

МОДЕЛЮВАННЯ СИСТЕМИ ЗБОРУ НОВИННОГО КОНТЕНТУ ДЛЯ АНАЛІЗУ РИТОРИКИ ПУБЛІЧНИХ ОСІБ

Вінницький національний технічний університет

Анотація

У роботі проведено порівняльний аналіз двох основних підходів до збирання новинного контенту – пакетної та потокової моделі. Здійснено вибір оптимальної моделі з урахуванням затримки, повноти даних і навантаження на систему. На основі формалізованих критеріїв доведено доцільність використання batch-моделі в задачах із невисокою частотою оновлення новин. Представлено аналітичну формулу зваженої ефективності моделі, що дозволяє адаптивно обирати архітектурне рішення під конкретні потреби системи моніторингу публічної риторики.

Ключові слова: веб-скрапінг, риторика, новини, batch-модель, потокова модель, системи підтримки рішень

Abstract

This work presents a comparative analysis of two common approaches to collecting news content: batch and stream models. Based on latency, data completeness, and system load, the batch model was identified as optimal for low-frequency update tasks. A weighted effectiveness criterion was proposed, which allows selecting an architectural solution adapted to public rhetoric monitoring systems.

Keywords: web scraping, rhetoric, news, batch model, stream model, decision support systems.

Вступ

У процесі побудови інтелектуальних систем підтримки прийняття рішень важливим завданням є організація ефективного збору новинних даних, що містять висловлювання публічних осіб. Це дозволяє виявляти зміни в риториці та потенційні суперечності. Одним з критичних рішень при проектуванні є вибір моделі обробки потоку даних.

Основна частина

У сучасному інформаційному просторі кількість новинних матеріалів, що з'являються щодня, є надзвичайно великою. Водночас публічні особи — політики, експерти, лідери думок — постійно формують свою позицію щодо суспільно значущих тем.

Ручний моніторинг таких висловлювань є надзвичайно трудомістким, а іноді й практично неможливим, особливо коли йдеться про довготривале відстеження змін риторики або виявлення суперечностей у позиціях. Саме тому виникає потреба у створенні автоматизованої підсистеми, яка здатна збирати, обробляти та зберігати новинний контент у структурованому вигляді — як передумову для подальшого аналізу на основі методів обробки природної мови. Збір інформації з веб-джерел є технічно складним завданням, адже потребує врахування частоти оновлення, форматів публікацій, можливих дублікатів та змін у структурі сторінок. Крім того, необхідно знайти баланс між оперативністю збору та навантаженням на систему, щоб забезпечити як стабільність функціонування, так і повноту зібраної інформації. Саме на цьому етапі постає питання вибору архітектурної моделі веб-парсингу — механізму, за яким система буде регулярно отримувати новини з відкритих джерел.

У роботі розглянуто дві основні моделі: пакетну (batch) та потокову (stream). Перша ґрунтується на періодичному запуску скрапера за розкладом, який за один сеанс збирає велику кількість новин і зберігає їх після перевірки на унікальність. Друга працює в режимі реального часу або наближеному до нього — вона постійно "слухає" джерело, отримуючи новину одразу після її публікації. Потокова модель забезпечує мінімальну затримку, однак вимагає складнішої інфраструктури, більших ресурсів та створює суттєве навантаження на сайт-джерело.

Для обґрунтованого вибору між цими підходами було проведено моделювання на прикладі ресурсу, який публікує в середньому 10 новин на добу. Було враховано три ключові параметри: середню затримку між появою новини та її обробкою, інтенсивність запитів до джерела як показник

навантаження на систему, а також повноту зібраної інформації — тобто здатність системи не пропускати публікації. Для порівняння результатів застосовано інтегральний показник ефективності, що враховує всі три критерії з відповідними ваговими коефіцієнтами, де найвищий пріоритет було надано ресурсній стабільності, а не миттєвій реакції.

Результати аналізу засвідчили, що для задач із невисокою частотою оновлення контенту — таких як моніторинг риторики політичних фігур — пакетна модель є більш збалансованим і практичним рішенням. Вона забезпечує високу повноту збору при мінімальному навантаженні на систему, має просту архітектуру та є стійкою до збоїв. Незначна затримка між моментом публікації та обробки новини не є критичною, що робить цей підхід доцільним для подальшої реалізації в системі підтримки прийняття рішень, орієнтованій на аналіз публічної риторики.

Висновки

У ході дослідження було обґрунтовано необхідність створення автоматизованої підсистеми збирання новинного контенту як ключового елементу системи аналізу риторики публічних осіб. Великий обсяг текстової інформації та динамічність її оновлення потребують надійного механізму збору, який би забезпечував баланс між актуальністю, повнотою даних і ефективністю використання ресурсів. Проведене моделювання дозволило порівняти два базові підходи — пакетний і потоковий — на основі формалізованих критеріїв. Результати аналізу показали, що за умов невисокої частоти оновлення новинного потоку доцільним є застосування саме пакетної моделі. Вона дозволяє забезпечити достатній рівень оперативності при мінімальному навантаженні на систему, зберігаючи повноту й структурованість зібраних матеріалів. Таким чином, обґрунтований вибір batch-моделі створює підґрунтя для подальшої реалізації інтегрованої платформи збору, аналізу та візуалізації змін у риториці публічних осіб у рамках систем підтримки прийняття рішень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Alieinyk V. Factors of the growing role of mass communication as a tool for shaping political consciousness in the modern world // *Society and Security*. 2024. № 4(4). С. 3–13. URL: [https://doi.org/10.26642/sas-2024-4\(4\)-3-13](https://doi.org/10.26642/sas-2024-4(4)-3-13) (дата звернення: 03.06.2025).
2. Roberts H. та ін. Media Cloud: Massive Open Source Collection of Global News on the Open Web // *Proceedings of the International AAAI Conference on Web and Social Media*. 2021. Т. 15. С. 1034–1045. URL: <https://doi.org/10.1609/icwsm.v15i1.18127> (дата звернення: 04.06.2025).
3. Shahrivari S. Beyond Batch Processing: Towards Real-Time and Streaming Big Data // *Computers*. 2014. Vol. 3. P. 117–129. DOI: <https://doi.org/10.3390/computers3040117>.

Король Іван Олександрович – студент групи СА-21б факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, м. Вінниця, email: vatestsyper@gmail.com

Жуков Сергій Олександрович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій Вінницького національного технічного університету, м. Вінниця, e-mail: sazhukov@vntu.edu.ua

Korol Ivan O. – student of group CA-21b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine, email: vatestsyper@gmail.com

Zhukov Sergiy O. – Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: sazhukov@vntu.edu.ua