УДК 004.65

Шевчук О.Ф.
Нечитайло А. О.

Алгоритм виявлення дублікатів вакансій на основі семантичної схожості

Вінницький національний технічний університет

Анотація

У роботі досліджено проблему інтелектуальної фільтрації та дедуплікації вакансій, зібраних із неструктурованих джерел. Розглянуто алгоритм семантичного аналізу текстів оголошень за допомогою багатовимірних числових векторів. Запропоновано рішення на основі хмарного API OpenAI для генерації векторних представлень та розширення pgvector для СУБД PostgreSQL, що дозволяє виявляти приховані копії вакансій за метрикою косинусної схожості. Експериментальна перевірка показала зниження рівня інформаційного шуму майже вдвічі.

Ключові слова: *алгоритм, дедуплікація даних, косинусна відстань, векторний пошук, обробка природної мови, агрегація вакансій.*

Abstracts

The paper explores the problem of intelligent filtering and data deduplication of job vacancies collected from unstructured sources. The algorithm for semantic analysis of job postings using high-dimensional numerical vectors is considered. A solution based on the OpenAI cloud API for generating embeddings and the pgvector extension for PostgreSQL DBMS is proposed, which allows detecting hidden duplicates of vacancies using the cosine similarity metric. Experimental validation showed a reduction in information noise by almost twice.

Keywords: *data deduplication, cosine distance, vector search, natural language processing, job aggregation.*

Вступ

Сучасний ринок праці характеризується стрімким переходом процесів рекрутингу на децентралізовані платформи, зокрема у спеціалізовані канали та чати Telegram. Проте текстові дані у месенджерах є абсолютно неструктурованими, перевантаженими професійним сленгом та емодзі, що унеможливлює використання класичних інструментів парсингу. Головною проблемою стає масове дублювання контенту, коли одне й те саме оголошення тиражується у десятках спільнот із частково зміненим формулюванням або структурою. Традиційні лексичні методи фільтрації (точний збіг, регулярні вирази) не здатні розпізнати такі приховані копії. Вирішенням проблеми є застосування методів обробки природної мови та векторного представлення текстів. Метою роботи є розробка та оптимізація алгоритму семантичної дедуплікації вакансій за допомогою метрики косинусної схожості в багатовимірному просторі.

Результати дослідження

У процесі дослідження спроєктовано та реалізовано програмний модуль оптимізації процесу агрегації вакансій. Кожне вхідне неструктуроване повідомлення проходить обробку через хмарне API OpenAI, яке трансформує текст у щільний числовий вектор розмірністю 1536 елементів [1]. Центральною ланкою системи є алгоритм фільтрації схожих вакансій, логічна схема якого представлена на рисунку 1.

Відповідно до розробленої логічної схеми, процес розпочинається з передачі структурованої нової вакансії та її згенерованого вектора до бази даних pgvector, де зберігаються вектори існуючих пропозицій [2]. Для оптимізації швидкодії на цьому етапі застосовується векторний пошук найближчих сусідів з використанням спеціалізованих індексів HNSW або IVFFLAT [3]. Ключовим кроком алгоритму є блок обчислення косинусної відстані. Як проілюстровано на графіку всередині цього блоку, метрика оцінює кут θ між напрямком вектора нової вакансії (V_{new}) та вектора вже збереженої ($V_{existing}$).

Обчислення косинусної відстані (d) здійснюється за формулою:

$$d = 1 - \frac{V_{new} \cdot V_{existing}}{|V_{new}| \cdot |V_{existing}|}, \quad (1)$$

де d – обчислена косинусна відстань між семантичними векторами вакансій;
V_{new} – числовий вектор (ембедінг) вмісту нового повідомлення;
$V_{existing}$ – семантичний вектор вакансії, що вже зберігається у базі даних;

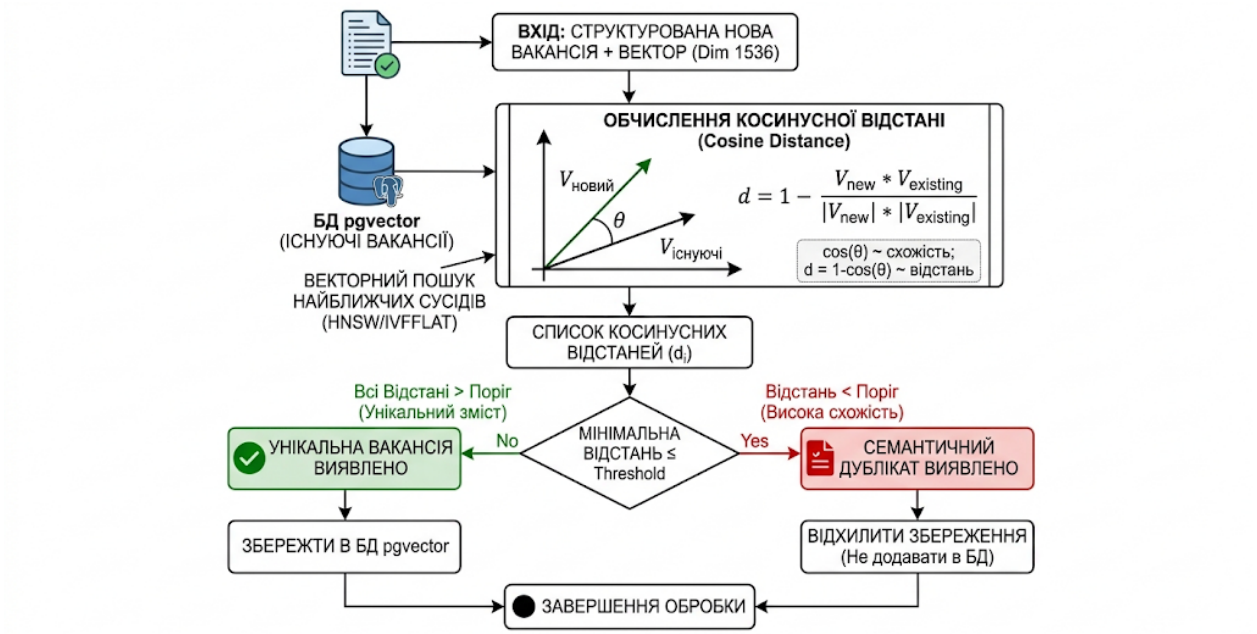

Рисунок 1 – Логічна схема алгоритму фільтрації семантичних дублікатів на основі обчислення косинусної відстані між векторами

Відповідно до наведеного на схемі пояснення, косинус кута cos(θ) відображає безпосередню семантичну схожість текстів, а його віднімання від одиниці трансформує цей показник у фактичну математичну відстань між оголошеннями. На наступному етапі сформований список косинусних відстаней (d_i) передається до блоку прийняття рішень, де мінімальна обчислена відстань порівнюється із заданим порогом чутливості системи, який зафіксовано в межах 0,91–0,93. Якщо мінімальна відстань є меншою за поріг або дорівнює йому (гілка «Yes»), це свідчить про критичну смислову схожість текстів – система маркує запис як семантичний дублікат і відхиляє його збереження. Натомість, якщо всі відстані перевищують поріг (гілка «No»), констатується унікальний зміст: алгоритм підтверджує виявлення нової вакансії, ініціює її збереження в БД і успішно завершує цикл обробки.

Експериментальна перевірка розробленого алгоритму на вибірці з 1000 реальних повідомлень показала, що близько 40% зібраного масиву становили дублікати. Застосування описаного векторного підходу дозволило ефективно їх усунути, знизивши загальний обсяг інформаційного шуму для кінцевого користувача майже вдвічі.

Висновки

У результаті проведеного дослідження поставлені науково-практичні завдання вирішено в повному обсязі. На основі аналізу процесів агрегації вакансій обґрунтовано доцільність застосування інтелектуальних методів обробки природної мови для зниження рівня інформаційного шуму та усунення дублювання контенту. Спроєктовано мікросервісну архітектуру системи та успішно реалізовано програмний модуль, що поєднує автономний збір даних через Telegram-юзербот, ШІ-структурування сирого тексту й генерацію багатовимірних векторів моделей OpenAI.

Головним результатом роботи є побудова й оптимізація алгоритму семантичної дедуплікації на основі обчислення косинусної відстані з використанням розширення pgvector для СУБД PostgreSQL та індексації HNSW. Експериментальна оцінка підтвердила високу швидкодію та відмовостійкість системи, яка дозволяє зменшити обсяг прихованих дублікатів у потоці оголошень майже вдвічі, що забезпечує суттєве розширення функціональних можливостей автоматизованого моніторингу ринку праці порівняно з існуючими аналогами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. OpenAI Platform Docs. Робота з векторними вбудовуваннями (Embeddings) за допомогою OpenAI API. Режим доступу: https://platform.openai.com/docs/guides/embeddings. – Дата звернення: 21.05.2026.
2. ServBay Support. Посібник з використання розширення pgvector PostgreSQL. Режим доступу: https://support.servbay.com/uk/database-management/postgresql-extensions/pgvector. – Дата звернення: 21.05.2026.
3. Malkov Y. A., Yashunin D. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, Vol. 42, no. 4. P. 824–836.

Шевчук Олександр Федорович – доцент кафедри комп'ютерних наук, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: shevchuk@vntu.edu.ua

Нечитайло Антон Олександрович – студент групи 5КН-22б факультету інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: anton.nechytailo@gmail.com

Shevchuk Oleksandr F. – Associate Professor of the Department of Computer Sciences, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: shevchuk@vntu.edu.ua

Nechytailo Anton O. – student, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: anton.nechytailo@gmail.com