УДК 004.77:004.8:004.6
Валігура І. – ст. гр. СП-41
Тернопільський національний технічний університет імені Івана Пулюя

РОЗРОБКА ТА ТЕСТУВАННЯ ВЕБ-ЗАСТОСУНКУ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ НАУКОВИХ ПУБЛІКАЦІЙ З ВИКОРИСТАННЯМ NCBI ENTREZ API

Науковий керівник: к.т.н., доцент Багрій-Заяць О. А.

Valihura I.
Ternopil Ivan Puluj National Technical University

DEVELOPMENT AND TESTING OF A WEB APPLICATION FOR INTELLIGENT ANALYSIS OF SCIENTIFIC PUBLICATIONS USING NCBI ENTREZ API

Supervisor: Bahrii-Zaiats O.

Ключові слова: PubMed, API, інтелектуальний аналіз, Python, агрегація даних
Keywords: PubMed, API, intelligent analysis, Python, data aggregation

Вступ. Експоненціальне зростання кількості наукових публікацій, зокрема в базі PubMed, налічує понад 36 млн записів із щорічним приростом у 1.5 млн статей, створючи критичне інформаційне перевантаження для дослідників [1, 2]. Наявні інструменти, зокрема PubMed Labs, використовують алгоритми які не забезпечують автоматичної фільтрації результатів за критеріями методологічної якості та рівня доказовості. Метою роботи є розробка веб-застосунку для автоматичного відбору публікацій із найвищим рівнем доказовості. На відміну від стандартних систем, програма аналізує зміст анотацій для пріоритезації статей за типом дизайну дослідження та обсягом вибірки учасників.

Матеріали та методи. Об’єктом дослідження є анотації та метадані з бази PubMed, отримані через NCBI Entrez API. Програмна частина системи написана на мові Python із використанням бібліотеки Biopython, автоматично розбирає технічний код від сервера PubMed і перетворює його на структуровані об'єкти даних [3]. Інтелектуальний аналіз анотацій реалізовано через алгоритми патерн-метчингу. Система автоматично детектує та структурує ключові параметри доказовості (тип дослідження, кількісні показники вибірки), що є вхідними даними для математичної моделі ранжування.

Ключовим компонентом системи є розроблений алгоритм ранжування, що базується на математичній моделі зваженого сумування критеріїв якості публікації. Для оцінки знайдених матеріалів алгоритм обчислює підсумковий рейтинг доказовості R за формулою:

$$R = (K_d \cdot w_1) + (\log(n) \cdot w_2) + (I_a \cdot w_3) \quad (1)$$

де K_d — ієрархічний коефіцієнт дизайну дослідження (від 1 для описів випадків до 10 для мета-аналізів), $\log(n)$ — логарифмічно нормалізований обсяг вибірки для згладжування статистичних розбіжностей, I_a — індекс актуальності на основі дати

публікації та цитованості, а w_1, w_2, w_3 — вагові коефіцієнти пріоритетності параметрів.

Процес розрахунку та аналізу реалізовано як асинхронну чергу задач для забезпечення швидкості роботи інтерфейсу. Для оцінювання системи використано методику порівняльного аналізу, де швидкість роботи вимірювалася часом відгуку при обробці до 500 анотацій одночасно. Розрахунок метрики точності (Precision) проводився шляхом зіставлення даних, знайдених алгоритмом, із результатами ручної перевірки 100 випадкових публікацій.

Результати. Розроблений веб-застосунок автоматизує процес вибору наукових публікацій, перетворюючи неструктуровані тексти анотацій на ранжований список за критеріями доказовості. Експериментальна перевірка швидкодії на масивах до 500 записів підтверджує ефективність асинхронної архітектури: час очікування користувача не перевищує 2 секунд, оскільки основне навантаження з парсингу XML-коду та виконання математичних обчислень розподіляється у фоновому режимі.

Аналіз точності алгоритму (Precision) демонструє стабільні показники: 94% для ідентифікації числових даних вибірки (n) та 92% для класифікації дизайну дослідження. Виявлено, що найвищу точність система показує при обробці анотацій рандомізованих клінічних досліджень (РКД) завдяки стандартизованій структурі тексту (CONSORT), тоді як у звітах про клінічні випадки точність дещо знижується через варіативність термінології.

Застосування логарифмічної нормалізації у формулі R дозволяє збалансувати видачу, запобігаючи домінуванню статей з екстремально великими вибірками над методологічно сильнішими роботами (наприклад, мета-аналізами). Встановлено закономірність: при стандартних вагових коефіцієнтах перші позиції результатів пошуку займають найбільш надійні типи публікацій — мета-аналізи та систематичні огляди.

Порівняльний аналіз із алгоритмом PubMed Best Match показав, що авторська система скорочує час на первинну селекцію матеріалів на 60%. У той час як стандартна видача PubMed вимагає від дослідника ручного перегляду десятків анотацій для пошуку обсягу вибірки, розроблений інтерфейс візуалізує ці параметри миттєво. Гнучке налаштування ваг w дозволяє адаптувати систему під різні задачі: від пошуку найбільш актуальних даних до відбору масштабних епідеміологічних досліджень.

Висновки. Розроблений веб-застосунок на основі асинхронної архітектури та математичної моделі ранжування забезпечує швидкий і об’єктивний відбір медичних публікацій за критеріями доказовості. Висока точність ідентифікації параметрів дослідження та можливість гнучкого налаштування вагових коефіцієнтів дозволяють скоротити час на первинну селекцію наукової інформації на 60%.

ЛІТЕРАТУРА

1. The strain on scientific publishing / M. A. Hanson, P. G. Barreiro, P. Crosetto, D. Brockington. *Quantitative Science Studies*. 2024. Vol. 5, No. 4. P. 823–844. DOI: 10.1162/qss_a_00327.
2. Growth rates of modern science: A bibliometric analysis based on the number of publications from 1880 to 2012 / L. Bornmann, R. Mutz. *Scientometrics*. 2015. Vol. 104, No. 1. P. 575–590. DOI: 10.1007/s11192-015-1575-3.
3. Cock P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009. Vol. 25, No. 11. P. 1422–1423.