

ІНТЕЛЕКТУАЛЬНА ТЕХНОЛОГІЯ ВИДОБУВАННЯ ТА ВЕРИФІКАЦІЇ ЧИСЛОВИХ ТА ТЕКСТОВИХ ДАНИХ У БАГАТОЗВ'ЯЗНИХ БАГАТОСТАДІЙНИХ ГЕОІНФОРМАЦІЙНИХ СИСТЕМАХ

¹Вінницький національний технічний університет

Запропоновано вирішення проблеми підвищення обсягів та рівня достовірності інформації про стан елементів багатозв'язних багатовстадійних геоінформаційних систем (БГІС) з використанням інформаційних та інтелектуальних технологій, що особливо актуально для моніторингу стану цих об'єктів. Для розв'язання задачі використано числові та текстові дані з різних джерел (звіти, кадастри, публікації у ЗМІ та у соціальних мережах тощо), які прив'язані та верифіковані до елементів багатовстадійних геоінформаційних систем з урахуванням просторово-часових метаданих. Застосовано критерій топологічної спостережуваності як міри повноти знань для верифікації числових даних у текстах.

Вперше запропоновано метод верифікації числових даних у текстах з використанням бази знань, повнота якої оцінюється за допомогою критерію топологічної спостережуваності, що дає можливість оцінити повноту та точність видобутої інформації. Удосконалено метод формування користувачьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембедингів до просторових об'єктів і часових меж, що уможливіє точніше знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень. Набув подальшого розвитку метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і точніше зіставляє ці факти з описами масивів вод. Запропоновано інтегровану технологію, що поєднує ці методи — автоматизоване видобування числових і текстових фактів з різних джерел, їх просторово-часову прив'язку до елементів ГІС та формалізовану верифікацію з метою зменшення інформаційної невизначеності та підвищення якості цифрових двійників для підтримки рішень.

Наведено приклад застосування цієї технології на прикладі задачі аналізу стану масиву вод басейну річки Південний Буг. Дослідження показало, що індекс достовірності новин про забруднення вод може бути збільшений від 0,3 до близько 0,99, тобто у 3,3 рази.

Ключові слова: штучний інтелект, база знань, модель, інформаційна система, датасет, моніторинг, водні ресурси, екосистема.

Вступ

В сучасному світі велика кількість систем різної природи формалізується як геоінформаційні системи з геометричними мережами (ГІС з ГМ) [1]—[4]. Як відомо, ГІС з погляду призначення — це інформаційна система, яка забезпечує збирання, збереження, оброблення, доступ, відображення та поширення просторово-орієнтованих даних (просторових даних), а з погляду програмно-інформаційної реалізації — це сукупність електронних карт з умовними позначеннями об'єктів на них, баз даних з інформацією про ці об'єкти та програмного забезпечення для зручної роботи з картами і базами, як з єдиним цілим [2], [5]. Водночас, геометрична мережа — це визначений набір класів векторних об'єктів, які утворюють частину нерозривної мережі, складеної з граничних елементів, переходів і поворотів [2]—[4]. Прикладами ГІС з ГМ є усі інженерні мережі (транспортні мережі, електричні та електроенергетичні мережі, трубопроводи тощо), природні мережі (річкові системи, екомережі тощо).

Але найважливішим і складнішим підкласом ГІС з ГМ є багатостадійні багатозв'язні системи, які є не просто статичними взаємопов'язаними об'єктами — між ними здійснюється активне послідовне передавання потоків речовини, енергії, інформації. Передавання може бути як в одному напрямку (поширення речовин у насосних трубопровідних системах, поширення забруднень у воді річкових мереж у напрямку течії річки, перенесення речовин з вітром в атмосферному повітрі між певними умовними елементарними чи контрольними об'єктами), так і в обох напрямках (перенесення речовин внаслідок молекулярної дифузії між умовними елементарними чи контрольними об'єктами, за відсутності вітру у повітрі чи течії у воді тощо). При цьому, це не є простим механічним передаванням, як це прийнято у класичній фізиці для умовних тіл. В реальному світі цей процес залежить від параметрів різних стадій, тобто зміна однієї стадії (ділянки, об'єму) залежить як від її початкового стану, так і від стану інших стадій, з якими вона може бути пов'язана у різний спосіб за багатьма маршрутами. Такі системи відіграють важливу роль у нашій діяльності і є складними для досліджень. Однією з ключових проблем є інформаційна невизначеність, оскільки в реальних системах, особливо у природних, важко вимірювати усі аспекти зміни стану в будь-якій точці в будь-який час. Відкриті дані державних систем моніторингу довкілля України, як правило, надають дані, середні за день, місяць, квартал, рік (<https://ecozagroza.gov.ua/>, <http://monitoring.davr.gov.ua/EcoWaterMon/GDKMap/Index> тощо). При цьому, мережа спостережень не є достатньо розгалуженою. Особливо відчутною є проблема, пов'язана з моніторингом та управлінням станом масивів вод України.

Серед основних цілей Водної стратегії України на період до 2050 року, відповідно до вимог Водної Рамкової Директиви ЄС, є «поліпшення якісного стану водних об'єктів шляхом досягнення та підтримання «доброго» екологічного та хімічного стану масивів поверхневих вод, екологічного потенціалу штучних або істотно змінених масивів поверхневих вод, кількісного та хімічного стану масивів підземних вод». Усі просторово зосереджені поверхневі та підземні води в Україні за спеціально затвердженою методикою, гармонізованою з вимогами ЄС, розбиті на масиви вод (ділянки, затверджені законодавчо — по суті, формалізовані як ГІС з ГМ. Сутність зазначеної вимоги зводиться до того, що у кожному масиві вод (МВ) варто забезпечити стан, не нижчий «доброго» за спеціальною класифікацією, для чого спочатку цей стан аналізують, прогнозують, потім розробляють та законодавчо затверджують програму заходів для забезпечення поліпшення чи дотримання цього стану, а потім контролюють хід її виконання [9].

Про ступінь невизначеності можна судити з того, що, до прикладу, у басейні річки Південний Буг є 1090 масивів поверхневих вод, а пунктів регулярної системи державних спостережень Мінприроди — лише 50. І на деяких вимірюється далеко не весь список показників якості вод. До того ж, як відомо, для моделювання та прогнозування змін якості вод потрібно знати витрати вод, але система спостережень витрат вод є ще меншою і станції спостережень якості та кількості вод нечасто розташовані в одному місці, оскільки їх обслуговують різні відомства (Мінприроди та ДСНС України). Основними забруднювальними джерелами річкових вод в Україні є джерела антропогенного характеру як просторово-зосереджені (скидання стічних і зворотних вод), так і дифузні (змив забруднень з сільськогосподарських полів, неупорядковані зливові стоки тощо), але їх регулярний моніторинг майже не проводиться. Лише існує ведення форми 2 ТП (водгосп), де підприємства, які мають дозвіл, раз на рік подають форму про те, скільки забрано і скинуто вод і якої якості та які у них очисні системи. Щодо системи моніторингу стану атмосферного повітря ситуація є ще складнішою, оскільки система державного моніторингу якості повітря слабо розвинена. Наприклад, станом на 2021 р. у Вінницькій області було лише 2 станції спостережень за якістю та метеопараметрами атмосферного повітря і обидві розташовані у м. Вінниці. Мережа громадського моніторингу якості повітря України є значно розвиненішою (https://ecozagroza.gov.ua/map?layer=air_pollution), але для стану вод поки такої не існує і немає їх даних на якомусь загальному порталі — поодинокі спроби реалізації оперативного моніторингу вод у басейнах річок Тиса, Десна, Дністер, Сіверський Донець були, на жаль, тимчасовими, через різні проблеми. Закордонні системи є значно розвиненішими, але дороговартісними і в розгортанні, і в обслуговуванні.

Отже, є необхідність знати сучасний та ретроспективний стан усіх елементів багатостадійних ГІС з ГМ та різних факторів, які на нього впливають. У цій статті розглядаються питання збирання цієї інформації не стільки шляхом прямих вимірювань у кількісному вигляді — скільки в результаті видобування (англ. — «Data mining») з текстової інформації як з різних офіційних звітів, када-

стрів, так і просто з соцмереж та ЗМІ. Видобування текстових даних як джерело кількісної інформації є загальновідомим і активно застосовується з використанням сучасного рівня розвитку технологій штучного інтелекту та NLP (з англ. «Natural Language Processing» — оброблення природної мови, хоча частіше під цим розуміється природномовний текст [10]). Проте ключовою проблемою є забезпечення її достовірності і саме її розв’язання розглянуто у цьому дослідженні.

Мета статті — підвищення обсягів та рівня достовірності інформації про стан елементів багатозв’язних багатостадійних геоінформаційних систем з використанням інформаційних та інтелектуальних технологій.

Огляд проблем наявних рішень

Для збирання великих обсягів інформації про стан багатозв’язних багатостадійних геоінформаційних систем найкраще використати прямі вимірювання, але для забезпечення високої точності формування їх цифрових двійників краще використовувати не автоматизовані системи за участю людини, а мобільні інформаційно-вимірювальні системи або системи на базі Інтернету речей (IoT-системи) [12]—[15]. Але IoT-системи потребують постійної верифікації за даними з інших джерел.

Сучасні технології веб-скрапінгу та інтелектуального видобування числової інформації з текстових джерел є добре розвиненими, але для багатостадійних геоінформаційних систем основною проблемою практичного використання такої інформації є геоприв’язка до їхніх елементів. Простіше кажучи, треба точно знати про який(і) елемент(и) (ділянку річки, трубопровода, регіон) йдеться у повідомленні. А нестационарний характер цих систем потребує ще й максимально точну прив’язку в часі. Наприклад, як зазначено у статті авторів [11], у сховищах знань нового покоління, таких як Wikidata, YAGO4, передбачено зберігання атрибутів про час і географічне розташування [16], проте, автоматичне визначення таких просторово-часових атрибутів безпосередньо в текстових документах досі досліджено недостатньо [17]. Визначення географічних назв у тексті (топонімів) та їх прив’язки до географічних координат називається геопарсингом [16]. Геопарсинг ускладнюється різноманітністю стилів, мовних зворотів, неоднозначністю географічних назв, визначенням загального контексту тощо [18], [19]. Використання великих мовних моделей (англ. «Large Language Model» — LLM) спрощують виявлення зв’язків між сутностями у форматі структурованих текстових даних [20], що спрощує виявлення зв’язків між інформацією про різні елементи ГІС з ГМ, але вони схильні до «галюцинацій», тобто до вигадкування чи спотворення фактів. А використання варіанта генерації з доповненим пошуком (англ. «RAG»), де спочатку у базі шукають релевантні факти, а потім їх використовують для аналізу має проблеми з цим пошуком, передусім з урахуванням часової і географічної прив’язки, а також з урахуванням складної структури текстів [21].

Таким чином, мають місце такі основні проблеми:

- 1) перевірка достовірності числових даних, зібраних інформаційно-вимірювальними системами;
- 2) автоматична прив’язка даних у часі та у просторі до конкретних елементів багатостадійних ГІС з ГМ.

Перевірка достовірності інформації про стан багатостадійних систем

Пропонується для перевірки достовірності використовувати статистичний підхід, тобто що більше є випадків підтвердження певного факту з різних джерел, то вищим буде його рівень достовірності або, у разі кількісної оцінки — індекс. Недоліком такого підходу є те, що один повірений точний вимірювальний прилад може давати значно достовірнішу інформацію, ніж 10 неповірених та застарілих, тому бажано застосовувати ще й вибірккову експертну верифікацію результатів цього методу для корегування результатів такого підходу.

Для оцінювання достовірності числових даних можна використати текстові дані з числовою інформацією і навпаки.

Використаємо багатозв’язність систем, які аналізуються. Якщо якісь дані є невідомими, то їх можна оцінити за іншими даними із заданої стадії чи сусідніх стадій системи з використанням відомих математичних чи інтелектуальних моделей. Усі такі залежності можна формалізувати у вигляді графу, де змінні є вершинами, а залежності — ребрами [1]. Потрібний показник, за яким можна оцінити повноту таких зв’язків, тобто чи можна за зміною значень одних змінних (вершин графа) визначити зміну значень усіх інших змінних у різних комбінаціях. Як такий показник пропонується використати повноту топологічної спостережуваності [1], [22].

Як відомо, топологічна спостережуваність визначається на базі біхроматичного графа. Біхроматичний граф, це граф, що складається з двох типів вузлів: вузлів змінних y_i ($i = \overline{1, 5}$) та вузлів, що представляють залежності між такими змінними w_j ($j = \overline{1, 4}$). На рис. 1 показано приклад такого графу та його максимальне паросполучення (ребра, що належать до множини збігання, виділені). На прикладі добре видно, що існує декілька розв'язків задачі вибору максимального паросполучення: наприклад, замість ребра $(y_3 - w_3)$ може бути ребро $(y_4 - w_3)$. За умови, що існує макси-

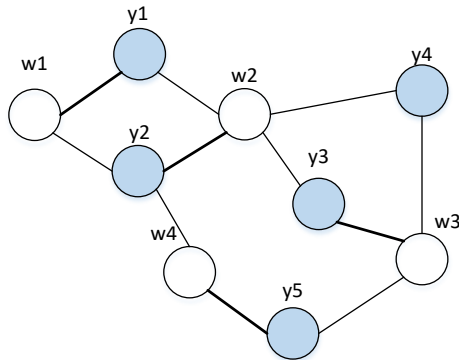


Рис. 1. Сильні ребра у біхроматичному графі для визначення топологічної спостережуваності складної системи

мальне паросполучення, де кожна змінна, що описує залежність, є інцидентною сильному ребру (тобто ребру, яке не належить до зазначеного паросполучення), то система, що описується системою таких залежностей, є повністю топологічно спостережуваною (рис. 1) [22].

Пропонується всю доступну числову інформацію максимально пов'язувати між собою з використанням математичних й інтелектуальних моделей, когнітивних карт та інших видів залежностей між змінними (ознаками). По суті, ця модель буде цифровим двійником заданої багатадійної ГІС. За методологію, викладену у роботах [1], [22], потрібно будувати граф цієї системи залежностей, по ньому будувати біхроматичний граф, аналізувати його на топологічну спостережуваність. Якщо вона не є повною, додавати ознаки чи залежності. Але, при цьому,

варто оцінювати достовірність щодо цих залежностей, оскільки деякі з них можуть бути доволі наближеними.

Аналогічно варто оцінювати індекс достовірності числових даних, які фігурують у текстовій інформації (у реченнях звітів, новин, повідомлень тощо у різний час у різних місцях).

Для оцінювання індексу достовірності J кожної новини (факту, елементу інформації у вигляді тексту чи числа) пропонується використовувати метод Байєса [26], ґрунтуючись на результатах зіставлення апіорних та апостеріорних ймовірностей того, що знайдений факт наближається до істинного

$$J_i = \frac{j_i \cdot p_i}{j_i \cdot p_{i-1} + (1 - j_i)(1 - p_{i-1})}, \quad i = 1, 2, 3, \dots, N, \quad (1)$$

де p_i — апіорна достовірність на кроці i , а j_i — індекс достовірності визначеної новини на кроці i ; $i = 1, 2, 3, \dots, N$ — кількість зіставлень новин.

А тоді відкриваються такі можливості:

1) якщо оцінка індексу великого текстового датасету вища за достовірність цифрового двійника, створеного на основі числової інформації, тоді цифровий двійник можна уточнювати, використовуючи датасет;

2) якщо індекс достовірності даних та залежностей щодо цифрового двійника має вищу оцінку індексу, ніж окремі текстові дані, тоді варто корегувати числові значення у тексті;

3) якщо зіставлення текстових даних з цифровим двійником виявило, що наявних залежностей недостатньо для порівняння, фактично, це означатиме появу нової вершини (змінної) у цифровому двійнику, для якої не вистачатиме залежностей, а отже його спостережуваність вже не є повною, тоді варто провести його удосконалення та знову забезпечити його повноту, після чого зіставлення варто повторити.

Цей метод дозволяє максимізувати індекс достовірності та адекватність і цифрового двійника кожного елемента (стадії, ділянки) багатозв'язної багатадійної ГІС з ГМ, побудованого на основі числових даних, і текстових даних з числовою інформацією про цей елемент. А потім цифровий двійник та текстові дані можна використати для аналізу та узагальнення інформації, для підтримки прийняття рішень щодо поліпшення стану цього елемента системи та для інших цілей.

Автоматична прив'язка даних у часі та просторі до елементів ГІС з ГМ

Пропонується застосовувати метод, викладений у роботах [23], [24] для автоматичної прив'язки

даних у часі та просторі до елементів ГІС з ГМ, який базується на вилученні ключових слів (n -грам), аугментації даних та оптимізованому геоприв'язуванню на основі коефіцієнта Жаккара. Оскільки раніше проведені дослідження показали складність узгодження, порівняння та недостатню точність загально відомих алгоритмів видобування ключових слів, розроблено уніфіковану нотацію та графічне представлення етапів кожного методу у ланцюжку видобування n -грам, завдяки чому вдалось проаналізувати відомі методи та підходи і синтезувати нові гібридні, що продемонстрували вищу ефективність.

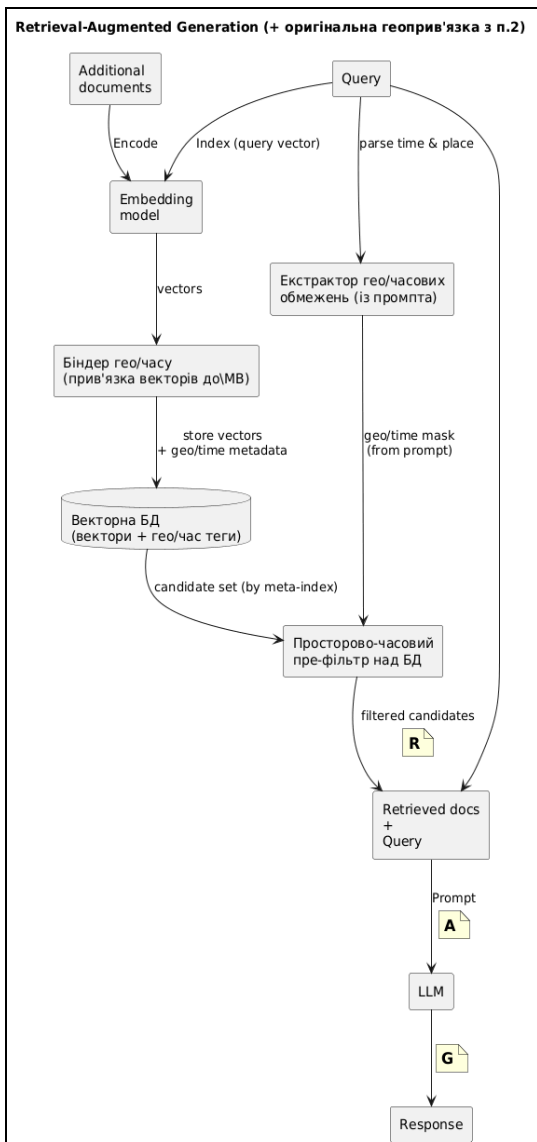


Рис. 2. Інтеграція Retrieval-Augmented Generation з геоприв'язкою

Для вирішення проблеми формування достатньої кількості навчальних даних україномовних текстів про стан МВ (елементів ГІС) зі встановленими зв'язками між ними формалізовано річкову мережу як багатозв'язну геоінформаційну систему (БГІС) та удосконалено алгоритм прив'язування даних до БГІС, запропоновано алгоритм процесу аугментації даних з верифікацією. Для підвищення точності автоматичної прив'язки даних до ГІС пропонується класифікувати видобуті ключові слова відповідно до рівнів: глобальні, регіональні та локальні, де кожному рівню визначено свої відповідні ваги.

Технологічно оперування векторною базою даних можна реалізувати за RAG-архітектурою (рис. 2). Перед основною обробкою користувацький запит аналізується для визначення часових та просторових обмежень (наприклад, якщо користувач написав «...за останні 5 років...»), то вихідні дані не будуть включати старіші дані). За допомогою ембедингів документи перетворюються на векторні дані, де кожен вектор містить корисну інформацію щодо конкретного масиву вод та метадані (гео-, часові). Може додаватись й інша атрибутивна інформація, наприклад, віднесення її до певних класів (для вод — це різні види моніторингу, факти виявлення забруднень, управлінські заходи тощо). Для виконання пошуку на першому етапі формується набір кандидатів, після чого просторово-часовий фільтр відсіює ті, що не відповідають встановленим користувачем обмеженням. На наступному етапі RAG-модуль передає релевантні записи до великої мовної моделі, яка формує остаточну відповідь, базуючись на цих записах та на знаннях отриманих під час навчання. В результаті цього підходу підвищується точність результатів.

Для вирішення проблеми формування достатньої кількості навчальних даних україномовних текстів про стан МВ (елементів ГІС) зі встановленими зв'язками між ними формалізовано річкову мережу як багатозв'язну геоінформаційну систему (БГІС) та удосконалено алгоритм прив'язування даних до БГІС, запропоновано алгоритм процесу аугментації даних з верифікацією. Для підвищення точності автоматичної прив'язки даних до ГІС пропонується класифікувати видобуті ключові слова відповідно до рівнів: глобальні, регіональні та локальні, де кожному рівню визначено свої відповідні ваги.

Інтелектуальна технологія видобування та верифікації числових та текстових даних у багатозв'язних багатостадійних ГІС та приклад її застосування

Пропонується структура інтелектуальної технології видобування та верифікації даних, що працює на базі автоматичного механізму збору текстових та числових даних з соцмереж, ЗМІ, звітів та інших джерел (рис. 3).

Система класифікації відносить кожний факт до певної категорії чи класу (стан, проблема, захід та ін.), після чого видобуває наявні доступні просторово-часові метадані для конкретного факту. Для геоприв'язки факту використано ГІС-довідники та коефіцієнти Жаккара і MLT-ISOF, що шукають збіги описів масивів вод зі згаданими фактами у текстах. Після чого дані верифікуються, зокрема, вибірково експертом та потрапляють до бази знань з подальшим оновленням RAG. Цей підхід підвищує точність геоприв'язки і дозволяє створити та підтримувати актуальність знань у базі.

Практично, до прикладу, для задачі моніторингу стану масивів вод басейну р. Південний Буг, це може бути реалізовано у такому вигляді [25]:

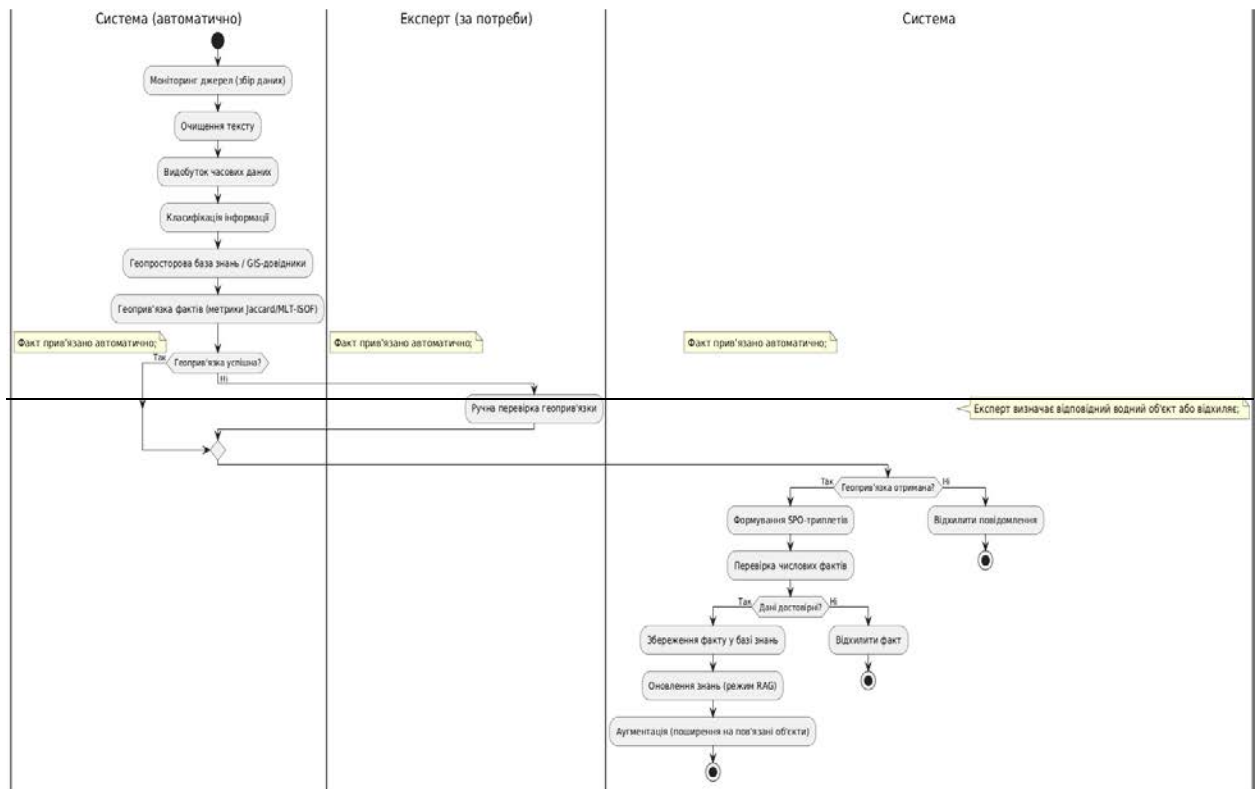


Рис. 3. Процес видобування та верифікації даних у системі

1. До відповідного модулю системи підвантажується узгоджена інформація щодо кожного масиву вод, після чого обробляється та створюються текстові описи кожного масиву вод.
2. Отримані з п.1 описи зв'язуються з відповідними об'єктами ГІС.
3. За допомогою веб-скрапінгу та доступних джерел даних до системи додаються нові записи.
4. Далі за допомогою розроблених методів відбувається автоматична класифікація за типом та геоприв'язка до масивів вод, у випадку, якщо вона неможлива — такі дані будуть перевірені експертом, також вибірково певні результати автоматичної прив'язки теж будуть перевірені експертом.
5. Оцінюється індекс достовірності кожного даного (потенційного факту) шляхом зіставлення з іншими даними за формулою (1). У разі збігу з іншими даними, оцінка індексу достовірності цього даного J зростає.
6. Сформовані дані зберігаються у базі знань та стають доступними до роботи у відповідних модулях системи (мапа, пошук тощо).

Ключовою перевагою запропонованої технології є те, що вона дозволяє зіставляти у часі та просторі багато різномірної числової та текстової інформації з різним індексом достовірності. За рахунок структури критерію (1), що більше потенційних фактів вдасться зіставити, то більше буде впевненості у них, а отже, можна припустити, що й вищим буде індекс їхньої достовірності. У разі присвоєння цього рівня J усім даним, які зіставлялись і збіг яких вдалось довести, вдасться досягти одночасно підвищення обсягів та індексу достовірності інформації про стан елементів БГІС, в цілому.

У роботі [25] авторами запропоновано оригінальну назву такої інформаційної системи: «WISEST-SBB» (англ.: Water Information System with Spatial and Temporal References for the Southern Bug Basin — «Водна інформаційна система з просторовою і часовою прив'язкою для басейну Південного Бугу»).

Приклад застосування технології

Продемонструємо приклад застосування запропонованої інформаційної технології. На рис. 4а зображено новину від 18 травня 2018 року, яка опублікована на сайті [27] і пов'язана з мором риби. Розроблена технологія уможливила здійснити геоприв'язку цієї інформації до масиву вод № UA_M5.4_00019 (Ладизинське водосховище, р. Південний Буг) та, відповідно — до числових даних моніторингу стану цього масиву вод Державним агентством водних ресурсів України на посту біля с. Маньківка (рис. 4б). На рис. 4в показано графове подання моделі показників з рис. 4б,

яке дозволяє прогнозувати їх зміну в часі і враховувати взаємовплив на основі рівнянь евтрофікації, внаслідок якої концентрація розчиненого у воді кисню зменшується, у разі зростання кількості біогенних речовин.



Рис. 4. Результат видобування та геоприв'язки інформації про потенційний мор риби у р. Південний Буг біля м. Ладизина 18.05.2018 р.: *а* — фрагмент скріншота з текстом новини [27]; *б* — фрагмент бази даних державного моніторингу; *в* — графове подання моделі показників з рис. 4б

На рис. 5а показано біхроматичний граф, який відповідає графовій моделі, наведеній на рис. 4в, побудований за описаною вище технологією. Як видно, з кількості сильних (зелених) ребер, модель є топологічно повністю спостережуваною, тобто рівень спостережуваності дорівнює 1. Однак, на моделі на рис. 4в відсутня вершина «мор риби». Пропонується її доповнити інтелектуальною моделлю, яка на основі аналізу даних, наприклад з рис. 4б, передбачатиме мор риби (якраз на рис. 4б добре видно, що з квітня по травень концентрація біогенних речовин («NO₃», «PO₄») різко зросла, а кисню — різко впала у декілька разів. Такі фактори дозволяють доволі впевнено прогнозувати мор риби). Ця нова модель на біхроматичному графі додана у вигляді червоних ліній (рис. 5б). Як легко переконатись, оптимізований біхроматичний граф теж є повністю топологічно спостережуваним (4 зелених «сильних» ребра — 4 вершини).

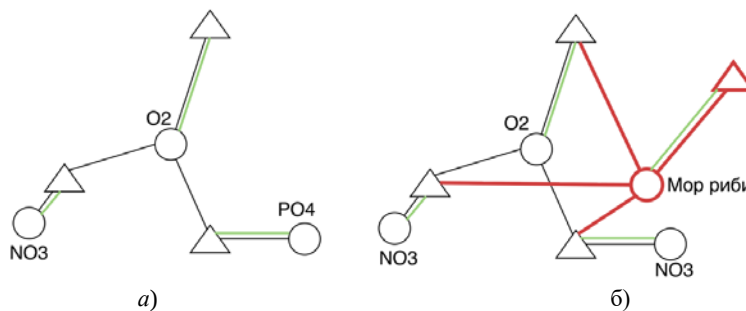


Рис. 5. Біхроматичні графи для моделей з рівнем топологічної спостережуваності 1: *а* — відповідає графовій моделі на рис. 4в; *б* — оптимізований (*а*) під додаткову вершину «мор риби»

Але, якщо даних для тренування дійсно адекватної інтелектуальної моделі для графа рис. 5б буде недостатньо, тоді важливо зіставити різні факти про наявність мору риби з інших джерел, які вдалось прив'язати до цієї ділянки (масиву вод) та до цього ж проміжку часу (рис. 6).

Достовірність події, J: 0.3

date	title	link	text	j
0	2018-05-15	-	https://www.facebook.com/plugins/post.php?href... Це П. Буг, район порохових складів (бакланів) ...	0.3

Рис. 6. Приклад свідчень очевидця про мор риби на Південному Бугу у масиві вод №UA_M5.4_00019 (біля м. Ладизина) у травні 2018 року

Експерти у сфері державного моніторингу вод пропонують брати такі базові індекси достовірності для різних видів джерел інформації: слова жителів-очевидців, блогерів та інших релевантних, але неперифікованих, джерел — 0,3; дані з офіційних сайтів чи звітів з акредитованих лабораторій моніторингу різних держуправлінь чи держекоінспекцій, судів або інших офіційних органів — 0,95; дані зі ЗМІ з посиланням на офіційні джерела, але зі своїми доповненнями — 0,7. Індикси достовірності оцінок за моделями можна оцінювати від 0,1 до 0,9, в залежності від обсягу трену-

вальних даних та досягнутої точності прогнозу на валідаційних даних з крос-валідацією. Візьмемо приклад свідчень очевидця в соціальній мережі про від 15 травня 2018 про мор риби біля м. Ладжин (рис. 6), без додаткових оброблень система оцінює індекс достовірності цієї події як 0,3. Далі на рис. 7 подано приклад знайдених новин про цей самий факт, які доповнюють оцінку, отриману за моделлю, для тренування якої не вистачало даних.

Достовірність події, J: 0.9972375690607735						
	date	title	link	text	j	p
0	2018-05-15	-	https://www.facebook.com/plugins/post.php?href...	Це П. Буг, район порохових складів (бакланів) ...	0.3	0.300000
1	2018-05-18	Біля Ладжини у Південному Бугу загинуло риби ...	https://vinbazar.com/news/podiyi/bilya-ladijin...	15 травня 2018 року на річці Південний Буг у м...	0.95	0.890625
2	2018-05-16	Екологи відібрали проби води в річці поблизу Л...	https://lada.fm/2018/05/16/ekologi-vdbrali-pro...	Екологи зафіксували загибель риби поблизу Лади...	0.95	0.993578
3	2018-12-06	Якісна вода для містян: ДТЕК Ладжинська ТЕС п...	https://dtek.com/media-center/news/yakisna-vod...	У прес-релізі ДТЕК зазначено, що Ладжинська т...	0.7	0.997238

Рис. 7. Приклад оцінювання індексу достовірності новини про мор риби на Південному Бугу у масиві вод №UA_M5.4_00019 (біля м. Ладжини) у травні 2018 року

Отже, для прогнозу за моделлю або для свідчень очевидців із соцмереж вдалось підвищити індекс достовірності з 0,3 до 0,99, за рахунок прив'язки у часі та просторі різних числових та текстових даних.

Висновки

У роботі розглянуто проблему підвищення обсягів та рівня достовірності інформації про стан елементів багатозв'язних багатостадійних геоінформаційних систем (БГІС) з використанням інформаційних та інтелектуальних технологій, що особливо актуально для моніторингу стану цих об'єктів. Для розв'язання задачі використовуються числові та текстові дані з різних джерел (звіти, кадастри, публікації у ЗМІ та у соціальних мережах тощо), що були прив'язані та верифіковані до елементів БГІС з урахуванням просторово-часових метаданих.

Вперше запропоновано метод верифікації числових даних у текстах з використанням бази знань, повнота якої оцінюється за допомогою критерію топологічної спостережуваності, що дає можливість оцінити повноту та точність видобутої інформації. Удосконалено метод формування користувацьких запитів до великої мовної моделі та RAG-архітектури шляхом прив'язки ембедингів до просторових об'єктів і часових меж, що уможливує точніше знайти релевантну для користувача інформацію з урахуванням встановлених просторово-часових обмежень. Набув подальшого розвитку метод геоприв'язки, який автоматично відбирає релевантні факти у текстах і точніше зіставляє ці факти з описами масивів вод. Створена інформаційна технологія, яка поєднує в собі усі ці методи та алгоритми, підвищує якість і повноту знань у ГІС та автоматизує обробку текстів.

Охарактеризовано особливості реалізації цієї технології на прикладі задачі моніторингу стану масивів вод басейну річки Південного Бугу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] В. Б. Мокін, І. В. Варчук, і С. М. Крижановський, *Інформаційна технологія аналізу та оптимізації топологічної спостережуваності багатозв'язних геоінформаційних систем*, моногр. Вінниця, Україна: ВНТУ, 2019, 121 с. ISBN 978-966-641-765-0.
- [2] В. Б. Мокін, В. Г. Сторчак, Є. М. Крижановський, О. В. Гавенко, і В. Ю. Балачук. *Інформаційні технології автоматизації обробки параметрів геоінформаційних систем з геометричними мережами*, моногр. Вінниця: ВНТУ, 2014, 196 с. ISBN: 978-966-641-946-3.
- [3] Zeiler Michael, *Modeling our World*, ESRI: Redlands, USA, 1999, 202 p.
- [4] *National Hydro Network. User Guide Series. Creating a Geometric Network in ArcGIS using NHN Data.* Edition 1.0.
- [5] *Комп'ютеризовані регіональні системи державного моніторингу поверхневих вод: моделі, алгоритми, програми*, моногр., В. Б. Мокін, Ред. Вінниця, Україна: вид-во ВНТУ "Універсум-Вінниця", 2005, 315 с.
- [6] «Водний Кодекс України», *Відомості Верховної Ради (ВВР)*, 1995, № 24, ст. 189, введений в дію Постановою ВР № 214/95-ВР від 06.06.95 [Електронний ресурс]. Режим доступу: <http://zakon2.rada.gov.ua/laws/show/213/95-%D0%B2%D1%80>.
- [7] *Водна стратегія України на період до 2050 року*. Розпорядженням Кабінету Міністрів України від 9 грудня 2022 р. № 1134-р. [Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/laws/show/1134-2022-%D1%80#Text>.
- [8] *Водна Рамкова Директива ЄС 2000/60/ЄС*. Основні терміни та їх визначення. Київ, 2006. 240 с.
- [9] Розпорядження Кабінету Міністрів України від 1 листопада 2024 р. № 1078-р «Про затвердження планів управління річковими басейнами Вісли, Дністра та Південного Бугу на 2025—2030 роки.» <https://zakon.rada.gov.ua/laws/show/1078-2024-%D1%80#Text>.
- [10] В. Б. Мокін, і М. В. Дратованій, *Наука про дані: машинне навчання та інтелектуальний аналіз даних*, електр. навч. посіб. комбінованого (локального та мережевого) використання. Вінниця, Україна: ВНТУ, 2024, 258 с. [Електрон-

ний ресурс]. Режим доступу: <https://docs.vntu.edu.ua/card.php?id=8163>.

[11] В. Б. Мокін, К. О. Бондалетов, І. М. Штельмах, О. В. Слободянюк, «Автоматичне видобування знань з екологічних звітів з прив'язкою до часу та до просторових координат масивів вод», *Вісник Вінницького політехнічного інституту*, вип. 3, с. 101-110, Черв. 2025, <https://doi.org/10.31649/1997-9266-2025-180-3-101-110>.

[12] Vitalii B. Mokin, et al., "Information measuring systems with mobile devices for identification of air pollution parameters caused by transport," *Proc. SPIE 10031, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments*, 2016, 1003128 (September 28, 2016), <https://doi.org/10.1117/12.2249202>.

[13] Vitalii Mokin, Ilona Varchuk, Kostiantyn Bondaleto, and Olena Slobodianiuk, "Method for Analyzing and Optimizing the Topologic Observability of Cognitive Maps of Complex Spatially Distributed Systems," in *Conference Proceedings 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (Ukrcon)* May 29 – June 2, 2017 Kyiv, pp. 865-869. <https://doi.org/10.1109/UKRCON.2017.8100371>.

[14] В. Б. Мокін, К. О. Бондалетов, Г. В. Горячев, і Д. Ю. Дзюняк, «Універсальна інформаційно-вимірювальна система оперативного екологічного моніторингу з використанням мобільних пристроїв», *Вісник Вінницького політехнічного інституту*, № 5 (122), с. 116-122, 2015.

[15] В. Б. Мокін, Д. Ю. Дзюняк, К. О. Бондалетов, і В. В. Олійник, «Метод і технологія моніторингу стану атмосферного повітря за допомогою універсальної інформаційно-вимірювальної системи з використанням мобільних пристроїв», *Наукові праці Вінницького національного технічного університету*, № 4, 2015. [Електронний ресурс]. Режим доступу: <http://praci.vntu.edu.ua/index.php/praci/article/view/456/454>.

[16] A. Halterman, "Mordecia 3: A Neural Geoparser," *arXiv*, 2023, [Electronic resource]. Available: <https://arxiv.org/pdf/2303.13675>.

[17] Hanwen Zheng, et al., "A Comprehensive Survey on Document-Level Information Extraction," in *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, 2024, pp. 58-72, USA: Association for Computational Linguistics, [Electronic resource]. Available: <https://aclanthology.org/2024.futurED-1.6.pdf>.

[18] J. Dagdelen, et al., "Structured information extraction from scientific text with large language models," *Nature Commun*, no. 15, 1418, 2024. <https://doi.org/10.1038/s41467-024-45563-x>.

[19] D. Dessi, et al., "CS-KG 2.0: A Large-scale Knowledge Graph of Computer Science," *Scientific Data*, no. 12, 964, 2025. <https://doi.org/10.1038/s41597-025-05200-8>.

[20] Haoran Luo, et al., "Text2NKG: Fine-Grained N-ary Relation Extraction for N-ary relational Knowledge Graph Construction," *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. [Electronic resource]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/Abstract-Conference.html. Accessed: 06.06.2025.

[21] R. Bommasani, et al. "On the Opportunities and Risks of Foundation Models," *Computer Science, Machine Learning*, 2021. [Electronic resource]. Available: <https://arxiv.org/abs/2108.07258>.

[22] Vitalii Mokin, Ilona Varchuk, Kostiantyn Bondaleto, and Olena Slobodianiuk, "Method for Analyzing and Optimizing the Topologic Observability of Cognitive Maps of Complex Spatially Distributed Systems," *Conference Proceedings 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON)* May 29 – June 2, 2017 Kyiv. pp. 865-869. <https://doi.org/10.1109/UKRCON.2017.8100371>.

[23] K. Bondaleto, and V. Mokin, "Notation System for Comparing and Synthesis of Intelligent Key Phrase Extraction Methods for Ontological Models in Information Systems," in *Information and Communication Technologies and Sustainable Development. ICT&SD 2022*, S. Dovgyi, O. Trofymchuk, V. Ustimenko, L. Globa., Eds., *Lecture Notes in Networks and Systems*, vol. 809. Springer, Cham, 2023. https://doi.org/10.1007/978-3-031-46880-3_11.

[24] В. Б. Мокін, К. О. Бондалетов, Є. М. Крижановський, і В. О. Караваєв, «Метод аугментації текстів про стан масивів вод на основі інтелектуальної прив'язки до багатозв'язних геоінформаційних систем іменованих сутностей», *Вісник Вінницького політехнічного інституту*, № 3, с. 55-65, Черв. 2023. <https://doi.org/10.31649/1997-9266-2023-168-3-55-65>.

[25] К. О. Бондалетов, В. Б. Мокін, Є. М. Крижановський, і В. О. Караваєв, «Розроблення інтелектуальної веб-системи з інформацією про екологічні проблеми та природоохоронні заходи у масивах вод басейну р. Південний Буг WISEST-SBV», *Актуальні проблеми та перспективи розвитку фундаментальних, прикладних, загальнотехнічних та безпекових наук*, в Матеріали IV всеукр. науково-практ. конф. пам'яті акад. Акад. наук вищ. освіти, проф. А. В. Касперського, Київ, 27 черв. 2024, с. 6-12. <https://kztdop.ipf.npu.edu.ua/?view=article&id=244>.

[26] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Third Edition, 2013. [Electronic resource]. Available: <https://sites.stat.columbia.edu/gelman/book/>.

[27] Vinbazar, «Біля Ладжана у Південному Бугу загинуло риби на 270 тисяч гривень», *Всі Новини Вінниці. Vinbazar.com*. [Електронний ресурс]. Режим доступу: <https://vinbazar.com/news/podiyi/bilya-ladijana-u-pivdenному-buzi-zaginulo-ribi-na-270tisyach-griven>.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 17.10.2025

Бондалетов Костянтин Олегович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: bondaleto.k@gmail.com ;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: vbmokin@vntu.edu.ua.

Вінницький національний технічний університет, Вінниця

K. O. Bondalietov¹V. B. Mokin¹

Intellectual Technology for Extraction and Verification of Numerical and Textual Data in Multi-Connected Multi-Stage Geoinformation Systems

¹Vinnitsia National Technical University

The article is devoted to solving the problem of increasing the volume and level of reliability of information regarding the state of elements of multi-connected multi-stage geographic information systems (GIS) using information and intellectual technologies, which is especially relevant when monitoring the state of these objects. To solve the problem, numerical and textual data from various sources (reports, cadasters, publications in the media and social networks, etc.) are used, which were bound and verified to the elements of the GIS taking into account spatial-temporal metadata. The criterion of topological observability is applied as a measure of the completeness of knowledge for the verification of numerical data in texts.

For the first time, a method of verification of numerical data in texts using a knowledge base is proposed, the completeness of which is assessed using the criterion of topological observability, which makes it possible to assess the completeness and accuracy of the extracted information. The method of forming user queries to a large language model and RAG architecture has been improved by binding embeddings to spatial objects and temporal boundaries, which provides an opportunity to more accurately find information relevant to the user, taking into account the established spatial-temporal restrictions. The georeferencing method has been further developed, which automatically selects relevant facts in texts and more accurately compares these facts with descriptions of water bodies. The integrated technology has been proposed that combines these methods — automated extraction of numerical and text facts from various sources, their spatial-temporal binding to GIS elements and formalized verification in order to reduce information uncertainty and improve the quality of digital twins for decision support. An example of the application of this technology is given on the example of the task of analyzing the state of the water body of the Southern Bug River basin. The study showed that the reliability index of news about water pollution can be increased from 0.3 to approximately 0.99, i.e. 3.3 times.

Keywords: Artificial Intelligence, Knowledge Base, Model, Information System, Dataset, Monitoring, Water Resources, Ecosystem.

Bondalietov Kostiantyn O. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: bondaletov.k@gmail.com ;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua