УДК 004:89

ЛОГІКО-АГЛОРИТМІЧНА КОНЦЕПЦІЯ ПОШУКУ ТА АНАЛІЗУ ДАНИХ ПРО ВИДАТНИХ УКРАЇНЦІВ

Ковальов Д.О.[1]; Шибаєв Д.С.[2]
[1]Міжрегіональна академія управління персоналом
[2]Приватний фаховий навчальний заклад Одеський коледж комп'ютерних технологій та дизайну «Сервер»
[1]Україна, Київ; [2]Україна, Одеса
shybaiev.ds@server.odessa.ua

Анотація. *У роботі розглянуто проблематику збереження, систематизації та аналітичної обробки даних про видатних українців у контексті цифровізації культурної спадщини. Запропоновано концепцію створення інтелектуальної інформаційної системи, що об'єднує технології семантичного пошуку, обробки природної мови та машинного навчання для побудови єдиного базису знань про діячів науки, культури, мистецтва й історії.*
Ключові слова: *інтелектуальна система; пошук даних; семантичний аналіз; машинне навчання; культурна спадщина.*

Вступна частина. Проблема цифрового збереження інформації про видатних українців набуває дедалі більшої актуальності в умовах інформаційного суспільства. Багато джерел, що містять дані про історичних діячів, науковців, митців і культурних постатей, залишаються розпорошеними, неструктурованими або представленими у фрагментарному вигляді [1]. Це значно ускладнює доступ до достовірної інформації, її верифікацію та використання для освітніх, наукових або культурних цілей.

Важливість створення сучасних інтелектуальних систем для пошуку та аналізу культурно-історичних даних визначається кількома чинниками. По-перше, це потреба у формуванні достовірних баз знань, що зберігають біографічну, наукову, мистецьку й суспільну інформацію в уніфікованому форматі. По-друге, розвиток цифрових технологій створює нові можливості для автоматизованого аналізу великих обсягів даних, які містяться у відкритих джерелах, архівах, медіа та електронних бібліотеках. По-третє, актуальною стає задача інтеграції культурної спадщини у світовий цифровий простір через розробку платформ, здатних об'єднувати дані різної структури — текстові, графічні, аудіовізуальні — в єдиному аналітичному середовищі [2,3].

Традиційні пошукові системи надають користувачам лише поверхневі результати без глибинного аналізу чи контекстного зв'язку між персоналіями, історичними подіями й артефактами. Тому створення інтелектуальної системи пошуку та аналізу даних про видатних українців, яка б забезпечувала семантичне розуміння запитів, автоматичну класифікацію інформації та формування аналітичних висновків, є науково й практично доцільним напрямом розвитку інформаційних технологій у сфері культури.

Метою дослідження є розробка концепції інтелектуальної системи пошуку та аналізу даних про видатних українців, яка дозволить здійснювати автоматизоване збирання, структуризацію, семантичний аналіз і візуалізацію культурно-історичної інформації з відкритих джерел. Досягнення цієї мети передбачає побудову алгоритмічної основи системи, що поєднує методи обробки природної мови, пошуку за змістом, розпізнавання іменованих сутностей і машинного навчання для аналізу взаємозв'язків між об'єктами бази знань.

Основна частина. Існуючі бази даних, присвячені українським діячам, мають ряд обмежень. Вони переважно виконують роль довідкових порталів і не підтримують аналітичних функцій. Відсутність уніфікованих форматів збереження ускладнює інтеграцію між різними ресурсами, такими як «Українська Вікіпедія», наукові бібліотеки, електронні енциклопедії та архівні фонди. Зі збільшенням обсягу інформації виникає потреба в системах, здатних самостійно відрізняти сутності, визначати контекст, групувати дані за темами та генерувати нові знання на основі статистичного й семантичного аналізу.

Ключовим викликом є необхідність розроблення системи, що працює не лише з фактами, а й з їх взаємозв'язками. Наприклад, пошук за запитом «українські математики кінця XIX століття» повинен не просто відображати перелік осіб, а й надавати контекстні зв'язки між ними, установами, науковими школами чи публікаціями. Для цього потрібні інструменти семантичного аналізу, побудовані на принципах інтелектуальної обробки текстів.

Запропонована інтелектуальна система базується на архітектурі, що об'єднує три основні рівні: рівень збору даних, рівень аналітичної обробки та рівень візуалізації.

На рівні збору даних система використовує веб-скрапінг і API-запити до відкритих джерел (наприклад, Wikipedia, Wikidata, Google Books, архівні каталоги). Отримані тексти проходять етапи попередньої обробки — нормалізацію, токенізацію, видалення шумів і морфологічний аналіз української мови. Для цього використовуються бібліотеки spaCy, Stanza або UDPipe, адаптовані для обробки українського тексту.

Аналітичне ядро виконує розпізнавання іменованих сутностей (Named Entity Recognition, NER) для визначення осіб, географічних назв, дат, наукових установ і культурних об'єктів. Далі за допомогою алгоритмів класифікації (наприклад, Support Vector Machines або Random Forest) система групує інформацію за категоріями: наука, культура, література, політика, військова справа тощо.

На наступному етапі здійснюється побудова семантичної мережі знань, у якій вузли відповідають окремим персоналіям або поняттям, а ребра — типам зв'язків між ними: спільна діяльність, вплив, географічна або історична належність. Цей підхід реалізується з використанням графових баз даних, таких як Neo4j, що дозволяє виконувати складні аналітичні запити та візуалізувати взаємозв'язки між елементами.

Візуалізаційний рівень системи представлений інтерактивним вебінтерфейсом, який забезпечує динамічний пошук і побудову графів зв'язків. Використання бібліотек D3.js та Cytoscape.js дозволяє реалізувати інтуїтивну графічну інтерпретацію результатів пошуку у вигляді інтелектуальних мап персоналій.

Користувач може вводити запити природною мовою, наприклад «жінки-науковці Харківського університету XX століття», і система автоматично формує релевантну вибірку, аналізуючи лінгвістичну структуру запиту та відповідність у базі знань.

Для побудови таких механізмів застосовуються нейромережеві моделі трансформерного типу (BERT, mBERT або ukr-roberta), які дозволяють враховувати контекст слів і підвищують точність семантичного пошуку. Зокрема, адаптація моделі Ukrainian BERT забезпечує розпізнавання морфологічних особливостей української мови, що підвищує релевантність результатів.

Інтелектуальна складова системи реалізує механізми кластеризації персоналій за схожими ознаками, наприклад за періодом діяльності, науковим напрямом чи належністю до певної школи. Для цього застосовуються алгоритми k-means та DBSCAN, які аналізують багатовимірні простори ознак і формують логічно пов'язані групи.

Додатковий модуль статистичного аналізу дозволяє виявляти тенденції, наприклад динаміку кількості українських науковців у різних галузях за періодами або просторовий розподіл культурних центрів. Таким чином, система набуває не лише пошукового, а й аналітичного характеру, що перетворює її на науково-дослідницький інструмент.

На рівні зберігання даних передбачається комбіноване використання реляційних (PostgreSQL) та документно-орієнтованих (MongoDB) баз, що забезпечує ефективне управління як структурованими, так і неструктурованими об’єктами. Метадані про джерела, авторів і час публікацій зберігаються для подальшої перевірки достовірності інформації.

Розробка системи виконується з дотриманням принципів FAIR (Findable, Accessible, Interoperable, Reusable), що гарантує відкритість і сумісність із міжнародними науковими базами.

Практична частина розробки передбачає реалізацію прототипу вебплатформи з можливістю реєстрації користувачів, персональних добірок і формування аналітичних звітів. Для прикладу, дослідник може сформувати дашборд із візуалізацією життєвих періодів діячів, впливових центрів української науки чи мистецтва, а також взаємозв’язків між ними.

Алгоритмічно система поєднує підходи пошуку за ключовими словами, лексико-семантичного порівняння та аналізу текстових патернів. Для підвищення точності використовується гібридний метод, що об’єднує статистичний підрахунок співпадінь і нейронну оцінку семантичної близькості. Наприклад, при запиті «відомі українські хіміки XIX століття» система обчислює ймовірність належності кожної знайденої особи до відповідної категорії, формуючи рейтинг достовірності.

Результати зберігаються у вигляді графових структур, що дає змогу формувати інтелектуальні шляхи між подіями — від навчання в певному університеті до участі в культурних ініціативах. Така аналітика відкриває нові можливості для візуального відображення еволюції української наукової й культурної думки.

Висновки. Розробка інтелектуальної системи пошуку та аналізу даних про видатних українців є важливим кроком у напрямі цифрової трансформації культурної спадщини. Створення такого інструменту забезпечує інтеграцію розрізнених джерел, формування єдиної бази знань і можливість глибинного аналізу історико-культурних взаємозв’язків.

Система здатна автоматизувати процеси збору, класифікації та інтерпретації інформації, що робить її корисною не лише для науковців і освітян, а й для широкої аудиторії користувачів. Запропонована архітектура об’єднує технології обробки природної мови, машинного навчання та графових баз даних, що забезпечує гнучкість, масштабованість і високу точність пошуку.

Її впровадження сприятиме створенню нових форм цифрових досліджень, популяризації української культури у світовому інформаційному просторі та розвитку національних інтелектуальних систем з відкритим доступом.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint* arXiv:1810.04805. https://arxiv.org/abs/1810.04805

[2] Leskovec, J., Rajaraman, A., & Ullman, J.D. (2020). *Mining of massive datasets* (3rd ed.). Cambridge University Press.

[3] Boychuk, M. (2021). *Digital humanities and knowledge graphs in cultural heritage*. Springer.

Kovalov D.O.[1], Shybaiev D.S.[2]

Logical-aglorithmic concept of searching and analyzing data about outstanding ukrainians

Abstract. *The paper examines the issues of preservation, systematization, and analytical processing of data about outstanding Ukrainians in the context of cultural heritage digitalization. A concept for creating an intelligent information system is proposed, integrating technologies of semantic search, natural language processing, and machine learning to build a unified knowledge base about figures of science, culture, art, and history.*

Keywords: *intelligent system; data mining; semantic analysis; machine learning; cultural heritage.*