https://doi.org/10.31649/1997-9266-2021-154-1-77-87

УДК 004.9+556

В. Б. Мокін[1]
А. М. Лучко[1]
О. М. Давидюк[1]
Т. Є. Вуж[2]

ТЕХНОЛОГІЯ ПОБУДОВИ ЕКСПЕРТНОЇ ІНФОРМАЦІЙНОЇ ВЕБ-СИСТЕМИ ВИЯВЛЕННЯ ТА ВЕРИФІКАЦІЇ ПРІОРИТЕТНИХ ЕКОЛОГІЧНИХ ПРОБЛЕМ У МАСИВАХ ВОД БАСЕЙНУ РІЧКИ

[1]Вінницький національний технічний університет;
[2]Вінницький національний медичний університет імені М. І. Пирогова

Розглянуто питання збирання, верифікації та узагальнення великої кількості експертних оцінок про наявний стан вод, наявні екологічні проблеми та впливові фактори, які збільшують ризик недосягнення екологічних цілей кожним масивом вод під час розроблення програм заходів планів управління річкового басейну (ПУРБ), спрямованих на досягнення чи стабілізацію доброго екологічного стану води у масивах вод цього басейну. Задача ускладнюється великою кількістю таких масивів вод, оскільки зібрати достовірну інформацію про об'єкти, розташовані у кожному з них, вкрай важко. Розв'язати це питання дозволить створення веб-системи з картою масивів вод і залученням великої кількості експертів із місцевих жителів, небайдужих до проблем свого довкілля. Проте, тоді виникає проблема перевірки достовірності оцінок цих експертів, враховуючи їх різну кваліфікацію, досвід роботи, різне бачення цілей ПУРБ, та проблема як їх зіставляти, щоб визначити найвразливіші регіони за різними критеріями. Для цього запропоновано вимагати від експертів не просто експертні оцінки на основі єдиних довідників варіантів можливих відповідей, а й посилання на текстові веб-ресурси, які підтверджують їх оцінки. А потім аналізувати наскільки ці джерела дійсно підтверджують кожну оцінку щодо відповідного типу проблеми для певного регіону. Авторами розглянуті різноманітні підходи для зіставлення експертних оцінок як на основі нечітких множин, так і за допомогою технологій машинного навчання та опрацювання природної мови (англ. — Natural Language Processing (NLP)). Розглянуто аналоги розроблювальної авторами системи.

Розроблено метод виявлення та верифікації пріоритетних екологічних проблем у масивах вод басейну річки за нечіткими експертними оцінками, з урахуванням ймовірностей того, що процитовані експертом текстові матеріали дійсно відповідають зазначеній проблемі. Ці ймовірності визначаються з використанням моделей NLP-технологій. Охарактеризовано етапи функціонування експертної інформаційної веб-системи для реалізації запропонованої технології, яка дозволить одночасно зібрати максимально достовірну і детальну інформацію про об'єкти масивів вод та прискорити її опрацювання і ранжування.

Наведено приклад реалізації інформаційної веб-системи виявлення пріоритетних екологічних проблем у масивах вод басейну річки Південний Буг. Наведено приклади обчислення достовірності експертних оцінок із застосуванням авторської програми на Python на основі NLP-моделі BERT і логістичної регресії до реальної текстової інформації.

Ключові слова: інформаційна веб-система, технології захисту водних ресурсів, експертні оцінки, нечіткі множини, екологічні проблеми, масив вод, ПУРБ, машинне навчання, класифікація текстових даних, технології опрацювання природної мови, NLP, BERT.

Постановка задачі та вихідні передумови

Етапу розроблення програм заходів планів управління річкового басейну (ПУРБ) в Європі, спрямованих на досягнення чи стабілізацію доброго екологічного стану води у масивах вод цього басейну, передують етапи збирання великої кількості інформації про наявний стан вод, наявні

екологічні проблеми та впливові фактори, які збільшують ризик недосягнення екологічних цілей кожним масивом вод [1], [2]. Задача ускладнюється великою кількістю таких масивів вод. В Україні їх десятки тисяч. Наприклад, у басейні р. Південний Буг — більше 1100. Зібрати достовірну інформацію про об'єкти, розташовані конкретно у кожному з таких вкрай важко. Цьому може зарадити створення веб-системи з картою масивів вод і залучення великої кількості експертів із місцевих жителів, небайдужих до проблем свого довкілля.

Можна знайти групу дійсно фахових експертів, але навряд чи вони орієнтуються в проблемах усіх цих регіонів та й, навіть, були там, хоч раз. З іншого боку, можна знайти небайдужих мешканців усіх населених пунктів цих регіонів, але як потім зіставляти їх оцінки? У другому підході виникнуть інші проблеми: як перевірити достовірність оцінок цих експертів, враховуючи їх різну кваліфікацію, досвід роботи, різне бачення цілей ПУРБ, та як їх зіставляти, щоб визначити найвразливіші регіони? Неможливість ефективного вирішення цих проблем спричиняє використання на практиці саме першого підходу, що часто призводить до неврахування важливих показників багатьох масивів вод, віддалених від великих населених пунктів та головної річки, і де немає стаціонарних постів державного моніторингу вод. Ми ж пропонуємо спробувати застосувати другий підхід, розробивши спеціальний математичний апарат.

Експертами можуть виступати як усі ті, хто збирає різну інформацію на різних етапах розроблення ПУРБ, так і просто широкі верстви населення, учні і вчителі шкіл, юні натуралісти, різні громадські організації тощо. Треба вимагати від них не просто експертні оцінки на основі єдиних довідників варіантів можливих відповідей, а й посилання на джерела (якісь текстові веб-ресурси), які підтверджують їх слова. А потім оцінювати наскільки ці джерела дійсно підтверджують експертну оцінку щодо певного регіону.

Для опрацьовування експертних оцінок з різним рівнем впевненості в інформаційних системах зазвичай використовується теорія нечітких множин [3]—[5]. Саме її математичний апарат дозволяє правильно зіставляти оцінки, ранжувати і в межах кожного масиву вод і межах басейну в цілому, що може бути важливим, наприклад, на етапі поділу бюджетного фінансування на запровадження відповідних природоохоронних заходів в окремих масивах вод. Але цей математичний апарат слід адаптувати до кожної задачі окремо.

Для розв'язання задачі в комплексі доцільно проєктувати одразу і метод аналізу інформації, і веб-систему для її збирання та опрацювання та перевірити їх на реальних прикладах.

Метою статті є розроблення методу виявлення та верифікації пріоритетних екологічних проблем у масивах вод басейну річки за експертними оцінками та підтверджуючими їх текстовими матеріалами і побудова експертної інформаційної веб-системи для його реалізації, що дозволить одночасно зібрати максимально достовірну і детальну інформацію про об'єкти масивів вод та прискорити її опрацювання і ранжування.

Ідея розв'язання задачі

Ключовою особливістю пропонованого методу аналізу експертних оцінок є опрацювання текстової інформації, якою експерт підтверджує кожну свою оцінку. Для її опрацювання пропонується використати сучасні технології.

Існують різноманітні підходи для зіставлення експертних оцінок як на основі нечітких множин, так і за допомогою технологій опрацювання природної мови (англ. — Natural Language Processing (NLP)). У дослідженні [5] автори запропонували новий нечіткий підхід до прийняття рішень з кількома атрибутами, що базується на надійності експертів та правилі доказових міркувань (ER) в інтервально-значущому нечіткому середовищі. Для визначення надійності експертів розробляється об'єктивний метод, що поєднує подібність оцінок, проведених до та після групового обговорення. У праці [6] розглянуто вирішення проблеми узагальнення відгуків експертів з використанням передових методів машинного навчання (англ. — «Machine Learning» (ML)). Запропоновані методи ґрунтуються на великій кількість методів NLP, що використовують вміст твітів та деяку довідкову контекстну інформацію.

Аналоги інформаційної веб-системи виявлення пріоритетних екологічних проблем у масивах вод басейну річки описані у працях [7], [8]. В роботі [7] автори описали розроблення системи прогнозування психічного здоров'я, яка оцінює ризик самогубства за інформацією форумів підтримки за допомогою технології BERT. Ця система здійснює аналіз, наприклад, повідомлень в Twitter, оновлення статусу або повідомлення на форумі, та на його основі оцінює психічне здоров'я люди-

ни. А в роботі [8] описана технологія аналізу та оцінки коментарів покупців для товарів за допомогою засобів NLP. Вона дозволяє виділяти нюанси в коментарях, до прикладу, визначає, чи він є позитивним чи негативним. Оцінка коментарів дозволяє виробникам коригувати речі та відправляти нові речі з кращими можливостями купівлі.

Основною метою технології побудови експертної інформаційної веб-системи виявлення пріоритетних екологічних проблем у масивах вод басейну річки є створення інформаційної веб-системи, яка дозволить усім бажаючим виступити в ролі експертів та оцінити які проблеми потребують першочергового вирішення, максимально усуне суб'єктивізм таких оцінок, дозволить порівняно швидко накопичити статистику по усіх масивах вод і візуалізувати наважливіші проблеми для масивів вод та для водного басейну в цілому.

В процесі побудови ПУРБ, як правило, розрізняють такі найрозповсюдженіші екологічні проблеми, які збільшують ризик недосягнення екологічних цілей масивами вод в Україні:

– забруднення органічними речовинами: дифузні (змив з полів, забруднені зливові води, змиті з доріг, магістралей, і газонів) та/або точкові джерела (місця скидання стічних та зворотних вод водоканалами тощо);

– забруднення біогенними речовинами: дифузні (змив з полів, магістралей, стоки виробничих площ підприємств) та/або точкові джерела (місця скидання підприємств хімічної промисловості);

– забруднення небезпечними речовинами: дифузні (змив з полів, забруднені зливові води, змиті з доріг, магістралей, і газонів) та/або точкові джерела (місця скидання підприємств металургійної, хімічної, харчової промисловості, підприємства целюлозно-паперової галузі);

– зміни гідрологічного режиму (підтоплення або, навпаки — маловоддя);

– зміни морфологічних показників водних об'єктів (руйнація берегів, замулення, зміна русла течії);

– забруднення підземних вод: дифузні та/або точкові джерела, біологічні (порушення санітарних норм в приповерхневих умовах: створення водовідстійників, поява стічних вод, зберігання сміття, техновідходів) та хімічні (засолення підземних вод);

– виснаження підземних вод;

– засмічення побутовими відходами басейну водних об'єктів та їх акваторії;

– вплив змін клімату (зокрема сильні зливи і повені або маловоддя, посуха тощо);

– інвазивні види, зменшення біорізноманіття тощо.

В розроблюваній інформаційній технології можна виділити три етапи роботи:

– етап збирання інформації про проблеми;

– етап оброблення зібраних даних;

– етап візуалізації результатів.

Рис. 1. Послідовність опрацювання оцінки проблеми для масиву вод

Етапи роботи запропонованої інформаційної технології можна продемонструвати за допомогою діаграми, показаній на рис. 1.

На етапі збирання інформації користувач веб-системи (експерт) описує проблему для певного масиву вод. Після цього здійснюється верифікація достовірності опису. Верифікація здійснюється за допомогою засобів NLP, зокрема, пропонується використовувати технологію BERT. В результаті верифікації, вказаному опису виставляється певна оцінка $k_{i,j}$ де i — індекс проблеми, j — індекс водного масиву, яка вказує на ступінь достовірності опису. Ця оцінка може набувати значення від 0 до 1. Також потрібно виділити певне число, яке означає поріг достовірності. Відповідно, якщо $k_{i,j}$ є меншим цього порогу, тоді опис проблеми i для певного водного масиву j вважається недостовірним і він ігнорується, в іншому випадку — достовірним.

На етапі оброблення даних беруться усі достовірні оцінки проблеми i та їх ступеня достовірності $k_{i,j}$, і за формулою обчислюється $K_{i,j}$ — вага важливості проблеми i для водного масиву j. Для водного масиву ваги важливості проблем $K_{i,j}$ сортуються і, відповідно, проблеми, для яких значенням $K_{i,j}$ є найбільшим у цьому водному масиві, вважаються найзначущими для водного масиву.

На етапі візуалізації на мапі відображаються масиви вод водного басейну та їх найзначущі проблеми — проблеми з найбільшим зна-

ченням ваги важливості. Тому, можна легко побачити, які проблеми є важливими для водного масиву.

Запропонуємо математичний апарат для реалізації цих ідей на основі теорії нечітких множин та поєднаємо його NLP-технологіями і методами машинного навчання.

Розроблення математичного апарату для аналізу експертних оцінок з використанням теорії нечітких множин

У статтях [9], [10] охарактеризовані методи та підходи нечіткого оцінювання основних параметрів (паспортних характеристик) малих річок за експертними оцінками та їх опрацювання на основі теорії нечітких множин, зокрема, метод «один об'єкт — один експерт», метод синхронізації інтервалів значень, коли кожен експерт проводить оцінювання на своєму інтервалі значень, підходи щодо мінімізації суб'єктивізму експертів та врахування їх фаху та досвіду. Адаптуємо ці підходи до нашої задачі і поєднаємо з сучасними можливостями, які надають методи машинного навчання та технологій NLP.

Як зазначено вище, експерти оцінюють наявність кожної проблеми із наперед заданої множини, складеної експертами, що розробляють ПУРБ, для вибраного на карті чи зі списку масиву вод. Оцінка формалізується на так званій універсальній множині значень (1 відповідає максимальному значенню інтервалу числових значень параметра, 0 — мінімальному) [3], [9], [10], де знак «+» означає не арифметичну операцію, а операцію приєднання елементів множини:

$$U = 0 + 0{,}1 + 0{,}2 + 0{,}3 + 0{,}4 + 0{,}5 + 0{,}6 + 0{,}7 + 0{,}8 + 0{,}9 + 1. \quad (1)$$

Сама оцінка набуває одного зі значень так званої терм-множини, сформованої для лінгвістичної змінної «Ступінь впливу на екологічний стан масиву вод»:

«НМ» — «Надзвичайно малий», «ДМ» — «Дуже малий», «М» — «Малий»,

«МС» — «Менший середнього», «С» — «Середній», «ВС» — «Вищий середнього», (2)

«В» — «Високий», «ДВ» — «Дуже високий», «НВ» — «Надзвичайно високий».

Кожне значення терм-множини (2) формалізується на універсальній множині (1) з використанням функції належності $\mu_X(u)$, заданої у вигляді дзвінницевої гауссової функції:

$$\mu_X(u) = \exp\left[-\frac{1}{2}\left(\frac{u - m_X}{\sigma_X}\right)^2\right], \quad (3)$$

де параметр m_X — це значення універсальної множини U, заданої виразом (1), взяті у тому ж порядку, а σ_X — параметр, що характеризує розкид значень навколо m_X.

Експертне оцінювання екологічних проблем кожного масиву вод полягає у виборі експертом значення з терм-множини (2), яке, як на його думку, є найвідповіднішим.

Для врахування впевненості експерта у своїх оцінці та рівня достовірності його тверджень пропонується використовувати широко відомий підхід, описаний, також, у роботі [10], згідно з яким це робиться через параметр σ_X у виразі для ФН (3) (для певної проблеми у певному масиві вод, тому індекси i та j опускаємо):

$$\sigma_X = \sigma_0 \beta k, \quad (4)$$

де σ_0 — базове значення параметра σ_X для усіх терм-множин; β — коефіцієнт, який враховує впевненість експерта у своїй оцінці $(0 < \beta \le 1)$, k — коефіцієнт, який враховує достовірність оцінки експерта $(0 \le k \le 1)$.

Чим вагоміша експертна оцінка, тим σ_X менше, а ФН має більш стиснену по ширині (вузьку) форму, а чим експертна оцінка менш вагома, тим σ_X більше, а ФН має розтягнутішу по ширині (широку) форму. Як відомо, ширина ФН має значення під час зіставлення різних експертних оцінок за формулою дефаззифікації. Узагальнене значення є ближчим до більш впевнених оцінок.

Можна по-різному зібрати інформацію щодо впевненості експерта у своїй оцінці, наприклад запропонувати йому вибрати один з трьох варіантів, звідки він знає про цю проблему:

1) бачив на власні очі ($k_c = 1$);

2) з достовірних джерел (наукові статті, довідники, звіти державних установ) ($k_c = 0{,}9$);

3) з соцмереж ($k_c = 0{,}5$ чи $0{,}25$).

Для визначення достовірності оцінки пропонується вимагати від кожного експерта вказувати посилання на текстові веб-ресурси, які містять підтвердження наявності зазначених проблем у заданому масиві вод. Далі методами NLP, по-перше, перевіряти чи текст містить географічні об'єкти (наприклад, назви адміністративних утворень, річок, водойм та населених пунктів, які містить відповідний масив вод у певній ГІС). У разі успішності цієї перевірки, перевіряти наскільки кожне речення цього тексту відповідає сформульованій проблемі. Для цього слід мати наперед треновані моделі для кожного виду проблем, за якими застосовувати для кожного речення тексту метод predict_proba, який оцінює ймовірність кожного класу моделі. У разі бінарної цільової ознаки, як в нашому випадку, повертає ймовірність значення 1, тобто того, що задане речення тексту містить опис заданої проблеми [11]. Для кожного речення тексту обчислюється така ймовірність p_r, а потім серед них визначається максимальне значення P, яке присвоюється всьому тексту

$$P = \max_r p_r, \quad r = \overline{1,r}. \tag{5}$$

У разі, якщо експерт надає m текстів на підтвердження своєї оцінки, тоді загальну для них достовірність k для формули (4) пропонується обчислювати за таким виразом:

$$k = 1 - \prod_{(q=1)}^{m} \left(1 - P_q\right). \tag{6}$$

Наприклад, якщо $P_1 = 0{,}7$; $P_2 = 0{,}6$, тоді $k = 1 - (1 - P_1)(1 - P_2) = 1 - 0{,}3 \cdot 0{,}4 = 0{,}88$.

Але, важливо враховувати випадок, коли ймовірність P_q жодного з текстів не подолає певний мінімальний поріг P^*, наприклад у 0,6, що означатиме негативний результат етапу верифікації оцінки, оскільки параметр σ_X (див. формулу (4)) дорівнюватиме 0. Для врахування такого обмеження пропонуємо (6) записати у вигляді

$$k = \begin{cases} 0, \text{ коли } \overset{m}{\underset{q=1}{\wedge}} P_q < P^*, \\ 1 - \prod_{q=1}^{m} \left(1 - P_q\right), \text{ коли } \overset{m}{\underset{q=1}{\vee}} P_q \geq P^*, \end{cases} \tag{7}$$

тобто перше значення має місце, коли зазначена умова виконується для усіх текстів, а друге значення — коли хоча б для одного тексту ймовірність не є нижчою порогу P^*.

Випадок $k = 0$ означає, що експертна оцінка не пройшла верифікацію і має бути відкинута, не збережена, а формули (1)—(4) для її опрацювання застосовувати не слід.

Побудова інформаційної веб-системи

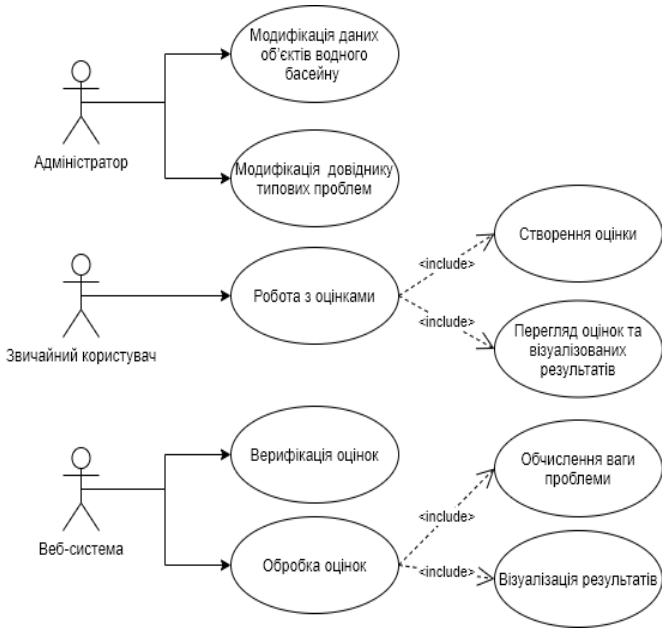

Рис. 2. Use Case-діаграма веб-системи

Охарактеризуємо інформаційну веб-систему, яка могла б реалізувати запропоновані підходи для опрацювання даних. Користувачів такої системи пропонуємо ділити, як це загальноприйнято, на дві групи: користувачі з правами адміністратора та звичайні користувачі (рис. 2). Звичайні користувачі можуть додавати власні оцінки і переглядати «узагальнений статус» масиву вод чи водного басейну загалом. Користувачі з правами адміністратора можуть додатково заповнювати довідник типових для масивів вод проблем та, наприклад, редагувати їх опис.

Доречно повний список екологічних проблем з детальним описом навести у Довідці чи у спеціальному вікні, а у самій веб-системі замінити їх на стисліші і зрозуміліші широким верствам населення формулювання, наприклад, замінити «Забруднення біогенними речовинами: дифузні та/або точкові

джерела» на «Цвітіння води», що зазвичай є наслідком такої проблеми. Інші варіанти: «Забруднення органікою», «Маловоддя», «Підтоплення», «Розораність берегів та долини», «Засмічення берегів та акваторії» тощо.

Один з авторів цієї статті А. М. Лучко розробив пілотну експертну інформаційну веб-систему RiversECO. Система є простою у використанні, з гарною візуалізацією, щоб потенційним експертам було цікаво та легко з нею працювати. На її головній сторінці (рис. 3) розташована мапа OpenStreetMaps з водними масивами, для яких користувачі можуть додавати оцінки та переглядати основні проблеми.

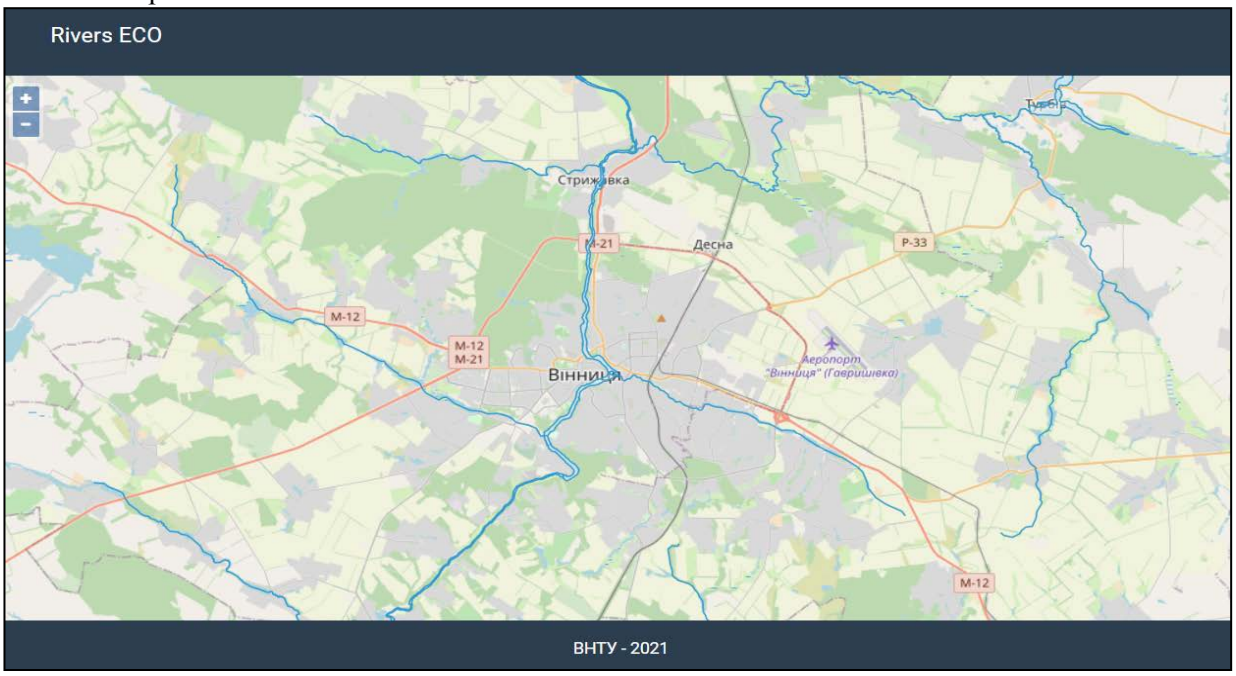

Рис. 3. Головна сторінка веб-системи

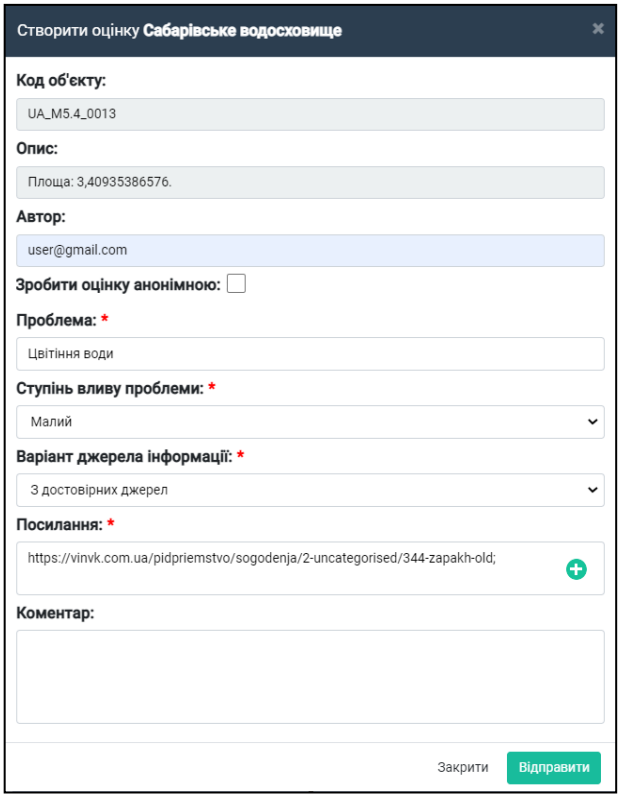

Рис. 4. Форма створення оцінки

Для того, щоб додати оцінку для певного масиву вод, користувачеві потрібно натиснути мишею на нього на мапі та натиснути на кнопку «Додати оцінку». Після цих дій він побачить форму створення оцінки (рис. 4). В ній користувач вводить таку інформацію:

– своє ПІБ та електронну пошту (за бажанням); якщо була поставлена помітка «Зробити оцінку анонімною», тоді персональна інформація не буде видима і доступна публічно;

– обирає екологічну проблему з випадного списку (обов’язково);

– вводить значення експертної оцінки з множини (2) щодо його бачення ступеню впливу цієї проблеми на екологічний стан обраного масиву вод (обов’язково);

– вибирає варіант джерела інформації (наприклад: власний досвід, з достовірних джерел, із соцмереж) (обов’язково);

– вказує посилання на відкрите(і) веб–джерело(а), де є текстова інформація про цю проблему для вибраного масиву вод (обов’язково);

– додає власний коментар (за бажанням).

Після того, як користувач відправить свою оцінку для водного об’єкта, її можна буде побачити у списку оцінок для вибраного водного масиву зі статусом результату верифікації (рис. 5):

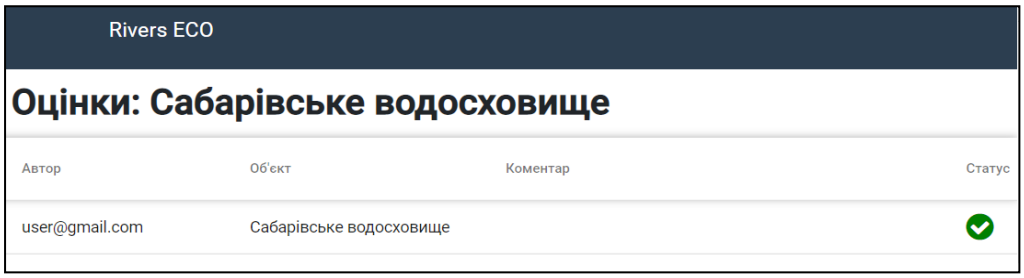

Рис. 5. Експертна оцінка зі статусом результату верифікації

Приклад використання розробленої веб-системи

Для випробування запропонованих підходів та створеної веб-системи вибрані масиви вод, розташовані на р. Південний Буг у Вінницькій області. Як підтверджуючі тексти використано окремі речення монографії одного з авторів статті В. Б. Мокіна [12], розташовані ним у датасеті [13] на платформі Kaggle. Повний текст монографії, написаний колективом експертів шведсько-українського проєкту, є доступним у багатьох веб-ресурсах, наприклад на сайті Регіонального офісу водних ресурсів у Миколаївській області: https://mk-vodres.davr.gov.ua/sites/default/files/Bug_plan_final_2.pdf.

Авторами розроблено програму-ноутбук на Python, яка застосовує модель DistilBertModel та логістичну регресію для опрацювання так званих ембедингів цієї BERT-моделі [14]. Для кожного речення розраховується p_r.

Наведемо приклади розрахунку параметра k для декількох оцінок (табл. 1, 2). Задамо мінімальне порогове значення ймовірностей, які слід вважати достовірними.

Обчислимо загальну достовірність k за формулою (6) для оцінки експерта проблеми «Забруднення води речовинами» у масиві вод м. Вінниця з використанням трьох текстових джерел (табл. 1), для яких у [14] порахована ймовірність відповідності цій проблеми.

Таблиця 1

Оцінка для масиву вод UA_M5.4_0013, де розташоване м. Вінниця

Тексти, що підтверджують проблему	P_r
«Однак, КОС великих міст (ЕН > 100000) Кіровограда, Вінниці і Хмельницького також працюють незадовільно і призводять до підвищеного надходження сполук азоту у природні водні об'єкти.»	0,79
«З листопада по квітень концентрація нітратних іонів варіювала в межах 2…9 мг/дм^3 а у вересні — 0,5…2 мг/дм^3, окрім створів нижче скидів основних забруднювачів — Хмельницького та Вінницького водоканалів.»	0,87
«У басейні Південного Бугу основна частка вказаних сполук (71 %) скидається через очисні споруди міст Вінниця, Хмельницький, Кіровоград, Умань та Первомайськ.»	0,10

$$k = 1-(1-P_1)(1-P_2)(1-P_3) = 1-0,21\cdot 0,13\cdot 0,9 \approx 0,98.$$

Як бачимо, загальна достовірність k цієї оцінки становить 0,98, що є дуже високим значенням. Це означає, що ця оцінка є достовірною.

Аналогічно обчислимо загальну достовірність k за формулою (6) для проблеми «Забруднення води речовинами» у масиві вод м. Ладижин з використанням трьох текстових джерел теж з [14] (табл. 2).

Таблиця 2

Оцінка для масиву вод UA_M5.4_0019, де розташоване м. Ладижин

Тексти, що підтверджують проблему	P_r
«Домінуючу частину (86 %) фосфатних іонів у скидних водах досліджуваних підприємств забезпечують Смолінська шахта, Ладижинська ТЕС Новокостянтинівська шахта, ВАТ «БОС» м. Вознесенськ.»	0,83
«В Ладижинському водосховищі, завдяки великому об'єму та скиду підігрітих вод з ТЕС відбувається інтенсифікація біологічних процесів та зростання біотопічного і біологічного різноманіття.»	0,31
«Більша частина населення басейну Південного Бугу, що становить 3,08 млн осіб (73 %), збирає стічні води в індивідуальні накопичувачі або скидає їх прямо на поверхню.»	0,07

$$k = 1-(1-P_1)(1-P_2)(1-P_3) = 1-0,17\cdot 0,69\cdot 0,93 \approx 0,89.$$

Загальна достовірність k такої оцінки становить 0,89. На основі цього можна зробити висновок,

що така оцінка також є достовірною.

Також проаналізуємо експертну оцінку для масиву вод з м. Вінниця, в якій вказані тексти не містять інформації про проблему (табл. 3).

Таблиця 3

Оцінка для масиву вод UA_M5.4_0014, де розташовано м. Вінниця

Тексти, що підтверджують проблему	P_r
«Сучасний рельєф території формувався під впливом геологічних процесів і складається з підвищених і понижених ділянок.»	0,30
«У травні починає розвиватися грозова діяльність.»	0,13
«Переважна більшість КОС у басейні Південного Бугу обладнана спорудами для біологічної очистки стічних вод.»	0,11

Аналіз ймовірностей P_r показує, що жодна з них не долає мінімальне порогове значення, а це означає оцінка не є достовірною, оскільки не містить фактичного підтвердження, а отже, вона не буде верифікована та не буде збережена в системі.

Отже, проілюстровано яким чином можна застосувати запропоновані у роботі підходи.

Висновки

Розглянуто питання збирання великої кількості експертних оцінок про наявний стан вод, наявні екологічні проблеми та впливові фактори, які збільшують ризик недосягнення екологічних цілей кожним масивом вод під час розроблення програм заходів планів управління річкового басейну (ПУРБ). Охарактеризована проблема збирання та опрацювання інформації для великої кількості масивів вод й її верифікація. Щоб розв’язати це питання, запропоновано створити веб-систему з картою масивів вод, яка дозволить залучати велику кількість експертів. Також запропоновано вимагати від експертів не просто оцінки на основі єдиних довідників варіантів можливих відповідей, а й посилання на джерела, які підтверджують їх слова. Це дозволить усунути проблему перевірки достовірності оцінок.

Розглянуто та проаналізовано різноманітні підходи для зіставлення експертних оцінок як на основі нечітких множин, так і за допомогою технологій опрацювання природної мови (англ. — Natural Language Processing (NLP)).

Розроблено технологію роботи експертної інформаційної веб-системи виявлення пріоритетних екологічних проблем у масивах вод басейну річки. Відзначено, що в цій технології можна виділити три етапи: етап збирання інформації про проблеми, етап оброблення зібраних даних та етап візуалізації результатів. На етапі збирання інформації про проблему користувач інформаційної технології (експерт) описує проблему для певного масиву вод. Після цього здійснюється верифікація опису на достовірність інформації за допомогою засобів NLP. В результаті верифікації, вказаному опису виставляється певна оцінка — ступінь достовірності. Якщо ступінь достовірності оцінки є меншою за певний поріг, тоді оцінка проблеми для водного масиву вважається недостовірною і вона ігнорується, в іншому випадку — достовірною і такою, що успішно пройшла етап верифікації. На етапі оброблення даних беруться усі достовірні оцінки проблеми водного масиву та обчислюється вага важливості проблеми для водного масиву. Для водного масиву вважаються найзначущими проблеми з найбільшими значеннями ваги важливості. На етапі візуалізації на мапі відображаються масиви вод водного басейну та їх найзначущі проблеми.

Розроблено та описано математичний апарат для аналізу експертних оцінок з використанням теорії нечітких множин, який адаптує існуючі методи та підходи нечіткого оцінювання основних параметрів малих річок за експертними оцінками та їх опрацювання на основі теорії нечітких множин і поєднує їх зі сучасними можливостями, які надають методи машинного навчання та технології NLP.

Наведено приклад реалізації інформаційної веб-системи виявлення пріоритетних екологічних проблем у масивах вод басейну річки та продемонстровано приклад використання розробленої системи на певних наборах текстових даних.

Запропонований математичний апарат та підходи до опрацювання експертних оцінок можуть бути застосовані не тільки для аналізу екологічного стану масивів вод водних об’єктів України, а й для інших подібних об’єктів, де доцільно залучення великої кількості експертів з різною кваліфікацією та досвідом, а кожна оцінка може мати підтвердження у текстових веб-ресурсах.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Кабінет Міністрів України, *Постанова № 336 від 18.05.2017 року «Про затвердження Порядку розроблення плану управління річковим басейном».* [Електронний ресурс]. Режим доступу: https://www.davr.gov.ua/postanova-kabinetu-ministriv-ukraini-vid-18-travnya-2017-roku--336-pro-zatverdzhennya-poryadku-rozroblennya-planu-upravlinnya-richkovim-basejnom .

[2] The EU Water Framework Directive – integrated river basin management for Europe Directive 2000/60/EC establishing a framework for the Community action of water policy (Water Framework Directive).

[3] Ю. І. Мітюшкін, Б. І. Мокін, і О. П. Ротштейн, *Soft Computing: ідентифікація закономірностей нечіткими базами знань*, моногр. Вінниця, Україна: УНІВЕРСУМ-Вінниця, 2002, 145 с. [Електронний ресурс]. Режим доступу: http://mokin.com.ua/files/articles/60/88/SoftComputing.pdf .

[4] В. М. Дубовой, Р. Н. Квєтний, О. І. Михальов, і А. В. Усов, *Моделювання та оптимізація систем*, підруч. Вінниця, Україна: ПП «ТД«Едельвейс», 2017, 804 с.

[5] Haining Ding, Xiaojian Hu, and Xiaoan Tang «Multiple-attribute group decision making for interval-valued intuitionistic fuzzy sets based on expert reliability and the evidential reasoning rule,» *Neural Computing and Applications*, vol. 32, pp. 5213-5234, 2020. [Electronic resource]. Available: https://link.springer.com/article/10.1007%2Fs00521-019-04016-z .

[6] Jean-Valère Cossu, Emmanuel Ferreira, Killian Janod, Julien Gaillard, and Marc El-Bèze «NLP-Based Classifiers to Generalize Expert Assessments in E-Reputation,» in *International Conference of the Cross-Language Evaluation Forum for European Languages*, vol. 9283, pp. 340-351, 2015. [Electronic resource]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24027-5_37 .

[7] Matthew Matero, et al., «Suicide Risk Assessment with Multi-level Dual-Context Language and BERT,» *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 39-44, 2019. [Electronic resource]. Available: https://www.aclweb.org/anthology/W19-3005.pdf .

[8] Er. Samadhan, U. Birajdar, and V. A. Losarwar, «Mining opinion targets and opinion words from reviews using natural language processing (NLP) techniques,» *Journal of Critical Reviews*, vol. 8, pp. 1029-1036, 2020. [Electronic resource]. Available: http://www.jcreview.com/?mno=95186 .

[9] В. Б. Мокін, «Ідентифікація параметрів малих річок на основі теорії нечітких множин по експертних оцінках та по їх геоінформаційній моделі,» *Вісник ЖДТУ*, с. 133-142, 2004.

[10] В. Б. Мокін, «Новий підхід до ідентифікації параметрів малих річок за нечіткими експертними оцінками,» *Вісник Вінницького політехнічного інституту*, № 4, с. 34-41, 2005.

[11] *Scikit-learn 0.24.1. Machine Learning in Python. User Guide.* [Electronic resource]. Available: https://scikit-learn.org/stable/user_guide.html .

[12] V. Mokin, D. Pasichniuk, O. Radetskyi, and M. Horash, *Kaggle Dataset NLP: Reports & News Classification. ENG & UKR Automatic Environmental Reports & News Classification*, 2020. [Electronic resource]. Available: https://www.kaggle.com/vbmokin/nlp-reports-news-classification .

[13] Афанасьєв С., та ін., *План управління річковим басейном Південного Бугу: аналіз стану та першочергові заходи*. Київ, Україна: ТОВ «НВП «Інтерсервіс», 2014, 188 с.

[14] V. Mokin, A. Luchko, and O. Davidyuk, *Kaggle. NLP for EN: BERT Predict_proba in Water Report*, [Electronic resource]. Available: https://www.kaggle.com/vbmokin/nlp-for-en-bert-predict-proba-in-water-report?scriptVersionId=53779986 .

Рекомендовано до друку кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 9.02.2021

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

Лучко Андрій Михайлович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: andriyluchko@gmail.com ;

Давидюк Оксана Миколаївна — аспірантка кафедри системного аналізу та інформаційних технологій, e-mail: davidyuk-ok@ukr.net .

Вінницький національний технічний університет, Вінниця;

Вуж Тетяна Євгенівна — канд. техн. наук, доцент кафедри біологічної фізики, медичної апаратури та інформатики, e-mail: tatiana.vuzh@gmail.com .

Вінницький національний медичний університет імені М. І. Пирогова, Вінниця

V. B. Mokin[1]
A. M. Luchko[1]
O. M. Davydiuk[1]
T. Ye. Vuzh[2]

Technology of Construction of Expert Information Web System of Identification and Verification of Priority Ecological Problems in Water Bodies of the River Basin

[1]Vinnytsia National Technical University;
[2]National Pirogov Memorial Medical University, Vinnytsia

The article considers the collection, verification and generalization of a large number of expert assessments of the current state of waters, existing environmental problems and influential factors that increase the risk of failure to achieve environmental goals of each volume of water during the development of river basin management plans (RBMPs) or stabilization of good ecological status of water in the water bodies of this basin. The task is complicated by the large number of such volumes of water, as it is extremely difficult to gather reliable information about the objects located in each of them. Creation of a web system with a map of water volumes and the involvement of a large number of experts from among locals who are not indifferent to the problems of their environment will help to solve this problem. However, then there is the problem of verifying the assessments of these experts, taking into account their different qualifications, experience, different views of the objectives of the RBMP, and the problem of how to compare them to identify the most vulnerable regions by different criteria. It is proposed that experts will be required not only expert assessments on the basis of single directories of possible answers, but also - links to text web resources that confirm their assessments to solve this problem. And then it will be analyzed if these sources really confirm each assessment of the appropriate type of problem for a given region. The authors consider various approaches for comparing expert assessments both based on the basis of fuzzy sets and with the help of machine learning and natural language processing (NLP) technologies. Analogues of the system developed by the authors are considered.

A method has been developed to identify and verify priority of environmental problems in water bodies of the river basin based on fuzzy expert estimates, taking into account the probabilities that the text materials cited by the expert do correspond to this problem. These probabilities are determined to use models of NLP technologies. The stages of functioning of the expert information web system for the implementation of the proposed technology are described, which will simultaneously collect the most reliable and detailed information about the objects of water bodies and accelerate its processing and ranking.

An example of the implementation of an information web system for identifying priority environmental problems in the water bodies of the Southern Bug River basin is given. Examples of calculating the reliability of expert estimates using the author's program in Python based on NLP-model BERT and logistic regression which were applied to real text information are given.

Keywords: information web system, water protection technologies, expert assessments, fuzzy sets, environmental problems, water body, RBMP, machine learning, text data classification, natural language processing, NLP, BERT.

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technology, e-mail: vbmokin@gmail.com ;

Luchko Andrii M. — Post-Graduate Student of the Chair of System Analysis and Information Technology, e-mail: andriyluchko@gmail.com ;

Davydiuk Oksana M. — Post-Graduate Student of the Chair of System Analysis and Information Technology, e-mail: davidyuk-ok@ukr.net ;

Vuzh Tetiana Ye. — Cand. Sc. (Eng.), Associate Professor of the Chair of Biophysics, Informatics and medical equipment, e-mail: tatiana.vuzh@gmail.com

В. Б. Мокин[1]
А. М. Лучко[1]
О. Н. Давидюк[1]
Т. Е. Вуж[2]

Технология построения экспертной информационной веб-системы выявления и верификации приоритетных экологических проблем в массиве вод бассейна реки

[1]Винницкий национальный технический университет;
[2]Винницкий национальный медицинский университет имени Н. И. Пирогова

Рассмотрены вопросы сбора, верификации и обобщения большого количества экспертных оценок о реальном состоянии вод, имеющихся экологических проблемах и влияющих факторах, увеличивающих риск недостижения экологических целей каждым массивом вод при разработке программ мероприятий планов управления речного бассейна (ПУРБ), направленных на достижение или стабилизацию хорошего экологического состояния воды в массивах вод этого бассейна. Задача усложняется большим количеством таких массивов вод, поскольку собрать достоверную информацию об объектах, расположенных в каждом из них, крайне трудно. Решить эту проблему позволит создание веб-системы с картой массивов вод и привлечением большого количества экспертов из числа местных жителей, неравнодушных к проблемам своей окружающей среды. Однако, тогда возникает проблема проверки достоверности оценок этих экспертов, учитывая их разную квалификацию, опыт работы, разное видение целей ПУРБ, и проблема как их сопоставлять, чтобы определить наиболее уязвимые регионы по различным критериям. Для этого предложено требовать от экспертов не просто экспертные оценки на основе единых справочников вариантов возможных ответов, но и ссылки на текстовые веб-ресурсы, подтверждающие их оценки. А потом анализировать насколько эти источники действительно подтверждают каждую оценку относительно соответствующего типа проблемы для этого региона. Авторами рассмотрены различные подходы для сопоставления экспертных оценок как на основе нечетких множеств, так и с помощью технологий машинного обучения и обработки естественного языка (англ. — Natural Language Processing (NLP)). Рассмотрены аналоги разрабатываемой авторами системы.

Разработан метод выявления и верификации приоритетных экологических проблем в массивах вод бассейна реки с нечеткими экспертными оценками, с учетом вероятностей того, что процитированные экспертом текстовые материалы действительно соответствуют указанной проблеме. Эти вероятности определяются с использованием моделей NLP-технологий. Охарактеризованы этапы функционирования экспертной информационной веб-системы для реализации предложенной технологии, которая позволит одновременно собрать максимально достоверную и подробную информацию об объектах массивов вод и ускорить ее обработку и ранжирование.

Приведен пример реализации информационной веб-системы выявления приоритетных экологических проблем в массивах вод бассейна реки Южный Буг. Приведены примеры вычисления достоверности экспертных оценок с применением авторской программы на Python на основе NLP-модели BERT и логистической регрессии к реальной текстовой информации.

Ключевые слова: информационная веб-система, технологии защиты водных ресурсов, экспертные оценки, нечеткие множества, экологические проблемы, массив вод, ПУРБ, машинное обучение, классификация текстовых данных, технологии обработки естественного языка, NLP, BERT.

Мокин Виталий Борисович — д-р техн. наук, профессор, заведующий кафедрой системного анализа и информационных технологий, e-mail: vbmokin@gmail.com ;

Лучко Андрей Михайлович — аспирант кафедры системного анализа и информационных технологий, e-mail: andriyluchko@gmail.com ;

Давидюк Оксана Николаевна — аспирант кафедры системного анализа и информационных технологий, e-mail: davidyuk-ok@ukr.net ;

Вуж Татьяна Евгеньевна — канд. техн. наук, доцент кафедры биологической физики, медицинской аппаратуры и информатики, e-mail: tatiana.vuzh@gmail.com