УДК 004.8

КОМПЛЕКСНАЯ СИСТЕМА СЕМАНТИЧЕСКОГО АНАЛИЗА И СТРУКТУРИРОВАНИЯ ЗНАНИЙ В ОБЛАСТИ ТЕХНОЛОГИЙ ПЕРЕРАБОТКИ ДРЕВЕСНЫХ ОТХОДОВ

Петухов С. А.[1]
Научный руководитель – канд. техн. наук, доцент Молодкина Н. Р.[1]
[1]Университет ИТМО
sergei_petukhov@itmo.ru

Введение

Современная индустрия переработки древесины характеризуется избыточностью неструктурированной информации. Ежегодный рост научных публикаций и патентов превышает 10%, что создает барьер для качественного стратегического планирования. Главная проблема - «семантическая пропасть»: одни и те же процессы переработки описываются авторами с использованием разной терминологии, а ключевые параметры технологий (методы, выходы продуктов, условия синтеза) скрыты в тексте аннотаций и выводов [1].

Существующие инструменты обеспечивают лишь навигацию по метаданным, не позволяя «видеть» структуру самой технологии. Требуется переход к комплексным системам, которые способны на входе принимать «сырой» текст, а на выходе выдавать структурированную базу знаний с выявленными связями между методами переработки и конечными продуктами [1, 2].

Основная часть

Разработанная система представляет собой конвейер, где на каждом этапе решается конкретная задача извлечения и верификации данных.

На вход системы подается массив данных, собранный из агрегаторов Lens, Scopus и TypeSet.io. Параллельно разработана отраслевая таксономия, включающая 7 классов методов и более 50 категорий продуктов, что служит «эталоном» для нормализации терминов [1].

Вместо прямого классификатора используется цепочка специализированных агентов:

2.1) Analyst Agent выполняет алгоритм Multi-source Extraction. Он сканирует заголовок, аннотацию и заключение статьи, вычленяя перечень всех упомянутых методов и продуктов. Это позволяет фиксировать связи «многие-ко-многим», характерные для переработки древесных отходов.

2.2) Validation Agent: проводит сопоставление извлеченных данных с таксономией. Он отсеивает информационный шум и объединяет синонимы (например, приводя «ферментативное расщепление» к классу «Ферментация»).

В качестве ядра выбрана модель Flan-T5 Large. Выбор обоснован возможностью её эффективной эксплуатации на стандартных CPU-серверах, что обеспечивает независимость от облачных API и сохранность корпоративных данных [2].

Заключительным этапом является конвертация базы знаний в интерактивный граф связей. Здесь платформа WoodMind выступает как инструмент навигации. Визуализация позволяет мгновенно определять центральные технологические «хабы» и периферийные разработки. Анализ 20 тысяч записей выявил, что технологии получения ксилита уже достигли стадии TRL 7 - 9 (промышленное внедрение), в то время как сегмент биопластиков и наноцеллюлозы всё еще доминирует в области прикладных НИР (TRL 3 - 4) [3].

Выводы

Реализованный комплексный подход доказывает, что автоматизация анализа технологических данных возможна даже в узкоспециализированных отраслях. Главным результатом стало создание системы, которая не просто находит статьи, а структурирует заложенные в них знания. Сокращение времени на патентный и технологический поиск в 10-15 раз открывает новые возможности для R&D-департаментов. Внедрение подобных систем позволяет объективно оценивать ландшафт технологий и фокусировать ресурсы на наиболее перспективных методах глубокой переработки древесной биомассы [3].

Литература

1. Shashi Kant Bhatia, Sang Hyoun Kim, Jeong Jun Yoon, Yung Hun Yang Current status and strategies for second generation biofuel production using microbial systems // Energy Conversion and Management. - 2017. - №148. - С. 1142-1156.
2. Leif J Jönsson, Carlos Martín Pretreatment of lignocellulose: Formation of inhibitory by-products and strategies for minimizing their effects // Bioresour Technol. - 2016. - №199. - С. 103-112.
3. Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.