https://doi.org/10.31649/1997-9266-2025-180-3-101-110

УДК 004.9+556

К. О. Бондалєтов[1]
В. Б. Мокін[1]
І. М. Штельмах[1]
О. В. Слободянюк[1]

АВТОМАТИЧНЕ ВИДОБУВАННЯ ЗНАНЬ З ЕКОЛОГІЧНИХ ЗВІТІВ З ПРИВ'ЯЗКОЮ ДО ЧАСУ ТА ДО ПРОСТОРОВИХ КООРДИНАТ МАСИВІВ ВОД

[1]Вінницький національний технічний університет

Запропоновано новий метод автоматичного видобування екологічних знань з текстів звітів та новин про факти щодо стану вод річок чи їхнього забруднення. Видобування знань здійснюється з урахуванням прив'язки отриманих фактів до просторових координат конкретних масивів вод і інтервалів часу. Актуальність роботи зумовлена значною доступністю таких екологічних даних у новинах, веб-сайтах установ та соціальних медіа, необхідністю їхнього швидкого та точного оброблення. Запропонований метод поєднує виявлення фактів про стан вод чи про їх забруднення, розпізнавання географічних назв з тексту та заголовків, а також визначення часових ознак за допомогою аналізу ієрархічної структури документа. Метод оптимізує контекстно-семантичний критерій, який максимізує повноту та ймовірність виявлення усіх наявних зв'язків між ключовими словосполученнями у тексті фактів, періодами часу і масивами вод та, одночасно, мінімізує кількість хибнопозитивних зв'язків між ними, за рахунок формалізації зв'язків у вигляді триплетів "subject–predicate–object" (SPO) та використання міри Жаккара для пошуку ступеня подібності між списками ключових словосполучень, що характеризують ці факти і масиви вод. Видобування знань основано на виявленні і використанні ієрархічної структури документа, використанні великих мовних моделей, на актуалізації бази знань інформацією з використанням методу генерації з доповненням через пошук (RAG) для регулярного оновлення знань та їхньої прив'язки до періоду часу і просторових координат. Результатом є структурована база знань у вигляді триплетів «факт–масив вод–інтервал часу», який може використовуватися для аналізу динаміки стану вод, виявлення тенденцій та ухвалення управлінських рішень щодо поліпшення стану поверхневих вод.

Наведено результат застосування запропонованого методу на прикладі річного звіту про діяльність Басейного управління водних ресурсів річки Південний Буг за 2019 рік, який проілюстрував його працездатність.

Ключові слова: видобування знань, SPO-триплети, штучний інтелект, геоприв'язка даних, масив вод, великі мовні моделі, генерація з доповненим пошуком.

Вступ

Одним з найпопулярніших і доступних джерел отримання даних про стан поверхневих вод є текстові звіти, новини, повідомлення із соцмереж про виявлені факти забруднення вод, проблеми зі станом води тощо. Особливо актуальним це стало з активним розвитком електронних засобів передачі інформації та соціальних медіа. Традиційно, аналіз подібних текстів вимагає ручного опрацювання, коли фахівці шукають таку інформацію, асоціюють її з тими чи іншими водоймами, опрацьовують дати, якісні показники впливу на екологічну ситуацію. Такий підхід є трудомістким та потребує багато часу, що стримує ефективність використання цієї інформації для екологічного моніторингу та швидкості реагування на факти негативного впливу.

Тому актуальною є задача створення методів, які здатні автоматично опрацьовувати неструктуровані текстові екологічні звіти та видобувати з них знання — факти про стан конкретних масивів вод у задані часові інтервали. Це дозволить систематизувати інформацію з багатьох джерел, яка

накопичувалася роками. Таким чином, екологічні організації зможуть швидше виявляти тенденції та ухвалювати обґрунтовані рішення, передусім, для досягнення не нижче доброго екологічного стану масивів вод, що вимагає законодавство України та міститься у вимогах ЄС [1]—[3].

Існує багато методів структурування природномовного тексту та видобування знань з нього, зокрема й про забруднення вод чи їхній стан. Як правило, вони зводяться до пошуку ключових словосполучень та семантичних зв'язків між ними на основі контексту. Але задача ускладнюється, коли необхідним є не просто знайти зв'язки, а й здійснити геоприв'язку фактів забруднення чи даних про результати моніторингу до масивів вод. За таких умов, є ризик знаходження великої кількості хибнопозитивних зв'язків, які, теоретично, мають місце, але їх неможливо прив'язати до жодного масиву вод, відтак їх пошук недоцільний. Важливо зазначити, що під геоприв'язкою до масиву вод розуміється прив'язка не стільки до координат самого масиву вод (частини поверхневих вод суші), скільки до координат водозбірної зони, з якої вода надходить у цей масив вод і безпосередньо впливає на його стан.

Метою роботи є підвищення точності прив'язки знань з екологічних звітів з природномовним текстом до часу та до просторових координат масивів вод. Запропонований метод має забезпечити автоматичне видобування знань з текстів (звітів, або новин, публікацій із соціальних мереж, тощо), покращити виявлення усіх релевантних зв'язків між ключовими словосполученнями у них, часовими періодами та масивами вод, а також мінімізувати кількість хибнопозитивних результатів. Також, метод повинен враховувати структурованість тексту, множинність і складність зв'язків між фактами, часовими відрізками та масивами вод, а також забезпечувати можливість динамічного оновлення бази знань і оперативного оброблення нової інформації.

Огляд наявних рішень

Автоматичне структурування знань з текстів часто здійснюється шляхом видобування фактів у вигляді трійок (триплетів) «Підмет–Присудок–Доповнення» (англ. "Subject–Predicate–Object" – SPO). Такий підхід лежить в основі побудови графів знань з неструктурованих даних [4]. Проте класичні SPO-триплети, зазвичай, не містять чіткої інформації про місце та час події. Як зазначено в [5], виділення фактів у форматі триплетів фактично ігнорує часовий вимір, тобто в результаті отримується набір статичних тверджень без інформації про те коли і де відбувалися ці події. Для врахування цих недоліків, такі методи стали розширювати, з'явилися підходи, які доповнюють SPO-триплети часовими мітками і географічними координатами [6]. У сховищах знань нового покоління, таких як Wikidata, YAGO4, також передбачено зберігання атрибутів про час та географічне розташування [7]. Проте автоматичне визначення таких просторово-часових атрибутів безпосередньо в текстових документах досі досліджено недостатньо [8]. Особливо актуальною є ця проблема для екологічних звітів, де факти, до прикладу, «вміст розчиненого кисню у воді для річки X становить Y мг/дм^3» мають сенс лише з урахуванням інформації де і коли вони зафіксовані.

Задача визначення географічних назв у тексті (топонімів) та їх прив'язки до географічних координат називається геопарсингом [7]. Дослідження показують, що геопарсинг залишається складною задачею через велике різноманіття стилів і мовних зворотів у текстах, неоднозначність назв населених пунктів і річок, а також складності визначення загального контексту під час їхнього розпізнавання [9]. Сучасні інструменти геокодування дозволяють дещо зменшити неоднозначність [10], [11], але це однаково залишається певною проблемою. Вони часто застосовують правила на зразок частоти згадування географічних назв у тексті, наприклад, якщо в ньому кілька разів згадано слово «Париж», то ймовірно весь текст відноситься саме до цієї локації [11].

Проте стандартні методи оброблення природної мови (NLP) та розпізнавання іменованих сутностей (NER) працюють, здебільшого, на рівні окремого речення і можуть не враховувати загальний контекст документа. Наприклад іменована сутність «Вінницька область» у заголовку задає прив'язку для всіх речень з тексту, адже ідеться про всі водойми із цього географічного регіону. Більшість наявних NLP-моделей не враховують таку ієрархію та заголовки у межах всього тексту, через що геопарсинг є не достатньо точним. Ця проблема виникає через необхідність оброблення великих текстів зі складною структурою та вирішується за допомогою алгоритмів видобування інформації з документів, проте і вони не вирішують проблеми повністю [12].

З появою великих мовних моделей (англ. "Large Language Model" — LLM) дослідники також почали застосовувати їх для побудови графів знань. LLM можна навчити знаходити зв'язки між сутностями у форматі структурованих даних [13], генерувати триплети на основі підказок, або

навіть створювати JSON-об'єкти згідно з заданою схемою. Проте цей підхід має низку обмежень. По-перше, LLM мають обмеження на розмір контексту, моделі не завжди можуть врахувати весь текст, особливо якщо він великий чи містить таблиці. По-друге, вони схильні до галюцинацій і можуть спотворювати факти. По-третє, ці моделі без підкріплення базою знань погано працюють з прив'язкою до офіційних ідентифікаторів (кодів об'єктів). Саме тому актуальним є підхід з використанням генерації з доповненим пошуком (RAG) — поєднання мовної моделі із зовнішньою базою знань для уточнення та збагачення контексту. Такий підхід підвищує точність і нівелює галюцинації [12]. Проте навіть RAG у стандартному вигляді не гарантує врахування складної структури текстів [14].

Таким чином, можна виділити такі обмеження сучасних методів, які варто намагатись подолати: 1) вузький контекст оброблення — моделі обмежуються окремим реченням чи абзацом і не враховують інформацію з сусідніх розділів; 2) проблеми із геоприв'язкою — недостатнє використання географічних довідників, в результаті чого знайдені у тексті назви локацій можуть залишатися неоднозначними; 3) відсутність ієрархічного спадкування заголовків — структура документа (рік, назва басейну чи області в заголовках тексту) не ототожнюється з рівнем окремих фактів у тексті, до якого стосується заголовок, що веде до втрати геочасових атрибутів; 4) складність зіставлення фактів з об'єктами — факти розподілені по тексту і не групуються з відповідними об'єктами, якщо використовувати лише базовий підхід до побудови триплетів; 5) проблеми із забезпеченням регулярного оновлення фактів та знань на їхній основі.

Постановка та ідея розв'язання задачі

Введемо систему позначень:

- S — множина речень з початкового корпусу C документів, які аналізуються;
- S_n — речення з нових документів, що надходять у режимі потокового оновлення (RAG);
- $S_{all} = S \cup S_n$ — повний корпус текстів;
- $W = \{w_1,...,w_m\}$ — множина масивів вод, які вважаються як їхні водозбірні зони, до території яких прив'язуються географічні назви — це, швидше, можна назвати «просторовий контекст» масивів вод;
- $G = \{g_1,...,g_k\}$ — географічні назви (словосполучення);
- $T = \{t_1,...,t_l\}$ — допустимі часові інтервали;
- N — абстрактні об'єкти (наприклад, «вода»), які можуть бути елементами SPO-триплетів, в яких фігурують географічні назви;
- P — множина предикатів SPO-триплетів;
- E — множина речень про «екологічні факти» (стан вод або факти забруднення вод, яке впливає на цей стан);
- E_f — підмножина речень з множини E, які вдасться точно прив'язати до інтервалу часу і координат хоча б одного масиву вод.

Множинні зв'язки:

- $G(w) \subseteq G$ — географічні назви, пов'язані з w з множини W;
- $G(s) \subseteq G$ — назви в реченні s з множини S_{all};
- $T(s) \subseteq T$ — час, прив'язаний до s з множини S_{all};
- $R_{GW} \subseteq G \times W$ — зв'язки між назвами й масивами вод;
- $R_{EW} \subseteq E_f \times W$ — зв'язки між фактами й масивами вод.

Вхідна інформація методу:

– корпус C україномовних документів, який містить багато інформації, зокрема дані про екологічні факти та згадку про час (наприклад, рік чи діапазон дат), за який подані ці дані, і дані про те, яких річок чи населених пунктів біля річок стосується інформація;

– Ψ — таблиця з кодами, назвами, координатами масивів вод як географічних площинних об'єктів, координатами їхніх водозбірних зон та списки пов'язаних з цими зонами географічних назв об'єктів.

Метою є пошук в C речень з множини S про стан чи забруднення вод річки та їхня прив'язка у часі (по суті, з множини E) та до масивів вод Ψ, щоб потім можна було узагальнити інформацію про кожен масив вод за заданий час.

Основні етапи запропонованого методу полягають в такому:

1) розбити вхідний текст (кожен документ корпусу текстів C) на речення і відібрати такі з них, де йдеться про забруднення або стан вод річки, що утворить множину речень з екологічними фактами E;

2) на основі таблиці Ψ побудувати множину усіх можливих кодів масивів вод W та відповідних їм усіх можливих географічних назв G (об'єднати усі можливі варіанти для кожного елемента з W як ключових словосполучень чи слів;

3) з використанням LLM у C знайти усі географічні назви з множини G (з урахуванням відмінків, різних закінчень) — усі знайдені назви сформують множину G_f, тобто назви, які реально є у наведених фактах у текстах C;

4) за використання LLM пов'язати кожне речення із E з кожною географічною назвою із G_f (географічна назва може бути заголовком, а текст про стан річки міститиметься в тексті у подальших реченнях) шляхом формалізації зв'язків у вигляді згаданих вище SPO-триплетів "subject–predicate–object", де "subject" та/або "object" — це елемент множини G_f, а "predicate" формується на основі речення з E;

5) за використання LLM у текстах C знайти прив'язку до різних часових інтервалів (рік, квартал, сезон року, місяць, дата, діапазон дат тощо) T та спробувати прив'язати за їхньою допомогою кожне речення з E у часі (весь текст може бути звітом за 2019 рік, тоді це означає, що всі речення будуть за 2019 р., але можуть бути вказані й конкретні дати; важливо, що LLM варто «зрозуміти» текст — дуже часто прив'язка до часу має місце не у самому реченні, а в попередньому, на іншій сторінці, або, навіть, на титульній сторінці, а може бути й у заголовках різного рівня, до яких відноситься задане речення);

6) на основі множини E формується множина E_f, яка містить тільки такі речення, які вдалось прив'язати до хоча б однієї географічної назви із G_f та хоча б одного інтервалу часу із T;

7) на основі відповідності G і W встановити відповідність E_f, T, W, тобто визначити які речення з E_f з прив'язкою у часі до елементів T відповідають певним масивам вод W;

8) прибрати дублікати рядків у таблиці W, E_f, T та об'єднати E_f для однакових пар значень W, T.

Для пояснення головної ідеї методу формалізуємо контекстуально-семантичний критерій повноти та коректності знань про задані просторові об'єкти. Для цього спочатку уточнимо одну з його складових. Нехай:

- $\Pr\left(s \to G(w)\right) \in [0,1]$ — модельна ймовірність того, що речення s стосується просторового контексту масиву вод w;
- $\Pr\left(s \to t\right) \in [0,1]$ — модельна ймовірність того, що речення s належить часовому інтервалу t;
- $FP(B)$ — оцінка кількості хибнопозитивних зв'язків у множині B.

Тоді повна ймовірність істинності зв'язку буде дорівнювати:

$$\Pr\left(s \to w,t\right) = \Pr\left(s \to G(w)\right) \cdot \Pr\left(s \to t\right).$$

А тоді як критерій оптимальності пропонуємо такий:

$$L(B) = \sum_{(s,w,t)\in B} \Pr\left(s \to w,t\right) - \lambda \cdot FP(B) \to \max,$$

де $\Pr\left[\left(e_i \to w_k, t_1\right)\right]$ — ймовірність того, що зв'язок є семантично та контекстуально правильним; $FP(B)$ — кількість хибнопозитивних (англ. "False Positive") зв'язків; $\lambda \in \mathbb{R}, \lambda > 0$ — коефіцієнт штрафу за хибну прив'язку знань.

Головна ідея методу полягає в тому, що він забезпечує максимізацію критерію (1), одночасно вирішуючи 3 задачі:

1) забезпечити пошук максимально великої кількості варіантів зв'язків між елементами множин

G, W та словами з речень із множини S;

2) обмежити кількість хибнопозитивних зв'язків між реченнями з множин S та G, T шляхом відбору тільки таких SPO-триплетів, які пов'язують кожне речення хоча б з одним триплетом та хоча б один елементом з множини G і хоча б одним — з T;

3) обмежити кількість хибнопозитивних зв'язків між елементами множин S, G, T та W, щоб уникнути ситуацій, коли новина чітко про конкретний *w* (де розташований Київський міст у Вінниці через р. Південний Буг у Вінницькій області), а її хибно прив'яжуть до інших *w* (м. Хмельницький на р. Південний Буг або м. Хмільник на р. Південний Буг у Вінницькій області), в яких теж є такі самі географічні назви $g_1 g_2$ із G — для цього використано міру Жаккарда для порівняння списку підмножин із G у реченні s та підмножин з G в масиві вод *w* (див. це у роботі авторів [15]).

Для цього важливо поєднати та припасувати відомі методи видобування і формалізації знань, побудови SPO-триплетів з використанням великих мовних моделей та RAG-режим оновлення векторної бази даних фактів.

Розв'язання задачі

Для розв'язання поставленої задачі пропонуємо такий алгоритм.

Будемо вважати, що вхідний документ з корпусу текстів C вже перетворено на текст з множиною речень S та заголовків H, яким відповідають ці речення.

$$S = \{s_1, s_2, ..., s_n\}; H = \{h_1, h_2, ..., h_k\}. \quad (1)$$

Важливо, що кожне речення відповідає як заголовку, в якому воно розміщене $h = \pi(s)$, так і усім заголовкам вищого рівня, які охоплюють цей заголовок

$$H(s) = \{\pi(s)\} \cup Ancestors(\pi(s)). \quad (2)$$

Введемо множину масивів вод з офіційного реєстру

$$W = \{w_1, w_2, ..., w_m\}, \quad (3)$$

де кожен об'єкт w_i має унікальний ідентифікатор та назву. Розглянемо функцію $G(w_i)$, що повертає множину географічних ключових термінів, асоційованих із об'єктом w_i

$$G: W \to 2^U, \quad (4)$$

де U — універсум усіх географічних назв (областей, річок, міст, тощо), а $G(w)$ — підмножина U, відповідна об'єкту w. Наприклад, $G(UA\_M5.4\_0011) = \{$"Вінницька область"$\}\{$"Південний Буг"$\}\{$"Хмільник"$\}$.

Визначимо функцію виділення географічних термінів в тексті речення. Нехай $X(s)$ — множина географічних назв, знайдених безпосередньо в тексті речення s. До того ж, визначимо розширену множину географічних назв у реченні з урахуванням контексту заголовків, та їхньої ієрархії (2)

$$G(s) = X(s) \cup \bigcup_{h \in H(s)} X(h), \quad (5)$$

де $X(h)$ — множина географічних назв, присутніх у заголовку розділу h. Таким чином, $G(s)$ містить усі географічні назви, які стосуються факту із речення s.

Виділяємо речення, що містять екологічні факти:

$$E = \{s \in S_{all} : \text{IsEcological}(s) = 1\}. \quad (6)$$

Знайдемо максимальну кількість SPO-триплетів, де "subject" та/або "object" — елементи з множини G, а "predicate" сформуємо на основі речень з E. При цьому, відбираються тільки такі триплети, які містять прив'язку у часі.

Витягуємо триплети (з максимально великої множини варіантів потенційних зв'язків між елементами)

$$T_{SPO} \subseteq (G \cup N) \times P \times (G \cup N), \text{ExtractSPO}: E \to 2^{T_{SPO}}, \quad (7)$$

$$\forall s \in E, T(s) = \text{ExtractSPO}(s). \tag{8}$$

Цей крок впливає на $\sum \Pr(s \to w,t)$ — чим більше правильних триплетів, тим вища сума ймовірностей у критерії.

Визначимо функцію, що задає часовий контекст речення $t(s)$:

$$t(s) = \begin{cases} TimeExplicit(s), & \text{if } TimeExplicit(s) \neq 0, \\ Time(h), & \exists h \in H(s) : Time(h) \neq 0, \\ TimeMeta(D), & \text{otherwise.} \end{cases} \tag{9}$$

Функція $TimeExplicit(s)$ визначає часовий інтервал зі змісту речення (наприклад, «2019»). Якщо ж така функція не знаходить значення, тоді шукаємо серед заголовків $H(s)$. Якщо серед усіх рівнів заголовків немає вказівки часу, тоді $TimeMeta(D)$ — часовий інтервал, пов'язаний з цілим документом (наприклад, рік публікації звіту). У практичних реалізаціях можна вдосконалити цю функцію, щоб вона враховувала різні формати дат і вибирала найспецифічніший час (якщо є рік і місяць — брати місяць, якщо є день — брати його). Для нашої задачі достатньо хоча б прив'язки фактів до року.

Прив'язка у часі *t* може бути й у метаданих документа *A* (з англ. "About"), тоді формула спрощується. За наявності метаданих $At(s)$ з опису документа (наприклад, рік звіту)

$$T(s) = \begin{cases} TimeExplicit(s), & \text{if specified explicitly,} \\ At(s), & \text{if specified in metadata,} \\ \varnothing, & \text{otherwise.} \end{cases} \tag{10}$$

Цей крок підвищує точність $\Pr(s \to t)$, тим самим збільшуючи $L(B)$.

Наступним кроком є формування множини E_f, яка містить тільки такі речення, які вдалось прив'язати до хоча б однієї географічної назви із G та хоча б до одного періоду часу із T

$$E_f = \{ s \in E \mid T(s) \neq \varnothing \wedge T(s) \cap P \neq \varnothing \}. \tag{11}$$

Це обмеження зменшує $FP(B)$, залишаючи лише речення з чіткими предикатами й часовими прив'язками.

Зробимо прив'язку R_w географічних назв як ключових словосполучень у реченні *s* до подібних слів в описі масиву вод з використанням Жаккард-міри [15]

$$R_W = \{ (s,w) \in E_f \times W \mid \text{Jaccard}(s,w) \geq \theta \}, \theta \in [0,1]. \tag{12}$$

$$\text{Jaccard}(s,w) = \frac{|G(s) \cap G(w)|}{|G(s) \cup G(w)|}. \tag{13}$$

Цей крок безпосередньо зменшує $FP(B)$, фільтруючи хибні просторові прив'язки.

Побудуємо триплети «Факт–Масив вод–Інтервал часу».

$$B \subseteq E_f \times W \times T. \tag{14}$$

Важливо, що зв'язки включаються до *B* лише, якщо $T(s) \in t$ та $(s,w) \in R_w$.

Проведемо видалення дублікатів та об'єднання фактів щодо однакових масивів вод за однаковий інтервал часу, для чого зведемо однакові (w,t) в одне значення та об'єднаємо всі відповідні речення s у поле *Text*. Формуємо фінальне відношення R_Σ:

$$R_\Sigma \subseteq W \times T \times \text{Text, where Text}(w,t) = \bigcup_{\substack{s \in E_f \\ (s,w) \in R_w \\ (s,t) \in R_T}} S \tag{15}$$

Отже, на першому етапі аналізується та обробляється вхідний корпус документів, які містять екологічну інформацію. Важливо, що документи мають певне форматування: заголовки, виділення, відступи та порожні рядки. За допомогою спеціального парсера алгоритм виконує структурний аналіз документа. Виявляються заголовки різних рівнів, короткі ізольовані рядки, логічно-відокремлені блоки тексту. У результаті цього етапу документи перетворюються на багаторівневий JSON-об'єкт.

Текст розбивається на окремі речення, які передаються на фільтрацію. За допомогою LLM-аналізу вибираються лише ті речення, які стосуються стану або забруднення вод — формується множина E. Нові тексти S_n обробляються шляхом перетворення на ембеддінги $\mathrm{Emb}(s_i) \in \mathbb{R}^d$, та індексації за допомогою бібліотеки FAISS. Завдяки цьому забезпечується швидкий семантичний пошук фактів у майбутньому за рахунок семантичного порівняння з новими фрагментами.

За допомогою LLM в тексті знаходяться всі географічні назви з довідника G. Алгоритм визначає, чи є у реченні пряма вказівка на часовий інтервал. Якщо прив'язка відсутня — використовується дата із заголовку або метаданих документу. Всі прив'язки формалізуються у вигляді множини T.

Створюється дві ключові бінарні матриці — матриця «речення–географічна назва» EG та матриця «речення–час» ET. Враховується ієрархічне наслідування від заголовків. Для кожної знайденої географічної назви визначається, до яких масивів вод W вона належить. Враховується, що одна назва може відповідати кільком масивам і один масив може бути пов'язаний з кількома назвами. Результат очищується від дублювання та подається у структурованому вигляді для подальшого аналізу.

Практична реалізація методу

Проведено експеримент, метою якого є створення бази знань на основі автоматично структурованих фрагментів тексту з екологічних звітів. Тобто розв'язуємо задачу отримання з неструктурованих текстів впорядкованих наборів фактів з метаданими (як-то прив'язка до масиву вод, року) з використанням запропонованого методу (1)—(15). Як вхідні дані використано фрагмент офіційного звіту про стан водних ресурсів, зокрема розділ IV, підрозділ 4.3, пункт 4.3.1 документу «Річний звіт про діяльність Басейного управління водних ресурсів річки Південний Буг за 2019 рік» [16]. У цей фрагмент (файл text_ua.csv) включено узагальнену за результатами моніторингу інформацію про якість поверхневих вод у 4 контрольних пунктах р. Південний Буг у Вінницькій області за 2019 рік. Другий датасет (файл water_bodies.csv) — це довідник, де зазначено деякі масиви вод басейну річки Південний Буг: офіційне кодування та назви географічних об'єктів в їхньому просторовому контексті, що дозволяє однозначно ідентифікувати кожен масив вод. На рис. 1 наведено приклад цих обох вхідних датасетів.

```
=== text_ua ===
 Узагальнена інформація про стан поверхневих вод Вінницької області
Відповідно до затвердженої програми моніторингу вод, у Вінницькій області, протягом 2019 року,
щомісячно здійснювалися спостереження масивів поверхневих вод, забір води з яких здійснюється д
ля задоволення питних і господарсько-побутових потреб населення, на чотирьох пунктах моніторинг
у, які розташовані в басейні Південного Бугу.
Річка Південний Буг.
Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах
4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3.
Жорсткість води середня, значення середньорічних показників становили 5,52-7,70 мг-екв/дм3( ГДК
7,0 мг-екв/дм3), перевищення у 3 пробах, максимальне у 1,1 рази в питному водозаборі м. Ладижи
н, м. Хмільник, с. Гущинці питний водозабір, м. Калинівка. ...

=== water_bodies ===
```

	code_water_body	list_geo_objects
0	UA_M5.4_0011	Вінницька область, річка Південний Буг, м. Хмільник
1	UA_M5.4_0013	Вінницька область, річка Південний Буг, м. Калинівка, м. Вінниця, с. Гущинці
2	UA_M5.4_0019	Вінницька область, річка Південний Буг, м. Ладижин

Рис. 1. Вхідні датасети з файлів text_ua.csv (множина C) та water_bodies.csv (множини W та G)

Розроблений метод реалізує автоматичне формування зв'язків за триплетом «факт — масив вод — інтервал часу». Суть методу полягає в поетапному аналізі тексту, контексту та зіставлення даних з множиною G. Ключові етапи:

1. Текстовий фрагмент розбиваємо на окремі речення, проводимо первинний аналіз для розпізнавання речень з потенційно важливими екологічними фактами. До таких речень відносимо ті, що містять метрики або спеціальні визначені поняття (наприклад, «концентрація речовин», «перевищення нормативних значень», «показники кисневого режиму» тощо).

2. Для визначення контексту речень аналізуємо ієрархічну структуру вхідного документа: якщо в реченні відсутня у явному вигляді інформація про географічний об'єкт чи час, то ці дані успадковуються із попередніх речень, розділів або контекст документа загалом (у нашому прикладі 2019 рік згадується у назві документа). Для визначення географічного контексту вважаємо, що текст відноситься до всієї р. Південний Буг, доки не знайдемо специфічніших вказівок.

3. Порівнюємо знайдені в тексті назви у географічному довіднику, порівнюючи підмножини словосполучень в тексті та в довіднику масивів вод на основі міри Жаккара. Наприклад, для «р. Південний Буг, м. Хмільник» алгоритм знаходить частковий збіг у довіднику з офіційною назвою «Вінницька область, річка Південний Буг, місто Хмільник — код UA_M5.4_0011», що дає можливість однозначно прив'язати факт до цього масиву вод.

4. Формуємо триплети за шаблоном «факт–масив вод–інтервал часу» на підставі аналізу пунктів 1—3. Вважаємо, що кожен триплет відображає одну конкретну подію, характеристику чи факт. Після виділення фактів з прив'язкою здійснюємо їх об'єднання про конкретний масив вод у певний інтервал часу в єдиний текст. На прикладі вхідного фрагменту звіту два окремих речення: про перевищення жорсткості води та про кисневий режим об'єднано в єдиний текст для кожного місця спостереження. Таким чином, отримуємо короткі, але інформативні записи з розрізнених фактів.

У результаті роботи розробленого методу отримано автоматично структурований датасет екологічних фактів R_{Σ}, де кожен запис складається з опису комплексного факту та його взаємозв'язку з масивами вод у часі (рис. 2).

	code_water_body	time_period	text
0	UA_M5.4_0011	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Хмільник. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
1	UA_M5.4_0013	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Калинівка. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
2	UA_M5.4_0013	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у с. Гущинці. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3
3	UA_M5.4_0019	2019	Перевищення середньорічних показників жорсткості мало місце максимально в 1.1 рази (ГДК 7,0 мг-екв/дм3) у м. Ладижин. Кисневий режим річки Південний Буг задовільний, значення розчиненого кисню знаходяться у межах 4,10-19,10мгО2/дм3, при нормі не менше 4,0мгО2/дм3

Рис. 2. Фрагмент описаних вихідних наборів даних

На нашому прикладі система сформувала 4 записи, по одному для кожного з 4 пунктів спостереження, успішно прив'язавши кожен екологічний факт до конкретних масивів вод та часових інтервалів.

Таким чином, проведений експеримент підтвердив працездатність методики автоматичного видобування екологічних знань з екологічних звітів з прив'язкою до масивів вод і часу. Вихідна база знань за шаблоном триплетів «факт – масив вод – інтервал часу» узгоджена з експертними даними і зручна для подальшого аналізу. Отже, завдяки запропонованому підходу з'явилась можливість суттєво автоматизувати агрегацію екологічних звітів, забезпечивши екологів та інших відповідальних експертів інструментами для швидкого виявлення ключових фактів і тенденцій у великих об'ємах розрізнених текстових документів.

Висновки

В роботі запропоновано метод автоматичного видобування екологічних фактів з текстових звітів. Метод використовує багаторівневий аналіз документа: враховує як інформацію з локальних речень, так і з глобальних заголовків, що дозволяє успадковувати необхідний контекст. Запропоновано формальну модель, основану на SPO-триплетах. Використано міру Жаккара для пошуку

подібності між географічною прив'язкою фактів до реєстру масивів вод. Доведено, що в межах свого класу алгоритм досягає оптимальної прив'язки контексту, перевершуючи класичні підходи, які не враховують ієрархічний контекст або зовнішні знання. За результатами практичного експерименту, запропонований метод дозволив успішно структурувати фрагмент екологічного звіту про стан річки Південний Буг у вінницькій області за 2019 рік, сформувавши 4 триплети «факт–масив вод–інтервал часу» для кожного з контрольних пунктів спостереження. Метод продемонстрував здатність автоматично розпізнавати екологічні факти в неструктурованому тексті, прив'язувати їх до конкретних географічних об'єктів через порівняння з довідником масивів вод. Очікується, що у разі масштабування запропонованого методу на велику кількість різних текстів (звітів, новин тощо) з фактами екологічного характеру збільшиться швидкість та ефективність обробляння цих фактів. Більше того, експерименти показали, що точність прив'язування до конкретних масивів вод фактів зростає до рівня 85…90 %, іноді досягаючи і 100 %. Все це, в цілому, дозволить підвищити якість та обґрунтованість рішень, які будуть ухвалюватись у сфері управління водними ресурсами та охорони вод.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Верховна Рада України, «Водний Кодекс України», *Постанова ВР № 214/95-ВР від 06.06.95, Відомості Верховної Ради (ВВР)*, 1995, № 24, ст. 189). [Електронний ресурс]. Режим доступу: http://zakon2.rada.gov.ua/laws/show /213/95-%D0%B2%D1%80 .

[2] Кабінет Міністрів України, *Водна стратегія України на період до 2050 року*. Розпорядження від 9 грудня 2022 р. № 1134-р. [Електронний ресурс]. Режим доступу: https://zakon.rada.gov.ua/laws/show/1134-2022-%D1%80#Text .

[3] *Водна Рамкова Директива ЄС 2000/60/ЄС. Основні терміни та їх визначення*. Київ, Україна, 2006, 240 с. [Електронний ресурс]. Режим доступу: http://dbuwr.com.ua/docs/Waterdirect.pdf .

[4] J. Zhu, "A Temporal Knowledge Graph Generation Dataset Supervised Distantly by Large Language Models," *Scientific Data*, no. 12, p. 734, 2025. [Electronic resource]. Available: https://doi.org/10.1038/s41597-025-05062-0 .

[5] K. Salmas et al., "Extracting Geographic Knowledge from Large Language Models: An Experiment," *Workshop LM-KBC*, 2023, [Electronic resource]. Available: https://lm-kbc.github.io/workshop2023/proceedings/13_Salmas.pdf .

[6] M. Gritta et al., "What's missing in geographical parsing?" *Springer Nature Link.* [Electronic resource]. Available: https://link.springer.com/article/10.1007/s10579-017-9385-8 .

[7] A. Halterman "Mordecai 3: A Neural Geoparser," *arXiv, 2023*, [Electronic resource]. Available: https://arxiv.org/pdf/2303.13675 .

[8] 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 .

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

[10] В. Б. Мокін, К. О. Бондалєтов, 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, 2023. https://doi.org/10.31649/1997-9266-2023-168-3-55-65 .

[11] D. Dessí, et al., "CS-KG 2.0: A Large-scale Knowledge Graph of Computer Science," *Scientific Data*, no. 12, pp. 964, 2025. [Electronic resource]. Available: https://doi.org/10.1038/s41597-025-05200-8 .

[12] Yunyi Zhang, "Automated Mining of Structured Knowledge from Text in the Era of Large Language Models," in *KDD'24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* [Electronic resource]. Available: https://doi.org/10.1145/3637528.3671469 .

[13] 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 (date of access: 06.06.2025) .

[14] 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 .

[15] К. Бондалєтов, і В. Мокін, « Інтелектуальна автоматизація геоприв'язки повідомлень з соцмереж до масивів вод за допомогою зваженої Jaccard-міри,» *ВНТКП ВНТУ*. Факультет інтелектуальних інформаційних технологій та автоматизації ВНТУ, Вінниця, 24-27 березня 2025. [Електронний ресурс]. Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2025/paper/view/23298/19275 .

[16] *Річний звіт про діяльність басейнового управління водних ресурсів річки Південний Буг з питань управління водними ресурсами за 2019 р*ік, Вінниця. Україна: БУВР, 2019.

Рекомендована до друку кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 18.06.2025

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

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

Штельмах Ігор Миколайович — канд. техн. наук, асистент кафедри системного аналізу та інформаційних технологій, e-mail: igor.shtelmakh@vntu.edu.ua ;

Слободянюк Олена Валеріївна — канд. пед. наук, доцент, доцент кафедри опору матеріалів, теоретичної механіки та інженерної графіки, e-mail: olenas8@gmail.com .

Вінницький національний технічний університет, Вінниця

K. O. Bondalietov[1]
V. B. Mokin[1]
I. M. Shtelmakh[1]
O. V. Slobodianiuk[1]

Automatic Knowledge Extraction from Environmental Reports with Reference to Time and Spatial Coordinates of Water Bodies

[1]Vinnytsia National Technical University

The paper presents a new method for automatically extracting environmental knowledge from reports and news texts related to facts about the state of river waters or their pollution. Knowledge extraction is carried out taking into account the binding of the obtained facts to the spatial coordinates of specific water bodies and time intervals. The relevance of the work is due to the significant availability of such environmental data in the news, websites of institutions, and social media, and the need for their quick and accurate processing. The proposed method combines the detection of facts about the state of waters or their pollution, recognition of geographical names from the text and headlines, as well as the determination of time features by analyzing the hierarchical structure of the document. The method optimizes the contextual-semantic criterion, which maximizes the completeness and probability of detecting all existing connections between key phrases in the text of facts, time periods and water bodies and, at the same time, minimizes the number of false positive connections between them, by formalizing the connections in the form of "subject–predicate–object" (SPO) triplets and using the Jaccard measure to find the degree of similarity between the lists of key phrases that characterize these facts and water bodies. Knowledge extraction is based on identifying and using the hierarchical structure of the document, using large language models, and actualization the knowledge base with information with Retrieval-Augmented Generation (RAG) for regular knowledge update and binding to the time intervals and spatial coordinates. The result is a structured knowledge base in the form of "fact – water body – time interval" triplets, which can be used to analyze the dynamics of water status, identify trends, and make management decisions to improve the state of surface waters.

The result of applying the proposed method is presented using the example of the annual report on the activities of the Southern Booh River Basin Water Resources Management for 2019, which illustrates its efficiency.

Keywords: knowledge mining, SPO-triplets, artificial intelligence, data georeferencing, water array, large language models, Retrieval-Augmented Generation.

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 ;

Shtelmakh, Igor M. — Cand. Sc. (Eng.), Assistant Professor of the Chair of System Analysis and Information Technologies, e-mail: igor.shtelmakh@vntu.edu.ua ;

Slobodianiuk Olena V. — Cand. Sc. (Ped.), Associate Professor, Associate Professor of the Chair of Strength of Materials, Theoretical Mechanics and Engineering Graphics, e-mail: olenas8@gmail.com