данных [8]. Разработка методов предварительной обработки текстовых данных для последующего их использования в процессе анализа данных является весьма актуальной задачей. Публикуются работы, ставящие своей целью систематизацию имеющихся знаний по этой области для английского языка [9]. Много работ посвящено влиянию методов предобработки на итоговый результат применительно к определенным структурам данных, как, например, сообщения в социальных сетях [10, 11]. В то же время научные коллективы из разных стран рассматривают способы обработки для анализа текстов на локальных языках [12]. В контексте управления организационными системами эти задачи относятся к процессу информационного управления – управления информацией, которой владеет ЛПР на момент принятия решения [13]. При работе с большими массивами данных, которые могут быть учтены при принятии решений, нужно стремиться к тому уровню информированности ЛПР, который побудит его принимать наиболее эффективные с точки зрения оперативности и обоснованности решения.

Однако инструментария для обработки слабоструктурированных текстовых данных на русском языке, особенно содержащих профессиональную лексику, недостаточно для качественной поддержки задачи эффективного информационного управления в системах, в которых информация обновляется и обрабатывается регулярно, без привлечения большого количества человеческих ресурсов. С учетом изложенного целью исследования является построение модели автоматизированной обработки слабоструктурированных текстовых данных на русском языке для использования в процессе информационного управления объектами динамических организационных систем.

Материалы и методы

1. Процесс принятия управленческих решений в динамических организационных системах с использованием слабоструктурированных данных

В данном разделе рассмотрен процесс принятия управленческих решений в динамических организационных системах с использованием слабоструктурированных данных, включая модель обработки подобных данных.

Предлагаемая модель основана на классическом контуре управления, отображающем связь между объектом и субъектом управления [6].

Модель процесса принятия управленческих решений в динамических организационных системах с использованием слабоструктурированных данных можно описать следующим образом:

$$A = \langle R,\ U,\ S,\ I,\ M,\ Z; O \rangle,$$

где R – многократно повторяющиеся задачи, которые необходимо решить ЛПР; M – множество многократно повторяющихся управленческих воздействий, реализующих принцип обратной связи; Z – обратная связь по управленческому решению; I – информация, доступная для анализа и прогнозирования будущих состояний; U – обновляющиеся во времени слабоструктурированные данные, состоящие из: U_i – данных из внутренних источников субъекта

управления, U_o – данных из внешних источников, описывающих субъект управления; S – обновляющиеся во времени структурированные данные; O – результаты обработки данных, информация для ЛПР.

Примерами внутренних источников данных могут служить: корпоративные базы данных, системы документооборота и т.д. В качестве внешних источников может выступать: информация из интернет-ресурсов, СМИ, социальных сетей и т.п.

Схема модели процесса принятия управленческих решений в динамических организационных системах с использованием слабоструктурированных текстовых данных представлена на рис. 1.

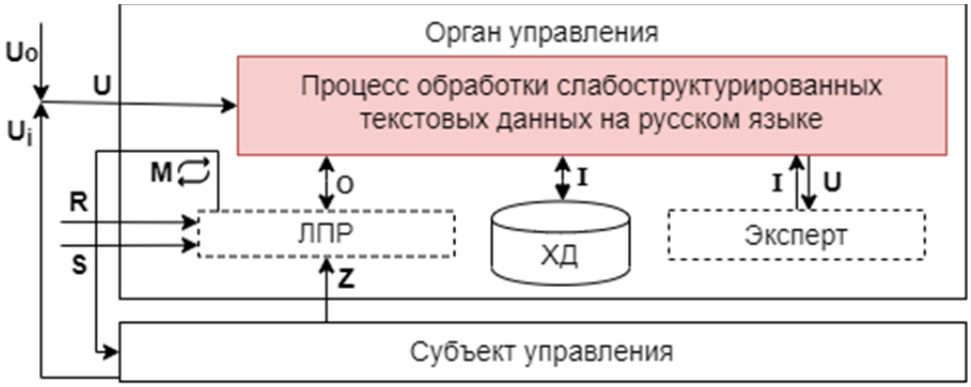

Рис. 1. Модель процесса принятия управленческих решений в динамических организационных системах с использованием слабоструктурированных текстовых данных

Исходя из низкого качества полностью автоматической оценки семантической и тональной составляющей текстов на естественном языке, процесс обработки подобных данных все еще требует привлечения эксперта. Задача организации обработки текстовых данных сводится к сокращению трудозатрат эксперта путем частичной автоматизации данного процесса. Рассматривая динамические организационные системы, важно также учесть повторяемость процесса обработки во времени. Использование результатов (в том числе промежуточных), полученных на предыдущих итерациях принятия решений, может дать дополнительный прирост скорости при подготовке данных для принятия последующих решений.

Описанная модель процесса принятия управленческих решений в динамических организационных системах позволяет включать слабоструктурированные текстовые данные в процесс анализа уже на этапе первичной настройки сбора данных, а также выполнять их обработку с привлечением ЛПР или эксперта в предметной области. Параметры модели, заданные на этом этапе, и результаты их применения сохраняются, что позволяет сократить время при решении повторяющихся задач, таких как, например, мониторинг состояния объекта управления.

Поскольку в статье рассматривается информационное управление динамической организационной системой с использованием слабоструктурированных текстовых данных, именно процессам сбора, обработки и представления информации будет уделено особое внимание. Представленный процесс

будет расширен на этапе обработки слабоструктурированных текстовых данных для более быстрого и эффективного получения необходимой информации о субъекте управления путем использования обработанных данных в моделях ИАД или передачи их напрямую ЛПР.

2. Модель обработки слабоструктурированных текстовых данных для использования в процессе информационного управления объектами динамических организационных систем

Расширение модели анализа данных за счет введения в нее дополнительных входов со слабоструктурированными данными существенно увеличивает трудозатраты и время на ее построение вследствие накладных затрат на получение новых данных, их предобработку, последующее обучение модели и анализа данных. Для того чтобы сократить трудозатраты и сэкономить время ЛПР и экспертов, предлагается автоматизировать типовые процессы обработки слабоструктурированных текстовых данных.

На рис. 2 представлено развитие процесса сбора и обработки слабоструктурированных текстовых данных для дальнейшего анализа, представленного в работе [15]. Этапы, в рамках которых целесообразно применение разработанных авторами моделей, методов и алгоритмов автоматизированной обработки данных, выделены зеленым фоном. На этих этапах также предусмотрена валидация с привлечением эксперта, которая будет подробнее рассмотрена далее.

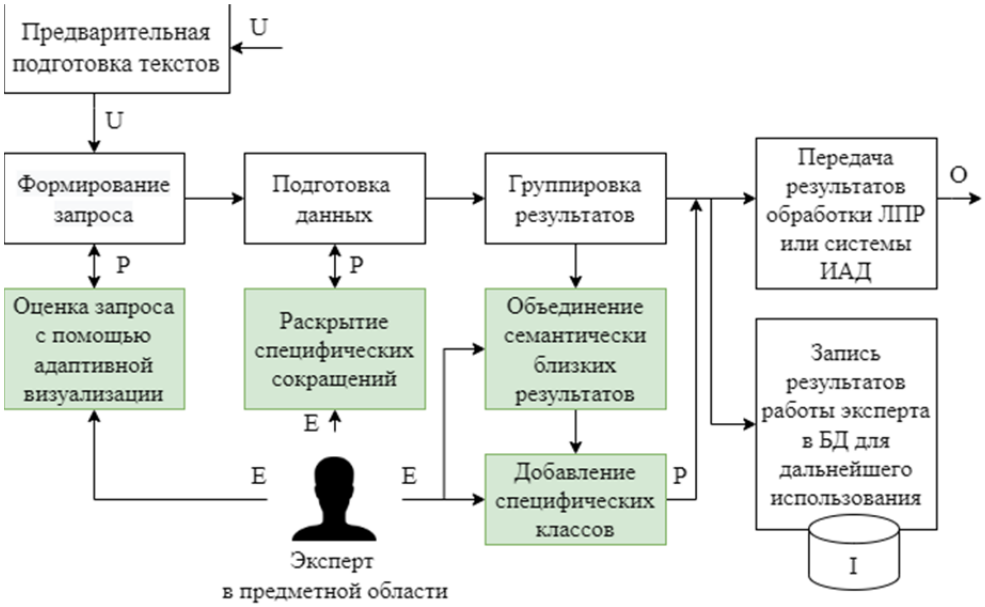

Рис. 2. Процесс обработки слабоструктурированных текстовых данных с привлечением эксперта

Модель обработки описывается следующим образом:

$$M = \{U, P, E, I; O\},$$

где U – слабоструктурированные текстовые данные на входе; P – множество автоматических воздействий по обработке текста, зависит от структуры

и специфики текстов; E – воздействия эксперта над данными – показатель, который в данной модели стремится к уменьшению; I – информация, доступная для анализа и прогнозирования будущих состояний, полученная в результате работы эксперта; O – результаты обработки данных, информация для ЛПР.

Предыдущие работы авторов [16–18] рассматривали данные процессы изолированно и не описывали очередность применения «улучшающих» воздействий над данными и целесообразность использования этих этапов с точки зрения баланса трудозатрат и качества. Благодаря разработанной модели возможно построение различных траекторий сбора и обработки данных, которые позволяют обрабатывать данные разной степени структурированности из различных предметных областей.

Для построения траектории обработки данных, позволяющей получить наиболее точную модель ИАД, необходимо из представленных выше этапов выбрать нужные. До начала работы эксперта с данными необходимо выполнить предобработку текста и создать модель языка предметной области в формате Word2Vec [19].

Остановимся на условиях, когда необходимо применение методов и алгоритмов обработки, предложенных в предыдущих статьях авторов.

Этап «Оценка запроса с помощью адаптивной визуализации». Необходимость привлечения эксперта может возникнуть для различных задач на этапе сбора данных. Так, например, оценка релевантности собранных данных необходима для текстовых документов, содержащих большие объемы текста, на просмотр которых потребуется много времени. В подобных случаях будет эффективно использование адаптивной визуализации больших массивов текстовых данных для оценки запроса [16]. Объем набора документов, с которого целесообразно применять визуализацию, вычисляется по формуле

$$\frac{\sum_{i=1}^{n} w_i}{n} > w_m,$$

где w_i – количество слов в документе (по итогам простой токенизации); n – количество документов, достаточное для оценки корректности сбора (по умолчанию 10); w_m – средняя скорость чтения (слов в минуту), по умолчанию взята в размере 150 исходя из скорости чтения взрослого человека на русском языке, которая лежит в диапазоне от 120 до 180 [20]. В качестве средней скорости чтения может быть принято другое значение, в зависимости от сложности текста.

Этап «Раскрытие специфических сокращений». Методы для поиска и раскрытия специфических сокращений детально описаны авторами в работе [17]. Исправлять ошибки или раскрывать специфические сокращения необходимо для данных, в которых их содержится большое количество – это можно выяснить при ручном просмотре части выборки или в результате автоматической проверки. Например, большое количество ошибок и специфических сокращений свойственно сферам, где требуется быстрая запись информации человеком или используется профессиональная терминология [21]. Сами по себе специфические сокращения не несут проблемы для дальнейшей обра-

ботки, однако, если в имеющихся данных одни и те же определения сокращены по-разному, это может уменьшить итоговое качество модели и даже привести к ее переобучению [22]. Условие необходимости раскрытия сокращений в наборе слабоструктурированных текстовых данных представлено в формуле

$$\frac{abbr}{w} > k_{abbr},$$

где *abbr* – количество сокращений, найденных в наборе текстовых данных; *w* – общее количество слов в наборе текстовых данных; k_{abbr} – коэффициент, отражающий насыщенность текста сокращениями. Определяется экспертным путем или подбирается эмпирически в результате серии экспериментов, в данной работе принимается за 0,05. При выборе коэффициента стоит учесть, насколько семантически значима сокращаемая информация. Сокращения вспомогательных слов в связанных текстах («т.к., т.о.») не влияют на качество текстов в контексте дальнейшего использования в моделях ИАД, в отличие от специфичных сокращений и описанных выше ситуаций различных сокращений одних и тех же терминов. В художественных и публицистических текстах описанный коэффициент часто не превышает 0,01, в то время как в записях из ИЭМК (интегрированных электронных медицинских карт) может достигать значения 0,2 [17].

Этап «Объединение семантически близких коротких текстовых сообщений». Использование предложенной модели обработки слабоструктурированных данных позволяет улучшить процесс за счет использования более «чистых» данных, с меньшим числом ошибок и сокращений, для классификации которых (по тональности, темам, тегам) необходимо привлечение эксперта. Привлечение может быть как на постоянной основе, так и с целью набора достаточного количества данных для построения модели машинного обучения и осуществления дальнейшей разметки автоматически. С целью избежать ошибок автоматической обработки привлекается эксперт для валидации результатов. Общая схема экспертной валидации процесса обработки данных представлена на рис. 3.

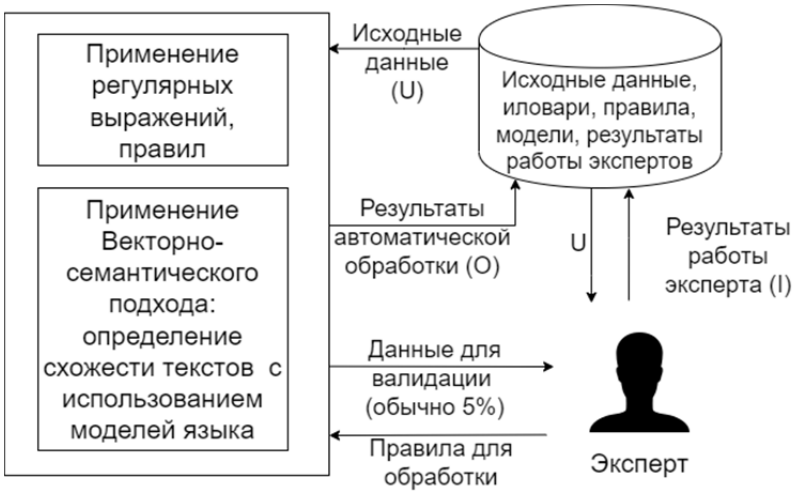

Рис. 3. Экспертная валидация результатов
процесса обработки слабоструктурированных текстовых данных

Эксперт задает параметры обработки, затем валидирует обработанные данные: часть примеров выбирается случайным образом, также эксперт должен рассмотреть «пограничные» случаи, в зависимости от заданных настроек. Результаты редактирования параметров обработки или ручной валидации сохраняются в модели для дальнейшего применения в аналогичных случаях.

Группировка результатов по семантической близости и удаление явных дублей позволяет сократить время эксперта по работе с выборкой. Группировка используется в описанной схеме для упрощения работы эксперта путем предоставления ему сгруппированных по схожести результатов.

Объединение семантически близких текстов в группы осуществляется исходя их косинусного расстояния между векторами исследуемых текстов:

$$\frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \geq k,$$

где A_i и B_i – компоненты векторов A и B соответственно, расстояние между которыми вычисляется, для решения задачи группировки текстов векторизация проводится методом «Bag of words»; n – размерность сравниваемых векторов; k – семантическая близость, по умолчанию 0,5.

После вычисления семантической близости между всеми парами текстов они объединяются в группы вокруг единицы текста, с которой больше всего других текстов имеют семантическую близость выше выбранного порога. Привлечение эксперта на данном этапе позволяет минимизировать риск ошибки первого рода, чтобы не отбросить, приняв за дубль, уникальную информацию. Риск ошибки второго рода при применении данной модели минимален, так как настройка сбора и выгрузки информации также происходит под контролем эксперта.

Использование данной модели обработки имеет положительное влияние на качество модели ИАД и, как следствие, на процесс информационного управления в целом при соблюдении следующих условий:

1) данные в организационной системе не могут быть обработаны полностью автоматически с достаточным уровнем точности и, соответственно, требуют привлечения эксперта в предметной области;

2) процесс принятия решений, касающихся обработки, повторяется во времени с появлением новых данных;

3) большое количество вариаций данных, которые модели ИАД должны корректно обрабатывать.

Только в описанной ситуации время, затраченное на подготовительный этап обработки, будет оправданно. Математически данные условия можно описать следующим образом:

$$T_a = P + \sum_{j=1}^{c} S_j T_E V,$$

$$T_m = \sum_{j=1}^{c} S_j T_E,$$

$$T_a < T_m,$$

где T_a – общее время обработки при полностью ручном подходе; P – время, необходимое на первичную настройку; c – количество циклов обработки; j – номер цикла предобработки; T_E – среднее время работы эксперта над одной единицей данных (строкой или документом в зависимости от задачи), включая время простоя; S_j – множество вариаций текстовых данных в цикле обработки j; V – доля данных для ручной валидации, зависит от структурированности и качества исходных текстов, выбирается пользователем на этапе настройки работы системы, по умолчанию взято за 0,05; T_m – общее время обработки при полностью ручном подходе;

Время, необходимое на первичную настройку, зависит от опыта эксперта по работе с моделью, а подсчет размера множества вариаций текстов осуществляется автоматически. Возможности для сокращения трудозатрат при применении описанного процесса для обработки слабоструктурированных данных будут также проверены экспериментально далее.

Результаты и обсуждение

Для реализации описанной выше модели обработки слабоструктурированных данных был разработан программный комплекс, включающий в себя сервис для сбора и обработки слабоструктурированных текстовых данных и веб-приложение для работы эксперта. Разработанная модель применима в различных предметных областях, где соблюдаются описанные ранее условия. Для иллюстрации этого будет рассмотрено применение разработанной модели в двух динамических организационных системах управления: кадрами в ИТ-компании и здравоохранением на уровне региона. Эти системы объединяет динамичность изменения объектов управления и необходимость учета слабоструктурированных данных для оценки этих изменений. Однако текстовые данные, используемые при управлении обеими системами, очень различаются по структуре. Обеспечивающие процессы, такие как сбор данных, их предобработка, в данных примерах не рассматриваются, упор сделан на расширенном процессе обработки, позволяющим использовать все преимущества разработанной модели.

Кейс 1. Оценка рынка труда в рамках процесса управления кадрами в ИТ-компании

Регулярная оценка рынка труда – задача, которая стоит перед большинством компаний. Анализ вакансий сферы и резюме соискателей может помочь решить следующие задачи:

1) определение актуальных технологий и инструментов, используемых конкурирующими компаниями;

2) оценка среднерыночной зарплаты сотрудников на различных позициях;

3) поиск и определение количества соискателей, обладающих необходимыми навыками.

Вопросы автоматизированной обработки вакансий с целью быстрого реагирования на потребности рынка рассматривались, например, в работе [23]. Для анализа обычно применяются такие подходы, как: выделение ключевых слов, визуализации, автореферирование и др. В рамках исследования будут использоваться разработанные ранее подходы для сбора и обработки слабоструктурированных данных с привлечением эксперта.

Для решения вышеописанных задач на основе разработанной модели использовались данные с сайта «Работа в России» в сфере информационных технологий [22]. Всего в наборе данных присутствует более 13 млн вакансий из разных сфер, описание которых представлено в виде слабоструктурированных текстовых данных. Задача работника кадровой службы – настроить выделение конкретных технологий и навыков из вакансий в автоматизированном режиме. Настроенный мониторинг поможет выявлять изменения трендов на рынке труда в автоматизированном режиме, экономя время работника кадровой службы и помогая ему оперативно их отслеживать. Для этого будут использоваться инструменты, предоставляемые в модели обработки данных.

В данном примере важную роль будет играть корректный сбор данных для анализа. Первый этап обработки – сбор и валидация корректности сбора данных. На этом этапе используется специальная визуализация. При попытке собрать данные по конкретной области разработки программного обеспечения аналитик может столкнуться с рядом сложностей. Например, если попытаться выделить нужный технологический стек по упоминанию в заголовке (в данном примере мы берем стек технологий Frontend), то вакансий будет выбрано мало, чтобы составить статистически достоверную картину о происходящем в отрасли. В то же время если мы используем все вакансии с упоминанием в описании, например, «JavaScript» (ключевой язык программирования в клиентской разработке), то в выборку попадает большое количество технологий, упоминаемых совместно с «JavaScript». При изучении избранных вакансий можно сделать вывод, что эти вакансии относятся к «Fullstack-разработке», объединяющей технологии Frontend и Backend. Используя адаптивный интерфейс визуализаций, эксперт может быстро настроить гибкий запрос. Исключить ключевые понятия из выборки можно двумя способами:

1) исключить из запроса: тогда вакансии, содержащие выбранное ключевое понятие, не будут попадать в выборку;

2) убрать ключевое слово из визуализации, но не из выборки, для возможности сосредоточить внимание на оставшихся ключевых словах.

В данном случае, используя адаптивную визуализацию, возможно исключить следующие упоминания из выборки: «php», «python» и т.д. Количество выбранных по разным типам запросов вакансий представлено в табл. 1. После изменения запроса визуализация будет перестроена на основе обновленной информации.

Таблица 1

Количество вакансий из области «Frontend-разработка», найденных по разным типам запроса

Тип запроса	2019 г.	2020 г.	2021 г.
Простой по заголовкам	15	12	18
Простой по требованиям	301	265	232
Сложный запрос, составленный с помощью визуального редактора	182	144	123

Использование описанных методов визуализации довольно трудоемко при первоначальном запуске и может не принести выигрыша во времени, если в выборке находится мало вакансий или если задача разовая и нет необхо-

димости ее повторного решения на новой порции данных. Например, из данной визуализации, помимо автоматически исключенных слов, которые встречаются в 95 % вакансий вне поискового запроса («опыт», «требования», «график», различные распространенные в русском языке слова), из самой визуализации возможно исключить слова, специфические для Frontend-разработки, но не приносящие дополнительной информации по используемым технологиям («проект», «сайт», «продукт», «HTML»). После того как сбор настроен и провалидирован, визуализация может использоваться для детектирования изменений во времени. Например, обзор технологий, используемых во Frontend в 2020 и 2021 гг.: на рис. 4 отображен экран работы с визуализацией для оценки изменений в сравнении с предыдущим анализом. Из данной визуализации можно сделать вывод, что в выборке исследуемых вакансий значительно увеличился спрос на разработку для высоконагруженных систем и применение инструментов Docker для развертывания Frontend-приложений. В то же время количество упоминаний фреймворка Angular существенно упало. Время ручного и автоматизированного анализа трендов упоминаний в вакансиях представлено в табл. 2.

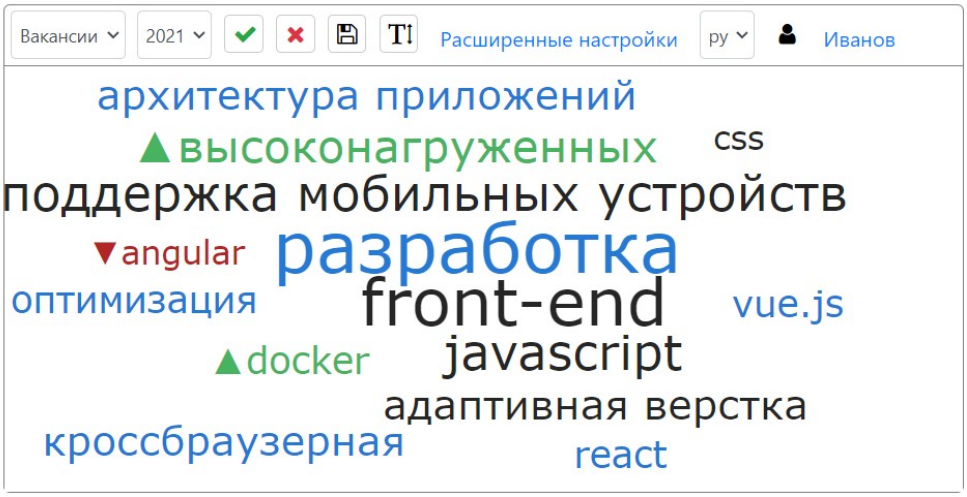

Рис. 4. Анализ трендов по направлению «Frontend» после настройки визуализации, по результатам обработки данных с сайта «Работа в России»

Таблица 2

Время ручного и автоматизированного анализа трендов упоминаний в вакансиях

Тип анализа	2019 г.	2020 г.	2021 г.
Ручной	4,6 ч	4,53 ч	4,73 ч
Автоматизированный	3 ч	0,5 ч	0,5 ч

Из выборки в 670 вакансий обнаружилось 149 полных дублей (повтор вакансии слово в слово) и 38 вакансий с высоким уровнем схожести по требованиям. Таким образом, выборка перед анализом или ручной классификацией отдельных вакансий (если таковая требуется), уменьшается на **28 %**.

Результаты проведенного анализа могут быть как переданы в кадровую службу для дальнейшего ручного анализа, так и использованы совместно с другими методами анализа текстов при построении различных моделей ИАД в контуре управления кадровыми ресурсами организации. Работник кадровой службы может использовать результаты анализа трендов для следующих целей: разработка или корректировка программ повышения квалификации сотрудников с учетом актуальных трендов развития ИТ-технологий; изменение требований к расширению штата с учетом покрытия недостающих компетенций; составление описания новых вакансий и корректировке существующих. Благодаря своевременной реакции на изменение технологий штат сотрудников будет с большей вероятностью покрывать компетенции, необходимые для соответствия продуктов компании современным стандартам. При необходимости ручной классификации пользователь может воспользоваться редактором, описание которого представлено во втором кейсе.

Сокращение времени работы эксперта на этапе обработки данных позволяет повысить оперативность управления динамической организационной системой – системой управления кадрами, в которой решения необходимо принимать быстро и регулярно, учитывая обновленную информацию. Управление информацией, доступной специалистам кадровой службы на момент принятия решений – удаление дублей, использование визуализаций, демонстрирующих тренды – позволяет быстрее строить и проверять гипотезы об изменениях на трудовом рынке, что также увеличивает оперативность принимаемых решений. При дальнейшем использовании полученных данных в СППР отсутствие дублирующейся информации в выборке способствует построению более качественных моделей ИАД, что в свою очередь способствует улучшению качества и обоснованности принятых на основе их анализа управленческих решений.

Кейс 2. Автоматизация обработки данных анамнеза пациента в рамках процесса управления здравоохранением на уровне региона

Система регионального здравоохранения является примером динамической организационной системы, в которой требуется регулярное распределение различных ресурсов (людских, финансовых, фармацевтических и др.) как в краткосрочном, так и в долгосрочном горизонте планирования. Для анализа использовались обезличенные данные из ИЭМК пациентов государственных учреждений здравоохранения Брянской области, предоставленные для проведения совместных научных исследований в рамках соглашения об информационном взаимодействии между ГАУЗ МИАЦ и ФГБОУ ВО БГТУ. В данном случае использовался набор данных, состоящий из 269 217 записей анамнезов пациентов с 2017 по 2021 г. С момента начала активной цифровизации здравоохранения исследователи имеют возможность использовать накопленные объемы данных для создания систем интеллектуальной поддержки управленческих решений. Однако серьезным препятствием для использования этих данных является их слабая структурированность, поскольку часто запись важной информации выполняется в полях свободного ввода. Кроме того, данные из информационных систем учреждений здравоохранения насыщены различными специфическими сокращениями и профессио-

нальной терминологией. И эта проблема актуальна не только для русскоязычных баз данных [25, 26].

Извлечение информации об анамнезе пациента и дальнейшее ее использование имеет большое значение при выстраивании прогностических моделей [27]. Улучшение качества прогнозирования заболеваний поможет эффективнее решать ряд важнейших управленческих задач: планирование ресурсов системы здравоохранения, закупок медикаментов, сезонной загрузки специалистов и т.д. Задача классификации анамнеза, описанного в свободной форме, в большинстве случаев не поддается быстрому автоматическому решению, поэтому будет целесообразно использовать предложенную выше модель автоматизированной обработки данных. Чем выше будет качество выборки, которую будет в дальнейшем размечать эксперт в предметной области, тем меньше времени пройдет до получения размеченного набора данных, который можно будет использовать в моделях ИАД и, соответственно, до использования результатов анализа в процессе принятия управленческих решений.

Одним из первых этапов обработки, касающихся этих данных, будет являться раскрытие специфических сокращений и исправление орфографических ошибок. Несмотря на то, что раскрытие сокращений является частной задачей улучшения качества выборки, без ее решения нельзя использовать эти данные в моделях ИАД из-за того, что различные токены (сочетания символов) обозначают одно и то же понятие, что вносит неоднозначность, ухудшающую качество анализа данных. Например, слово «хронический» сокращается в выборке как «хр», «хрон», «хрнч», с точкой и без. Необходимость данного вмешательства программный сервис определяет автоматически, также эксперт в процессе работы видит случайную выборку записей на экране и может принять решение о необходимых этапах. Часть сокращений возможно раскрыть автоматически, используя алгоритм, предложенный авторами в предыдущей работе [17], при решении данной задачи не будем заострять внимание на этом процессе, результаты представлены в табл. 3.

Таблица 3

Автоматически раскрытые сокращения

Сокращения (использованные с точкой и без)	Количество упоминаний в выборке	Раскрытие
стац	378	Стационарное
хр	4265	Хронический
тубер	15	Туберкулез
бр	163	Бронхиальная

Аналогичный алгоритм используется для поиска ошибочных написаний. Самые популярные ошибки, обнаруженные в данных, представлены в табл. 4. Корректность найденного автоматически исправления проверяется экспертом в предметной области. Так как одни и те же ошибки встречались в наборе данных много раз, выигрыш по времени работы эксперта достигался даже с учетом высокого уровня валидации.

Таблица 4

Результат автоматизированного исправления ошибок

Варианты написания	Правильный вариант
отягащен, отгощен	отягощен
кровточивость, кровточивсоть	кровоточивость
ботуина, ботткина	(болезнь) Боткина
туберкулз, тубекрулез	туберкулез

Чтобы объединить одинаковые записи, используем подсистему поиска дублей [18]. Данное вмешательство на стандартных «осторожных» настройках позволяет уменьшить количество вариантов, которые необходимо будет обработать, на 3,2 %.

Далее нам необходимо обработать сами строки с информацией об анамнезе пациента. Для этого мы используем модуль «классификации данных». Задача программного комплекса, которая достигается реализацией предложенной модели: как можно меньшим количеством обращений к эксперту обработать большее количество записей с максимальным сохранением качества. Для этого записи ранжируются, объединяясь в группы по полному совпадению или высокой семантической близости (настроенной пользователем), исходя из наиболее частой встречаемости группы в выборке: таким образом, потратив адекватное количество времени, эксперт может обработать большее количество данных. Кроме того, формировать группы эксперту предлагается и самостоятельно, используя интерактивные инструменты: варианты для объединения подбираются исходя из семантической близости записей (>0,5), что позволяет сэкономить время на их классификацию. Для удобства сравнения семантически близких результатов несовпадающие части в редакторе выделяются, давая возможность при обработке тратить меньше времени за счет фокусировки внимания на похожих объектах (рис. 5). Особенно это актуально для развернутого описания анамнеза, часть которого может совпадать с другим описанием.

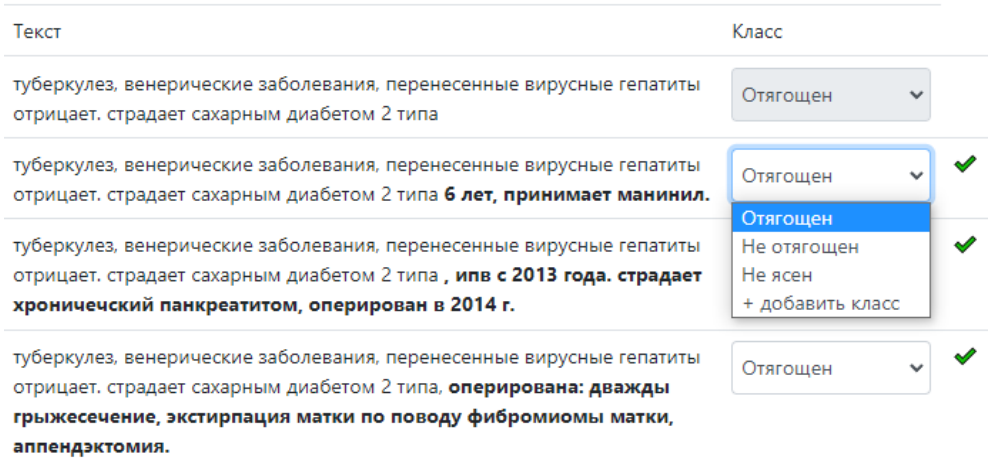

Рис. 5. Интерфейс для классификации описаний анамнезов

Примеры несовпадения результатов классификации близких по тексту строк представлены в табл. 5. Накопление достаточного количества данных ручной разметки в перспективе позволит создавать модели автоматической классификации.

Таблица 5

Примеры анамнезов с несовпадающим результатом классификации

Строки	Результат
Туберкулез, болезнь боткина	Отягощен
Туберкулез, болезнь боткина **отрицает**.	**Не отягощен**
Из перенесенных заболеваний отмечает орви. болезнь боткина, туберкулез, венерические . заболевания, вич, вирусный гепатиты у себя отрицает. операций в анамнезе нет. кровь и ее компоненты ранее не переливались.	Не отягощен
Из перенесенных заболеваний отмечает орви. болезнь боткина, туберкулез, венерические заболевания, вич, вирусный гепатиты у себя отрицает. операций в анамнезе нет. кровь и ее компоненты ранее не переливались. **длительное время страдает сах. диабетом 2 тип(принимает диабетон 60 мг), артериальная гиперетензия**	**Отягощен**

Кроме того, все решения эксперта, принятые в процессе настройки, такие как: раскрытие специфических сокращений, определение порога близости или объединение записей в группы, сохраняются в базе данных программного комплекса и при появлении новых данных могут быть использованы в автоматическом режиме. Таким образом, достигается такое качество предлагаемой модели, как ***оперативность*** обработки, что позволяет быстро переобучать модели ИАД, которые будут использовать результат обработки в дальнейшем. Полученные результаты могут быть использованы в моделях машинного обучения оценки уровня рисков различных заболеваний пациентов при прохождении диспансеризации, для обучения на размеченных данных и полностью автоматической дальнейшей разметки. Результаты сокращения стартовой выборки уникальных записей в количестве 74 291 представлены в табл. 6.

Таблица 6

Результаты операций по обработке данных

Операция над данными	Количество уникальных сокращений после обработки	Сокращение размерности
Подготовка данных	72 597	2,2
Раскрытие сокращений и исправление ошибок	65 696	9,3
Удаление достоверных дублей	63 315	3,2

Суммарное сокращение выборки перед классификацией: **на 14,7 %**, что позволит сократить общее время с учетом валидации экспертом не менее 5 % записей **на 14,1 %**. Сокращение времени спрогнозировано исходя из того, что эксперт сможет разметить большую часть выборки, классифицировав меньшее количество строк. Кроме того, раскрытие специфических сокраще-

ний и исправление ошибок дает возможность более эффективного использования полученных данных в моделях ИАД. Количество записей с уровнем совпадения более 50 %, которые могут быть обработаны в редакторе быстрее, чем полностью уникальные записи: 6660, что может сократить дальнейшее время обработки вплоть **до 4,4 %**.

Дополнительный эксперимент показал, что если для данной выборки пропустить этап раскрытия сокращений и ошибок, то количество найденных достоверных дублей на следующем этапе уменьшается более чем в два раза – с 2301 до 1112. Этот результат свидетельствует о том, что порядок выполнения этапов обработки важен для достижения максимального результата.

Для расчета сокращения времени при динамической обработке было проверено, сколько записей за последующие годы будет обработано автоматически после обработки первого массива записей и сохранения результатов в хранилище. Расчеты показали, что **от 8 до 12 %** записей в последующие годы будет обработано автоматически. Ускорение этапа обработки данных позволяет быстрее начать использовать обновленные данные в моделях ИАД и, соответственно, передавать результат анализа ЛПР. Это важно для ускорения принятия решений и повышения их обоснованности при планировании деятельности системы здравоохранения региона, что позволит добиться большей оперативности и эффективности управления динамической организационной системой.

Обсуждение

Предложенная авторами модель обработки слабоструктурированных данных была апробирована на двух примерах, которые объединяет место использования слабоструктурированных текстовых данных в процессе информационного управления и динамический тип организационных систем. При этом характеристики самих текстовых данных различаются для каждого примера. В первом случае это достаточно объемные тексты из открытых источников. Во втором случае – короткие тексты из внутренней информационной системы, наполненные специфическими сокращениями и ошибками. Эти тексты изначально создавались для передачи информации от одних профессионалов к другим. Несмотря на лингвистические различия, процесс обработки текстов не меняется – варьируются лишь дополнительные этапы, позволяющие наиболее корректно собрать и подготовить тексты для оценки экспертом или дальнейшей автоматической обработки.

Методы, используемые при реализации модели, были опробованы на текстах из различных социально-экономических областей: оценка кредитных рисков, анализ рынка труда, здравоохранение, образование и т.д. [28]. Дальнейшие усилия планируется направить на расширение возможностей модели для поддержки информационного управления, а также изучение возможностей оптимизации и развития отдельных методов и инструментов для работы эксперта с большими объемами данных.

Результаты проведенных экспериментов показывают, что разработанная модель эффективна для обработки слабоструктурированных текстовых данных в динамических организационных системах из различных предметных областей. Наибольшая оперативность достигается при принятии повторных решений, благодаря сохранению параметров модели. Ограничением

предложенной модели является обязательный этап создания языковой модели предметной области и необходимость настройки модели под выполнение конкретных задач. В связи с этим ограничением применение модели является наиболее эффективным и целесообразным в случаях, требующих многократной обработки большого количества текстовых данных.

Заключение

Интеллектуальный анализ данных ускоряет процесс принятия решений в динамических организационных системах. Если в системе используется много слабоструктурированных текстовых данных, целесообразно включать их в модели анализа после предварительной обработки. Авторами предложена модель обработки слабоструктурированных текстовых данных на русском языке для использования в моделях интеллектуального анализа данных в процессе информационного управления в динамических организационных системах.

Данная модель позволяет автоматизировать сбор и обработку текстовых данных, сочетая сильные стороны человеческой и машинной обработки. Модель возможно использовать для оценки корректности сбора данных и их обработки. В случае, если данные насыщенны сокращениями и ошибками, обработка позволит использовать эти данные в моделях ИАД и способствует повышению точности и скорости работы. Для контроля корректности сбора и обработки используются визуализация результатов и валидация части выборки экспертом. Прирост скорости обработки информации при использовании модели может быть достижим в задачах, где важную роль играют постоянно обновляющиеся массивы текстовой информации, при этом прирост может быть достижим только со второй итерации использования из-за времени, необходимого на первичную настройку. Описанная модель является универсальной по отношению к таким показателям текстов, как: размер, корректность, наличие профессиональной лексики.

По результатам проведенных экспериментов за счет визуализаций был достигнут прирост оперативности на этапе сбора для текстов, получаемых из обширных массивов открытых данных, таких как вакансии. Применение модели обработки к обоим примерам позволило выполнить сокращение выборки для дальнейшей обработки экспертом – от 14 до 28 %, в зависимости от задачи, несмотря на необходимость валидации результатов, а благодаря сохранению результатов обработки дополнительно автоматически сократить выборку от 8 до 12 % при последующей обработке (от года к году).

Полученные результаты демонстрируют целесообразность применения предложенной модели в СППР, использующих модели ИАД для информационного управления в организационных системах, в которых текстовые данные регулярно обновляются и являются одним из информационных входов для ЛПР. При этом сокращается нагрузка на эксперта, а в дальнейшем, при накоплении достаточного количества данных, возможен переход на полностью автоматическую классификацию текстов при обучении соответствующих моделей машинного обучения на достаточном объеме данных.

Дальнейшие исследования планируется направить на развитие инструментов эксперта по работе с текстами с использованием программного сервиса, а также на исследование возможности применения разработанной модели для решения задач информационного управления в других предметных областях.

Список литературы

1. Российский рынок M&A: в трендах IPO, возобновляемая энергетика и медицина. URL: https://sber.pro/publication/rossiiskii-rynok-m-a-v-trendakh-ipo-vozobnovliaemaia-energetika-i-meditsina (дата обращения: 22.04.2022).
2. Digital 2021 // We Are Social. URL: https://wearesocial.com/uk/blog/2021/01/digital-2021-uk/ (дата обращения: 22.04.2022).
3. Mai F., Tian S., Lee Ch. [et al.]. Deep Learning Models for Bankruptcy Prediction using Textual Disclosures // European Journal of Operational Research. 2019. Vol. 274. P. 743–758.
4. Балашова И. Ю., Волынская К. И., Макарычев П. П. Методы и средства генерации тестовых заданий из текстов на естественном языке // Модели, системы, сети в экономике, технике, природе и обществе. 2016. № 1. С. 195–202.
5. Pérez J., Iturbide E., Olivares V. [et al.]. A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases // Journal of Medical Systems. 2015. Vol. 39. P. 152.
6. Новиков Д. А. Теория управления организационными системами. 3-е изд. М. : Издательство физико-математической литературы, 2012. С. 30–32.
7. Петрова Д. А., Трунин П. В. Анализ влияния пресс-релизов ЦБ РФ на показатели денежного рынка // Бизнес-информатика. 2021. Т. 15, № 3. С. 24–34. doi:10.17323/2587-814X.2021.3.24.34
8. Dorfleitner G., Priberny C., Schuster S. [et al.]. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms // Journal of Banking & Finance. 2015. № 64. P. 169–187. doi:10.1016/j.jbankfin.2015.11.009
9. Hickman L., Thapa St., Tay L. [et al.]. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations // Organizational Research Methods. 2022. № 25. doi:10.1177/1094428120971683
10. Ebrahimi A., Barforoush A. Preprocessing Role in Analyzing Tweets Towards Requirement Engineering // 27th Iranian Conference on Electrical Engineering (ICEE). 2019. doi:10.1109/IranianCEE.2019.8786652
11. Yanwei B., Quan Ch., Wang L. [et al.]. The Role of Pre-processing in Twitter Sentiment Analysis // Procedia Computer Science. 2014. № 89. P. 549–554. doi:10.1016/j.procs.2016.06.095
12. Hasanah U., Astuti Tr., Wahyudi R. [et al.]. An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian // 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE). 2018. P. 230–234. doi:10.1109/ICITISEE.2018.8720957
13. Ожерельева Т. А. Информационное управление подвижными объектами // Государственный советник. 2018. № 4. С. 29–37
14. Rashid A., Shoaib U. Knowledge discovery in database using intention mining // Science International. 2016. № 28. P. 5145–5151.
15. Makarova E. A., Lagerev D. G., Lozbinev F. Y. Approaches to visualizing big text data at the stage of collection and pre-processing // Scientific Visualization. 2019. № 4.
16. Лагерев Д. Г., Макарова Е. А. Поиск и раскрытие сокращений в русскоязычных данных медицинских информационных систем // Вестник компьютерных и информационных технологий. 2020. № 7. С. 44–54.
17. Макарова Е. А., Лагерев Д. Г. Оценка семантической близости новостных сообщений на основе анализа заголовков // Вестник компьютерных и информационных технологий. 2021. Vol. 18, № 7. С. 46–56. doi:10.14489/vkit.2021.07.pp.046-056
18. Аверченков В. И., Будыльский Д. В., Подвесовский А. Г. Анализ применения моделей векторного представления текстовой информации для русскоязычных текстов // Вестник компьютерных и информационных технологий. 2016. № 3. С. 31–37. doi:10.14489/vkit.2016.03.pp.031-037

19. Душков Б. А., Королев А. В., Смирнов Б. А. Быстрое чтение // Энциклопедический словарь: Психология труда, управления, инженерная психология и эргономика. М. : Академический проект ; Фонд «Мир», 2005.
20. Rand S., Lall R. Development of a Custom Spell-Checker for Emergency Department Data // Online Journal of Public Health Informatics. 2019. Vol 11. doi:10.5210/ojphi.v11i1.9745
21. Zhang H., Qiu Sh., Duan X. [et al.]. Token Drop mechanism for Neural Machine Translation // Proceedings of the 28th International Conference on Computational Linguistics. 2020. P. 4298–4303. doi:10.18653/v1/2020.coling-main.379
22. Mytna Kurekova L., Beblavý M., Thum-Thysen A. Online job vacancy data as a source for micro-level analysis of employers' preferences. A methodological enquiry // IZA Journal of Labor Economics. 2015. Vol. 4. P. 1–20.
23. «Работа в России»: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru // Роструд / отв. Бабушкина В. О., Тимошенко А. Ш. ; Инфраструктура научно-исследовательских данных, АНО «ЦПУР», 2021. URL: http://data-in.ru/data-catalog/datasets/186/ (дата обращения: 22.04.2022).
24. Kreuzthaler M., Oleynik M., Avian A., Schulz S. Unsupervised Abbreviation Detection in Clinical Narratives // Studies in Health Technology and Informatics. 2016. № 245. P. 539–543.
25. Mykowiecka Ag., Marciniak M. Experiments with ad hoc ambiguous abbreviation expansion // Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis. 2019. P. 44–53. doi:10.18653/v1/D19-6207
26. Ramos A., Allende-Cid H., Taramasco C. [et al.]. Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis // IEEE Access. 2020. doi:10.1109/ACCESS.2020.3000075
27. Макарова Е. А., Лагерев Д. Г. Применение автоматизированной системы интеллектуального анализа текстовых данных для управления процессом формирования индивидуальных образовательных траекторий // Информационные системы и технологии ИСТ-2020 : сб. материалов XXVI Междунар. науч.-техн. конф. 2020. С. 362–367.

References

1. *Rossiyskiy rynok M&A: v trendakh IPO, vozobnovlyaemaya energetika i meditsina = Russian M&A market: IPO trends, renewable energy and medicine*. (In Russ.). Available at: https://sber.pro/publication/rossiiskii-rynok-m-a-v-trendakh-ipo-vozobnovliaemaia-energetika-i-meditsina (accessed 22.04.2022).
2. Digital 2021. *We Are Social*. Available at: https://wearesocial.com/uk/blog/2021/01/digital-2021-uk/ (accessed 22.04.2022).
3. Mai F., Tian S., Lee Ch. et al. Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. *European Journal of Operational Research*. 2019;274:743–758.
4. Balashova I.Yu., Volynskaya K.I., Makarychev P.P. Methods and means of generating test tasks from texts in natural language. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve = Models, systems, systems in economics, technology, nature and society*. 2016;(1):195–202. (In Russ.)
5. Pérez J., Iturbide E., Olivares V. et al. A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases. *Journal of Medical Systems*. 2015;39:152.
6. Novikov D.A. *Teoriya upravleniya organizatsionnymi sistemami. 3-e izd = Theory of management of organizational systems. 3rd ed*. Moscow: Izdatel'stvo fiziko-matematicheskoy literatury, 2012:30–32. (In Russ.)
7. Petrova D.A., Trunin P.V. Analysis of the impact of press releases of the Central Bank of the Russian Federation on money market indicators. *Biznes-informatika = Business Informatics*. 2021;15(3):24–34. (In Russ.). doi:10.17323/2587-814X.2021.3.24.34

8. Dorfleitner G., Priberny C., Schuster S. et al. Description-text related soft infor-mation in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking & Finance*. 2015;(64):169–187. doi:10.1016/j.jbankfin.2015.11.009
9. Hickman L., Thapa St., Tay L. et al. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organizational Research Methods*. 2022;(25). doi:10.1177/1094428120971683
10. Ebrahimi A., Barforoush A. Preprocessing Role in Analyzing Tweets Towards Requirement Engineering. *27th Iranian Conference on Electrical Engineering (ICEE)*. 2019. doi:10.1109/IranianCEE.2019.8786652
11. Yanwei B., Quan Ch., Wang L. et al. The Role of Pre-processing in Twitter Sentiment Analysis. *Procedia Computer Science*. 2014;(89):549–554. doi:10.1016/j.procs.2016.06.095
12. Hasanah U., Astuti Tr., Wahyudi R. et al. An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian. *3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*. 2018:230–234. doi:10.1109/ICITISEE.2018.8720957
13. Ozherel'eva T.A. Information management of mobile objects. *Gosudarstvennyy sovetnik = State Advisor*. 2018;(4):29–37 (In Russ.)
14. Rashid A., Shoaib U. Knowledge discovery in database using intention mining. *Science International*. 2016;(28):5145–5151.
15. Makarova E.A., Lagerev D.G., Lozbinev F.Y. Approaches to visualizing big text data at the stage of collection and pre-processing. *Scientific Visualization*. 2019;(4).
16. Lagerev D.G., Makarova E.A. Search and disclosure of abbreviations in Russian-language data of medical information systems. *Vestnik komp'yuternykh i informatsionnykh tekhnologiy = Bulletin of Computer and Information Technologies*. 2020;(7):44–54. (In Russ.)
17. Makarova E.A., Lagerev D.G. Assessment of semantic proximity of news reports based on headline analysis. *Vestnik komp'yuternykh i informatsionnykh tekhnologiy = Bulletin of Computer and Information Technologies*. 2021;18(7):46–56. (In Russ.). doi:10.14489/vkit.2021.07.pp.046-056
18. Averchenkov V.I., Budyl'skiy D.V., Podvesovskiy A.G. Analysis of the application of models of vector representation of textual information for Russian-language texts. *Vestnik komp'yuternykh i informatsionnykh tekhnologiy = Bulletin of Computer and information technologies*. 2016;(3):31–37. (In Russ.). doi:10.14489/vkit.2016.03.pp.031-037
19. Dushkov B.A., Korolev A.V., Smirnov B.A. Quick reading. *Entsiklopedicheskiy slovar': Psikhologiya truda, upravleniya, inzhenernaya psikhologiya i ergonomika = Encyclopedic dictionary: Psychology of labor, management, engineering psychology and Ergonomics*. Moscow: Akademicheskiy proekt; Fond «Mir», 2005.
20. Rand S., Lall R. Development of a Custom Spell-Checker for Emergency Department Data. *Online Journal of Public Health Informatics*. 2019;11. doi:10.5210/ojphi.v11i1.9745
21. Zhang H., Qiu Sh., Duan X. et al. Token Drop mechanism for Neural Machine Translation. *Proceedings of the 28th International Conference on Computational Linguistics*. 2020:4298–4303. doi:10.18653/v1/2020.coling-main.379
22. Mytna Kurekova L., Beblavý M., Thum-Thysen A. Online job vacancy data as a source for micro-level analysis of employers' preferences. A methodological enquiry. *IZA Journal of Labor Economics*. 2015;4:1–20.
23. Babushkina V.O., Timoshenko A.Sh. (resp.). "Work in Russia": processed and combined information about vacancies, resumes, responses and invitations of the portal trudvsem.ru. *Rostrud = Rostrud*. Infrastruktura nauchno-issledovatel'skikh dannykh, ANO «TsPUR», 2021. Available at: http://data-in.ru/data-catalog/datasets/186/ (accessed 22.04.2022).

24. Kreuzthaler M., Oleynik M., Avian A., Schulz S. Unsupervised Abbreviation Detection in Clinical Narratives. *Studies in Health Technology and Informatics*. 2016;245:539–543.
25. Mykowiecka Ag., Marciniak M. Experiments with ad hoc ambiguous abbreviation expansion. *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis*. 2019:44–53. doi:10.18653/v1/D19-6207
26. Ramos A., Allende-Cid H., Taramasco C. et al. Application of Machine Learning and Word Embeddings in the Classification of Cancer Diagnosis Using Patient Anamnesis. IEEE Access. 2020. doi:10.1109/ACCESS.2020.3000075
27. Makarova E.A., Lagerev D.G. Application of an automated system of intellectual analysis of text data to control the process of formation of individual educational trajectories. *Informatsionnye sistemy i tekhnologii IST-2020: sb. materialov XXVI Mezhdunar. nauch.-tekhn. konf. = Information systems and technologies IST-2020 : collection of materials XXVI International Scientific-technical. conf.*. 2020:362–367. (In Russ.)

Информация об авторах / Information about the authors

Елена Андреевна Макарова
ассистент преподавателя
кафедры информатики
и программного обеспечения,
Брянский государственный
технический университет
(Россия, г. Брянск, б-р 50 лет Октября, 7)
E-mail: m4karova.e@yandex.ru

Elena A. Makarova
Teaching assistant of the sub-departments
of computer science and software,
Bryansk State Technical University
(7 50 years of October boulevard,
Bryansk, Russia)

Дмитрий Григорьевич Лагерев
кандидат технических наук, доцент,
доцент кафедры информатики
и программного обеспечения,
Брянский государственный
технический университет
(Россия, г. Брянск, б-р 50 лет Октября, 7)
E-mail: lagerevdg@yandex.ru

Dmitriy G. Lagerev
Candidate of technical sciences,
associate professor,
associate professor of the sub-department
of computer science and software,
Bryansk State Technical University
(7 50 years of October boulevard,
Bryansk, Russia)

Авторы заявляют об отсутствии конфликта интересов /
The authors declare no conflicts of interests.

Поступила в редакцию/Received 31.05.2022
Поступила после рецензирования/Revised 03.08.2022
Принята к публикации/Accepted 02.09.2022