УДК 004.8 DOI: 10.26583/npe.2022.4.11

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ НАПОЛНЕНИЯ И АКТУАЛИЗАЦИИ БАЗ ЯДЕРНЫХ ЗНАНИЙ

В.П. Тельнов, Ю.А. Коровин
ИАТЭ НИЯУ МИФИ
249039, Калужская обл., г. Обнинск, Студгородок, 1

Рассматриваются вопросы проектирования и создания баз знаний в области ядерной науки и техники. Приводятся результаты поиска и исследования оптимальных алгоритмов классификации и семантического аннотирования текстового сетевого контента в интересах автоматизированного наполнения и актуализации масштабируемых семантических репозиториев (баз знаний) в области ядерной физики и атомной энергетики, а в перспективе и для иных предметных областей на русском и английском языках. Предложенные алгоритмы обеспечат методическую и технологическую основу для создания проблемно-ориентированных баз знаний как систем искусственного интеллекта, а также предпосылки для развития семантических технологий приобретения новых знаний в интернете без непосредственного участия человека. Тестирование исследуемых алгоритмов машинного обучения осуществляется методом скользящего контроля (cross-validation) на профильных корпусах текстов. Новизна представленного исследования обусловлена применением принципа оптимальности Парето для многокритериальной оценки и ранжирования исследуемых алгоритмов при отсутствии априорной информации о сравнительной значимости критериев. Проект реализуется в соответствии со стандартами семантического веба (RDF, OWL, SPARQL, др.). Не существует технологических ограничений для интеграции создаваемых баз знаний со сторонними хранилищами данных, с метапоисковыми, библиотечными, справочно-информационными и вопросно-ответными системами. Предлагаемые программные решения основаны на облачных вычислениях с использованием сервисных моделей DBaaS и PaaS для обеспечения масштабируемости хранилищ данных и сетевых сервисов. Созданное программное обеспечение находится в открытом доступе и может свободно тиражироваться.

Ключевые слова: семантический веб, база знаний, машинное обучение, классификация, семантическое аннотирование, облачные вычисления.

Тельнов В.П., Коровин Ю.А. Применение методов машинного обучения для наполнения и актуализации баз ядерных знаний. // Известия вузов. Ядерная энергетика. – 2022. – № 4. – С. 122-133. DOI: https://doi.org/10.26583/npe.2022.4.11 .

ВВЕДЕНИЕ

Ядерная наука и техника относятся к областям с высокой интенсивностью информационного обмена и генерации знаний. Исследования, которые выполняются на ускорителях элементарных частиц, ежегодно продуцируют сотни терабайтов новых экспери-

ментальных результатов [1]. Мировые центры ядерных данных аккумулируют и систематизируют информацию о тысячах ядерных реакций и ядерных констант [2]. МАГАТЭ [3] и профильные национальные агентства [4] создают и сопровождают базы данных и базы знаний по ядерным технологиям и радиационной безопасности.

Практический вклад авторов статьи в развитие баз знаний состоит в создании рабочих прототипов, а затем масштабируемых семантических веб-порталов, которые развернуты на облачных платформах и предназначены для использования в образовательной деятельности университетов [5 – 10]. Первый проект [11] связан с обучением в области ядерной физики и атомной энергетики. Второй проект [12] связан с изучением компьютерных дисциплин и программирования. Оба проекта имеют дело с моделями и методами представления и обработки проблемно-ориентированных знаний для конкретных предметных областей. Создаются и апробируются технологии накопления, интеграции знаний и повышения уровня компетенции баз знаний как систем искусственного интеллекта.

Актуальность первого проекта обусловлена тем обстоятельством, что он направлен на создание и автоматизированное наполнение семантических репозиториев (баз знаний) по ядерной физике и по атомной энергетике. Это области, в которых Россия способна достигать конкурентных преимуществ и мирового лидерства. По состоянию на 2022 г. образовательные веб-порталы университетов, центры ядерных данных, системы управления ядерными знаниями МАГАТЭ и Госкорпорации «Росатом» не используют в достаточной мере возможности семантической паутины и методы машинного обучения.

Целью исследования является поиск и тестирование оптимальных алгоритмов классификации и семантического аннотирования текстового сетевого контента для автоматизированного наполнения и актуализации графов ядерных знаний на русском и английском языках. Соответствующая оптимизационная задача формулируется и решается далее в разделе «Результаты вычислительных экспериментов». Выявленные алгоритмы обеспечат методическую и технологическую основу для непрерывного наполнения и актуализации проблемно-ориентированных баз знаний как систем искусственного интеллекта, а также необходимые предпосылки для развития семантических технологий приобретения новых знаний в интернете без непосредственного участия человека.

С практической точки зрения осуществляется программное воплощение эффективных алгоритмов классификации и семантического аннотирования как части масштабируемого семантического веб-портала, размещенного на облачной платформе. На рисунке 1 показана панель управления, которая используется для настройки параметров семантического аннотирования (классификации) произвольного текста в интернете. Любые графы знаний, формирующие базу знаний, могут быть задействованы в любом количестве и в любой комбинации. Семантическое аннотирование и классификацию можно выполнять, используя только классы онтологии (TBox), только объекты онтологии (ABox) или то и другое совместно. Выбор предпочтительных технологий обработки данных (традиционный анализ текста или методы машинного обучения) является прерогативой инженера по знаниям, который управляет процессом. Поиск и извлечение исходных текстовых данных в интернете и их первичная кластеризация осуществляются агентом «Контекстно-зависимый поиск», который является неотъемлемой частью семантического портала [11].

Созданные онлайн-решения находятся в открытом доступе (исключая сведения конфиденциального характера) и могут свободно тиражироваться. Проект реализуется в соответствии со стандартами семантического веба (RDF, OWL, SPARQL, др.) [13 – 15]. По этой причине не существует технологических ограничений для интеграции создаваемых баз знаний со сторонними хранилищами данных, с метапоисковыми, библиотечными, справочно-информационными и вопросно-ответными системами.

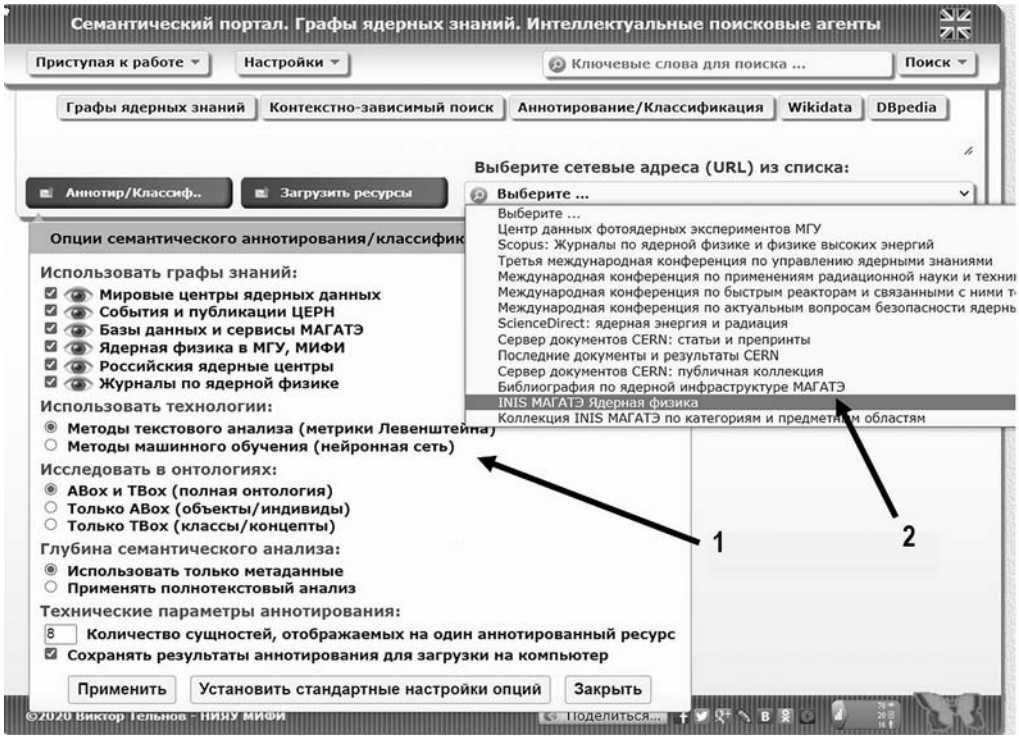

Рис. 1. Настройка параметров процесса семантического аннотирования/классификации: 1 – выбор технологии семантического аннотирования (классификации); 2 – сетевые адреса (URL) документов, подлежащих аннотированию (классификации)

Масштабируемость семантических репозиториев (баз знаний) осуществляется непосредственно средствами используемой облачной платформы. Научная новизна подходов, которые применяются в настоящем проекте, обусловлена использованием принципа оптимальности Парето, который позволяет проводить многокритериальную оценку и ранжирование исследуемых алгоритмов машинного обучения при отсутствии априорной информации о сравнительной важности критериев.

МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДАННЫХ

Классификация текстовых данных относится к задачам машинного обучения (Machine Learning – ML) в области обработки естественных языков (Natural Language Processing – NLP). К 2022 г. создано не менее дюжины методов машинного обучения, потенциально пригодных для решения задач классификации и семантического аннотирования текстов [16]. Существуют десятки программных реализаций этих методов [17].

Метод Naive Bayes Classifier. Наивный байесовский классификатор [18] считается одним из самых простых классификационных алгоритмов. Теорема Байеса инвариантна относительно причин и следствий событий. Зная, с какой вероятностью конкретная причина приводит к некоторому следствию, теорема Байеса позволяет рассчитать вероятность того, что именно эта причина привела к наблюдаемому событию. Данная идея лежит в основе байесовского классификатора, при этом для определения наиболее вероятного класса используется принцип максимума правдоподобия.

Для естественных языков вероятность появления очередного слова или фразы в тексте сильно зависит от текущего контекста. Байесовский классификатор игнорирует данное обстоятельство и представляет документ как набор слов, вероятности появления которых условно не зависят друг от друга. Этот подход иногда еще называется

«модель мешка слов» (bag of words model). Несмотря на сильные упрощающие предположения наивный байесовский классификатор хорошо работает во многих реальных задачах. Он не требует большого объема обучающих данных и на корпусах текстов умеренного объема зачастую не уступает более изощренным алгоритмам.

Метод Maxent Classifier (Softmax). Если наивный байесовский классификатор игнорирует корреляции между словами, то стохастический классификатор Maxent эти корреляции допускает и учитывает. Из моделей логистической регрессии, соответствующих обучающим данным, выбирается та модель, которая содержит наименьшее количество предположений об истинном вероятностном распределении текстовых данных. Другими словами, выбирается эмпирическое распределение вероятности с максимальной информационной энтропией. Такой подход особенно продуктивен именно для решения задач классификации текстов, когда слова в тексте очевидно не являются независимыми.

Функция Softmax, или нормализованная экспоненциальная функция, есть обобщение логистической функции для многомерного случая. В задачах многоклассовой классификации функция Softmax строится таким образом, чтобы на последнем слое нейронной сети количество нейронов оказалось равным количеству искомых классов. При этом каждый нейрон должен выдавать значение вероятности принадлежности объекта к классу, а значения всех нейронов в сумме должны дать единицу.

Классификатор Maxent обычно требует больше времени для обучения по сравнению с классификатором *Naive Bayes* из-за оптимизации, которую необходимо выполнить для оценки параметров модели. После вычисления этих параметров метод выдает весьма надежные результаты и является конкурентоспособным с точки зрения потребления вычислительного ресурса и памяти.

Метод SVM Classifier with SGD. Метод опорных векторов (Support Vector Machine – SVM) относится к линейным бинарным методам классификации с простой и ясной интерпретацией. Ставится задача отыскать в многомерном пространстве такую поверхность, в простейшем случае гиперплоскость, которая разделяет объекты на два класса с наибольшим зазором. Метод опорных векторов эквивалентен двухслойной нейронной сети, где число нейронов в скрытом слое определяется как число опорных векторов.

Стохастический градиентный спуск (Stochastic Gradient Descent – SGD) широко практикуется в моделях глубокого обучения. Здесь градиент оптимизируемой функции вычисляется не как сумма градиентов от каждого элемента выборки, а как градиент от одного, случайно выбранного подмножества элементов. Более медленная сходимость алгоритма может компенсироваться высокой скоростью выполнения итераций на больших наборах данных.

Оценка качества алгоритмов классификации. В качестве функционалов качества алгоритмов машинного обучения здесь применяются три апробированные метрики бинарной классификации [19].

1. Precision – точность классификации. Вычисляется как доля объектов, которые действительно принадлежат к некоторому положительному классу и при этом были классифицированы верно.
2. Recall – полнота классификации. Вычисляется как доля объектов, которые были отнесены алгоритмом к некоторому положительному классу и при этом были классифицированы верно.
3. F1-score. Агрегированная метрика, вычисляется как среднее гармоническое точности и полноты классификации.

Первые две метрики не зависят от наполнения классов объектами и потому применимы в условиях несбалансированных выборок. Метрика Precision характеризует способность алгоритма отличать классы друг от друга, а метрика Recall показывает способ-

ность алгоритма обнаруживать конкретный класс вообще. Третья метрика F1-score наиболее информативна в тех случаях, когда значения первых двух метрик значительно разнятся между собой. Для оценки качества алгоритмов мультиклассовой классификации используются так называемые макросредние значения, когда значения метрик усредняются по всем классам независимо от количества объектов в этих классах.

Для повышения достоверности результатов тестирования алгоритмов классификации применяется скользящий контроль (cross-validation). Исходное обучающее множество случайным образом *N* раз разбивается на *N* выборок примерно одинаковой длины. Каждая из *N* выборок поочерёдно объявляется контрольной выборкой, остальные *N* – 1 выборок объединяются в обучающую выборку. Алгоритм настраивается по обучающей выборке и затем классифицирует объекты контрольной выборки. Описанная процедура повторяется *N* раз, значение *N* изменяется в диапазоне от трёх до десяти.

РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Для выявления наиболее эффективных методов классификации текстовых данных с целью автоматизированного наполнения и актуализации графов ядерных знаний были проведены серии тестов на корпусах специфических текстов по ядерной физике и атомной энергетике. Всего было задействовано семь графов ядерных знаний [11], перечисленных в табл. 1. Для каждого исследованного метода классификации и каждого графа знаний были вычислены три метрики: Precision, Recall и F1-score.

Таблица 1

Метрики для трех методов классификации текстовых данных, вычисленные с использованием семи графов ядерных знаний

Графы ядерных знаний [11] в роли обучающих и контрольных выборок	Метод								
	Naive Bayes Classifier, %			Maxent Classifier (Softmax), %			SVM Classifier with SGD, %		
	P	R	F	P	R	F	P	R	F
Мировые центры ядерных данных	55	33	42	46	51	48	87	83	85
События и публикации ЦЕРН	94	72	82	72	71	71	42	43	43
Базы данных и сервисы МАГАТЭ	97	46	62	46	51	48	56	57	56
Ядерная физика в МГУ, МИФИ	96	57	71	78	59	67	87	82	84
Российские ядерные исследовательские центры	75	13	21	82	68	74	95	94	95
Журналы по ядерной физике	86	57	69	83	100	91	17	25	20
Объединенный граф ядерных знаний	99	25	39	63	37	46	88	85	86
Примечание. P – метрика Precision, R – метрика Recall, F – метрика F1-score									

Наглядное представление результатов вычислений из табл. 1 в виде объемных графиков дано на рис. 2. Как видно из данных таблицы и рисунка, каждый их трех исследованных методов классификации текстовых данных характеризуется четырнадцатью независимыми показателями качества Precision и Recall и семью производными пока-

зателями качества F1-score.

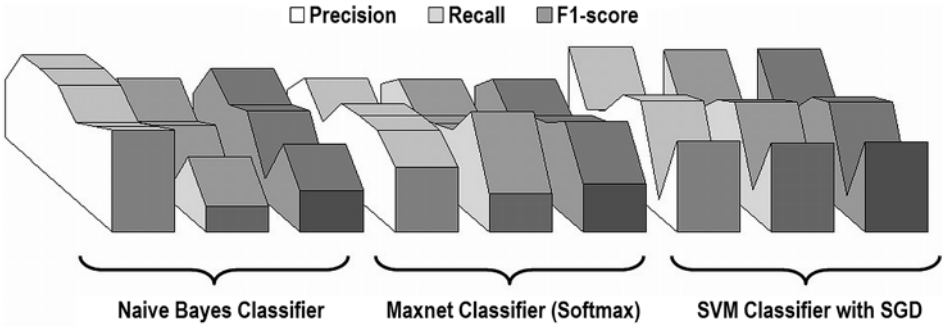

Рис. 2. Визуальное представление данных из табл. 1 в виде объемных графиков

Оптимизационная задача определяется следующим образом. Требуется выбрать наилучший метод классификации с учетом всех вычисленных показателей качества, не делая никаких предположений о сравнительной важности этих показателей. Для этого в классе транзитивных антирефлексивных бинарных отношений рассмотрим отношение Парето в евклидовом пространстве. Для любых двух элементов x и y из множества Ω отношение Парето P определяется следующим образом:

$$(\forall x, y \in \Omega)[xPy] \Leftrightarrow \{(\forall j = 1,\ldots,m)[x_j \geq y_j] \& (\exists j_0 \in \{1,\ldots,m\})[x_{j_0} > y_{j_0}]\}. \quad (1)$$

Множеством P-оптимальных элементов на Ω является множество Парето Ω^P:

$$\Omega^P = \{x \in \Omega : \ (\forall y \in \Omega)\ [y\overline{P}x]\}. \quad (2)$$

Отношение Парето обеспечивает универсальную математическую модель многокритериального контекстно-независимого выбора в евклидовом пространстве. Обозначим $d(y, x)$ количество критериев, по которым элемент y превосходит элемент x. Тогда значение

$$D_\Omega(x) = \max_{y \in \Omega} d(y, x) \quad (3)$$

называется показателем доминирования элемента x при предъявлении множества Ω. Говоря упрощенно, показатель доминирования равен количеству критериев, по которым элемент x не превосходит все прочие элементы из множества Ω. Определим функцию $C^D(\Omega)$ для выбора наилучших элементов следующим образом:

$$C^D(\Omega) = \{x \in \Omega : \ D_\Omega(x) = \min_{x \in \Omega} D_\Omega(x)\}. \quad (4)$$

Значение D_Ω есть показатель доминирования всего множества Ω. Элементы с минимальным значением показателя доминирования образуют так называемое множество Парето. Множество Парето включает в себя элементы, наилучшие по совокупности всех учтенных критериев, без каких-либо априорных предположений о сравнительной значимости этих критериев. В условиях реального выбора множество Парето нередко содержит в себе более одного элемента.

Возвращаясь к исходной задаче поиска наиболее эффективного метода классификации текстовых данных, обратимся к данным табл. 2. Там для каждого исследованного метода рассчитаны три показателя доминирования по 7, 14 и 21 метрикам соответственно. Исходные данные для расчетов взяты из табл. 1.

Как видно из данных табл. 2, лидером оказывается метод SVM Classifier with SGD со значениями показателей доминирования 3, 7 и 10. Метод Naive Bayes Classifier немногим ему уступает. Метод Maxent Classifier (Softmax) выглядит аутсайдером на фоне двух других методов. Данный вывод получен путем проведения вычислительных экспериментов на корпусах специфических текстов по ядерной физике и атомной энергетике. При этом использовались умеренные объемы исходных данных.

Таблица 2

Показатели доминирования для трех методов классификации текстовых данных, вычисленные с использованием семи графов ядерных знаний

Метод классификации текстовых данных	Показатель доминирования		
	Расчёт по метрике F1–score	Расчёт по метрикам Precision и Recall	Расчёт по метрикам F1–score, Precision и Recall
Naive Bayes Classifier	4	7	11
Maxent Classifier (Softmax)	5	10	15
SVM Classifier with SGD	3	7	10

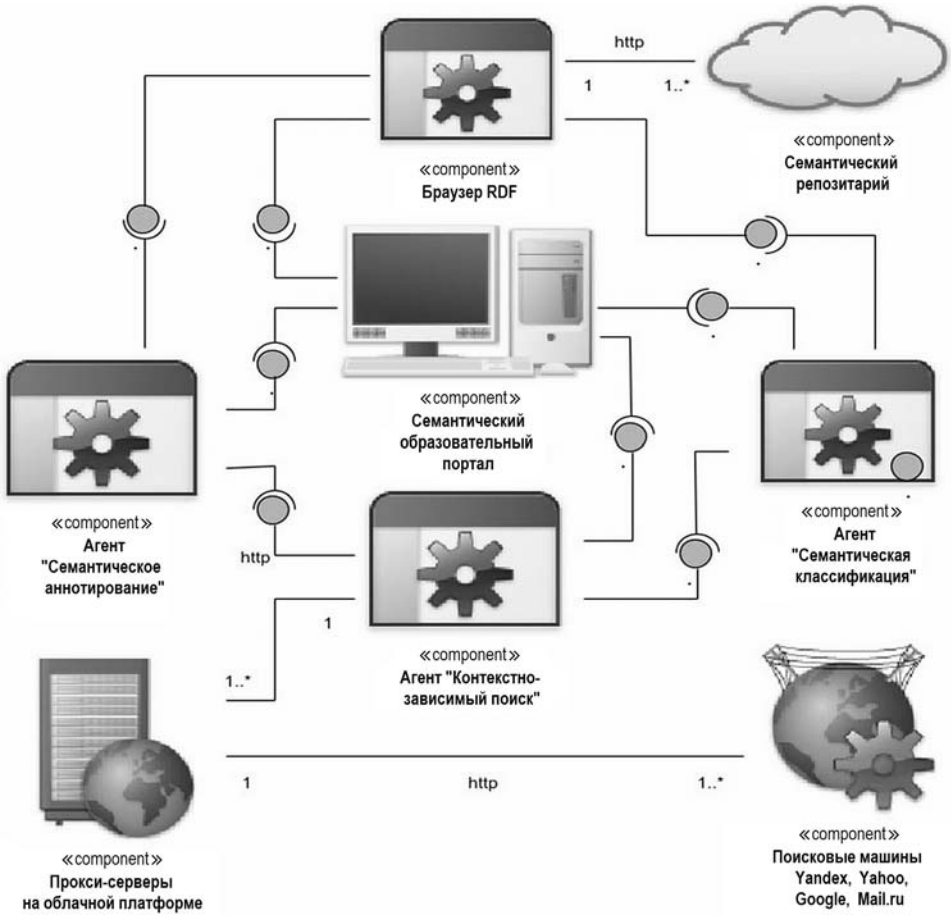

Рис. 3. Диаграмма компонентов масштабируемого семантического веб-портала.

Каждый из семи задействованных графов знаний содержал не более одной тысячи объектов и не более одной сотни классов. Следует отметить, что метод SVM Classifier with SGD является бинарным, т.е. позволяет распределять элементы всего по двум классам. Это техническое ограничение преодолевается, например, путем многократной классификации по принципу «один против всех» или «один против одного». Два других метода Maxent Classifier (Softmax) и Naive Bayes Classifier изначально являются мультиклассо-

выми, что делает их более удобными в применении.

АРХИТЕКТУРА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ПРОЕКТА

Реализуемые в проекте программные решения основаны на облачных вычислениях с использованием сервисных моделей DBaaS и PaaS для обеспечения масштабируемости хранилищ данных и сетевых сервисов. Серверные скрипты рабочего прототипа программного обеспечения работают на облачной платформе Jelastic в среде выполнения Java и Python.

На рисунке 3 представлена диаграмма компонентов, выполненная по стандарту UML 2 [20], из которой видны особенности программного воплощения агента «Семантическое аннотирование» и агента «Семантическая классификация» в составе масштабируемого семантического веб-портала [11]. Тестирование созданного программного обеспечения осуществляется на корпусах текстов по ядерной физике и атомной энергетике, включая релевантные тексты МАГАТЭ, ЦЕРН, НИЯУ МИФИ, физфак МГУ, а также тексты профильных журналов, публикации ядерных исследовательских центров и центров ядерных данных.

СМЕЖНЫЕ РАБОТЫ. ЗАКЛЮЧЕНИЕ

На проблемах развития семантического веба, связанных вопросах машинного обучения и обработки естественных языков концентрируются научные группы из Стэнфордского университета [21], Массачусетского института технологий, Университета Бари, Университета Лейпцига, Университета Манчестера. Мировые гиганты ИТ-индустрии активно развивают модели представления знаний и технологии машинного обучения, среди них IBM Watson Studio, Google AI and Machine Learning, Amazon Comprehend NLP, AWS Machine Learning, Yandex DataSphere (Jupyter Notebook) и др. Программные средства для исследований в области искусственного интеллекта и обработки естественных языков предоставляют Matlab [22], Stanford NLP [21], Scikit-learn [17], др. В России профильные исследования осуществляются в Центре компетенций НТИ МФТИ, Университете ИТМО, на факультете ВМК МГУ, в ИСП РАН им. В.П. Иванникова [23], российских подразделениях Hyawei.

В настоящем исследовании на семи корпусах профильных текстов по ядерной физике и атомной энергетике показана эффективность сравнительно простых, интуитивно ясных методов машинного обучения для решения задачи непрерывного наполнения из интернета и актуализации баз ядерных знаний без непосредственного участия человека. Метод опорных векторов и наивный байесовский классификатор обеспечивают компетентность семантизированных баз знаний как систем искусственного интеллекта. За рамками этой статьи остались некоторые результаты, которые были получены авторами при исследовании таких классификаторов, как метод «*K* ближайших соседей (*kNN*)» и терминологические деревья решений, которые строятся по результатам синтаксического анализа текста.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-21-00182, https://rscf.ru/project/22-21-00182/.

Литература

1. CERN Document Server. Электронный ресурс: https://cds.cern.ch (дата доступа 26.06.2022).

2. Центр данных фотоядерных экспериментов. Электронный ресурс: http://cdfe.sinp.msu.ru/index.ru.html (дата доступа 26.06.2022).

3. Международное агентство по атомной энергии. Управление ядерными знаниями.

Электронный ресурс: https://www.iaea.org/ru/temy/upravlenie-yadernymi-znaniyami (дата доступа 26.06.2022).

4. Госкорпорация «Росатом». Система управления знаниями (СУЗ). Электронный ресурс: http://www.innov-rosatom.ru/suz-rosatoma/ (дата доступа 26.06.2022).

5. *Telnov V., Korovin Yu*. Machine learning and text analysis in the tasks of knowledge graphs refinement and enrichment. / CEUR Workshop Proceedings, 2020, v. 2790, pp. 48-62. Supplementary Proceedings of the XXII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2020), Voronezh, Russia, October 13-16, 2020, EID: 2-s2.0-85098723055, ISBN: 16130073. Электронный ресурс: http://ceur-ws.org/Vol-2790/paper06.pdf (дата доступа 26.06.2022).

6. *Telnov V., Korovin Yu*. Semantic Web and Interactive Knowledge Graphs as Educational Technology. In: Cloud Computing Security, ed. Dinesh G. Harkut, IntechOpen, London, 2020, ISBN: 978-1-83880-703-0, DOI: https://doi.org/10.5772/intechopen.83221 .

7. *Telnov V., Korovin Yu*. Semantic web and knowledge graphs as an educational technology of personnel training for nuclear power engineering. // Nuclear Energy and Technology. – 2019. – No. 5(3). – PP. 273-280. DOI: https://doi.org/10.3897/nucet.5.39226 .

8. *Тельнов В., Коровин Ю*. Семантический веб и графы знаний как образовательная технология подготовки кажров для ядерной энергетики. // Известия вузов. Ядерная энергетика. – 2019. – № 2. – С. 219-229. DOI: https://doi.org/10.26583/npe.2019.2.19 .

9. *Тельнов В., Коровин Ю*. Программирование графов знаний, рассуждения на графах. // Программная инженерия. – 2019. – № 2. – С. 59-68. DOI: https://doi.org/10.17587/prin.10.59-68 .

10. *Telnov V*. Semantic Educational Web Portal. / Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), Moscow, Russia, October 9-13, 2017. Электронный ресурс: http://ceur-ws.org/Vol-2022, online http://ceur-ws.org/Vol-2022/paper11.pdf (дата доступа 26.06.2022).

11. Семантический портал. Графы ядерных знаний. Интеллектуальные поисковые агенты, Электронный ресурс: http://vt.obninsk.ru/x/ (дата доступа 26.06.2022).

12. Графы знаний по компьютерным дисциплинам. Интеллектуальные поисковые агенты. Электронный ресурс: http://vt.obninsk.ru/s/ (дата доступа 26.06.2022).

13. W3C Semantic Web. Электронный ресурс: https://www.w3.org/standards/semanticweb/ (дата доступа 26.06.2022).

14. W3C RDF Schema 1.1. Электронный ресурс: https://www.w3.org/TR/rdf-schema/ (дата доступа 26.06.2022).

15. W3C OWL 2 Web Ontology Language. Электронный ресурс: https://www.w3.org/TR/owl2-overview/ (дата доступа 26.06.2022).

16. *Geron A*. Hands-on ML with Scikit-Learn, Keras & TensorFlow. 2nd edn. – O'Reilly Media Inc., Boston. – 2019.

17. Scikit-learn. Machine Learning in Python. Электронный ресурс: https://scikit-learn.org/stable/ (дата доступа 26.06.2022).

18. Naive Bayes Classifier. Электронный ресурс:

https://scikit-learn.org/stable/modules/naive_bayes.html (дата доступа 26.06.2022).

19. Classification Metrics. Электронный ресурс: https://github.com/turi-code/userguide/blob/master/evaluation/classification.md (дата доступа 26.06.2022).

20. ISO/IEC 19505-2:2012(E) Information technology – Object Management Group Unified Modeling Language (OMG UML) – Part 2: Superstructure. ISO/IEC, Geneva, 2012.

21. *Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., McClosky D*. The Stanford CoreNLP Natural Language Processing Toolkit. / Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association

for Computational Linguistics, 2014, pp. 55-60. Электронный ресурс: https://aclanthology.org/P14-5010.pdf (дата доступа 26.06.2022). DOI: https://doi.org/10.3115/v1/P14-5010 .

22. Machine Learning with MATLAB & Simulink. Электронный ресурс: https://www.mathworks.com/solutions/machine-learning.html (дата доступа 26.06.2022).

23. *Stupnikov S., Kalinichenko A.* Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. / Proceedings of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018). – Springer, 2019. – PP. 17-39. DOI: https://doi.org/10.1007/978-3-030-23584-0_2 .

Поступила в редакцию 06.07.2022 г.

Авторы

Тельнов Виктор Петрович, доцент, канд. техн. наук
E-mail: telnov@bk.ru

Коровин Юрий Александрович, профессор, доктор физ.–мат. наук
E-mail: korovinyu@mail.ru

UDC 004.8

APPLICATION OF MACHINE LEARNING METHODS FOR FILLING AND UPDATING NUCLEAR KNOWLEDGE DATABASES

Telnov V.P., Korovin Yu.A.

IATE MEPhI
1 Studgorodok, 249039 Obninsk, Kaluga Reg., Russia

ABSTRACT

The paper considers issues involved in design and creation of knowledge databases in the field of nuclear science and technology. Results from searching for and investigating suitable algorithms to classify and semantically annotate the textual network content for the convenience of computer-aided filling and updating of scalable semantic repositories (knowledge bases) in the field of nuclear physics and nuclear power are presented in Russian and English. The proposed algorithms will provide a methodological and technological basis for creating problem-oriented knowledge databases as artificial intelligence systems, as well as prerequisites for developing semantic technologies to acquire new knowledge via the Internet without direct human participation. The machine learning algorithms under investigation are tested by cross-validation method using field-specific text corpora. The novelty of the presented study is defined by the application of the Pareto's optimality principle for multi-criteria evaluation and ranking of the algorithms under investigation in the absence of a priori information about the comparative significance of the criteria. The project is implemented in accordance with semantic web standards (RDF, OWL, SPARQL, etc.). There are no technological limits for integrating the knowledge bases created with third-party data repositories, with meta-search, library, reference and question-answering systems. The proposed software solutions are based on cloud computing using the DBaaS and PaaS service models to ensure that data repositories and network services are scalable. The software built is publicly available and free to copy.

Key words*:* semantic web, knowledge database, machine learning, classification, semantic annotation, cloud computing.

Telnov V.P., Korovin Yu.A. Application of Machine Learning Methods for Filling and Updating Nuclear Knowledge Databases. *Izvestiya vuzov. Yadernaya Energetika*. 2022, no. 4, pp. 122-133; DOI: https://doi.org/10.26583/npe.2022.4.11 (in Russian).

REFERENCES

1. CERN Document Server. Available at: https://cds.cern.ch (accessed Jun. 26, 2022).

2. Centre for Photonuclear Experiments Data. Available at: http://cdfe.sinp.msu.ru/index.en.html (accessed Jun. 26, 2022)

3. IAEA Nuclear Knowledge Management. Available at: https://www.iaea.org/topics/nuclear-knowledge-management (accessed Jun. 26, 2022)

4. Rosatom State Corporation. Knowledge Management System. Available at: http://www.innov-rosatom.ru/suz-rosatoma/ (accessed Jun. 26, 2022)

5. Telnov V., Korovin Yu. Machine Learning and Text Analysis in the Tasks of Knowledge Graphs Refinement and Enrichment. *CEUR Workshop Proceedings*. 2020, v. 2790, pp. 48-62. *Supplementary Proceedings of the XXII International Conference on Data Analytics and Management in Data Intensive Domains* (DAMDID/RCDL 2020), Voronezh, Russia, October 13-16, 2020, EID: 2-s2.0-85098723055, ISBN: 16130073. Available at: http://ceur-ws.org/Vol-2790/paper06.pdf (accessed Jun. 26, 2022)

6. Telnov V., Korovin Yu. Semantic Web and Interactive Knowledge Graphs as Educational Technology. In: *Cloud Computing Security*, ed. Dinesh G. Harkut, IntechOpen, London, 2020, ISBN: 978-1-83880-703-0; DOI: https://doi.org/10.5772/intechopen.83221 .

7. Telnov V., Korovin Yu. Semantic Web and Knowledge Graphs as an Educational Technology of Personnel Training for Nuclear Power Engineering. *Nuclear Energy and Technology*. 2019, no. 5 (3), pp. 273-280; DOI: https://doi.org/10.3897/nucet.5.39226 .

8. Telnov V., Korovin Yu. Semantic Web and Knowledge Graphs as an Educational Technology of Personnel Training for Nuclear Power Engineering. *Izvestiya vuzov. Yadernaya Energetika*. 2019, no. 2, p. 219-229; DOI: http://doi.org/10.26583/npe.2019.2.19 .

9. Telnov V., Korovin Yu. Programming Knowledge Graphs, Reasoning on Graphs. *Software Engineering*. 2019, no. 2, pp. 59-68; DOI: https://doi.org/10.17587/prin.10.59-68 .

10. Telnov V. Semantic Educational Web Portal. *Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains* (DAMDID/RCDL 2017), Moscow, Russia, October 9-13, 2017, http://ceur-ws.org/Vol-2022. Available at: http://ceur-ws.org/Vol-2022/paper11.pdf (accessed Jun. 26, 2022).

11. Semantic Educational Portal. Nuclear Knowledge Graphs. Intelligent Search Agents. Available at: http://vt.obninsk.ru/x/ (accessed Jun. 26, 2022).

12. Knowledge Graphs on Computer Science. Intelligent Search Agents. Available at: http://vt.obninsk.ru/s/ (accessed Jun. 26, 2022).

13. W3C Semantic Web. Available at: https://www.w3.org/standards/semanticweb/ (accessed Jun. 26, 2022).

14. W3C RDF Schema 1.1. Available at: https://www.w3.org/TR/rdf-schema/ (accessed Jun. 26, 2022).

15. W3C OWL 2 Web Ontology Language. Available at: https://www.w3.org/TR/owl2-overview/ (accessed Jun. 26, 2022).

16. Geron A. *Hands-on ML with Scikit-Learn, Keras & TensorFlow*. 2nd edn. O'Reilly Media, Inc. Boston, 2019.

17. Scikit-learn. Machine Learning in Python. Available at: https://scikit-learn.org/stable/ (accessed Jun. 26, 2022).

18. Naive Bayes Classifier. Available at: https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Jun. 26, 2022).

19. Classification Metrics. Available at: https://github.com/turi-code/userguide/blob/master/evaluation/classification.md (accessed Jun. 26, 2022)

20. ISO/IEC 19505–2:2012(E) Information technology – Object Management Group Unified Modeling Language (OMG UML) – Part 2: Superstructure. ISO/IEC, Geneva (2012).

21. Manning C., Surdeanu M., Bauer J., Finkel J., Bethard S., McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the LIInd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics*, 2014, pp. 55-60. Available at: https://aclanthology.org/P14-5010.pdf (accessed Jun. 26, 2022); DOI: https://doi.org/10.3115/v1/P14-5010 .

22. Machine Learning with MATLAB & Simulink. Available at: https://www.mathworks.com/solutions/machine-learning.html (accessed Jun. 26, 2022)

23. Stupnikov S., Kalinichenko A. Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. *Proceedings of the XX International Conference on Data Analytics and Management in Data Intensive Domains* (DAMDID/RCDL 2018), Springer, 2019, pp. 17-39; DOI: https://doi.org/10.1007/978-3-030-23584-0_2 .

Authors

Telnov Viktor Petrovich, Associate Professor, Cand. Sci. (Engineering)
E-mail: telnov@bk.ru

Korovin Yury Aleksandrovich, Professor, Dr. Sci. (Phys.-Math.)
Email: korovinyu@mail.ru