

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АВТОМАТИЗИРОВАННЫХ МЕТОДОВ СБОРА И ОБРАБОТКИ ДАННЫХ О ПЕРЕРАБОТКЕ ЛИГНОЦЕЛЛЮЛОЗНОЙ БИОМАССЫ

Петухов С.А. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Молодкина Н.Р.
(ИТМО)

Введение. Объем научных данных непрерывно растет, что требует разработки специальных подходов для сбора и анализа полученной информации. В области экологических изысканий особенностью данного поиска является также мультидисциплинарность, что часто осложняет поиск. В работе рассмотрен подход к решению этой задачи для темы микробиологической переработки лигноцеллюлозной биомассы из отходов древесины березы и осины. Однако значительная разрозненность этих данных, представленных в научных статьях, патентах и других источниках с 1898 по 2025 годы, усложняет их систематизацию и анализ для практического применения. В качестве специфических критериев требовалось представить информацию о продуктах, методах их получения, экологической и экономической эффективности. В этой связи возникает необходимость разработки эффективных методов автоматического сбора и обработки информации. В работе были использованы современные инструменты, такие как TypeSet.io, нейросети и машинное обучение (ML). Данное исследование направлено на формирование базы данных о микробиологической переработке биомассы путем поиска и анализа релевантных источников, определения критериев структурирования и сравнения различных подходов к обработке данных, с целью создания читаемого и полезного научного отчета.

Основная часть. Данное исследование направлено на поиск и выбор максимально эффективных способов автоматического сбора и анализа данных о микробиологической переработке биомассы березы и осины с определением критериев их структурирования [3]. Работа включает следующие этапы:

- 1) Составление комплексного запроса, включающего переработку биомассы березы и осины для получения продукции, включая список исключений
- 2) Определение критериев структурирования данных. На основе технического задания были выделены ключевые параметры: название продукта, способ получения, технология, экономическая эффективность и экологическая безопасность. Эти критерии легли в основу формирования запросов.
- 3) Сбор данных с использованием TypeSet.io. Инструмент позволил собрать данные из сотен статей, патентов и книг за период с 1898 по 2025 годы. Было обработано 224 файла по запросам “переработка древесины и целлюлозы” и “переработка березового баланса”, общее количество строк составило 20700 строк и 17 столбцов.
- 4) На следующем этапе нужно было из текста, описывающего технологию сделать заключение об уровне TRL, а также идентифицировать конкретное название получаемого продукта и метода, с помощью которого оно получается. На этом этапе рассматривали возможность применения машинного обучения и нейросетей [2].

Сравнительный анализ платформ показал, что TypeSet.io обладает преимуществами в скорости обработки и точности структурирования по сравнению с аналогами, такими как Base.net или Google Scholar, благодаря интеграции с широким спектром источников.

Выводы. В ходе исследования была сформирована база данных о микробиологической переработке биомассы березы и осины, содержащая более 20 700 строк, на основе анализа 224 файлов из научных статей, патентов и учебных пособий за период с 1898 по 2025 годы, выгруженных с использованием TypeSet.io, что позволило автоматизировать сбор данных, обеспечив их структурирование по ключевым критериям: название продукта, способ получения, технология, экологическая и экономическая эффективность [1]. Сравнительный

анализ показал преимущества TypeSet.io над аналогами (например, Base.net) в скорости и качестве структурирования, а также выявил различия в эффективности нейросетей и ML: нейросети лучше справились с контекстным анализом текстов, тогда как ML показал ограничения в обработке неструктурированных данных. Полученная база данных стала основой для научного отчета, а на следующем этапе планируется её очистка от нерелевантных значений и адаптация подхода для анализа патентных выгрузок, что открывает перспективы для дальнейших исследований и визуализации результатов.

Список использованных источников:

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. Akshay R. Mankar, Ashish Pandey, Arindam Modak, K.K. Pant Pretreatment of lignocellulosic biomass: A review on recent advances // Bioresource Technology. - 2022. - №334.