171. АВТОМАТИЧЕСКИЙ АНАЛИЗ ВЕБ-РЕСУРСОВ БЕЛОРУССКИХ УНИВЕРСИТЕТОВ НА ОСНОВЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ И НЕЙРОСЕТЕВЫХ ЭМБЕДДИНГОВ

Усова В.А, студент гр. 372302

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Федосенко В.А. – к.т.н., доцент кафедры ЭИ

В работе выполнен автоматический анализ и ранжирование сайтов белорусских университетов методами машинного обучения. На основе данных, собранных парсингом HTML главных страниц, вычислены матрицы сходства по коэффициентам Сёренсена-Дайса и косинусному расстоянию TF-IDF. Для группировки проведена кластеризация методами K-Means и иерархическая кластеризация, для визуализации использовался метод главных компонент, для семантического анализа – нейросетевые эмбеддинги. По результатам взвешенной оценки структурных признаков построен финальный рейтинг исследуемых веб-ресурсов.

Ключевые слова: кластеризация, коэффициент Сёренсена-Дайса, TF-IDF, K-Means, иерархическая кластеризация, метод главных компонент, нейросетевые эмбеддиинги.

Цифровое присутствие учреждения высшего образования во многом определяется качеством его официального интернет-ресурса. Будущие студенты изучают условия приёма, нынешние – получают доступ к расписанию и материалам, а потенциальные партнёры формируют первоначальное суждение о вузе именно по сайту. Это превращает характеристики веб-платформы из сугубо технической задачи в элемент конкурентного позиционирования учреждения.

Привычная практика оценки сайтов силами экспертов имеет два принципиальных недостатка: зависимость от индивидуального суждения проверяющего и трудозатратность, ограничивающая охват. Переход к вычислительным методам снимает оба ограничения: формализованные алгоритмы извлекают числовые характеристики страниц, устанавливают количественную меру схожести между ресурсами и выстраивают упорядоченный рейтинг без участия субъективного фактора.

Для анализа отобраны десять белорусских учреждений высшего образования – БГУИР, БГУ, БНТУ, БГМУ, БГТУ, ВГМУ, БРУ, ГрГУ, ПГУ, ГГУ – охватывающих инженерный, медицинский и гуманитарный профили. Разнородность специализаций придаёт сравнению содержательность. Задача исследования – создать и проверить на реальных данных инструментарий объективной оценки университетских веб-платформ.

Текстовое наполнение главных страниц получено путём автоматического разбора HTML-разметки. Исключение составил сайт БГУ, поскольку его содержимое формируется динамически средствами JavaScript и недоступно при статическом запросе, соответствующие данные зафиксированы вручную по отображению в браузере. Область анализа намеренно ограничена статически доступными элементами стартовой страницы – внутренние разделы, параметры производительности и мобильная адаптивность остаются за рамками настоящей работы.

Из доступного множества атрибутов к дальнейшей обработке допущены шесть, обладающих различающей способностью: число гиперссылок на странице, глубина основного меню, мощность уникального словаря, отношение словарного объёма к количеству изображений, сводный индикатор навигационного охвата, а также бинарный признак наличия поискового модуля. Атрибуты с нулевой дисперсией по выборке были исключены. Приведение признаков к единому диапазону [0,1] осуществлено нормировкой Min-Max; БГУИР по итогам нормировки занял лидирующую позицию по показателям навигационной структуры и охвата.

Степень лексического перекрытия между парами сайтов измерялась посредством коэффициента Дайса. Значения метрики лежат в единичном отрезке. Нулевое значение соответствует полному отсутствию общей лексики. Построенная матрица близости указала на слабое межсайтовое пересечение словарного состава: все ненулевые внедиагональные элементы уложились в диапазон 0,04–0,34, что свидетельствует о преимущественно уникальной лексике каждого ресурса. Наибольшее перекрытие зафиксировано для пары БГУИР–БГТУ (0,34) в силу сходства технических предметных областей.

Частотно-взвешенный подход TF-IDF кодирует каждый текст числовым вектором, где значимость термина пропорциональна его встречаемости в документе и обратно пропорциональна распространённости в коллекции. Близость пары документов выражается через косинус угла между векторами, что делает меру нечувствительной к длине текста. Подавление роли частотных общеупотребительных слов и акцент на специфической терминологии обусловливают заметно более низкие значения по сравнению с коэффициентом Дайса: большинство пар укладывается в интервал 0,05–0,19. Лидерство пары БГУИР–БГТУ сохраняется и здесь (0,19).

Агломеративный алгоритм Варда последовательно объединял ближайшие пары объектов, минимизируя суммарный прирост внутригрупповой дисперсии, и сформировал иерархию кластеров в виде дендрограммы без фиксации их числа заранее. При пороге 0,65–0,8 выделяются три кластера: первый объединяет ГрГУ, БРУ, БГУИР и БГТУ; второй – БНТУ, БГМУ, ВГМУ и ГГУ; третий образуют БГУ и ПГУ, которые сливаются лишь при расстоянии около 0,92, что указывает на наибольшее словарное разнообразие этих двух ресурсов.

Выбор числа кластеров для алгоритма K-Means основывался на графике зависимости инерции от k. Переход от двух групп (инерция 3,98) к трём (2,29) дал наибольший абсолютный выигрыш – 1,69; все последующие шаги приносили существенно меньший эффект, что однозначно указало на k=3 как оптимальное решение. Алгоритм выделил группу лидеров – БГУИР, БГМУ, ВГМУ с развитой навигацией и поиском, – промежуточную группу из БГУ, БГТУ, БРУ, ГрГУ и ГГУ, а также слабую группу – БНТУ и ПГУ.

Понижение размерности методом главных компонент проецировало шестимерное пространство признаков в плоскость для визуализации взаимного расположения ресурсов. Первая главная ось аккумулирует 53,6% суммарной дисперсии и трактуется как интегральная насыщенность страницы; вторая (19,5%) характеризует соотношение текстовой и структурной составляющих. Совместно они сохраняют 73,2% информации, достаточно для содержательных выводов. Тройка лидеров – БГУИР, БГМУ, ВГМУ – устойчиво концентрируется в зоне высоких координат по первой оси. Профильная диаграмма по шести измерениям дополнительно подтверждает, что БГУИР формирует наиболее равномерно развитый профиль среди пяти ведущих ресурсов.

Семантический анализ реализован посредством многоязычной нейронной модели paraphrase-multilingual-MiniLM-L12-v2, основанной на архитектуре Transformer и обученной на корпусе текстов 50 языков. Каждая страница отображалась в вещественный вектор размерности 384, а близость пар ресурсов оценивалась тем же косинусным критерием. В отличие от частотных методов, нейронная модель улавливает предметно-тематическое единство всей коллекции, что проявляется в существенно более высоких значениях близости: 0,6–0,88. Среднее семантическое сходство БГУИР со всеми прочими ресурсами оказалось максимальным (0,82), а наибольшие попарные значения зафиксированы с БГТУ и ГрГУ (по 0,88). Иерархическая кластеризация по семантическим векторам выделяет компактную группу технических вузов БНТУ, БГУИР, БГТУ (расстояние 0,12–0,15) и пару медицинских БГМУ, ВГМУ.

Итоговый балл каждого ресурса вычислялся как взвешенная сумма нормированных структурных показателей. Удельный вес навигационного меню принят равным 0,4 как наиболее показательного индикатора организованности ресурса, поисковый модуль и информативность текста учтены с весами по 0,2, плотность ссылок и сводный навигационный индекс – по 0,1. В итоговой таблице БГУИР занял первое место (0,889), следом – БГМУ (0,834) и ВГМУ (0,781).

Первенство БГУИР воспроизводится при четырёх различных аналитических подходах: структурном профилировании, обоих алгоритмах кластеризации, проекции на главные компоненты и семантическом нейросетевом сравнении. Согласованность разнородных методов свидетельствует о достоверности и устойчивости полученного рейтинга.

Исследование подтверждает, что вычислительные методы способны обеспечить объективную, воспроизводимую и масштабируемую оценку университетских веб-платформ. Нейронные языковые модели образуют хорошее дополнение к классическим частотным метрикам: они показывают тематические закономерности группировки, которые остаются скрытыми при анализе поверхностного лексического состава.

Список использованных источников:

Жилов Р. А. Интеллектуальные методы кластеризации данных // Молодой учёный. 2023. № 45. URL: https://cyberleninka.ru/article/n/intellektualnye-metody-klasterizatsii-dannyh (дата обращения: 14.05.2026).

Иванов И. П., Вишняков И. Э., Каркин И. А. Выявление и кластеризация шаблонных текстов в больших массивах сообщений // Труды СПИИРАН. 2022. № 3. URL: https://cyberleninka.ru/article/n/vyyavlenie-i-klasterizatsiya-shablonnyh-tekstov-v-bolshih-massivah-soobscheniy (дата обращения: 14.05.2026).

Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. URL: https://arxiv.org/abs/1908.10084 (дата обращения: 14.05.2026).

Scikit-learn: Machine Learning in Python. Текст : электронный // Scikit-learn : офиц. сайт. 2024. URL: https://scikit-learn.org/stable/ (дата обращения: 14.05.2026).

Sentence Transformers Documentation. Текст : электронный // SBERT : офиц. сайт. 2024. URL: https://www.sbert.net (дата обращения: 14.05.2026).