На правах рукописи

ДУДАРИН Павел Владимирович

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ КОРОТКИХ ТЕКСТОВ

Специальность: 05.13.01 - Системный анализ, управление и обработка информации (информационные технологии и промышленность)

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Ульяновск - 2021

Работа выполнена на кафедре “Информационные системы” федерального государственного бюджетного образовательного учреждения высшего образования «Ульяновский государственный технический университет».

Научный руководитель: **Ярушкина Надежда Глебовна,** *доктор технических наук, профессор, профессор кафедры “Информационные системы” федерального государственного бюджетного образовательного учреждения высшего образования «Ульяновский государственный технический университет».*

Официальные оппоненты: **Борисов Вадим Владимирович**, *доктор технических наук, профессор, профессор кафедры «Вычислительной техники», филиал ФГБОУ ВПО «Национальный исследовательский университет «МЭИ» в г. Смоленске.*

Сухов Сергей Владимирович, *кандидат физико-математических наук, старший научный сотрудник, Ульяновский филиал федерального государственного бюджетного учреждения науки «Институт радиотехники и электроники им. В.А. Котельникова».*

Ведущая организация: *Федеральное государственное бюджетное учреждение науки «Институт автоматики и процессов управления Дальневосточного отделения Российской академии наук».*

Защита состоится «29» сентября 2021 года в 12:00 часов на заседании диссертационного совета Д212.277.04 при ФГБОУ ВО «Ульяновский государственный технический университет» по адресу: 432027, г. Ульяновск, ул. Северный Венец, 32 (ауд. 211, Главный корпус).

С диссертацией можно ознакомиться в библиотеке и на сайте ФГБОУ ВО «Ульяновский государственный технический университет». Диссертация и автореферат размещены на сайте http://www.ulstu.ru/.

Автореферат разослан « » _____ 2021 г.

Ученый секретарь
диссертационного совета,
д.т.н., доцент  Наместников А.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Кластерный анализ является одним из методов системного анализа, активно применяемым при анализе больших объемов структурированных и неструктурированных данных. Исследованиям в данной области посвящены работы известных зарубежных и российских ученых: Basu S., Hinton G., Manning Ch.D., Hastie T., Kaufman L., Picard R.W., Воронцова К.В., Хорошевского В.Ф., Ярушкиной Н.Г. и др.

Кластеризация текстов допускает значительное число возможных принципов для разбиения на классы: тематика, автор, стиль, эмоциональная окраска, правовой статус и комбинация различных факторов. Методы, не позволяющие учесть интенцию эксперта, оказываются в общем случае не эффективными для решения описанной задачи. Альтернативным является подход, при котором эксперт включается в процесс кластеризации и на различных ее этапах задает ограничения на основе промежуточных результатов, которые учитываются на дальнейших стадиях кластеризации. Такие методы классифицируются как методы интерактивной кластеризации с использованием обратной связи от эксперта. Интерактивные методы обеспечивают сокращение суммарных затрат времени эксперта на обработку результатов кластеризации и позволяют повысить точность кластеризации за счет выявления скрытого знания эксперта на ранних этапах кластеризации. Учет дополнительной информации позволяет алгоритму выбрать правильное направление хода процесса разбиения на кластеры.

Актуальность диссертационного исследования

Стремительный рост массивов информации, состоящих из наборов коротких текстовых фрагментов, способствует интенсификации исследований в области развития методов обработки текстов с применением машинного обучения. Проблеме ежегодно посвящается значительно число исследований. Большая часть проводимых исследований относится к текстам на английском языке. Исследований в области русского языка значительно меньше, что объясняется не только меньшим числом исследователей, занимающихся вопросами русского языка, но и объективно большей сложностью русского языка для автоматизированной обработки. Недостаточная разработанность стандартных средств кластеризации для коротких текстов и низкая эффективность существующих методов на русскоязычных текстах затрудняет их использование в российских автоматизированных системах поддержки принятия решений и управления.

Это подтверждается отсутствием стандартных средств кластеризации для коротких текстов в ведущих NLP-пакетах (Natural Language Processing, обработка естественного языка) – например, в NLTK.

В данной работе рассматривается пример системы, в которой происходит генерация большого количества коротких текстов, – системы стратегического планирования Российской Федерации. В ней участники формируют документы стратегического планирования, в рамках которых определяются ключевые показатели эффективности. Формулировки ключевых показателей эффективности образуют набор данных, состоящий из коротких текстов. В рамках данной системы остро стоит задача формирования и актуализации классификатора основанного на данном наборе. Эта задача может быть решена с помощью кластеризации.

На основании вышеизложенного можно сформулировать вывод о том, что исследования в области интерактивной нечеткой кластеризации коротких текстов на русском языке являются важной и актуальной задачей.

Объектом исследования в диссертационной работе является кластеризация наборов данных, состоящих из коротких текстов на русском языке и экспертная информация, поступающая в ходе интерактивной обработки текстов.

Предметом исследования являются модели и методы нечеткой кластеризации коротких текстов и обработки экспертной информации.

Цель диссертационной работы

Повышение эффективности нечеткой кластеризации коротких текстов путем разработки модели, метода и алгоритма в системе поддержки принятия решений для кластеризации коротких текстов на русском языке с учетом экспертной информации. Эффективность определяется точностью кластеризации и сокращением времени и трудоемкости работы выполняемой экспертом при использовании предложенного решения.

Для достижения поставленной цели необходимо решить следующие задачи:

- провести исследование моделей и методов машинного обучения для обработки текстов для выявления новых подходов к повышению эффективности четкой и нечеткой кластеризации коротких текстов;

- разработать метод расширения словаря языковой модели на базе нейронной сети;

- разработать метод для обработки экспертной информации в ходе нечеткой интерактивной кластеризации коротких текстов;

- сформулировать перечень этапов программы проведения испытаний метода нечеткой интерактивной кластеризации коротких текстов;

- составить алгоритм автоматизации работ по нечеткой интерактивной кластеризации коротких текстов в системе поддержки принятия решений;

- провести апробацию разработанных модели, методов и алгоритма нечеткой интерактивной кластеризации коротких текстов в качестве элементов функционирующей системы поддержки принятия решений.

Методы исследования

При решении задач исследования были применены методы теории вероятностей, математической статистики, методы машинного обучения, кластерный анализ, теория нечетких множеств, численные методы. При разработке программного модуля были использованы методы объектно-ориентированного программирования.

Область исследования

Область исследования соответствует паспорту специальности 05.13.01. – «Системный анализ, управление и обработка информации (технические науки)», а именно:

п. 4 – разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации;

п. 13 – методы получения, анализа и обработки экспертной информации.

Научной новизной обладают**:**

- предложенная архитектура искусственной нейронной сети, отличающаяся от известных тем, что позволяет решать задачу кластеризации на базе скрытого пространства признаков языковой модели;

- предложенный метод обработки текстов для расширения словаря языковой модели на базе нейронной сети с использованием нечеткого иерархического классификатора, отличающийся от известных тем, что позволяет учитывать семантическую близость слов;

- предложенный метод обработки обратной связи от эксперта, отличающийся от известных тем, что позволяет корректировать весовые коэффициенты нейронной сети и проводить интерактивную кластеризацию наборов коротких текстов;

- разработанный алгоритм, автоматизирующий применение предложенных модели и методов для выполнения нечеткой интерактивной кластеризации наборов коротких текстов, интегрированный в систему поддержки принятия решений (СППР).

Достоверность результатов исследования. Достоверность полученных результатов обеспечена математически строгим выполнением расчетов, подтверждена вычислительными экспериментами и результатами практического использования.

Теоретическая значимость диссертационной работы заключается в разработке новых моделей и методов с использованием нейронных сетей и языковых моделей для решения задачи нечеткой кластеризации наборов данных состоящих из коротких текстов.

Практическая значимость диссертационной работы заключается в разработке программного модуля системы поддержки принятия решений на языке Python, позволяющего осуществлять интерактивную нечеткую кластеризацию наборов данных, состоящих из коротких текстов, и применение его в задаче анализа набора коротких текстов в рамках НИР в интересах Министерства экономического развития РФ для Системы стратегического планирования РФ.

Основные научные положения, выносимые на защиту:

1. Предложенная архитектура искусственной нейронной сети позволяет эффективно решать задачу кластеризации на базе пространства признаков языковой модели русского языка.
2. Предложенный метод обработки текстов для расширения словаря языковой модели на базе нейронной сети с использованием нечеткого иерархического классификатора повышает точность кластеризации.
3. Предложенный метод учета обратной связи от эксперта, используемый для корректировки весовых коэффициентов нейронной сети, позволяет проводить интерактивную кластеризацию наборов коротких текстов.
4. Разработанный алгоритм на основе предложенных моделей и методов реализован в системе поддержки принятия решений и автоматизирует применение предложенных моделей и методов для выполнения нечеткой интерактивной кластеризации наборов коротких текстов.

Реализация и внедрение результатов работы

Основные теоретические и практические результаты диссертационной работы использованы в рамках фундаментальных и прикладных научных исследований Министерства экономического развития РФ по темам: “Разработка рекомендаций по совершенствованию информационного обеспечения участников стратегического планирования в части осуществления мониторинга и контроля реализации документов стратегического планирования с использованием Федеральной информационной системы стратегического планирования (ФИС СП)” и “Разработка методического обеспечения интеллектуальной системы проверки уведомления об утверждении (одобрении) документа стратегического планирования или внесении в него изменений при ведении федерального государственного реестра документов стратегического планирования Федеральной информационной системы стратегического планирования (ФИС СП)”. Результаты НИР внедрены в системе ГАС “Управление”.

Архитектура искусственной нейронной сети и алгоритм нечеткой кластеризации коротких текстов, методы расширения словаря языковой модели и корректировки весов нейронной сети для учета обратной связи эксперта в интерактивной кластеризации, а также программная реализация метода нечеткой интерактивной кластеризации на языке Python внедрены в системе Планета.Аналитика 4.0 (включена в реестр отечественного ПО) компании ООО “ИБС “Экспертиза”.

Апробация работы. Основные положения и результаты диссертационной работы доложены и обсуждены на конференциях и конгрессах:

- Всероссийская научно-практическая конференция “Нечеткие системы и мягкие вычисления” (Санкт-Петербург, 2017);
- Международная конференция “Интеллектуальные информационные технологии в технике и на производстве” IITI (Варна, 2017; Сочи, 2018; Острава, 2019);
- Всероссийская научная конференции «Нечеткая логика и мягкие вычисления в промышленности» (Ульяновск, 2017, 2018, 2019);
- Национальная Конференция по Искусственному Интеллекту (Москва, 2018; Ульяновск, 2019);

- Международная конференция “World Conference on Soft Computing” (Баку, 2018);
- Международная конференция “Mexican International Conference on Artificial Intelligence” (Гвадалахара, 2018);
- Международная конференция “European Society for Fuzzy Logic and Technology” (Прага, 2019).
- Международная конференция по компьютерной лингвистике и интеллектуальным технологиям “Диалог” (Москва, 2019).
- I Национальный конгресс по когнитивным исследованиям, искусственному интеллекту и нейроинформатике (Москва, 2020).

Публикации по теме диссертации. Основные результаты диссертационного исследования опубликованы в 19 печатных работах, в том числе 6 статей в российских рецензируемых научных журналах из перечня, рекомендованного ВАК РФ, 7 публикаций в изданиях, индексируемых в Scopus и Web of Science, 6 – в материалах научных конференций.

Сведения о личном вкладе автора. Постановка задач исследования осуществлялась совместно с научным руководителем. Все основные теоретические и практические исследования диссертационной работы проведены лично автором. Подготовка к публикации некоторых результатов проводилась совместно с соавторами, вклад соискателя был определяющим.

Структура и объем работы. Диссертация изложена на 136 страницах машинописного текста, содержит 46 рисунка, 9 таблиц, состоит из введения, четырех глав, заключения, списка использованной литературы из 128 наименований на 15 страницах и 3 приложений на 6 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы диссертационного исследования, сформулированы цель и задачи работы, отражена научная новизна, практическая значимость, достоверность и обоснованность результатов исследований диссертации, приведены основные положения, выносимые на защиту, указана степень апробации и реализации результатов исследования, кратко раскрыто содержание глав диссертации.

Первая глава посвящена сравнительному анализу моделей и методов нечеткой кластеризации коротких текстов и формулировке цели и задач исследования.

Проведена систематизация методов кластеризации с участием исследователя, отмечено, что такие методы относятся к большому семейству методов кластеризации с привлечением учителя (semi-supervised clustering). Среди прочих выделены методы интерактивной кластеризации, особенно актуальные при обработке текстовой информации. Без участия эксперта, без выявления его скрытых интенций невозможно заранее определить, какое именно разбиение ожидается в результате работы алгоритма. Помимо очевидной группировки по тематике, тексты могут быть сгруппированы по лицу, от которого ведется повествование, по целевой аудитории текста, по правовому статусу текста или комбинации различных признаков. При этом участие эксперта не подразумевает сведение кластеризации к методам обучения с учителем. Доля объектов, с которыми взаимодействует эксперт, не превышает 10%.

Показано, что для получения качественного результата работы алгоритма кластеризации требуется включение эксперта в процесс кластеризации как органичной части алгоритма кластеризации. При этом желательно, чтобы от эксперта не требовалось понимания внутренних деталей работы алгоритма и причинно-следственная связь между действиями эксперта и результатами работы алгоритма была явной. Наилучшим образом в таком случае подходят методы кластеризации с обратной связью в виде оценки результата кластеризации.

Задача кластеризации коротких текстов рассматривается отдельно по причине того, что ряд особенностей коротких текстов не позволяют без специальной адаптации использовать традиционные методы кластеризации текстов. К таким особенностям относятся: разреженность векторов признаков; полисемия; синонимия; более частое по сравнению с обычными текстами использование аббревиатур, сленговых слов, неологизмов; проблема опечаток, грамматические и пунктуационные ошибки; частичное или полное отсутствие контекста.

Отмечено, что наибольших успехов в области обработки текстов на сегодня удается добиться при использовании языковых моделей на основе нейронных сетей. Задача языкового моделирования в узком смысле — спрогнозировать следующее слово в тексте, имея наблюдения по предшествующим словам. При этом отмечается, что короткие тексты не позволяют провести качественно обучение языковой модели, особенно, если короткие тексты относятся к специфичному домену. Решить эту проблему

призваны получившие в последнее время широкое распространение предварительно обученные нейронные сети. Они позволяют осуществить перенос знаний (“transfer learning”) одной нейронной сети в другую, решающую иную задачу. В диссертационной работы было проведено сравнение наиболее распространенных предварительно обученных моделей (ULMFit (русский и английский языки); ELMo (мультиязычная) и RuBERT (русскоязычная адаптация модели BERT от корпорации Google)) и обосновано использование ULMFit в качестве базовой модели для предлагаемой архитектуры нейронной сети.

Таким образом, целью настоящего исследования является построение эффективного алгоритма нечеткой интерактивной кластеризации коротких текстов на базе искусственной нейронной сети с современной архитектурой, включающей в себя предварительно обученную языковую модель, как блок, отвечающий за преобразование текста в сжатое векторное представление. Предлагаемый к разработке алгоритм должен обеспечивать возможность учета обратной связи от эксперта в форме оценки результатов кластеризации. В целях проверки работоспособности и оценки эффективности алгоритма необходимо разработать программный модуль для системы поддержки принятия решений с целью автоматизации процесса нечеткой интерактивной кластеризации коротких текстов.

Во второй главе разработаны: архитектура искусственной нейронной сети, позволяющая решить задачу кластеризации коротких текстов на базе языковой модели; метод расширения словаря языковой модели на базе нейронной сети с использованием иерархического классификатора и метод корректировки весов нейронной сети для обработки ограничений, задаваемых экспертом при решении задачи кластеризации.

В результате проведенных исследований разработана модель, представленная в виде искусственной нейронной сети, позволяющая решать задачу кластеризации коротких текстов, используя сжатые векторные представления, получаемые с помощью кодера языковой модели. Архитектурно нейронная сеть состоит из двух последовательных блоков (рис. 1): кодер языковой модели и блок кластеризации. Объединение языковой модели и блока кластеризации в рамках единой нейронной сети позволяет производить тонкую настройку весов слоев нейронной сети, отвечающих за языковую модель непосредственно в ходе процесса кластеризации. Благодаря этому повышается качество формируемых сжатых векторных

представлений обрабатываемых коротких текстов и, как следствие, улучшается результат кластеризации.

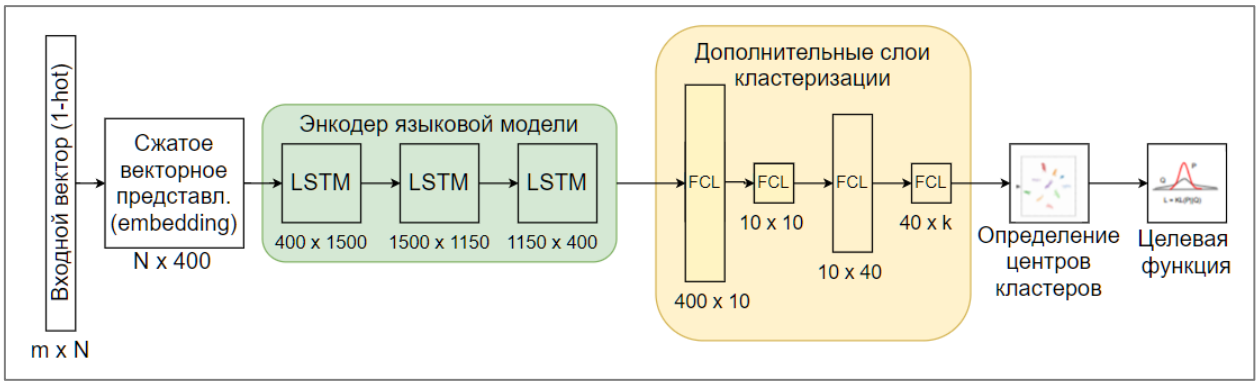

Рис. 1. Архитектура нейронной сети для кластеризации коротких текстов

В ходе исследований предложен метод расширения словаря языковой модели за счет добавления промежуточного слоя между входом нейронной сети и первым слоем нейронной сети, реализующей языковую модель.

Узким местом применения предварительно обученных языковых моделей является словарь, потому как словари исходной языковой модели и исследуемого набора данных могут существенно отличаться друг от друга. При этом существующие методы сопоставляют словам, не входящим в исходный словарь языковой модели (неизвестным, незнакомым), либо нулевой вектор, либо вектор, равный среднему всех векторов словаря.

Для решения задачи расширения словаря языковой модели в нейронную сеть добавляется дополнительный слой (рис. 2.), в котором происходит сопоставление “неизвестным” словам (словам, входящим в словарь исследуемого набора, но не входящим в словарь языковой модели) вектора, являющегося линейной комбинацией наиболее семантически близких “известных” слов (слов, входящих в словарь языковой модели).

$$новое\_слово = вес_1 \times слово_1 + вес_2 \times слово_2 + \ldots + вес_N \times слово_N$$

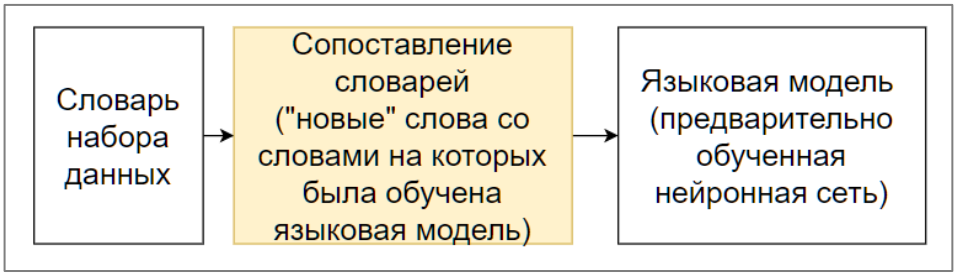

Рис. 2. Схема расширения словаря языковой модели

Таким образом, задача расширения словаря языковой модели свелась к задаче определения метода расчета весовых коэффициентов в линейной комбинации. В данной работе предлагается выполнять сопоставление в два этапа. На первом этапе для множества семантически близких слов к

целевому (новому, незнакомому) слову строится нечеткий граф (веса на ребрах равны семантической близости между словами). Далее при помощи иерархической модификации алгоритма ε-кластеризации нечеткого графа проводится построение иерархического классификатора, позволяющего выделить и классифицировать смысловые оттенки слова. На рисунке 3 представлен граф и классификатор для слова “график”. Слева представлены все слова в словаре нейронной сети, семантически близкие к слову “график” по модели Word2Vec, обученной на корпусе “Тайга”. Справа представлен результат работы алгоритма построения иерархического классификатора.

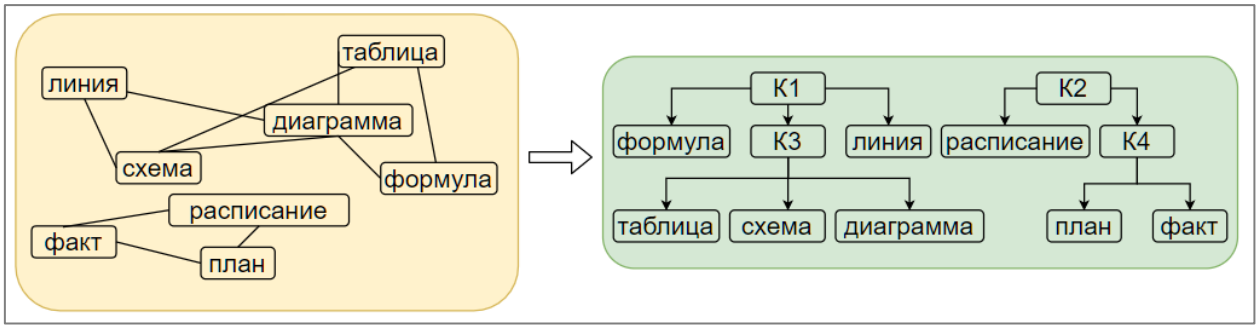

Рис. 3. Преобразование нечеткого графа в иерархический классификатор

На втором этапе при помощи предложенного в диссертационной работе алгоритма производится расчет весов для линейной комбинации векторов слов для “неизвестного” слова по построенному для него иерархическому классификатору. Пример работы алгоритма представлен на рис. 4 для слова “график”. Таким образом, все слова целевого словаря получают в соответствие векторы, которые могут быть использованы в предобученной языковой модели.

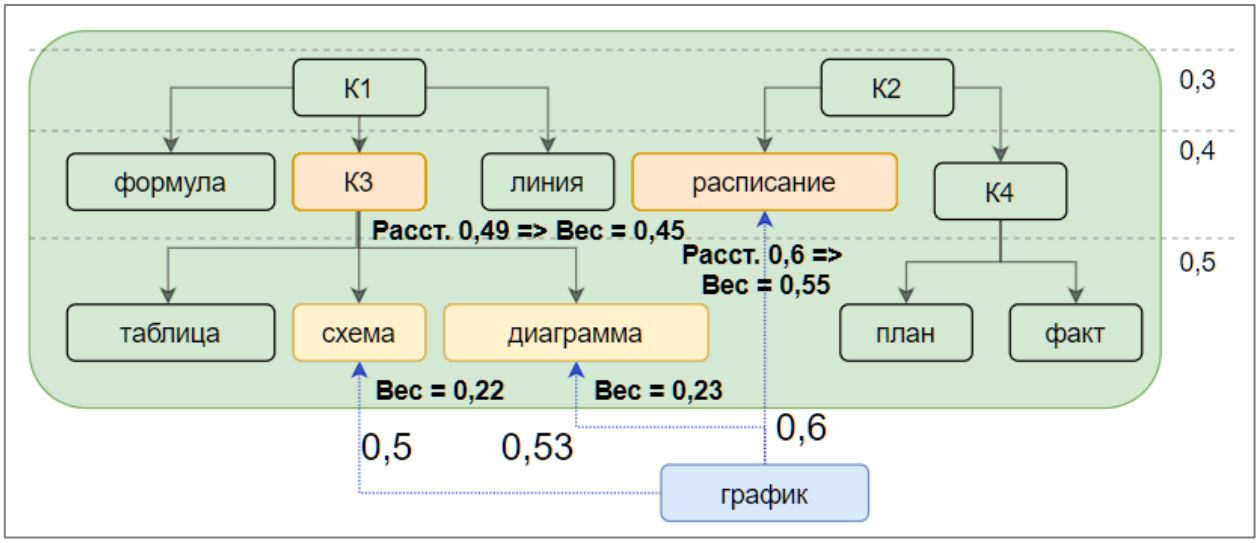

Рис. 4. Пример работы алгоритма расчета весовых коэффициентов для векторов слов

В третьем разделе второй главы представлен метод корректировки весов нейронной сети, позволяющий учитывать ограничения, задаваемые экспертом при решении задачи кластеризации.

В предложенной искусственной нейронной сети процессы определения оптимального расположения центров кластеров и построения пространства признаков происходят одновременно за счет определения общей функции потерь. Для этого в качестве меры расстояния между элементом и центром кластера используется метрика, основанная на распределении Стьюдента с одной степенью свободы (метод кластеризации DEC).

$$q_{ij} = \frac{\left(1 + \left|\left|z_i - \mu_j\right|\right|^2\right)^{-1}}{\sum_{l=0}^{k}\left(1 + \left|\left|z_i - \mu_l\right|\right|^2\right)^{-1}}). \quad (1)$$

Целевая функция (функция потерь, loss function), или штрафная функция, строится как метрика Кульбака-Лейблера (Kullback-Leibler divergence) между фактическим и целевым распределением.

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (2)$$

В качестве целевого распределения используется следующее распределение:

$$p_{ij} = \frac{{q_{ij}}^2/f_j}{\sum_{l=0}^{k} {q_{il}}^2/f_l}, \qquad f_j = \sum_j q_{ij}. \quad (3)$$

Инициализация весов в блоке кластеризации нейронной сети происходит для каждого слоя последовательно. Для этого используется стандартная техника послойного обучения, в которой каждый слой рассматривается как автоэнкодер (тождественная функция). Для инициализации центров кластеров используется алгоритм k-means с евклидовым расстоянием.

Для учета обратной связи большинство методов кластеризации предлагают эксперту манипулировать метапараметрами алгоритма кластеризации, которые в большинстве своем не имеют явной связи с конкретными парами “объект – центр кластера”. В данной работе предполагается обратная связь двух видов: “элемент X_i должен принадлежать кластеру C_j” и “элементу X_i не следует находиться в кластере

X_j”. Такой вид обратной связи легко формулируется экспертами любого уровня подготовки и не требует знаний технических и математических деталей конкретного метода кластеризации. Первый тип ограничений соответствует ограничениям, используемым в алгоритмах классификации, а также часто применяется в интерактивной классификации. Второй тип ограничений предложен впервые. Особенностью данного типа обратной связи является возможность формулировки менее жестких ограничений. Такие ограничения снижают трудоемкость работы эксперта и требования к уровню экспертизы. Одновременно может быть получено произвольное количество таких ограничений.

Метод, разработанный в данном исследовании для включения обратной связи от эксперта в процесс обучения нейронной сети, предлагает модификацию целевой функции нейронной сети через добавление в нее дополнительного множителя.

$$L = KL(P||Q) = \sum_i \sum_j t_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (4)$$

При корректировке весов нейронной сети методом обратного распространения ошибки используются следующие формулы для расчета градиентов целевой функции:

$$\frac{\partial L}{\partial z_i} = 2 \sum_j \left(1 + \left|\left|z_i - \mu_j\right|\right|^2\right)^{-1} * \left|p_{ij} - q_{ij}\right| * t_{ij} * \left(z_i - \mu_j\right), \quad (5)$$

$$\frac{\partial L}{\partial \mu_j} = -2 \sum_i \left(1 + \left|\left|z_i - \mu_j\right|\right|^2\right)^{-1} * \left|p_{ij} - q_{ij}\right| * t_{ij} * \left(z_i - \mu_j\right), \quad (6)$$

где T = $\{t_{ij}\}$ – матрица обратной связи (подсказок пользователя, tips), в которой:

$$t_{ij} = \begin{cases} > 0, & \text{для включения элемента } i \text{ в кластер } j \\ < 0, & \text{для исключения элемента } i \text{ из кластера } j \\ 1, & \text{иначе.} \end{cases}$$

Абсолютное значение t_{ij} определяет скорость, с которой элементы и центры кластера будут стремиться друг к другу или отталкиваться друг от друга.

Важно подчеркнуть, что одним из основных достоинств предлагаемого метода кластеризации является то, что добавляемые ограничения не являются жесткими, не приводят к необходимости решать системы уравнений, которые могут потенциально оказаться несовместными. Любые ограничения пользователя, вне зависимости от их внутренних противоречий будут учтены в штрафах со стороны целевой функции.

В третьей главе дается описание федеральной информационной системы “Стратегическое Планирование” (ФИС СП), которая является неотъемлемой частью системы стратегического планирования РФ и реализована в рамках государственной автоматизированной системы “Управление” (ГАСУ). В поступающих документах участники системы стратегического планирования самостоятельно проставляют классификацию ключевых показателей эффективности из краткого классификатора ЕМИСС (Единой межведомственной информационно-статистической системы), содержащего 11 классов (тематик). В этом процессе имеется два основных недостатка: малое число классов, не отражающее всего многообразия вводимых показателей и не пригодное для дальнейшего анализа, а также большое число случаев ошибочной классификации, вызванное тем, что классификацию проводят не эксперты. Дополнительным недостатком процесса является длительность этапа проверки и корректировки ошибочной классификации показателей.

Предложенный в данной работе метод нечеткой интерактивной кластеризации коротких текстов позволил усовершенствовать архитектуру ФИС СП. В результате проведенной интерактивной кластеризации экспертами методологами был обучен классификатор, который используется для автоматизации классификации показателей стратегического планирования и содержит несколько сотен классов. Помимо более точного классификатора, снизилась нагрузка на экспертов, проводящих проверку входящих документов. На рис. 5 представлена архитектура до и после проведенных изменений.

Для автоматизации предложенного метода кластеризации был составлен алгоритм нечеткой интерактивной кластеризации коротких текстов (НИККТ), состоящий из следующих шагов:

1) предобработка текстов: исправление орфографии, токенизация;
2) расширение словаря языковой модели;
3) тонкая настройка языковой модели (дополнительное обучение);

4) инициализация слоев нейронной сети блока кластеризации (обучение автоэнкодера);

5) инициализация центров кластеров (с использованием k-means);

6) первичная кластеризация (синхронное обучение нейронной сети кластеризации коротких текстов и определение центров кластеров);

7) шаги интерактивной кластеризации (цикл до достижения приемлемого для эксперта качества кластеризации):

a. анализ результатов кластеризации экспертом;

b. получение матрицы обратной связи от эксперта;

c. корректировка весов нейронной сети и корректировка центров кластеров на основании матрицы множителей обратной связи.

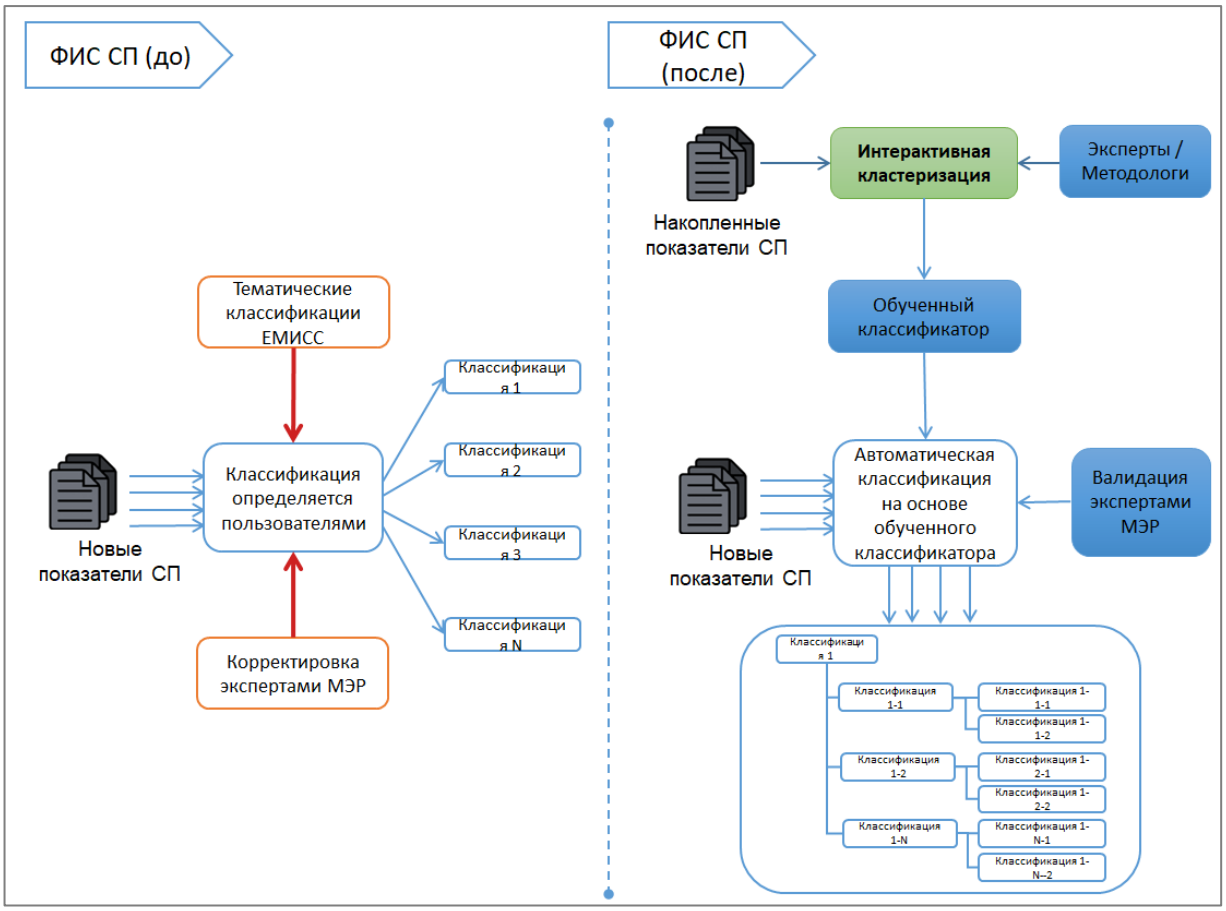

Рис. 5. Архитектура ФИС СП до и после проведенных изменений

Данный алгоритм был реализован в разработанных программных модулях. В качестве инструмента разработки был выбран язык Python, позволяющий реализовать алгоритмическую часть с помощью фреймворка FastAI. На рис. 6 представлена архитектура программного модуля. Алгоритмическая часть метода кластеризации находится в блоке “Машинное обучение”, и ее схема представлена на рис. 7.

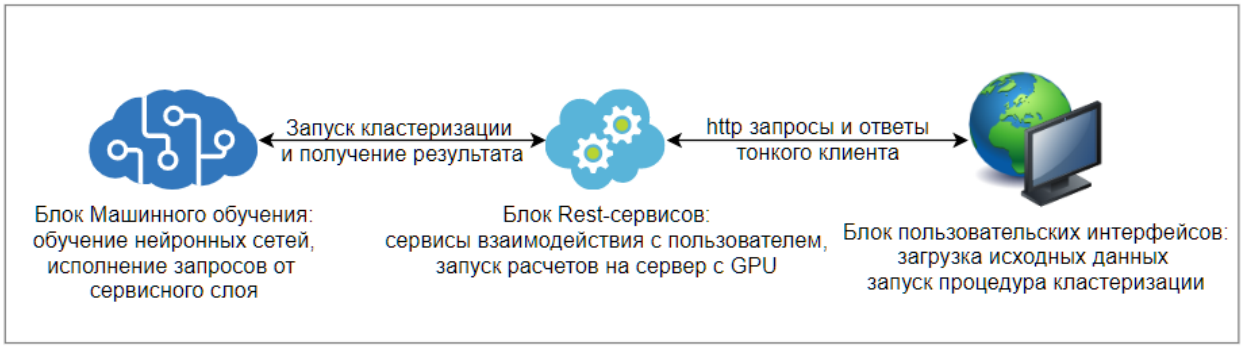

Рис. 6. Модульная архитектура программного комплекса для кластеризации коротких текстов

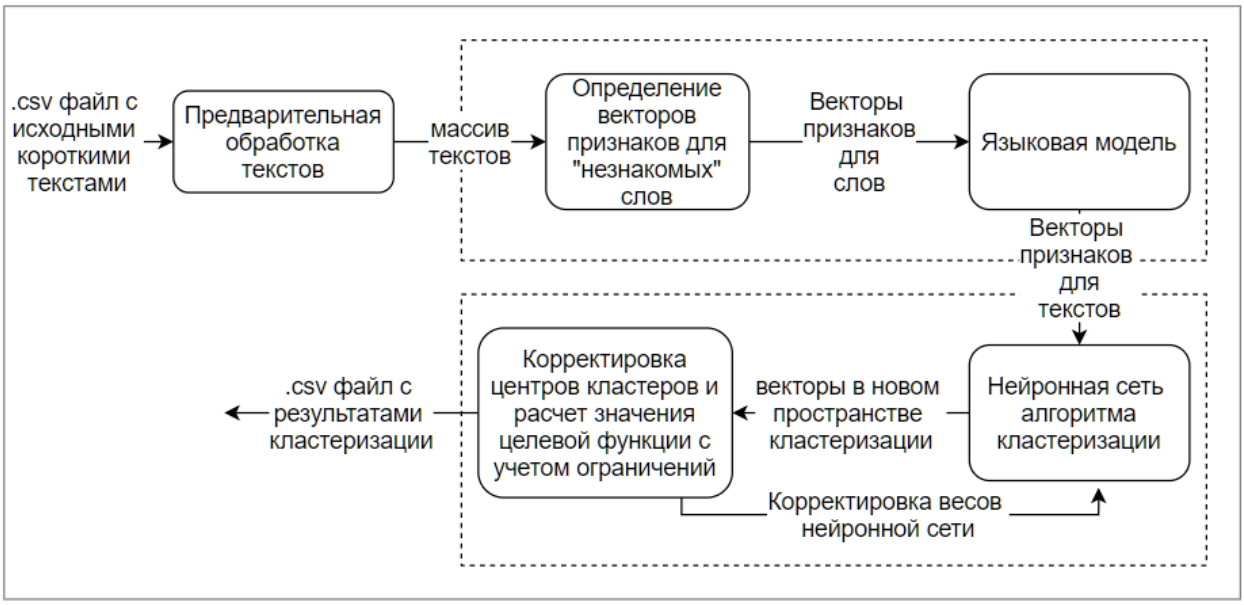

Рис. 7. Схема модуля “Машинное обучение”

На разработанный программный модуль, использованный для проведения экспериментов, получено свидетельство о государственной регистрации программы для ЭВМ № 2021615642.

В четвертой главе разработана программа для проведения испытаний по нечеткой интерактивной кластеризации коротких текстов, представлены результаты ряда экспериментов, подтверждающие работоспособность и эффективность предлагаемого метода. В первой группе экспериментов показана работа на сгенерированном наборе данных из простых векторов. Эксперименты наглядно демонстрируют процесс уточнения своих намерений экспертом и то, как кластеризация подстраивается под полученные ограничения.

В следующем разделе эксперимент проводится на каноническом наборе данных “Ирисы Фишера” для демонстрации положительного влияния на точность кластеризации дополнительной информации, не содержащейся в данных. Кластеры двух классов – ‘versicolor’ и ‘virginica’ –

разделены не достаточно четко, и точность составляет 95%. При добавлении двух ограничений количество неверно соотнесенных объектов уменьшилось с 26 до 10, и точность составила 98,3%.

В третьем разделе четвертой главы проводится эксперимент по кластеризации набора данных от компании Avito для соревнования по классификации рекламных объявлений. Все объявления в наборе разбиты на 4 категории: “Бытовая электроника”, “Для дома и дачи”, “Личные вещи”, “Хобби и отдых”. Количество объявлений: 489 000. Выполнение шага расширения словаря позволяет достичь 50-52% точности языковой модели. И первичная кластеризация, выполненная после шагов инициализации весов нейронной сети в блоке кластеризации и инициализации центров кластеров, в среднем (по 10 прогонам) позволяет достичь уровня точности 52,3%

В фазе интерактивной кластеризации эксперт добавлял по 10 ограничений на каждой итерации. График на рис. 10 показывает, что с ростом количества ограничений точность повышается. При добавлении количества ограничений более 5% от общего числа примеров график точности достигает своего максимума. Проведено сравнение с результатами схожего эксперимента на англоязычном корпусе Reuters RCV1-v2/LYRL2004, отмечены преимущества предлагаемого подхода.

В ходе экспериментальных исследований установлено положительное влияние дополнительного обучения слоев языковой модели на точность кластеризации. Процедура дополнительного обучения позволяет увеличить точность на 10%, понижая при этом количество необходимых ограничений.

Дополнительно на графике на рис. 10 приведены результаты работы алгоритмов классификации победителя соревнований (88%) и алгоритма классификации, настроенного с использованием языковой модели, обученной в настоящем исследовании (91%). Также на рис. 8 показаны результаты кластеризации сжатых векторов признаков, полученных с помощью обученной языковой модели стандартными алгоритмами кластеризации: k-means, DBScan, BIRCH. Графики демонстрируют значительное преимущество применения интерактивной кластеризации.

В целях установления границ применимости предлагаемого алгоритма в наборе данных Avito были выбраны группы текстов с различными размерами: 5-20 слов, 20-50 слов, 50-100 слов, 100-200 слов, 200-300 слов, 300-500 слов, 500-1000 слов. Рис. 9 с результатами эксперимента показывает обратную зависимость точности кластеризации от размеров текста (замер

проводился после первого прогона без добавления экспертных ограничений).

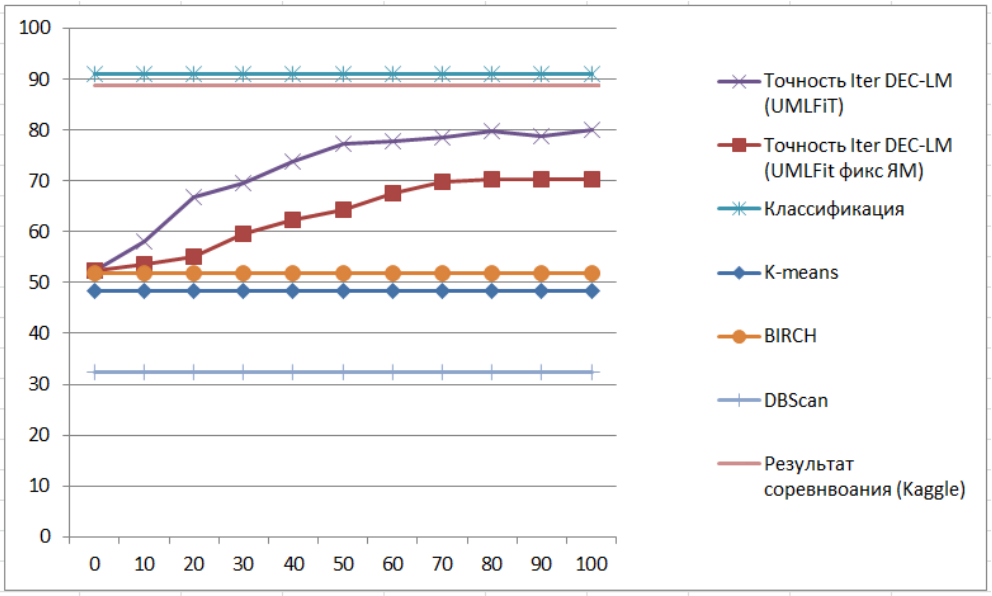

Рис. 8. Изменение качества с ростом количества ограничений от эксперта

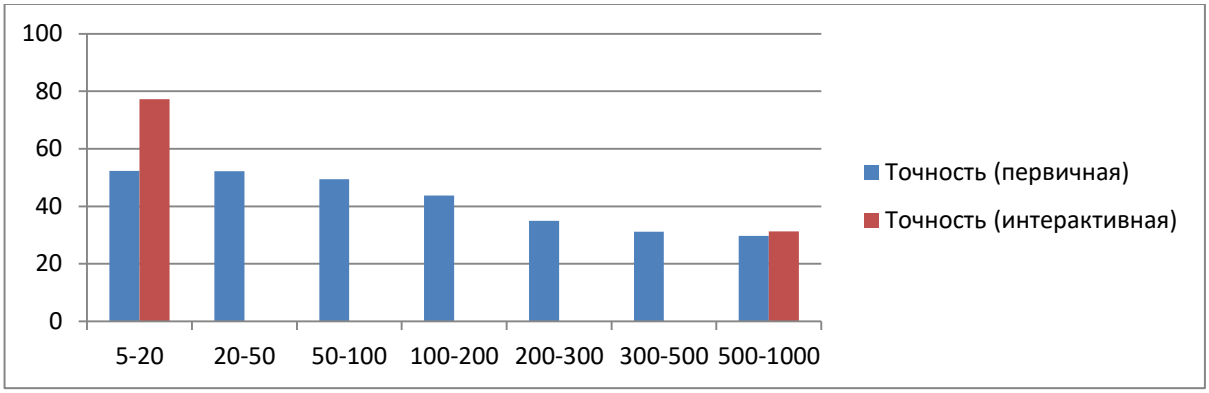

Рис. 9. График зависимости точности кластеризации от длины коротких текстов

В четвертой части приводятся результаты решения задачи кластеризации коротких текстов, содержащих ключевые показатели эффективности (КПЭ, KPI) системы стратегического планирования Российской Федерации, закрепленной федеральным законом № 172-ФЗ от 28.07.2014 г. (изм. от 22.05.2019 N 641,от 18.11.2019 N 1468). Исследование проводилось в рамках НИР в интересах Министерства экономического развития РФ. Общее количество документов – около 600 000. По 11-ти нормативно установленным категориям предлагалось построить классификатор в рамках каждой категории на базе результатов кластеризации. Для финального эксперимента была выбрана тематика “Образование (общее образование)”. Примеры показателей из документов, отнесенных к данной тематике, приведены на рис. 12. Общий объем набора данных для кластеризации составил 6 975 показателей. Все 6 975 показателей были кластеризованы предложенной моделью. Далее эксперт точечно вносил корректировки в

кластеры в виде дополнительных параметров кластеризации и запускал очередную итерацию алгоритма кластеризации. В частности, при знакомстве с содержимым кластера №14, который при первичной кластеризации был назван как “Аттестация”, эксперт отметил две принципиально разных группы показателей: аттестация преподавателей и аттестация учащихся. Примеры показателей приведены на рис. 10.

ДОЛЯ ПЕДАГОГИЧЕСКИХ РАБОТНИКОВ ОБЩЕОБРАЗОВАТЕЛЬНЫХ УЧРЕЖДЕНИЙ КОТОРЫМ ПРИ ПРОХОЖДЕНИИ АТТЕСТАЦИИ ПРИСВОЕНА ПЕРВАЯ И ВЫСШАЯ КАТЕГОРИИ
ДОЛЯ ПЕДАГОГИЧЕСКИХ РАБОТНИКОВ ОБЩЕОБРАЗОВАТЕЛЬНЫХ ОРГАНИЗАЦИЙ КОТОРЫМ ПРИ ПРОХОЖДЕНИИ АТТЕСТАЦИИ ПРИСВОЕНА ПЕРВАЯ ИЛИ ВЫСШАЯ КВАЛИФИКАЦИОННАЯ КАТЕГОРИЯ
ДОЛЯ ПЕДАГОГИЧЕСКИХ РАБОТНИКОВ ПРОШЕДШИХ АТТЕСТАЦИЮ С ПРИСВОЕНИЕМ ПЕРВОЙ КВАЛИФИКАЦИОННОЙ КАТЕГОРИИ ОТ ОБЩЕГО ЧИСЛА ПЕДАГОГИЧЕСКИХ РАБОТНИКОВ ОБЩЕОБРАЗОВАТЕЛЬНЫХ УЧРЕЖДЕНИЙ
СОХРАНЕНИЕ ДОЛИ ПЕДАГОГИЧЕСКИХ РАБОТНИКОВ ОБЩЕОБРАЗОВАТЕЛЬНЫХ ОРГАНИЗАЦИЙ КОТОРЫМ ПРИ ПРОХОЖДЕНИИ АТТЕСТАЦИИ ПРИСВОЕНА ПЕРВАЯ ИЛИ ВЫСШАЯ КАТЕГОРИЯ ЕЖЕГОДНО
ДОЛЯ ОБУЧАЮЩИХСЯ КЛАССОВ НЕ ПРОШЕДШИХ ГОСУДАРСТВЕННУЮ ИТОГОВУЮ АТТЕСТАЦИЮ В ФОРМЕ ЕГЭ В ОБЩЕЙ ЧИСЛЕННОСТИ ОБУЧАЮЩИХСЯ КЛАССОВ
ДОЛЯ ОБУЧАЮЩИХСЯ КЛАССОВ НЕ ПРОШЕДШИХ ГОСУДАРСТВЕННУЮ ИТОГОВУЮ АТТЕСТАЦИЮ В ФОРМЕ В ОБЩЕЙ ЧИСЛЕННОСТИ ОБУЧАЮЩИХСЯ КЛАССОВ МУНИЦИПАЛЬНЫХ ОБЩЕОБРАЗОВАТЕЛЬНЫХ ОРГАНИЗАЦИЙ
ДОЛЯ ОБУЧАЮЩИХСЯ КЛАССОВ НЕ ПРОШЕДШИХ ГОСУДАРСТВЕННУЮ ИТОГОВУЮ АТТЕСТАЦИЮ В ОБЩЕЙ ЧИСЛЕННОСТИ ОБУЧАЮЩИХСЯ КЛАССОВ МУНИЦИПАЛЬНЫХ ОБЩЕОБРАЗОВАТЕЛЬНЫХ

Рис. 10. Примеры КПЭ кластера №14 (“Аттестация”)

Для изменения кластеризации первые три показателя из нового класса “Аттестация персонала” помечены для вынесения из состава кластера. В результате образован новый кластер для аттестации персонала, а кластер №14 переименован в “Аттестацию учащихся”.

Аналогичным образом проведена кластеризация всех остальных тематик. В ходе экспериментов экспериментально подтверждено, что предлагаемый метод позволяет сократить временные затраты экспертов на этапе кластеризации и построения классификатора в 3 раза по сравнению с полностью ручной кластеризацией и в 2 раза – в случае использования методов тематической кластеризации. В абсолютных значениях экономия составила более 180 человеко-дней высококвалифицированного специалиста. Кроме того, установлено, что благодаря высокой точности полученного классификатора в 2-3 раза сократилось время, затрачиваемое экспертами на проверку корректности классификации ключевых показателей эффективности во входящих документах.

Таким образом, проведенные экспериментальные исследования подтвердили целесообразность и эффективность использования предложенных модели, методов и алгоритма нечеткой интерактивной кластеризации коротких текстов в рамках СППР “Федеральная система стратегического планирования РФ”.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Проведено исследование моделей и методов машинного обучения для обработки текстов и разработана архитектура искусственной нейронной сети, реализующей кластеризацию на базе пространства признаков языковой модели русского языка.

2. Разработан метод обработки текстов для расширения словаря языковой модели на базе нейронной сети с использованием нечеткого иерархического классификатора, который позволяет повысить точность кластеризации в среднем на 10%.

3. Разработан метод для обработки обратной связи от эксперта, используемый для корректировки весовых коэффициентов нейронной сети, что позволяет проводить интерактивную кластеризацию наборов коротких текстов.

4. Сформулирован перечень этапов программы проведения испытаний метода нечеткой интерактивной кластеризации коротких текстов. Проведенные исследования по данной программе позволили установить границы применения предлагаемого метода. Метод наиболее эффективен для текстов с количеством слов от 10 до 100. В ходе проведенных исследований была достигнута средняя точность кластеризации 80% при более низком числе дополнительных ограничений по сравнению с аналогичными методами. На разработанный программный модуль, использованный для проведения численных экспериментов, получено свидетельство о государственной регистрации программы для ЭВМ № 2021615642.

5. Составлен алгоритм автоматизации работ по нечеткой интерактивной кластеризации коротких текстов в СППР “Федеральная система стратегического планирования РФ (ФИС СП)”.

6. Проведена апробация разработанных модели, методов и алгоритма нечеткой интерактивной кластеризации коротких текстов в качестве элементов СППР “ФИС СП”. Внедрение алгоритма позволило решить задачу составления автоматического классификатора ключевых показателей эффективности документов стратегического планирования, при этом экономия трудозатрат эксперта оценивается более чем в 9 человеко-месяцев. Наличие автоматического классификатора позволяет снизить временные затраты экспертов при проведении процедуры проверки корректности классификации входящих (новых) документов.

7. Разработанный метод интерактивной кластеризации является универсальным и может быть применен для различных наборов коротких текстов. Предложенный метод может быть доработан для совместного использования с различными языковыми моделями и обобщен на случай совместной работы ряда экспертов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Основное содержание диссертации опубликовано в следующих работах: в российских рецензируемых научных журналах из Перечня, рекомендованного ВАК РФ по специальности:

1. Дударин П.В. МЕТОДИКА И АЛГОРИТМ КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ ЭКОНОМИЧЕСКОЙ АНАЛИТИКИ / Дударин П.В., Пинков А.П., Ярушкина Н.Г. // Автоматизация процессов управления. 2017. № 1 (47). С. 85-93.

2. Дударин П.В. АЛГОРИТМ ПОСТРОЕНИЯ ИЕРАРХИЧЕСКОГО КЛАССИФИКАТОРА КОРОТКИХ ТЕКСТОВЫХ ФРАГМЕНТОВ НА ОСНОВЕ КЛАСТЕРИЗАЦИИ НЕЧЕТКОГО ГРАФА / Дударин П.В., Ярушкина Н.Г. // Радиотехника. 2017. № 6. С. 114-121.

3. Дударин П.В. ФОРМИРОВАНИЕ ПРИЗНАКОВ ИЗ ИЕРАРХИЧЕСКОГО КЛАССИФИКАТОРА ДЛЯ КЛАСТЕРИЗАЦИИ КОРОТКИХ ТЕКСТОВЫХ ФРАГМЕНТОВ / Дударин П.В., Ярушкина Н.Г. // Нечеткие системы и мягкие вычисления. 2017. Т. 12. № 2. С. 87-96.

4. Дударин П.В. ПОДХОД К ТРАНСФОРМАЦИИ КЛАСТЕРНОГО ДЕРЕВА ПРИЗНАКОВ В ВЕКТОРНОЕ ПРОСТРАНСТВО ПРИЗНАКОВ. / Дударин П.В., Ярушкина Н.Г. // Радиотехника. 2018. № 6. С. 63-73.

5. Дударин П.В. ПОДХОД К ОЦЕНКЕ ТРУДОЕМКОСТИ ЗАДАЧ В ПРОЦЕССЕ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ. / Дударин П.В., Тронин В.Г., Святов К.В., Белов В.А., Шакуров Р.А. // Автоматизация процессов управления. 2019. № 3 (57). С. 65-72.

6. Дударин П.В. ПОДХОД К ОБРАБОТКЕ ОБРАТНОЙ СВЯЗИ ПОЛЬЗОВАТЕЛЯ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА РЕЗУЛЬТАТОВ РАБОТЫ АЛГОРИТМА КЛАСТЕРИЗАЦИИ. / Дударин П.В., Тронин В.Г., Ярушкина Н.Г. // Известия Самарского Научного Центра Российской Академии Наук. 2020. № 5 (97). С. 94-105.

в зарубежных рецензируемых журналах:

7. Дударин П.В. AN APPROACH TO FUZZY HIERARCHICAL CLUSTERING OF SHORT TEXT FRAGMENTS BASED ON FUZZY GRAPH CLUSTERING / Dudarin P.V., Yarushkina N.G. / Advances in Intelligent Systems and Computing. 2018. Т. 679. С. 295-304.

8. Дударин П.В. AN APPROACH TO FEATURE SPACE CONSTRUCTION FROM CLUSTERING FEATURE TREE / Dudarin P., Samokhvalov M., Yarushkina N. // Kuznetsov S., Osipov G., Stefanuk V. (eds) Artificial Intelligence. RCAI 2018. Communications in Computer and Information Science. 2018. Т. 934. С. 176-189.

9. Дударин П.В. AN APPROACH TO CUSTOMIZATION OF PRE-TRAINED NEURAL NETWORK LANGUAGE MODEL TO SPECIFIC DOMAIN / Dudarin P. V., Tronin V. G., Svyatov K. V. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. 2019. С. 1-9.

10. Дударин П.В. A TECHNIQUE TO PRE-TRAINED NEURAL NETWORK LANGUAGE MODEL CUSTOMIZATION TO SOFTWARE DEVELOPMENT DOMAIN / Dudarin P. V., Tronin V. G., Svyatov K. V. // Kuznetsov S., Panov A. (eds) Artificial Intelligence. RCAI 2019. Communications in Computer and Information Science. 2019. Т. 1093. С. 169-176.

11. Дударин П.В. LABOR INTENSITY EVALUATION TECHNIQUE IN SOFTWARE DEVELOPMENT PROCESS BASED ON NEURAL NETWORKS / Dudarin P. V., Tronin V.G., Svatov K. V., Belov V.A., Shakurov R.A. // Kovalev S., Tarassov V., Snasel V., Sukhanov A. Proceedings of the Fourth International Scientific Conference “Intelligent Information Technologies for Industry”. 2020. С. 75-84.

12. Дударин П.В. AN APPROACH TO USER FEEDBACK PROCESSING IN ORDER TO INCREASE CLUSTERING RESULTS QUALITY / Dudarin P., Yarushina N. // Selected Contributions of the "Russian Advances in Artificial Intelligence" Track at RCAI 2020 co-located with 18th Russian Conference on Artificial Intelligence. 2020. С. 48-58.

13. Дударин П.В. TWO PHASE APPROACH TO DETECTION OF SOFTWARE PROJECTS WITH SIMILAR ARCHITECTURE BASED ON CLUSTERING AND ONTOLOGICAL METHODS / Yarushina N., Guskov G., Dudarin P. // Recent Developments and the New Direction in Soft-Computing Foundations and Applications. 2021. С. 131-145.

в материалах научных семинаров и конференций:

14. Дударин П.В. ПОСТРОЕНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ НАБОРА КОРОТКИХ ТЕКСТОВЫХ ФРАГМЕНТОВ ПРИ ПОМОЩИ КЛАСТЕРИЗАЦИИ НЕЧЕТКОГО ГРАФА / Дударин П.В., Ярушкина Н.Г. // В сборнике: Нечеткие системы, мягкие вычисления и интеллектуальные технология (НСМВИТ-2017) труды VII всероссийской научно-практической конференции. 2017. С. 26-34.

15. Дударин П.В. ПОДХОДЫ К НЕЧЕТКОЙ И ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ КЛЮЧЕВЫХ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ СИСТЕМЫ СТРАТЕГИЧЕСКОГО ПЛАНИРОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ / Дударин П.В., Ярушкина Н.Г. // В сборнике: Нечеткие системы, мягкие вычисления и интеллектуальные технология (НСМВИТ-2017) труды VII всероссийской научно-практической конференции. 2017.С.65-74.

16. Дударин П.В. ПОДХОД К КЛАСТЕРИЗАЦИИ КОРОТКИХ ТЕКСТОВЫХ ФРАГМЕНТОВ ПО ИЕРАРХИЧЕСКОМУ КЛАССИФИКАТОРУ / Дударин П.В., Ярушкина Н.Г. // В сборнике: Нечеткие системы и мягкие вычисления. Промышленные применения материалы Первой всероссийской научно-практической конференции. 2017. С. 367-375.

17. Дударин П.В. AN APPROACH TO SIMILAR SOFTWARE PROJECTS SEARCHING AND ARCHITECTURE ANALYSIS BASED ON ARTIFICIAL INTELLIGENCE / Yarushina N., Guskov G., Dudarin P. // Advances in Intelligent Systems and Computing. 2019. Т. 679. С. 295-304.

18. Дударин П.В. AN APPROACH TO FEATURE SPACE CONSTRUCTION FOR SHORT TEXTS CLUSTERING / Dudarin P.V., Yarushkina N.G. // 17th Mexican International Conference on Artificial Intelligence. 2018.

19. Дударин П.В. AN APPROACH TO VOCABULARY EXPANSION FOR NEURAL NETWORK LANGUAGE MODEL BY MEANS OF HIERARCHICAL CLUSTERING / Dudarin P.V., Yarushkina N.G. // Proceedings of the 2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). 2019.

Дударин Павел Владимирович
Исследование и разработка моделей и методов нечеткой кластеризации коротких текстов

Автореферат

Подписано в печать __.__.2021. Формат 60х84/24.
Усл. печ. л. ___.
Тираж 100 экз. Заказ _____

Типография УлГТУ, 432027, г. Ульяновск, Северный Венец, 32.