УДК 004.4; 004.6; 004.9

ЭЛЕМЕНТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА НАУЧНО-ОБРАЗОВАТЕЛЬНОМ ПОРТАЛЕ ЯДЕРНЫХ ЗНАНИЙ BELNET

С. Н. Сытова, В. В. Гавриловец, А. П. Дунец, А. Н. Коваленко, С. В. Черепица
Институт ядерных проблем БГУ, Минск, Беларусь

Представлен краткий обзор методов и алгоритмов искусственного интеллекта (ИИ) на белорусском научно-образовательном электронном портале ядерных знаний BelNET.

1. Цели работы

Главной целью работ по внедрению элементов ИИ на научно-образовательном электронном портале ядерных знаний Республики Беларусь BelNET (*Bel*arusian *N*uclear *E*ducation and *T*raining), расположенном в Интернете по адресам https://belnet.bsu.by/, https://belnet.by/, является дальнейшее развитие данного портала, существенное повышение его функциональности и перерастание в полноценный Интеллектуальный семантический портал. Это в конечном итоге позволит эффективно привлекать дополнительных пользователей на портал, обеспечивать широкий доступ к ядерным знаниям, а также огромному спектру сопутствующих естественно-научных и технических знаний, стимулировать обмен результатами научных исследований и разработок, пропагандировать научные и научно-технические ядерные знания, повышать эффективность научной деятельности, улучшать коммуникации и обеспечивать прозрачность, а также активизировать информационный обмен в области широкого спектра научных знаний для интеграции Республики Беларусь в глобальное информационное научное пространство.

Функции Интеллектуального семантического научно-образовательного электронного портала ядерных знаний BelNET, включающего специализированную информационную архивную онлайн-систему управления ядерными знаниями, следующие:

– обеспечение открытого доступа к научным результатам для широкой аудитории, включая ученых, исследователей, студентов, преподавателей и представителей различных секторов экономики;

– быстрое обнаружение необходимой информации с использованием мощных инструментов поиска по ключевым словам, авторам, темам и другим критериям;

– поддержка научных исследований через обеспечение повторного использования данных для проведения новых исследований, проверки результатов и разработки новых гипотез;

– воспроизводимость научных исследований через предоставление доступа к исходным данным, коду и методологиям, использованным в исследованиях;

– обеспечение сохранности публикаций и данных, работа с метаданными объектов с целью облегчения поиска и понимания контекста.

В настоящий момент научно-образовательный электронный портал ядерных знаний BelNET под эгидой Министерства образования Республики Беларусь, является единственным полноценным порталом ядерных знаний в Республике Беларусь. С начала 2022 г. по апрель 2025 г. на портале количество записей в области фундаментальных и прикладных ядерных знаний, актуальной новостной мировой информации в ядерной области увеличилось в три раза до 6 500. С начала 2025 г. только на зеркале

https://belnet.by/ по данным счетчиков было осуществлено 5 180 переходов из Google и 3 400 из Яндекса.

Работа портала BelNET обеспечивается оригинальным белорусским программным обеспечением (ПО) на основе свободного программного обеспечения (СПО). На BelNET в настоящее время внедрены элементы семантических технологий [1]. Однако в современных условиях взрывного роста использования элементов ИИ во всем мире становится очевидной необходимость и реальная возможность дальнейшего развития функционала портала на основе передовых технологий ИИ и широкого использования семантических технологий, преобразовав портал BelNET в Интеллектуальный семантический интернет-портал.

Кроме этого, развитие портала BelNET в рамках данного проекта, осуществляемое через его научное сопровождение, научно-методическое обеспечение, техническую поддержку и разработку оригинальных материалов контента портала, является необходимым условием развития системы управления ядерными знаниями, которую активно продвигает и поддерживает Международное агентство по атомной энергии (МАГАТЭ) во всем мире, особенно в странах со сложившейся атомной энергетикой, которой является Республика Беларусь.

2. Обзор «краеугольных» камней в работе над порталом ядерных знаний

Компьютеры отлично справляются с обработкой и хранением данных. Однако если в организации отсутствует эффективная система управления данными, это часто приводит к появлению разрозненных и несовместимых хранилищ, несогласованности данных и их низкому качеству. В условиях экспоненциального роста объемов и разнообразия данных, возникает острая потребность в развитии информационных технологий, способных обеспечить эффективную работу с информацией.

При этом важная составляющая научных исследований – процесс управления знаниями – часто остается за кадром, в тени конкретных научных результатов. Это подчеркивает необходимость популяризации и сохранения научных знаний, а также разработки специализированных информационных систем для этой цели. Данное направление получило название «управление знаниями».

Стандарт ISO 30401:2018 определяет и конкретизирует основные принципы управления знаниями, включая процессы создания и использования знаний в организации, а также необходимые системы для их эффективной реализации.

ЮНЕСКО в своей «Декларации о науке и использовании научных знаний» акцентирует внимание на важности свободного распространения результатов научных исследований, что требует развития электронных научных архивов и порталов, основанных на передовых информационных технологиях. Успех разработки таких масштабных информационных систем напрямую зависит от правильного применения принципов управления знаниями, особенно в отношении научных знаний. Это подразумевает процессы создания, обмена, использования и управления знаниями и информацией. Примером успешного применения управления знаниями является деятельность МАГАТЭ в области управления ядерными знаниями, активно реализуемая с начала XXI в. [2].

В современной ИТ-сфере СПО занимает прочные позиции, предлагая пользователям значительные преимущества перед проприетарным (коммерческим) ПО. Главное из них – четыре фундаментальные свободы: неограниченная установка, бесплатное использование, возможность модификации и свободная передача программ. В то же время, коммерческое ПО часто не гарантирует полной прозрачности и может содержать незадокументированные возможности. Российское правительство осознало важность

СПО и в 2010 г. утвердило план перехода федеральных органов и бюджетных организаций на его использование. Этот процесс активно развивается с 2014 г., а с 2022 г. наблюдается ускоренный перевод на СПО ключевых сервисов российского электронного правительства. Цель – заменить дорогостоящее проприетарное ПО (Oracle, IBM, Microsoft) и зависимое от него аппаратное обеспечение (Solaris, VMware, Symantec, SPARC и др.) на более безопасные и открытые альтернативы – СПО и доверенное оборудование, лишенное скрытых уязвимостей.

В Республике Беларусь в целях обеспечения безопасности информационных систем принят ряд Указов Президента Республики Беларусь, постановлений Совета Министров Республики Беларусь, Оперативно-аналитического центра при Президенте Республики Беларусь, других министерств и ведомств. В соответствии с Приказом Министра обороны Республики Беларусь от 18.12.2011 № 112 «Об утверждении перечня форматов представления и протоколов передачи данных, используемых в информационных системах Вооруженных Сил и транспортных войск», ПО, работающее под ОС Linux, и собственно СПО является приоритетным при использовании в Вооруженных силах Республики Беларусь.

С другой стороны, системный подход, ориентированный на процессы, является ключевым элементом для достижения организационной эффективности и успешной компьютерной автоматизации. Он подразумевает глубокое понимание и грамотное управление всеми взаимосвязанными рабочими процессами. Внедрение автоматизации неизбежно влечет за собой ревизию и оптимизацию существующих технических, технологических и организационных процедур. Цель – выявить и устранить неэффективные звенья, такие как дублирование функций, неоптимальные последовательности действий или вовсе ненужные этапы. Соответствие этому подходу подтверждается международными стандартами ISO 9001, 9004 и 17025. Поэтому разработка информационных систем, поддерживающих полный цикл работ, должна строиться на его принципах.

Современные интернет-технологии сделали информацию беспрецедентно доступной, а поисковые системы стали основным инструментом для ее поиска. За последнее десятилетие мы стали свидетелями колоссальных усовершенствований в области сбора, управления, анализа и распространения знаний, извлеченных из мировых данных. Доступ к актуальной информации теперь сводится к простому интернет-поиску. Парадоксально, но даже на сегодняшний день многие официальные и корпоративные веб-ресурсы оснащены некачественными встроенными поисковыми механизмами, что мешает пользователям находить необходимую информацию.

Согласно материалам МАГАТЭ, семантические технологии в современном мире должны лежать в основе веб-поиска и управления онлайн-информацией, позволяя комплексно взглянуть на предметную область [3].

Таким образом, СПО, процессный системный подход и семантические технологии являются основой, на которой создаются современные информационные системы. Сюда также в обязательном порядке следует добавить методы и алгоритмы ИИ.

3. Используемые методы и алгоритмы ИИ

Бурный рост в последнее время инструментов ИИ и тесно связанных с ними семантических технологий предоставляет для интернет-порталов различной направленности такие преимущества их использования, как персонализацию и улучшение пользовательского опыта, оптимизацию поисковых функций, углубление аналитики в части принятия решений. Интернет-ресурсы, находящиеся на переднем крае в части использования ИИ и семантических технологий, получают неоспоримые выгоды, делая их работу эффективнее, удобнее и безопаснее для пользователей.

Бесспорно, что предлагаемые к разработке методы и алгоритмы должны быть экономичными в реализации, требующими минимальных вычислительных ресурсов, для возможности их эффективной реализации на доступном серверном оборудовании. Это требование накладывает дополнительные ограничения на условия реализации проекта. Одним из предлагаемых к развитию методов в данном проекте является векторизация в рамках вычислительной лингвистики и моделей машинного обучения на базе статистических методов [4, 5].

Чтобы обучить модели машинного обучения на текстовых данных, текст необходимо преобразовать в числовую форму. Этот процесс, называемый векторизацией текста, позволяет алгоритмам понимать и обрабатывать текстовую информацию, поскольку сами по себе тексты не структурированы и не подходят для прямого использования в моделях. Использование техник векторизации позволяет значительно повысить эффективность обработки больших объемов научных данных и документов, способствует усилению информационного обмена, улучшению аналитического инструментария и созданию удобных интерфейсов для пользователей. Векторизация помогает преобразовать тексты статей, препринтов, отчетов и книг в компактное числовое представление (векторы), которое упрощает быстрый точный поиск нужных материалов среди огромного массива информации. Использование методов векторизации также позволяет эффективно проводить сравнительный анализ различных публикаций для формирования персонализированных рекомендаций пользователям портала.

Для организации текстов в структуру используются методы классификации [6] и кластерного анализа. Это требует подбора подходящих метрик сравнения текстов по похожести с учетом особенностей контента документов предметной области. Необходимо отметить, что подходы классификации и кластерного анализа значительно отличаются между собой. Классификация оперирует заранее заданной структурой для организации текстов – глоссарием. И соответствующий алгоритм должен соотнести текст с этой структурой и определить место текста в структуре: к какому классу (или классам) следует отнести исследуемый текст.

Кластерный анализ (или кластеризация) – это метод анализа данных, который разделяет объекты на группы (кластеры) на основе их сходства [7]. Цель состоит в том, чтобы объекты внутри одного кластера были как можно более похожими, а объекты из разных кластеров – как можно более разными. Это задача обучения без учителя, поскольку нет заранее заданных меток для объектов. Другими словами, кластерный анализ основан на алгоритмах, которые формируют перечень классов в процессе работы с конкретным множеством документов, автоматически группируя документы по степени похожести. При этом перечень групп заранее не известен.

В качестве примера можно привести перспективную идею группировать результаты полнотекстового поиска при большом числе документов. Это позволит пользователю исключать из рассмотрения группы документов и обращать внимание на интересующие группы. При этом оба подхода используют векторизацию.

На основе указанных алгоритмов формируются интеллектуальные инструменты для автоматического семантического анализа содержания текстов и классификации научных публикаций по различным категориям (дисциплинам, типам исследований, уровням значимости) с целью обеспечения качественной навигации по ресурсам портала, извлечение ключевых терминов и концептов, анализа цитирований и установления авторского вклада, помощи в принятии этических решений, предварительной обработки материалов портала на предмет выявления генерации материала с помощью того или иного инструмента ИИ, контроль уникальности и оригинальности публикуемых материалов. Это в совокупности является реализацией принципов комплексной мето-

дологии цифровой интеграции ядерных знаний с использованием искусственного интеллекта и семантических технологий.

Отметим, что разработанные за 10 лет функционирования портала BelNET оригинальные материалы контента портала (корпус текстов) (препринты, рукописи, специально созданные для контента портала научные, новостные и учебные материалы, включая лекции и лабораторные работы с тестами к ним) являются хорошей основой, на которой будет обучаться создаваемая большая языковая модель.

4. Заключение

Активное развитие и совершенствование белорусского портала ядерных знаний BelNET под эгидой Министерства образования с привлечением пользователей из-за рубежа позволяет поднять рейтинг белорусских вузов в рейтинге Webometrics Ranking of World Universities.

Выполняемые работы призваны стимулировать создание и эффективное использование ядерных знаний, адаптированных к специфике страны, а также сбор, передачу, хранение и обмен существующими знаниями, минимизируя риски их утраты и обеспечивая широкий доступ к информации. Результатом является повышение качества коммуникации, прозрачности и активизация информационного обмена в ядерной области.

Новизна предлагаемой работы заключается в создании и внедрении стабильно работающего, безопасного и функционального решения с использованием элементов искусственного интеллекта и семантических технологий, разработанного на базе отечественного ПО с открытым исходным кодом на основе СПО, обеспечивающего доступность и экономическую эффективность, применимого, в том числе для свободной публикации широкого спектра ядерных знаний. Целевая аудитория портала – десятки тысяч специалистов, включая студентов, ученых, преподавателей, работников госсектора и бизнеса.

Работа выполняется в рамках мероприятия 3.1 Сводного перечня научных исследований и разработок по развитию государственной системы научно-технической информации Республики Беларусь на 2021–2025 годы.

Список литературы

1. Основы функционирования семантического портала ядерных знаний BelNET / С. Н. Сытова [и др.] // Информатика. – 2024. – Т. 21, № 2. – С. 7–23.

2. Maintaining knowledge, training and infrastructure for research and development in nuclear safety / IAEA. – Vienna : IAEA, 2003. – 19 p.

3. Exploring semantic technologies and their application to nuclear knowledge management / IAEA Nuclear Energy Series No. NG-T-6.15. – Vienna : IAEA, 2021. – 62 p.

4. Palmer, D. D. Text preprocessing / D. D. Palmer // Handbook of Natural Language Processing, Second Edition. – Chapman and Hall/CRC, 2010. – P. 9–30.

5. A Comparison of Semantic Similarity Methods for Maximum Human Interpretability / P. Sitikhu [et al.] // Proc. 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal. – 2019. – Vol. 1. – P. 1–4.

6. Efficient Estimation of Word Representations in Vector Space / T. Mikolov [et al.] // Proc. Int. Conf. on Learning Representations, 2013. – Mode of access: https://arxiv.org/abs/1301.3781v3. – Date of access 11.07.2025. – 12 p.

7. Jain, A. Data Clustering: A Review / A. Jain A., M. Murty, P. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, №. 3. – P. 1–69.