ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2025.156.57

ВНЕДРЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ МОДУЛЕЙ В ИНФОРМАЦИОННУЮ СИСТЕМУ ОБРАБОТКИ И СТРУКТУРИРОВАНИЯ ДОКУМЕНТОВ ОБРАЗОВАТЕЛЬНОЙ ОРГАНИЗАЦИИ

Научная статья

Тюшкевич Н.М.[1, *], Розов А.С.[2]
[1] ORCID : 0009-0008-1982-0486;
[2] ORCID : 0009-0007-2235-1038;
[1, 2] Российский технологический университет – МИРЭА, Москва, Российская Федерация

* Корреспондирующий автор (nikolayvita[at]gmail.com)

Аннотация

В данной статье рассматриваются информационные системы для обработки и структурирования документов образовательной организации, в частности системы для работы с научными изданиями в рамках кафедр, а также варианты внедрения в них интеллектуальных модулей с целью сокращения количества рутинных задач. С позиции аналитического метода и системного анализа изучены существующие программные решения в данной предметной области, а также методы машинного обучения, наиболее подходящие для создания интеллектуального модуля. Помимо этого, проведено сравнение различных вариантов его интеграции в архитектуру системы. В результате исследования установлено, что наиболее эффективным подходом является использование модели RoBERTa при разработке интеллектуального модуля и его внедрение в систему в качестве выделенного сервиса.

Ключевые слова: Интеллектуальные модули, машинное обучение, обработка научных публикаций, информационные системы, NLP.

IMPLEMENTATION OF INTELLIGENT MODULES IN THE INFORMATION SYSTEM FOR PROCESSING AND STRUCTURING DOCUMENTS OF AN EDUCATIONAL ORGANISATION

Research article

Tyushkevich N.M.[1, *], Rozov A.S.[2]
[1] ORCID : 0009-0008-1982-0486;
[2] ORCID : 0009-0007-2235-1038;
[1, 2] Russian Technological University – MIREA, Moscow, Russian Federation

* Corresponding author (nikolayvita[at]gmail.com)

Abstract

This article reviews information systems for processing and structuring documents of an educational organisation, in particular systems for working with scientific publications within departments, as well as options for implementing intelligent modules in them in order to reduce the number of routine tasks. From the position of analytical method and system analysis, the existing software solutions in this subject area, as well as machine learning methods most suitable for the creation of an intelligent module, have been studied. In addition, a comparison of various options for its integration into the system architecture was carried out. As a result of the research, it was found that the most effective approach is to use the RoBERTa model in the development of the intelligent module and its implementation in the system as a dedicated service.

Keywords: intelligent modules, machine learning, processing of scientific publications, information systems, NLP.

Введение

Широкое применение цифровизации в образовательных организациях затрагивает не только учебную сферу деятельности, но и работу внутренних структур. К подобным структурам можно отнести институты, кафедры и иные отделы прямо или косвенно задействованные в образовательном процессе. В рамках своего функционирования, данные подразделения формируют, используют, обрабатывают, структурируют и хранят большое количество различных документов. Внедрение цифровизации позволило упростить и автоматизировать большую часть этих процессов, путем использования специализированных информационных систем. Подобный подход позволил сократить количество рутинных задач, связанных с использованием аналогового способа хранения информации, однако полностью исключить ручной труд сотрудников не удалось. Зачастую сбор, структурирование и формирование документов происходит вручную, ввиду разнородности стандартов их оформления. В таком случае информационная система является не более чем местом хранения и быстрого доступа к данным в цифровом формате, сохраняя при этом проблему рутинности аналогового подхода.

Использование интеллектуальных модулей может существенно облегчить работу с извлечением данных, дальнейшим их структурированием и генерацией необходимых документов по заранее указанным шаблонам. При разработке подобных модулей подразумевается использование различных способов машинного обучения для задач категоризации документов, выявления скрытых зависимостей, группировки документов по смысловому содержанию и генерации выходных документов на основе полученных данных. Также машинное обучение можно использовать для экстракции информации из оцифрованных источников, например отсканированных документов.

Целью данной научной работы является исследование применения способов машинного обучения при разработке интеллектуальных модулей извлечения и систематизации данных, а также дальнейшего их внедрения в архитектуру информационной системы обработки и структурирования документов образовательной организации.

В качестве методологии исследования были выбраны аналитический метод и системный анализ. С позиции аналитического подхода были рассмотрены существующие решения в рамках предметной области, а также наиболее эффективные методы машинного обучения для решения поставленных задач. В ходе системного анализа были выявлены сильные и слабые стороны существующих систем и методов. На основе полученных данных были выделены наиболее эффективные методы машинного обучения и варианты их внедрения в информационную систему.

Анализ предметной области и существующих решений

Предметной областью исследования являются узконаправленные информационные системы для обработки и структурирования документов, а именно системы для работы с научными изданиями и публикационной деятельностью преподавателей. Использование подобных средств в процессе создания планов и отчетности по научным публикациям может значительно упростить и ускорить работу ответственного персонала, позволяя отказаться от часто используемых для этого электронных таблиц.

Основной проблемой при работе с подобного рода документами является большое количество метаданных, сопровождающих любые научные труды. К таким метаданных относятся: библиографические данные (название статьи, авторы, аффилиация, ключевые слова, аннотация, дата публикации, идентификатор статьи), информационные данные о публикации (название журнала или конференции, том, номер, страницы, редакторы, тип публикации), цитирования и ссылки (список литературы, внутренние ссылки, индекс цитирования), а также технические метаданные (формат документа, размер файла, язык статьи, лицензия, права и авторство). С точки зрения информационной системы, данную информацию должно быть возможно рассматривать как вкупе с основным текстом публикации, так и обособлено, для быстрого доступа к ней и возможности редактирования.

Существует несколько типов систем, используемых для решения вышеописанных задач. Системы для хранения и управления научными публикациями, представителями которых являются DSpace [1] и EPrints [2], позволяют создать институциональный репозиторий для хранения данных о публикациях, а также предоставляют инструменты текстового поиска, архивирования и экспорта данных. У таких систем есть и ряд недостатков, к примеру DSpace требует наличия веб-сервера и базы данных, а у EPrints имеется ограниченная поддержка современных методов NLP и нет встроенных инструментов машинного обучения.

Также можно воспользоваться системами, которые подходят для управления библиографией и ссылками, такими как Zotero [3] и JabRef [4]. Они предоставляют функционал для управления библиографическими ссылками, возможность хранения данных и простую организацию публикаций. К недостаткам этих систем можно отнести локальное хранение информации и ограниченные возможности автоматического анализа текста.

Для анализа публикаций и извлечения данных могут быть использованы такие инструменты, как GROBID [5] и Local Scholar (локальная версия Semantic Scholar) [6]. С их помощью можно автоматически извлекать метаданные публикаций, проводить категоризацию и анализ данных. Однако для использования GROBID требует навыков работы с Python и настройкой серверов, а для работы Local Scholar требуется предварительное обучения модели на кафедральных публикациях.

Представленные программные продукты могут быть использованы для решения узкого ряда поставленных задач, а также имеют существенные ограничения в использовании. При создании проприетарной системы обработки и структурирования документов необходимо учесть преимущества и недостатки каждой отдельно описанной выше системы, чтобы конечный продукт удовлетворял выявленным требованиям.

Анализ методов машинного обучения, используемых для разработки интеллектуальных модулей системы

При разработке интеллектуальных модулей системы обработки и структурирования документов предполагается использование методов машинного обучения для реализации функционала извлечения и структурирования данных. Для решения этих задач наиболее универсальным и эффективным подходом является использование архитектуры «Трансформер» [7]. Данную модель можно задействовать при анализе естественного текста в информационном поиске, а также в извлечении необходимой информации из текста.

Для улучшения производительности следует использовать предварительно обученные модели, такие как RoBERTa [8], BART [9] и T5 [10]. В статье «Анализ эффективности трансформеров для решения некоторых задач NLP» Прошиной М.В. и Виноградова А.Н. [11], подробно рассмотрены результаты применения данных моделей при решении различных задач NLP. В ходе испытаний наилучшую эффективность во всех видах задач продемонстрировала модель RoBERTa. Авторы отмечают, что трансформеры на данный момент еще недостаточно совершенные модели, однако при наличии должной оптимизации можно сократить вычислительные затраты, путем использования комбинированных архитектур или повышения объема обучающих данных.

Архитектура информационной системы и внедрение интеллектуальных модулей

Предполагается, что разрабатываемая система будет иметь монолитную клиент-серверную архитектуру с клиентской и серверной частями, а также базой данных. Внедрение интеллектуального модуля в клиент-серверную систему может осуществляться различными способами в зависимости от требований к производительности, масштабируемости и задержке обработки данных. Один из наиболее простых подходов – встроенный интеллектуальный модуль, когда алгоритмы машинного обучения интегрируются непосредственно в серверное приложение. В этом случае клиент отправляет запрос на сервер, который выполняет вычисления и возвращает результат. Такой подход удобен в развертывании, минимизирует задержки и обеспечивает обработку данных без передачи их в сторонние сервисы. Однако он ограничен в масштабируемости и требует значительных вычислительных ресурсов на стороне сервера.

Более гибким решением является выделенный AI-сервис, в котором интеллектуальный модуль работает как отдельный сервис и взаимодействует с основным сервером через API. Клиент отправляет запрос на основной сервер, который передает его в интеллектуальный модуль, где выполняется анализ данных. Затем обработанный результат возвращается клиенту. Это решение повышает масштабируемость, так как интеллектуальный модуль можно развернуть на мощных серверах, а также использовать его в нескольких системах одновременно. Однако добавляется дополнительная задержка из-за необходимости сетевого взаимодействия, и усложняется управление инфраструктурой.

Другим вариантом является обработка данных непосредственно на клиенте. В этом случае предварительно обученная модель загружается в веб-приложение или мобильное приложение, а вычисления выполняются локально. Это снижает нагрузку на сервер и позволяет работать в оффлайн-режиме, но требует мощных клиентских устройств и ограничивает сложность используемых моделей.

Компромиссным вариантом является гибридная архитектура, в которой базовая обработка выполняется на клиенте, а сложные вычисления – на сервере. Например, клиент может извлекать текст из документа, а сервер – выполнять семантический анализ и классификацию. Это позволяет оптимизировать передачу данных, снизить задержки и распределить нагрузку между клиентом и сервером. Однако такая архитектура требует сложной синхронизации процессов и тщательного планирования работы системы.

Результаты сравнения вариантов внедрения интеллектуальных модулей представлены в таблице 1.

Таблица 1 - Сравнение вариантов внедрения интеллектуального модуля

DOI: https://doi.org/10.60797/IRJ.2025.156.57.1

Критерий	Встроенный модуль (на сервере)	Выделенный сервис	Модуль на клиенте	Гибридная архитектура
Производительность	Высокая	Средняя	Низкая	Высокая
Масштабируемость	Плохая	Отличная	Средняя	Хорошая
Простота развертывания	Простая	Средняя	Простая	Сложная
Задержка обработки	Минимальная	Умеренная	Минимальная	Низкая
Объем передаваемых данных	Малый	Средний	Отсутствует	Оптимизированный
Требуемые вычислительные ресурсы	Высокие (на сервере)	Высокие (на выделенном сервере)	Низкие (на клиенте)	Средние (разделены)

Проанализировав достоинства и недостатки каждого из подходов, авторы выявили, что наиболее оптимальным вариантом внедрения модуля в клиент-серверную систему является использования выделенного сервиса и взаимодействие с ним через API. Такой подход позволит настраивать и тестировать интеллектуальный модуль, не изменяя логику основного приложения. Также это позволит сохранить работоспособность приложения при возможных сбоях в работе модуля.

Заключение

В ходе исследования были проанализированы существующие решения для обработки и структурирования документов, а также методы машинного обучения, применяемые в данной предметной области. Традиционные системы не обеспечивают достаточной автоматизации, а интеллектуальные решения требуют адаптации под образовательные задачи.

Оптимальным вариантом является разработка интеллектуального модуля, интегрируемого в клиент-серверную систему. Анализ архитектурных решений показал, что наиболее эффективным является выделенный сервис, взаимодействующий с основным сервером через API. Такой подход обеспечивает масштабируемость, гибкость и удобство тестирования.

Внедрение интеллектуальных модулей позволит автоматизировать обработку научных публикаций, ускорить работу с документами и снизить нагрузку на сотрудников. Дальнейшие исследования могут быть направлены на адаптацию моделей к образовательным данным и оптимизацию их вычислительной эффективности.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Негуляев Е.А. Анализ использования электронных библиотек, построенных на системе DSpace (на примере «Электронной библиотеки Белинки») / Е.А. Негуляев // Моргенштерновские чтения — 2020. Информационно-библиографическая деятельность библиотек: тенденции, современные проекты и инициативы. — Челябинск : Челябинский государственный институт культуры, 2020. — С. 78–80.

2. Федотова О.А. Цифровой репозиторий в информационных научно-образовательных системах / О.А. Федотова, А.М. Федотов, О.Л. Жижимов [и др.] // ТрудыГПНТБ СО РАН. — 2019. — 3. — С. 23–28. DOI: 10.20913/2618-7515-2019-3-23-28

3. Коденев М.А. Библиографические менеджеры как средство повышения качества подготовки курсовых и дипломных работ / М.А. Коденев // Проблемы и перспективы развития высшего образования в сфере культуры и искусств. — Минск : Белорусский государственный университет культуры и искусств, 2023. — С. 249–253.

4. Kopp O. JabRef: BibTeX-based literature management software / O. Kopp, C.Ch. Snethlage, Ch. Schwentker // TUGBOAT. — 2023. — 3. — P. 441–447. DOI: 10.47397/tb/44-3/tb138kopp-jabref

5. Foppiano L. Automatic extraction of materials and properties from superconductors scientific literature / L. Foppiano, P.B. Castro, P. Ortiz Suarez [et al.] // Science and Technology of Advanced Materials: Methods. — 2023. — 1. DOI: 10.1080/27660400.2022.2153633

6. Jelodar H. Recommendation system based on semantic scholar mining and topic modeling on conference publications / H. Jelodar, Y. Wang, M. Rabbani [et al.] // Soft Computing - A Fusion of Foundations, Methodologies and Applications. — 2020. — 1. DOI: 10.1007/s00500-020-05397-3

7. Лисовец Я.В. Нейросеть на архитектуре трансформер / Я.В. Лисовец, М.Д. Тропин // Сборник лучших докладов студенческой научно-технической конференции, посвященной 100-летию Отечественной гражданской авиации. — Москва : ИД Академии Н.Е. Жуковского, 2023. — С. 71–76.

8. Putri Masaling N.A. Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis / N.A. Putri Masaling, D. Suhartono // International Journal of Informatics and Communication Technology. — 2024. — 3. DOI: 10.11591/ijict.v13i3.pp410-421

9. Ермоленко Т.В. Разработка алгоритмов и языковых моделей длямультиязычной системы автоматического аннотирования текстов разных жанров / Т.В. Ермоленко, В.И. Бондаренко, Я.С. Пикалёв // Вестник Донецкого национального университета. — 2023. — 2. — С. 22–43.

10. Брылев Д.В. Исследование предобученных моделей нейронных сетей для генерации текста / Д.В. Брылев // Молодая наука Сибири. — 2023. — 3. — С. 140–149.

11. Прошина М.В. Анализ эффективности трансформеров для решения некоторых задач NLP / М.В. Прошина, А.Н. Виноградов // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. — Москва : Российский университет дружбы народов (РУДН), 2023. — С. 153–157.

Список литературы на английском языке / References in English

1. Negulyaev Ye.A. Analiz ispolzovaniya elektronnikh bibliotek, postroennikh na sisteme DSpace (na primere «Elektronnoi biblioteki Belinki») [Analysis of the use of electronic libraries built on the DSpace system (using the example of the "Belinka Electronic Library")] / Ye.A. Negulyaev // Morgenstern Readings — 2020. Information and bibliographic activities of libraries: trends, modern projects and initiatives. — Chelyabinsk : Chelyabinsk State Institute of Culture, 2020. — P. 78–80. [in Russian]

2. Fedotova O.A. Tsifrovoi repozitorii v informatsionnikh nauchno-obrazovatelnikh sistemakh [Digital repository for research and education information systems] / O.A. Fedotova, A.M. Fedotov, O.L. Zhizhimov [et al.] // Proceedings of the GPNTB SB RAS. — 2019. — 3. — P. 23–28. DOI: 10.20913/2618-7515-2019-3-23-28 [in Russian]

3. Kodenev M.A. Bibliograficheskie menedzheri kak sredstvo povisheniya kachestva podgotovki kursovikh i diplomnikh rabot [Bibliographical managers as a means of improving the quality of the preparation of course works and final works] / M.A. Kodenev // Problems and prospects of development of higher education in the sphere of culture and arts.— Minsk : Belarusian State University of Culture and Arts, 2023. — P. 249–253. [in Russian]

4. Kopp O. JabRef: BibTeX-based literature management software / O. Kopp, C.Ch. Snethlage, Ch. Schwentker // TUGBOAT. — 2023. — 3. — P. 441–447. DOI: 10.47397/tb/44-3/tb138kopp-jabref

5. Foppiano L. Automatic extraction of materials and properties from superconductors scientific literature / L. Foppiano, P.B. Castro, P. Ortiz Suarez [et al.] // Science and Technology of Advanced Materials: Methods. — 2023. — 1. DOI: 10.1080/27660400.2022.2153633

6. Jelodar H. Recommendation system based on semantic scholar mining and topic modeling on conference publications / H. Jelodar, Y. Wang, M. Rabbani [et al.] // Soft Computing - A Fusion of Foundations, Methodologies and Applications. — 2020. — 1. DOI: 10.1007/s00500-020-05397-3

7. Lisovets Ya.V. Neiroset na arkhitekture transformer [Neural network on the transformer architecture] / Ya.V. Lisovets, M.D. Tropin // Collection of the best reports of the Student Scientific and Technical Conference dedicated to the 100th anniversary of Russian civil aviation. — Moscow : ID Akademii N.E. Zhukovskogo, 2023. — P. 71–76. [in Russian]

8. Putri Masaling N.A. Utilizing RoBERTa and XLM-RoBERTa pre-trained model for structured sentiment analysis / N.A. Putri Masaling, D. Suhartono // International Journal of Informatics and Communication Technology. — 2024. — 3. DOI: 10.11591/ijict.v13i3.pp410-421

9. Yermolenko T.V. Razrabotka algoritmov i yazikovikh modelei dlyamultiyazichnoi sistemi avtomaticheskogo annotirovaniya tekstov raznikh zhanrov [Development of algorithms and language models for a multi-language system of automatic summary of texts of different genres] / T.V. Yermolenko, V.I. Bondarenko, Ya.S. Pikalyov // Bulletin of Donetsk National University. — 2023. — 2. — P. 22–43. [in Russian]

10. Brilev D.V. Issledovanie predobuchennikh modelei neironnikh setei dlya generatsii teksta [Research of pre-tained neural networks models for text generation] / D.V. Brilev // Young Science of Siberia. — 2023. — 3. — P. 140–149. [in Russian]

11. Proshina M.V. Analiz effektivnosti transformerov dlya resheniya nekotorikh zadach NLP [Efficiency analysis of transformers for some NLP tasks] / M.V. Proshina, A.N. Vinogradov // Information and telecommunication technologies and mathematical modeling of high-tech systems. — Moscow : Peoples' Friendship University of Russia (RUDN), 2023. — P. 153–157.[in Russian]