

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

БАРБАРУК ЛІНА ВІКТОРІВНА

УДК 004.067:639.311+639.312

**МОДЕЛІ ТА МЕТОД ОБРОБКИ ВЕЛИКИХ ДАНИХ В
ІНФОРМАЦІЙНО-АНАЛІТИЧНИХ СИСТЕМАХ МОНІТОРИНГУ
ВОДНИХ ОБ'ЄКТІВ**

05.13.06 – інформаційні технології

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Сєвєродонецьк – 2021

Дисертацією є рукопис.

Робота виконана на кафедрі комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Скарга-Бандурова Інна Сергіївна,
Східноукраїнський національний університет ім. В. Даля,
професор кафедри комп'ютерних наук та інженерії.

Офіційні опоненти: доктор технічних наук, професор
Петров Костянтин Едуардович,
Харківський національний університет радіоелектроніки,
завідувач кафедри інформаційних управляючих систем;

доктор технічних наук, доцент
Доценко Сергій Ілліч,
Український державний університет залізничного
транспорту,
доцент кафедри спеціалізованих комп'ютерних систем.

Захист відбудеться «22» квітня 2021р. о 13⁰⁰ годині на засіданні спеціалізованої вченої ради К 29.051.16 у Східноукраїнському національному університеті імені Володимира Даля за адресою: 93400, м. Сєвєродонецьк, вул. Донецька, 43.

З дисертацією можна ознайомитись у науково-технічній бібліотеці Східноукраїнського національного університету імені Володимира Даля за адресою: 93400, м. Сєвєродонецьк, вул. Донецька, 43.

Автореферат розісланий «20» березня 2021р.

Вчений секретар
спеціалізованої вченої ради К 29.051.16,
кандидат технічних наук, доцент



С.О. Сафонова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми дослідження. Якість води є одним з найважливіших факторів здорової екосистеми. Чиста вода підтримує різноманітність рослин та дикої природи. Особливі потреби до якості води в аквакультурі, оскільки неякісна вода може впливати на здоров'я та ріст риб. У ставках і штучних водоймах хімічні та фізичні показники води можуть швидко змінюватись, оскільки риби використовують воду для проживання, харчування, розмноження, тощо. Отже, ефективне використання води, її якість, потреби аквакультури та способи управління факторами якості води є основними чинниками екологічного благополуччя. Найкращим рішенням для ефективного використання води є можливість точного розуміння в реальному часі її фізико-хімічних параметрів, обсягів споживання та попиту. Таку інформацію можна отримати лише за допомогою моніторингових систем і засобів потужної аналітики.

Дослідженням обробки даних у водній галузі займаються вчені всього світу, зокрема, значний внесок у розвиток систем обробки та аналізу даних зробили видатні вітчизняні та зарубіжні вчені: В.П. Ковальчук, О.М. Клименко, В.Б. Мокін, А.М. Прищеп, А.В. Яцишин, R. Allen, C. Briesse, W. Wagner, G. Wong та ін. Стрімкий розвиток інформаційних технологій, систем моніторингу та обробки великих даних показав необхідність організації підтримки рішень в умовах нечіткої інформації. Серед вчених, які займалися розвитком теорії та методів обробки нечіткої інформації варто відзначити L. Zadeh, E. Mamdani, M. Sugeno, В.Д. Шапіро, Є. В. Бодянський, А.О. Каргін, Ю.П. Кондратенко, О.В. Леоненков, Д.О. Поспелов, А.П. Ротштейн та ін. Більшість світових експертів з питань води та водного господарства, наголошують на необхідності інвестувати в аналітичні інструменти та рішення з великими даними як наріжний камінь, для зменшення навантаження на водні ресурси. Разом з тим, будь-яка моніторингова система, що працює в режимі реального часу, зокрема система моніторингу водного середовища, стикається з великим тиском при обробці даних. З одного боку, традиційні аналітичні інструменти не здатні підтримувати обробку великих обсягів даних, з іншого боку - засоби аналітики Big Data, у більшості випадків, не орієнтовані на обробку даних систем моніторингу водних об'єктів. Хоча багато галузей промисловості вже використовують потужність великих даних та додатків аналітики для безперебійної роботи, є багато областей, куди аналітика все ще не змогла повноцінно проникнути та заощадити природні ресурси, до таких областей відноситься і вода. Таким чином, існує об'єктивне протиріччя між зростаючою кількістю впроваджень високотехнологічних систем он-лайн моніторингу водних об'єктів і наявними технологіями управління та обробки великих наборів даних, які ці системи здатні збирати.

Виходячи з вищевикладеного, *актуальним науковим завданням* є розроблення моделей та методу інформаційної технології, здатних забезпечити ефективну обробку та використання великих даних, отримуваних від систем моніторингу водних об'єктів.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у відповідності з науково-технічними планами кафедри

комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, у відповідності до державних програм і планів НДР, а також міжнародних проектів і програм:

- НДР “Дослідження стратегій та механізмів прийняття рішень для інтегрованого управління водними ресурсами” (Східноукраїнський національний університет імені Володимира Даля, № ДР 0116U005784, 12.2016-07.2020 pp.);
- НДР “Проектування системи моніторингу та контролю водних об’єктів на основі технології інтернет речей”, (№ ДР 0120U100421, 01.2020-12.2023 pp.);
- Проект Європейського Союзу ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP “Internet of Things: Emerging Curriculum for Industry and Human Applications” (2016-2020 pp.).

Роль автора в зазначених науково-дослідних роботах і проектах, у яких дисертант був безпосереднім виконавцем, полягає у проведенні аналізу моделей та методів обробки великих даних в системах моніторингу водних об’єктів; виборі набору базових компонент для прийняття рішень при обробці моніторингових даних якості води; розробленні методів і засобів моніторингу для автоматизації завдань аналітичної обробки великих даних; розробці та імплементації моделей для роботи з великими даними; практичній реалізації розроблених моделей та методів в інструментальний засіб, призначений для підтримки рішень при управлінні водним господарством.

Мета й завдання дослідження. *Метою* дисертації є підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об’єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології обробки великих даних.

Для досягнення мети дослідження в дисертації сформульовані та вирішені наступні *задачі*:

- аналіз моделей та методів обробки великих даних в системах моніторингу водних об’єктів;
- розроблення математичних моделей для аналітичної обробки великих даних системи моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних;
- розроблення методу нечіткої кластеризації с-середніх для великих даних, шляхом узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних;
- розроблення інформаційної технології на основі запропонованих моделей і методу для обробки великих даних та підтримки прийняття рішень в інформаційно-аналітичній системі моніторингу водних об’єктів;
- розроблення засобів візуалізації великих даних;
- реалізація програмних засобів та елементів інформаційної технології обробки великих даних в інформаційно-аналітичній системі моніторингу водних об’єктів та впровадження отриманих результатів.

Об’єкт дослідження – процеси обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів.

Предмет дослідження – моделі, методи та програмні засоби інформаційної технології забезпечення обробки великих даних.

Методи дослідження. В основу методології дослідження були покладені принципи системного аналізу для декомпозиції процесу обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів; методи пошукового аналізу, технології роботи з літературними джерелами для аналізу перспективних технологій обробки великих даних, оцінювання потенціалу інноваційних технологій моніторингу; методи теоретико-множинного опису, теорії ймовірностей, теорії нечітких множин – для розробки моделі для аналітичної обробки великих даних; нечітка евристична кластеризація с-середніх, інтуїтивістська теорія нечітких множин, міри подібності для нечітких множин з множинними атрибутами, метод множника Лагранжа – для удосконалення методу обробки великих даних та автоматичного маркування нечітких кластерів у контексті аналізу якості водойм рибогосподарського призначення; методи візуалізації кореляційних матриць, зокрема хордові діаграми, криві Безьє, технології спрощення полігональних ланцюгів – для побудови моделей візуалізації великих даних. Для розробки інформаційної технології використовувалися методи Data Mining, MapReduce, OLAP, технології сховищ даних.

Наукова новизна результатів досліджень полягає у наступному.

1) *вперше розроблено* модель для аналітичної обробки великих даних системи моніторингу водних об'єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об'єктів, для подальшого прийняття рішення;

2) *удосконалено метод* нечіткої кластеризації, в якому, на відміну від існуючих, виконано узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних, що дозволило застосовувати автоматичну розмітку при обробці великих даних і покращити конвергенцію алгоритму кластеризації;

3) *удосконалено* технологію та засоби візуалізації великих даних за рахунок використання хордових діаграм та технології спрощення полігональних ланцюгів, що дозволяють зберегти баланс між вимогами до деталізації вихідного зображення та отриманням мінімальної кількості точок на екрані систем он-лайн моніторингу водних об'єктів, знижуючи час на пошук невідповідностей і ретроспективний аналіз великих даних;

4) *набула подальшого розвитку* інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації водних господарств у частині реалізації основних аналітичних функцій.

Практичне значення одержаних результатів полягає в розробленні програмного забезпечення інформаційної технології для обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів. Усі теоретичні положення дисертації доведено до конкретних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-

аналітичних системах моніторингу водних об'єктів і вибору варіантів її використання в системі підтримки прийняття рішень. Використання моделей та інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів, зокрема у виробничих процесах рибогосподарських господарств та підприємств аквакультури, дозволило зробити висновок, про їх ефективність в частині підвищення точності інтегральної оцінки якості вод, зменшення часу на прийняття рішень щодо якості вод водойм рибогосподарського призначення, зниження витрат на виробництво (до 15% під час тестової експлуатації) за рахунок зменшення часу на пошук невідповідностей і ретроспективний аналіз великих даних.

Теоретичні та практичні результати дисертаційної роботи впроваджено:

- в Управлінні Державного агентства рибного господарства у Луганській області (акт впровадження від 12.01.2021 р.);
- у Східноукраїнському національному університеті імені Володимира Даля при розробці навчальних курсів “Комп’ютерне моделювання процесів та систем”, “Методи машинного навчання”, “Data-mining та бізнес-аналітика” на кафедрі комп’ютерних наук та інженерії, виконанні міжнародного проекту ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP “Internet of Things: Emerging Curriculum for Industry and Human Applications” - при розробленні магістерського курсу “Development and implementation of IoT based systems” (акт впровадження від 10.02.2021 р.).

Особистий внесок здобувача полягає в розробці нової моделі, методу, елементів інформаційної технології та програмних засобів, які забезпечують вирішення поставлених у дисертації завдань. Всі наукові результати, наведені в дисертаційній роботі та сформульовані в положеннях, що виносять на захист, отримані автором особисто. Роботи [9, 15, 18] опубліковано без співавторів.

У друкованих працях, опублікованих у співавторстві, здобувачеві належать: [7, 8] проведено аналіз моделей та методів обробки великих даних, [18] виконано класифікацію моделей обробки великих даних, [3] розроблено і досліджено нечіткі моделі для різних кліматичних умов, [10, 11] розроблено модель комплексної оцінки параметрів, [2] проведено узагальнення процедури автоматичного маркування нечітких кластерів, [2, 5] розроблено елементи інформаційної технології обробки великих даних та підтримки прийняття рішень в інформаційно-аналітичній системі, [6, 9, 15, 16] розроблено моделі багатовимірних структур даних, [4] запропоновано підходи до ефективного спрощення та візуалізації великих наборів даних, [1, 12, 13, 14, 17] досліджено результати практичної реалізації технології обробки великих даних в системі моніторингу водних об'єктів.

Апробація результатів наукових досліджень. Основні положення та результати дисертаційної роботи обговорювалися на 12 науково-технічних конференціях і тематичних семінарах, у тому числі: Міжнародній науковій конференції міжнародній науково-технічній конференції “Future Internet of Things and Cloud (FiCloudW)” (м. Стамбул, Турція, 2019 р.); міжнародній науково-технічній конференції “Проблеми інформатики і моделювання (ПІМ)” (м. Харків-Кароліно Бугаз, 2019 р.); міжнародних науково-технічних конференціях “Технологія” (м. Сєвєродонецьк, 2016-2018 рр.); III міжнародній науково-технічній конференції

студентів, магістрів та аспірантів «Інформатика, управління та штучний інтелект» (м. Харків, 2016 р.); міжнародних науково-технічних конференціях “Theoretical and Applied Computer Science and Information Technology (TACSIT)” (м. Сєверодонецьк, 2015, 2017 рр.); всеукраїнських науково-практичних конференціях “Електронні апарати та системи. Проблеми створення. Перспективи розвитку”, “Майбутній науковець” (м. Сєверодонецьк, 2017); міжнародній науково-технічній конференції “Комп’ютерне моделювання в хімії, технологіях і системах сталого розвитку” (м. Київ, 2014 р.); міжнародній науково-технічній конференції “Strategy of Quality in Industry and Education” (м. Варна, Болгарія, 2013 р.), наукових семінарах кафедри комп’ютерних наук та інженерії СНУ ім. В. Даля (м. Сєверодонецьк, 2011-2020 рр.).

Публікації. За темою дисертації з викладенням її основних результатів опубліковано 18 наукових праць, з яких 8 статей [3-10] у фахових наукових виданнях України, включених до міжнародних наукометричних баз, 1 стаття [2] у закордонному виданні, що внесено до міжнародної наукометричної бази Scopus; 1 стаття [1] у фаховому виданні іншої держави, що входить до Європейського Союзу, 8 публікацій [11-18] у збірниках матеріалів і праць міжнародних та всеукраїнських конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу, 4 розділів, висновків, списку використаних джерел з 177 найменувань на 20 сторінках та 3 додатків. Загальний обсяг дисертації становить 166 сторінок, у тому числі: анотації на 9 сторінках, зміст на 4 сторінках, перелік умовних позначень на 1 сторінці, основний текст на 156 сторінках. Робота містить 36 рисунків і 19 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

Вступ дисертації містить: обґрунтування актуальності теми й наукових задач; мету й задачі дослідження; об’єкт, предмет і методи дослідження; характеристику наукової новизни й практичного значення результатів, а також особистого внеску здобувача; дані щодо реалізації, апробації та публікації результатів.

У **першому розділі** проведено аналіз моделей та методів обробки великих даних, розглянуто потенціал інноваційних технологій моніторингу для кращого управління водними ресурсами, проаналізовано особливості великих даних в системах моніторингу водних об’єктів, виконано аналіз нечітких моделей, особливостей кластеризації великих даних, розглянуто парадигму MapReduce. Виконано огляд проблем візуалізації великих даних та підходів до їх вирішення.

Встановлено, що незважаючи на потужні інструменти для обробки та візуалізації, обсяги даних можуть стати важливим питанням в системах моніторингу водних об’єктів. Для українських господарств, що займаються розведенням риби, та інших об’єктів аквакультури, системи, що контролюють параметри води та навколишнього середовища дозволяють оптимізувати їх ефективність за рахунок зниження витрат на оплату праці та комунальних послуг. Безумовними перевагами використання інформаційно-аналітичних систем в процесах аквакультури є підвищення ефективності технологічних процесів, зменшення витрат енергії та води, зменшення витрат на оплату праці, знижений стрес і захворювання риби, покращений

облік, поліпшення розуміння процесу. Разом з тим, враховуючи обсяги та характер великих даних, виникає низка питань щодо організації обробки, зберігання та візуалізації даних, а саме: обсяги даних які можливо і потрібно збирати та зберігати; обсяги та типи даних, необхідні для побудови точних моделей прогнозування; засоби аналізу та візуалізації даних, здатні забезпечити найкраще розуміння ситуації та прийняття рішень; витрати на зберігання, обчислення та візуалізацію великої кількості даних, тощо. Результати проведеного аналізу нормативної бази, моделей, методів й інструментальних засобів обробки великих даних показали необхідність подальшої розробки моделей, методів та інформаційних технологій для систем моніторингу водних об'єктів. Зокрема, виділено наступні три напрями дослідження: розробка нечітких моделей якості води, розробка підходів до зменшення кількості та розмірності даних, зокрема кластеризації даних, розробка засобів, що дозволять візуально дослідити зміни та поведінку окремих параметрів та інтегрального індексу якості вод; що у випадку візуалізації часових рядів може бути досягнуто за допомогою спрощення полігональних ланцюгів та динамічного визначення сили спрощення в залежності від типу і характеру вхідних даних.

За результатами аналізу сформульоване основне науково-прикладне завдання роботи, яке полягає в розробці моделей та методу інформаційної технології для інформаційно-аналітичних систем моніторингу водних об'єктів. Визначені задачі досліджень, обрано стратегію та обґрунтовано методику та математичний апарат досліджень. Основні результати розділу опубліковані у [1, 2, 11, 17, 18].

Другий розділ присвячений дослідженню та розробленню нечітких моделей для оцінювання якості води на прикладі поверхневих вод та вод рибогосподарського призначення. Нечіткий логічний формалізм використаний для опису якості води шляхом розробки індексу якості води на основі нечітких міркувань. Нечіткі умови описані у вигляді відображення від заданого вхідного детермінанта до детермінанта виходу, використовуючи нечіткі логічні міркування.

Модель дозволяє приймати рішення на підставі відображення та розпізнаних шаблонів. Процес нечіткого висновку містить вибір функції належності, операції нечіткого набору та правила висновку.

У загальному виді, нечітка модель аналізу якості води оснований на функціональній залежності виду:

$$f: \bar{P} = \{P_1, P_2, \dots, P_n\} \rightarrow Z, \quad (1)$$

де $\bar{P}$ – вектор вхідних лінгвістичних змінних, які містять оцінки кожного з вимірюваних параметрів якості води, Z - вихідна змінна, значення якої відповідає інтегральній оцінці якості вод.

Щоб отримати інтегральну оцінку якості вод, модель (1) поєднує в собі набір нечітких моделей окремих параметрів моніторингу:

$$P = \{(p_1, \Omega_1^j), (p_2, \Omega_2^j), (p_3, \Omega_3^j), \dots, (p_k, \Omega_k^j)\}, p_k \in A \quad (2)$$

$$\Omega_i^j = \{\omega_i^j \mid \mu_{\Omega_i^j}(\omega_i^j)\}, \omega_i^j \in \Omega_i^j; \quad (3)$$

де $A = \{p_k\}$ – множина ознак якості води, $k \in [1; n]$; k – індекс ознаки якості води; $\Omega_i^j = \{\omega_i^j\}$ – множина значень ознаки p_k , що представляють собою найменування нечітких змінних; i значення індексів $j \in [1; m]$ які позначають множину значень ознаки. Опис ознак p_k , характеризує якість води. Таким чином, відомості про якість води можна отримати якщо відома P_i – поточна якість, де i є відповідний ідентифікатор якості води.

Побудова функцій належності термів лінгвістичних змінних нечіткої моделі аналізу якості вод рибогосподарського призначення виконувалася з використанням наступної методології:

- визначення параметрів, що використовуються в моделі якості вод рибогосподарського призначення;
- формування терм-множин лінгвістичних змінних;
- аналіз наборів даних моніторингу води та аналіз діапазонів гранично-допустимих значень кількісних змінних для кожного зразка, що відповідають лінгвістичним змінним;
- визначення форми та меж інтервалів функцій належності;
- побудова функцій належності для параметрів моніторингу.

Для визначення множини «якість води» $\{Z\}$ використовуються мінімальний набір параметрів, а саме: рН; кількість розчинного кисню; БСК, ХСК, аміак, азот амонійний, які є базовими лінгвістичними змінними моделі.

Після побудови для кожного параметра таблично-заданих функцій належності, було проведено їх апроксимацію класичними функціями належностей: трикутною, трапецієподібною та симетричною функцією Гауса для вибору виду функцій належності кожної терм-множини вхідних лінгвістичних змінних якості води.

В результаті, для побудови функцій належності для всіх параметрів якості води було обрано трапецієподібну функцію, що в загальному виді описується ф. (4).

$$\mu^f(x; a, b, c, d) = \left\{ \begin{array}{ll} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c}, & c \leq x \leq d \\ 0, & d \leq x. \end{array} \right\}, \quad (4)$$

Для визначення меж для інтервалів $X = [X_{\min}, X_{\max}]$ кожного інтервального розбиття параметрів що перекриваються (рис. 1) і отримання значень лінгвістичних змінних a, b, c, d функції належності в місті перетину визначених термів використано ф. (5).

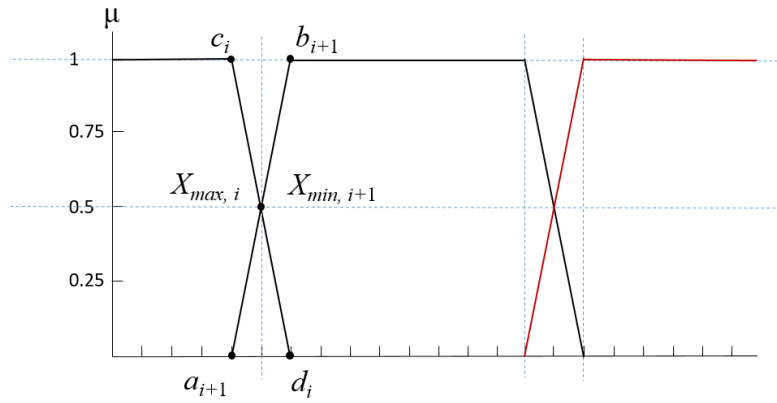


Рисунок 1 – Схема розбиття на межі інтервалів $X_i = [X_{min,i}, X_{max,i}]$ і $X_{i+1} = [X_{min,i+1}, X_{max,i+1}]$.

$$\begin{aligned}
 a_{i+1} &= \frac{X_{min,i+1} + X_{max,i}}{2} - k, \\
 b_{i+1} &= \frac{X_{min,i+1} + X_{max,i}}{2} + k, \\
 c_i &= \frac{X_{min,i+1} + X_{max,i}}{2} - k, \\
 d_i &= \frac{X_{min,i+1} + X_{max,i}}{2} + k,
 \end{aligned} \tag{5}$$

де X_{min}, X_{max} – мінімальні та максимальні значення, для кожного з показників інтервального розбиття;

В якості правил нечіткого виведення для локальних моделей використовуються правила *modus ponens*:

$$I : x = A \Rightarrow y = B, P : x = A', C : y = B' \tag{6}$$

де I позначає імплікацію, P - передумова, C - висновок. Це правило дає можливість зробити висновок про наступника на основі логічного значення попередника. Імплікація трактується як нечітке відношення, що означає, що A' не повинно бути рівним A , а B' не повинно бути рівним B .

З метою уніфікації підходу, виходи (значення інтегрального індексу) представлено п'ятьма класами, а саме a_1 – відмінно, a_2 – добре, a_3 – задовільно, a_4 – погано, a_5 – дуже погано (рис. 2). Нечіткі правила для водної екосистеми розроблені за допомогою знання експертів та ретельного розгляду параметрів отриманих від системи моніторингу, що мають потенційний ризик для здоров'я риби (рН, РК, БСК, NH_3 , ХСК). Представлені моделі є універсальними прикладами в яких правила і межі можуть бути легко адаптовані до прийнятих в господарстві стандартів щодо водорибогосподарського призначення відносно технологічних вимог до вирощування відповідного виду риби.

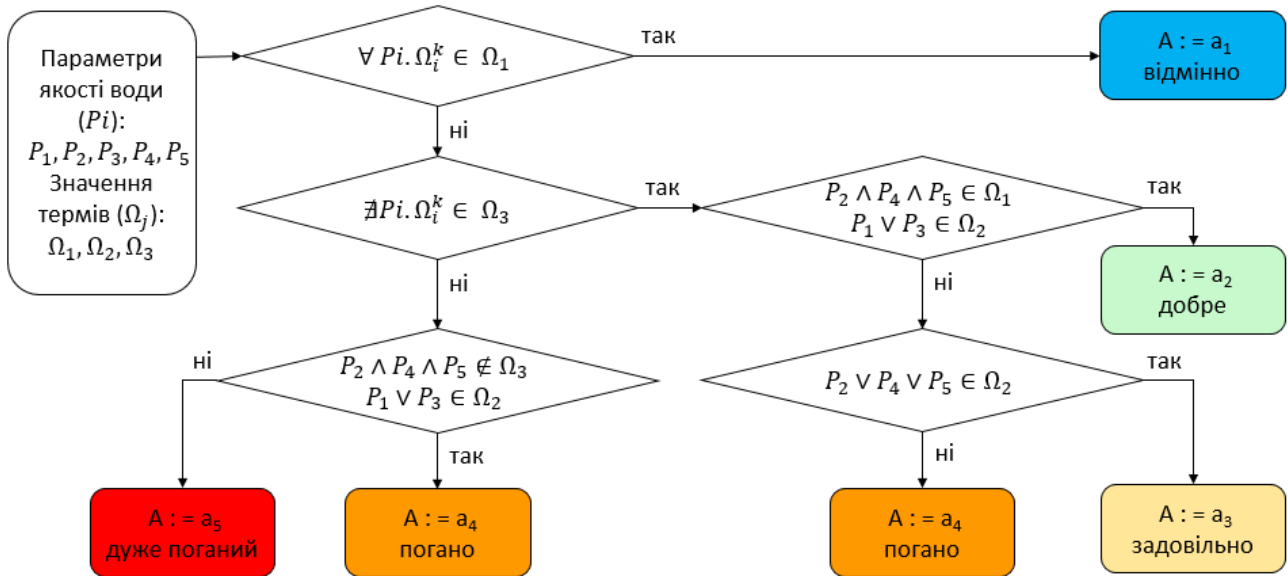


Рисунок 2 – Розширена структура технології використання нечітких правил для оцінки якості вод рибогосподарського призначення

У якості схеми нечіткого висновку використано алгоритм Мамдані. Алгоритм використовує метод $\min$ -активації $\mu'_i(x) = \min \{d_i, \mu_i(x)\}$, де $\mu'_i(x)$ – активізована функція належності; $\mu_i(x)$ – функція належності; d_i – ступінь істинності i -го часткового висновку. Для обчислення об'єднаного нечіткого висновку виконано агрегацію усіх правил нечіткого висновку за допомогою $\min$ -кон'юнкції $c_j = \min\{b_i\}$, де $j = 1..n$; i – число з множини підумов, в яких бере участь j -а вхідна змінна. Метод $\max$ -диз'юнкції був використаний для акумуляції висновків правил, як процедура агрегування, що застосовує операцію об'єднання на всіх усічених нечітких наборах вихідних даних $\mu'_i(x) = \max \{\mu_1(x), \mu_2(x)\}$, де $\mu_1(x), \mu_2(x)$ – функції належності поєднуваних множин.

Дефазифікація сукупного вихідного значення для перетворення нечітких наборів у числове значення досягається центроїдним методом. Математично, дефазифікований вихід u_i представлений як

$$X^* = \frac{\int_{Min}^{Max} \mu_i(x) \cdot x dx}{\int_{Min}^{Max} \mu_i(x) \cdot dx} \quad (7)$$

де $\mu_i(x)$ – функція належності відповідної нечіткої множини E_i ; Min і Max – кордони універсуму нечітких змінних; X^* – результат дефазифікації.

Остаточний числовий бал є нечіткий індекс якості вод рибогосподарського призначення. З метою розширення сфери використання підходу, розроблено також моделі комплексної оцінки поверхневих вод, основою яких є відповідні функції належності, для кожної з яких визначена матриця оцінки нечітких відношень M :

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{15} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n1} & \mu_{n2} & \mu_{n3} & \mu_{n4} & \mu_{n5} \end{bmatrix} \quad (8)$$

Оскільки різні фактори впливають на якість води по-різному, необхідно обчислювати вагу кожного фактору, щоб зробити модель оцінки більш прийнятною до аналізу. Етапи визначення коефіцієнтів ваги містять стандартизацію, нормалізацію та розрахунок матриці судження. При цьому приймається, що дані складаються з n індексів оцінки та m об'єктів оцінювання, які утворюють матрицю X :

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \quad (9)$$

Щоб усунути вплив розмірності та порядок величин, вихідна матриця даних X нормалізується за допомогою методу Z-Score.

$$Z = \frac{x_{ij} - \bar{x}_j}{S_j} \quad (10)$$

де $\bar{x}_j$ - середня кількість j -х зразків, а S_j - стандартне відхилення j -х зразків. Формули для обчислення $\bar{x}_j$ та S_j :

$$\bar{x}_j = \frac{1}{m} \sum_{i=1}^m x_{ij} \quad (11)$$

$$S_j = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (x_{ij} - \bar{x}_j)^2} \quad (12)$$

Визначена таким чином нормована матриця оцінки нечітких відношень використовується для розпізнавання потенційних забруднювачів.

Основні результати розділу опубліковані у [3, 5, 7, 12].

Третій розділ дисертації присвячений розробленню методу обробки великих даних для водойм рибогосподарського призначення. Розглянуто проблему інтерпретації результатів кластеризації на випадок застосування евристичних алгоритмів кластеризації для інтуїтивістської нечіткої обробки даних. Наведено приклад використання методу для аналізу якості вод водойм рибогосподарського призначення. Схема, що описує основні етапи вдосконаленого методу нечіткої кластеризації c -середніх на випадок великих даних, наведена на рис. 3.

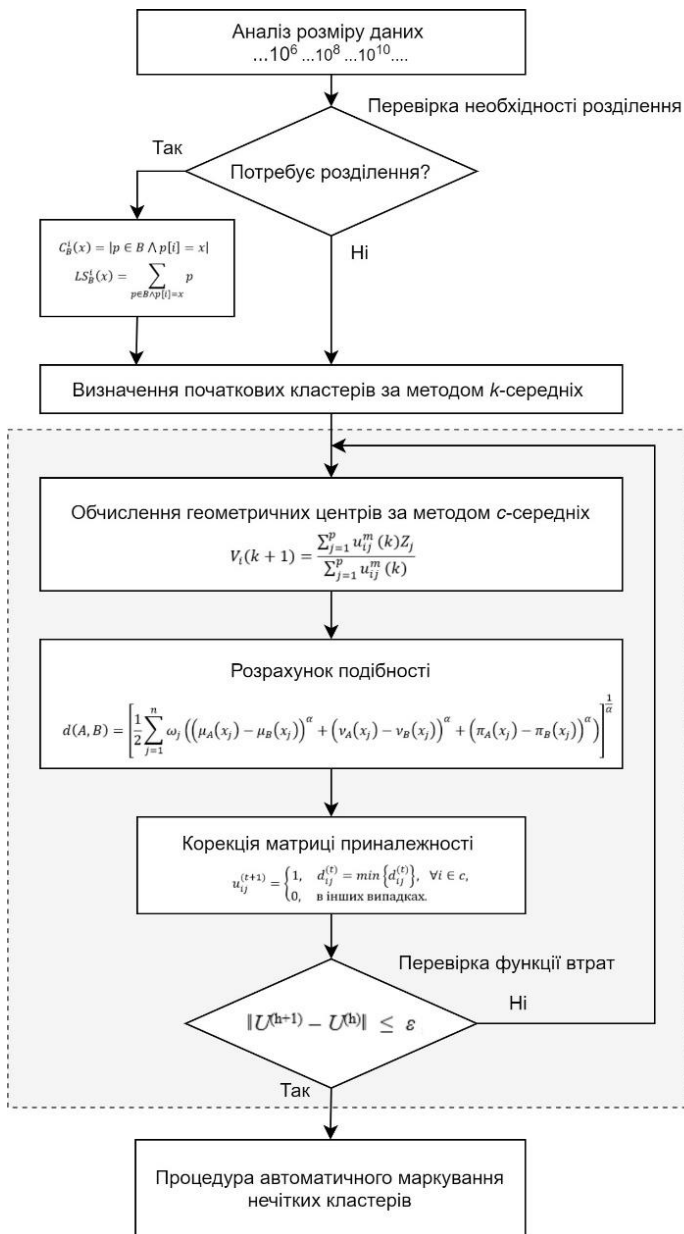


Рисунок 3 – Основні етапи вдосконаленого методу нечіткої кластеризації на випадок великих даних

На етапі 1 проводиться аналіз обсягів даних і приймається рішення щодо потреби в розділенні та розпаралелюванні даних. Для кластеризації даних вище 10^{10} розділення є абсолютно необхідним. Розрахунок починається з одного кластера S , що відповідає цілому набору даних. S вводиться в чергу пріоритетів Q . Q містить блок розділу B , який ітеративно будується над набором даних. Поки Q не порожній, блок B видаляється і розбивається на пару блоків. Якщо розбиття ефективно, пари блоків замінюють B в Q , інакше B стає „остаточним” блоком для цієї фази. Для того, щоб ефективно знайти найкращий розділ для блоку, слід обчислити граничні розподіли блоку. Зокрема, для кожної розмірності i d -вимірного блоку B необхідно обчислити функції $C_B^i: \mathbb{R} \rightarrow \mathbb{N} \subset \mathbb{R}^d$, визначені наступним чином

$$C_B^i(x) = |p \in B \wedge p[i] = x|$$

$$LS_B^i(x) = \sum_{p \in B \wedge p[i] = x} p$$

де підсумовуються всі значення точок p , що належать до блоку B , які мають значення i -ї координати, рівної x .

На етапі 2 застосовується класичний жорсткий метод кластеризації для швидкого визначення центру кластера вхідних даних: на цьому етапі визначається кількість кластерів ($1 < c < n$), та обираються початкові центри кластерів $V(0) = \{V_1, V_2, \dots, V_c\}$. Отримані центри жорсткого кластера використовуються як початковий центр кластера для ітерацій з використанням нечіткої кластеризації. Такий підхід дозволяє прискорити конвергенцію алгоритму.

Основні кроки:

- 1) Визначити кількість кластерів ($1 < c < n$), вибрати константу ε ($\varepsilon > 0$), встановити кількість ітерацій $t = 0$
- 2) Обрати початкові центри кластерів $V(0) = \{V_1, V_2, \dots, V_c\}$
- 3) Розрахувати початкову відстань d між точками даних та кластерами за ф.

(13)

$$d(A, B) = \left[\frac{1}{2} \sum_{j=1}^n \omega_j \left((\mu_A(x_j) - \mu_B(x_j))^\alpha + (\nu_A(x_j) - \nu_B(x_j))^\alpha + (\pi_A(x_j) - \pi_B(x_j))^\alpha \right) \right]^{\frac{1}{\alpha}}, \quad (13)$$

де $\omega = (\omega_1, \omega_2, \dots, \omega_n)^T$ – ваговий вектор елементів x_j ($j = 1, 2, \dots, n$), $\alpha \geq 0$.

4) Створити матрицю належності $U(0)$ за ф. (14)

$$u_{ij}^{(0)} = \begin{cases} 1, & d_{ij}^{(t)} = \min \{d_{ij}^{(t)}\}, \quad \forall i \in c, \\ 0, & \text{в інших випадках.} \end{cases} \quad (14)$$

5) Відкоригувати центр кластера $V(1)$ відповідно до $U(0)$ за ф. (15). Перейти до етапу 2.

$$V_i(k+1) = \frac{\sum_{j=1}^p u_{ij}^m(k) Z_j}{\sum_{j=1}^p u_{ij}^m(k)} \quad (15)$$

Етап 3 запускає швидкий алгоритм кластеризації, що містить наступні кроки.

1) Розрахувати початкову відстань d між точками даних та відкоригованими центрами кластерів за ф. (13)

2) Для t -ї ітерації обчислити матрицю належності $U(t)$ відповідно до $V(t)$ за ф. (16)

$$u_{ij}(k) = \frac{1}{\sum_{r=1}^c \left(\frac{d_{ij}(k)}{d_{rj}(k)} \right)^{\frac{2}{m-1}}} \quad (16)$$

3) Розрахувати функцію втрат.

Якщо $\|U^{(t+1)} - U^{(t)}\| \leq \varepsilon$, закінчити обчислення. Вивести центри кластера та матрицю належності U .

Якщо $\|U^{(t+1)} - U^{(t)}\| > \varepsilon$, встановити $t = t + 1$ і перейти до кроку (4).

4) Відкоригувати центр кластера $V(t+1)$ відповідно до $U(t)$ за ф. (15) і повернутися до кроку 1.

Обчислена матриця належності може визначати належність кожної точки даних до кожного кластеру. Серед них для однієї точки даних ступінь належності відображає кластер, до якого належить точка даних. Таким чином, центр жорсткого кластера визначається за допомогою традиційного методу кластеризації від кроку 1 до кроку 4. Оскільки центр жорсткого кластера дуже близький до центру нечіткого кластера, центр жорсткого кластера, який отриманий на кроці (5), буде вважатися початковим центром кластера методу кластеризації FCM. Тоді вибірками даних керуватиме кластеризація FCM через наступні кроки, щоб швидкість зближення алгоритму нечіткої кластеризації могла бути прискорена. Кількість ітерацій алгоритму нечіткої кластеризації буде зменшено, а час аналізу кластеризації значно скоротиться. Для моніторингу водного середовища дані про якість води можна швидко згрупувати за допомогою цього підходу, який дозволяє оцінити вимірювані параметри стану води.

Головною відмінністю даної роботи є узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів

кластеризації у випадку інтуїтивістських нечітких множин $\{IA_{(\alpha,\beta)}^1, \dots, IA_{(\alpha,\beta)}^n\}$, для яких $\alpha \in (0,1]$ і $\beta \in [0,1)$. Для набору типових точок $\{\tau^1, \dots, \tau^c\}$ інтуїтивістського нечіткого кластеру $IA_{(\alpha,\beta)}^l$, де $\tau_e^l = \arg \max_{x_i} \mu_{li}$, $\forall x_i \in IA_{(\alpha,\beta)}^l$, та набору міток $\{label\ 1, \dots, label\ c\}$ визначено процедуру автоматичного маркування, в результаті виконання якої, мітка призначається всім об'єктам, що відповідають кластеру або є на межі двох кластерів.

Для перевірки ефективності пропонованого методу обробки великих даних та його застосування до обробки даних моніторингу водного середовища, в дисертаційному дослідженні було використано дані, зібрані IoT системою SmartWater, розробленою на кафедрі комп'ютерних наук та інженерії СНУ ім. В. Даля, з якої було обрано 2026 наборів вибірових даних для імітаційного тесту. Результати аналізу (рис. 4) показали, що найкращий результат дає розподіл набору даних на 5 кластерів, разом з тим, 3 кластери також задовольняють умовам поставленої задачі. Отже, згідно отриманого графіку та попередніх припущень щодо розподілу даних якості вод водойм рибогосподарського призначення, визначено три класи, $c = 3$.

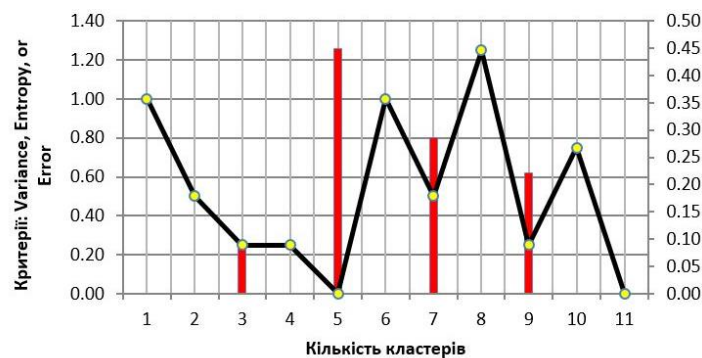


Рисунок 4 – Результати попереднього підрахунку кількості кластерів для набору даних якості вод водойм рибогосподарського призначення

Для оцінювання ефективності, проведено аналіз збіжності на вибірових даних. Порівняння проводилося з трьома відомими алгоритмами кластеризації: k-середніх, FCM та FFCM (рис. 5).

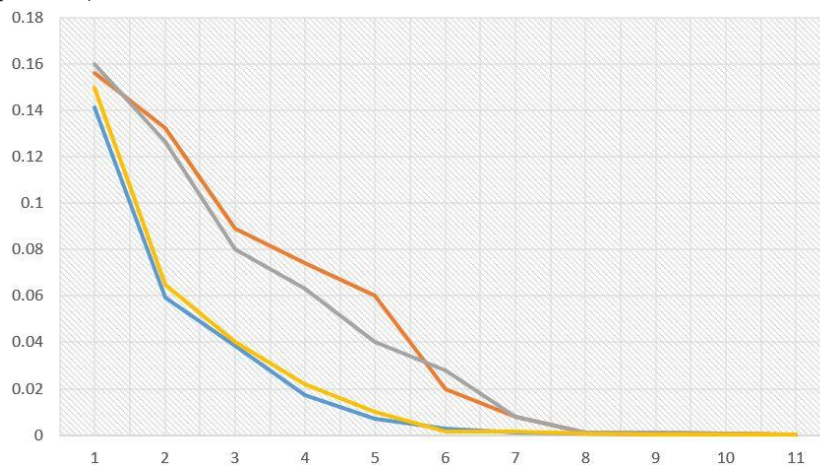


Рисунок 5 – Зближення цільової функції в залежності від кількості ітерацій (синій – пропонований метод, сірий – FCM, жовтий – FFCM, жовтогарячий – метод k-середніх).

Результати моделювання показують, що порівняно з кластеризацією k-means та FCM, запропонований метод дозволяє не тільки ефективно реалізувати нечітку кластеризацію даних, але й мати більш швидку конвергенцію. Швидкість конвергенції порівняна з відомим методом кластеризації FFCM, але останній не розрахований на роботу з великими даними. Метод може бути застосований як до обробки великих даних, отриманих з IoT систем моніторингу води, так і в інших областях при обробці великих даних. Разом з тим, відзначено, що для інших застосувань процедура автоматичного маркування класів має бути переглянута, щоб в найбільшій мірі відповідати розподілу.

Основні результати розділу опубліковані у [2, 3, 5, 10, 12].

Четвертий розділ присвячений розробці засобів та інформаційної технології для обробки великих даних, отриманих від станцій моніторингу водних об'єктів. Розкрито етапи розробки та реалізації інформаційної технології у вигляді системи підтримки прийняття рішень (СППР), надано архітектуру сховища даних, визначено критерії якості видобутку даних, представлено структуру бази даних та основні інструменти для аналітичної обробки даних у вигляді гібридного інструмента інтелектуального аналізу даних та управління непрямыми знаннями баз даних спеціалізованої аналітичної системи водних об'єктів. Наведено результати удосконалення та адаптації методів візуалізації великих даних на випадок он-лайн моніторингу водних об'єктів.

СППР з управління водними об'єктами визначена як інформаційна аналітична система, що використовується для накопичення екологічних даних діяльності підприємств аквакультури та водного господарства, видобутку знань з даних моніторингу водного середовища, зменшення часу на прийняття рішень та підвищення якості рішень в системах управління виробництвом рибної продукції у повністю або частково контрольованих умовах. Система використовує набір методів, алгоритмів та технологічних рішень для отримання необхідних знань із екологічних та рибпромислових інформаційних систем через наявні бази даних. Особливістю СППР є вибірка та застосування накопичених знань, правил і методик, що вочевидь переводить такі системи до рангу інтелектуальних аналітичних систем.

Система має багаторівневу архітектуру з 4 рівнів, яка з'єднує користувача з інформаційно-аналітичною системою: (1) Фільтрація даних; (2) Формування рекомендацій та управління знаннями; (3) Знаходження знань; (4) Управління знаннями (рис. 6). Основні завдання, що вирішувалися в роботі представлені в СППР у вигляді трьох компонентів: (1) Портал даних - візуалізує параметри якості води з різних датчиків на кожен день; (2) Аналіз параметрів – із заданою періодичністю (щодобово, щомісяця, щорічно, тощо) консолідує параметри якості води та показує порівняння між різними станціями моніторингу по всій ланці, де встановлено відповідні засоби; (3) Оперативний аналіз якості води – якість води розраховується за допомогою різноманітних параметрів, таких як розчинений кисень, БСК, рН, помутніння, солоність, фосфати, нітрати тощо. Індекс якості, що розраховується за методами запропонованими у розділах 2, 3. СППР є класичним рішенням, орієнтованим на вироблення рекомендацій як у сфері охорони навколишнього середовища загалом, так і для управління діяльністю об'єктів аквакультури.

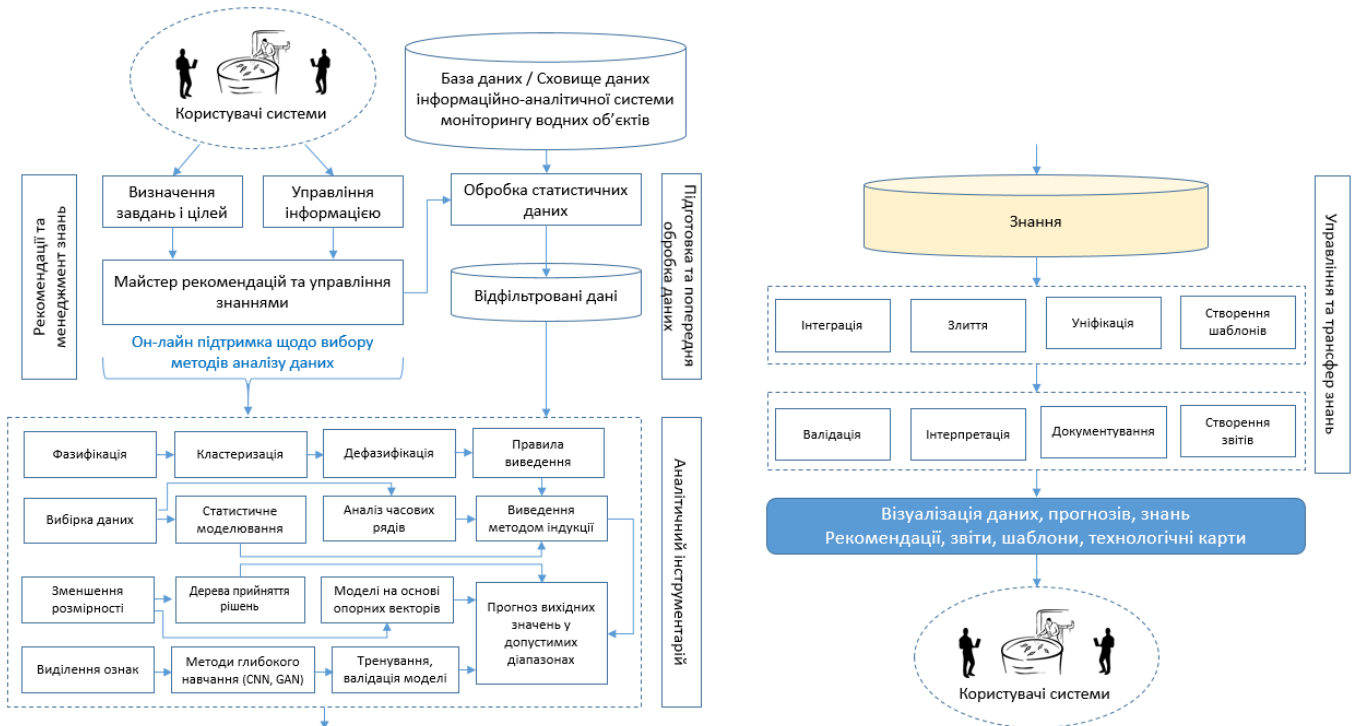


Рисунок 6 – Архітектура СППР

За результатами аналізу, представленого в розділі 1, для великих даних виділено також задачу їх візуалізації. Запропоновані у даному підрозділі рішення представлено у вигляді хордових діаграм і методів спрощення полігональних ланцюгів. Перший використовується для візуалізації залежностей, другий для отримання найбільш збалансованого відношення рівня деталізації до кількості точок ліній, що описують динаміку параметрів водного середовища.

Хордові діаграми є одним з найбільш прийнятних та зрозумілих засобів візуалізації кореляційних матриць, разом з тим, в системах екологічного моніторингу вони поки що не знайшли широкого застосування. Заповнюючи цю прогалину, для системи підтримки прийняття рішень, що є частиною запропонованої інформаційної технології, в роботі розглянуто технологію побудови відображення кореляцій параметрів якості води з використанням кривих Безьє, що задаються виразом:

$$B(t) = \sum_{k=0}^n P_k b_{k,n}(t), 0 \leq t \leq 1, \quad (17)$$

де P_k – функція-компонент векторів опорних вершин, а $b_{k,n}(t)$ – базисні функції кривої Безьє, що також називаються поліномами Бернштейна. $b_{k,n}(t) = \binom{n}{k} t^k (1-t)^{n-k}$, де $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ – кількість комбінацій із n до k , де n – ступінь полінома, k – порядковий номер опорної вершини.

В роботі наведено приклад візуалізації матриці кореляції параметрів якості води та відзначено, що найбільш серйозною проблемою відсутності моніторингових даних є значний рівень відсіву в довготривалих дослідженнях, який обмежує потужність інструментів аналітики. Рішення візуальної аналітики у вигляді хордових

діаграм може використовуватися для аналізу відсутніх значень, що значно покращує потенціал СППР.

Другою задачею візуалізації було спрощення полігональних ланцюгів. Проблема оптимізації відображення даних сформульована у вигляді задачі ε -мінімізації. Для наданого полігонального ланцюга P , $P = \langle p_1, \dots, p_n \rangle$ і цілого числа $k \leq n$, знайти серед усіх апроксимацій P , $P' = \langle p_{i_1}, \dots, p_{i_k} \rangle$ з $1 = i_1 < \dots < i_k = n$, з не більше ніж k вершинами апроксимацію з мінімальною погрішністю $\delta_F(p_i p_j, P)$

$$\delta_F(P', P) = \max_{1 \leq j \leq k} \delta_F(p_i p_{j+1}, P), \quad (18)$$

де P' є ε -спрощенням P , якщо $\delta_F(P', P) \leq \varepsilon$.

Для оцінки застосовності алгоритмів спрощення до мети роботи було розраховано позиційні помилки. Спосіб вимірювання відхилення спрощення полягає в знаходженні відстаней між вихідними точками, що були спрощені, та спрощеним ланцюгом. Для визначення якості спрощення було використано наступні показники:

– максимальне відхилення (максимальна відстань від оригінальної точки, що була спрощена, до спрощеної лінії)

$$\max_{i=1, N} (p_i - p_i^m) \quad (19)$$

– мінімальне відхилення (максимальна відстань від оригінальної точки, що була спрощена, до спрощеної лінії)

$$\min_{i=1, N} (p_i - p_i^m) \quad (20)$$

– середньоквадратичне відхилення

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (p_i - p_i^m)^2}{N}} \quad (21)$$

– сума відхилень (сума відстаней від оригінальних точок, що були спрощені, до спрощеної лінії)

$$S = \sum_{i=1}^N (p_i - p_i^m). \quad (22)$$

За результатами порівняльного аналізу, для використання в СППР та системі моніторингу було обрано алгоритм Дугласа-Рамера-Пекера. У межах поставленої задачі, баланс між вимогами до деталізації вихідного зображення та отриманням мінімальної кількості точок був зсунутий у бік зменшення сили спрощення та збільшення деталізації.

Остання частина роботи присвячена аналізу індикаторів ефективності інформаційної технології. Як зазначалося в аналітичній частині роботи, більша частина рибогосподарських об'єктів та підприємств аквакультури України

управляється вручну, водночас існує великий потенціал технологічних систем, зокрема для аналізу факторів витрат та пропозицій щодо вдосконаленого моніторингу, архітектури управління та підтримки прийняття рішень, що дозволяє не тільки автоматизувати всі процеси аквакультури але й забезпечити умови для інноваційного перетворення галузі. У цьому контексті, основні фактори витрат на ведення аквакультури представлено у вигляді трьох складових: (1) фіксована вартість, що включає витрати на технологічне обладнання, систему моніторингу і управління, датчики, резервуари, воду, корма, насоси, клапани, труби; (2) вартість робочої сили; і (3) змінні витрати. СППР з управління водними об'єктами, що використовує сучасні технології керування і моніторингу, у тому числі SCADA (наглядний контроль та збір даних), засоби аналізу та візуалізації даних, людино-машинної взаємодії, портативні мобільні пристрої розглядається як найбільш практичне рішення, здатне задовольнити потреби в екологічному регулюванні. Така високотехнологічна система в Луганській області ще повністю не розроблена, але базові компоненти, розроблені та представлені в дисертаційному дослідженні вже знаходяться на стадії дослідної експлуатації. Наразі виконано аналіз і налаштування найбільш важливих параметрів, які контролюються системою моніторингу: температура, рН, кисень, азот, нітрати, інтенсивність світла, ріст та загальний вихід риби, тощо. На даний момент перевірено наступні можливості СППР: Визначення індексу якості води; Аналіз залежностей споживання електроенергії, води, кисню та рівня рН від щільності риби, типу та кількості кормів; Автоматичне визначення дати і номерів резервуарів які необхідно поєднувати, виходячи з щільності риби внаслідок різного зростання; Прогноз кількості біологічних відходів в залежності від щільності, типу кормів та сезону. Наступним етапом роботи є точне визначення інших керованих змінних таких як норми та час подачі корму, параметри переробки біологічних відходів та вторинної води. Оскільки на деякі параметри витрат впливає масштаб підприємства, аналіз умов вирощування, попиту та пропозицій за допомогою СППР може істотно усунути зайві витрати. Так, наприклад, ціна на рідкий кисень в Україні обумовлює обмежене використання аерації, лише для забезпечення потреб у кисні риб. Це в свою чергу вимагає додаткових розрахунків і точного контролю та дозування кисню, постійного контролю кількості та щільності риб.

Результати роботи, моделі, метод та інформаційна технологія повністю відповідають глобальному руху за розширення технологічної складової у виробництво аквакультури, підкреслюючи баланс, який необхідно досягти для підтримання здоров'я екосистем, одночасно підштовхуючи систему до постійного вдосконалення. Основні результати розділу опубліковані у [6, 8, 11, 13-16].

У **додатках** наведено список публікацій здобувача, довідникові дані та показники якості вод рибогосподарських підприємств, використані при розрахунках, а також акти впровадження результатів дисертаційного дослідження.

ВИСНОВКИ

У дисертації поставлено і розв'язано актуальну науково-прикладну задачу розроблення моделей та методу інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів.

Отримано нові теоретичні та практичні результати:

1. На основі проведеного аналізу існуючих підходів і методів обробки великих даних встановлено, що існує об'єктивне протиріччя між зростаючою кількістю впроваджень високотехнологічних систем он-лайн моніторингу водних об'єктів і рівнем адаптації технологій управління та обробки великих наборів даних в системах моніторингу водних об'єктів.

2. З метою забезпечення швидкого та ефективного визначення якості води по даним, отриманим через моніторингову систему, у дисертації представлено нову модель обробки великих даних на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об'єктів, для подальшого прийняття рішення.

3. Проведено удосконалення методу нечіткої кластеризації с-середніх на випадок великих даних, шляхом узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних, що дозволило застосовувати автоматичну розмітку при обробці великих даних. Метод швидкої кластеризації заснований на методі нечітких с-середніх та традиційному жорсткому методі кластеризації. Результат жорсткої кластеризації використовується для орієнтування на початкове значення нечіткої кластеризації. Доведено, що запропонований метод дозволяє прискорити швидкість конвергенції. Як теоретичні результати, так і результати імітаційного моделювання показують, що підвищення ефективності кластеризації показників якості води, в свою чергу, дозволяє підвищити ефективність обробки великих даних.

4. Проведено удосконалення технології візуалізації великих даних за рахунок застосування кореляційних матриць та технології спрощення полігональних ланцюгів, що дозволило проводити динамічну візуалізацію та зменшити час на аналіз даних, отриманих з системи моніторингу, зокрема, записів довготривалого моніторингу.

5. Дістала подальшого розвитку інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації водних господарств у частині реалізації основних аналітичних функцій.

6. Усі теоретичні положення дисертації доведено до відповідних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів. Практичні результати дисертаційної роботи апробовано та впроваджено в Управлінні Державного агентства рибного господарства у Луганській області.

7. Матеріали дисертації використані і впроваджені в навчальний процес на кафедрі комп'ютерних наук та інженерії Східноукраїнського національного університету ім. В. Даля.

Тим самим доведено, що поставлені завдання виконано, мету роботи досягнуто.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Праці, які відображають основні наукові результати дисертації

1. Skarga-Bandurova I., Krytska Y., Velykzhanin A., Barbaruk L., Suvorin O., Shorohov M. "Emerging Tools for Design and Implementation of Water Quality Monitoring Based on IoT", *Complex Systems Informatics and Modeling Quarterly*. Published online by RTU Press, <https://csimq-journals.rtu.lv> Article 138, Issue 24, September/October 2020, pp. 1–14, 2020 <https://doi.org/10.7250/csimq.2020-24.01>.
2. Skarga-Bandurova I., Krytska Y., Shorohov M., Suvorin O., Barbaruk L., Ozheredova M. "Towards Development IoT-based Water Quality Monitoring System", *Proceedings of the 7th IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 140-145, 2019. doi: 10.1109/FiCloudW.2019.00038 (Індексується в SCOPUS).
3. Барбарук Л.В., Зубарев Д.В., Медведєв Є.М., Барбарук В.М. "Використання нечітких моделей для планування збиральних робіт у різних кліматичних умовах", *Вісник Східноукраїнського національного університету ім. В. Даля*, №6 (247), С. 175-179, 2018.
4. Скарга-Бандурова І.С., Грушка М.О., Барбарук Л.В. "Підходи до ефективного спрощення та візуалізації великих наборів даних", *Вісник Національного технічного університету "Харківський політехнічний інститут". Зб. наукових праць. Серія: Інформатика та моделювання*. Харків: НТУ "ХПІ", №. 50 (1271), С. 55-65, 2017. doi: 10.20998/2411-0558.2017.50.10.
5. Барбарук Л.В., Суворін О.В. "Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів", *Вісник Східноукраїнського національного університету ім. В. Даля*, №8 (238). С. 5-12. 2017.
6. Скарга-Бандурова І.С., Барбарук Л.В., Сіряк Р.В. "Моделі багатовимірних структур даних для аналізу природоохоронної діяльності промислових підприємств", *Зб. наукових статей Комп'ютерне моделювання в хімії, технологіях і системах сталого розвитку*. Київ: НТУУ "КПІ", С. 185-190, 2014.
7. Барбарук В.М., Барбарук Л.В. "Методи моделювання складних систем", *Вісник Східноукраїнського національного університету ім. В. Даля*, № 15(186), С. 132-136, 2012.
8. Барбарук В.М., Барбарук Л.В. "Засоби представлення багатомірних даних в інформаційно-аналітичних системах", *Міжвузівський збірник "Комп'ютерно-інтегровані технології: освіта, наука, виробництво"* Луцьк, №8. С. 5-10, 2012.
9. Барбарук Л.В. "Применение многомерных моделей данных в системах аналитической обработки данных", *Вісник Східноукраїнського національного університету ім. В. Даля*, №11, Ч. 2, С. 180-184, 2011.
10. Рязанцев О.І., Барбарук В.М., Барбарук Л.В. "Комплексна оцінка екологічної небезпеки території промислового виробництва", *Вісник Східноукраїнського національного університету ім. В. Даля*, №7(154), Ч.2, С. 136-140, 2010.

Тези доповідей у збірниках праць конференцій

11. Skarga-Bandurova I., Krytska Y.O., Barbaruk L.V. “Application of Internet of Things for long term water quality monitoring”, *Проблеми інформатики і моделювання. Тезиси дев'ятнадцятої міжнар. наук.-техн. конф.*, Харків: НТУ “ХПІ”, С. 77, 2019.
12. Барбарук Л.В., Михайличенко С.О. Бездротова система віддаленого моніторингу сільськогосподарських параметрів. *Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку»*, Сєверодонецьк : Східноукр. нац. ун-т ім. В. Даля, 2017. С. 206-208.
13. Михайличенко С.О., Барбарук Л.В. “Вибір протоколів маршрутизації для організації бездротових сенсорних мереж”, *Матеріали всеукр. наук.-практ. конф. “Майбутній науковець – 2017” 1 груд. 2017 р.*, м. Сєверодонецьк. укл. В. Ю. Тарасов. Сєверодонецьк : Східноукр. нац. ун-т ім. В. Даля, С. 296-297, 2017.
14. Барбарук Л.В., Добрецова А.О. “Застосування алгоритму Краскала для синтезу древоподібних глобальних мереж”, *Технологія-2016 : матеріали міжнар.наук.-техн. конф.*, 22-23 квіт. 2016 р., м. Сєверодонецьк. Ч. II [укл. : Тарасов В.Ю.] Сєверодонецьк : [Східноукр. нац. ун-т ім. В. Даля], С. 152-154, 2016.
15. Барбарук Л. “Методы формирования многомерных отчетов для задач учета ресурсопотребления”, *Theoretical and Applied Computer Science and Information Technology: Proceedings of the first International Conference TACSIT-2015*. Severodonetsk: East Ukrainian National University, P. 35-40, 2015.
16. Барбарук Л.В., Скарга-Бандурова І.С. Методы оперирования многомерными данными в хранилище данных опасных промышленных отходов. *Системи обробки інформації*. Вип. 2 (118). Т.2 Проблеми і перспективи розвитку ІТ-індустрії. С. 223, 2014.
17. Барбарук В.М., Барбарук Л.В. “Використання технології багатовимірних баз даних при проектуванні складних інформаційних систем”, *Матеріали IX міжнародної конференції “Strategy of Quality in Industry and Education: Part 3.”* Varna, Bulgaria: International Scientific Journal Acta Universitatis Pontica Euxinus. Special number, С. 440-443, 2013.
18. Барбарук Л.В. “Сложные информационные системы”, *Матеріали XIV всеукраїнської науково-практичної конференції “Технологія”*: ч. 2. Сєверодонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєверодонецьк), С. 129-131, 2011.

АНОТАЦІЇ

Барбарук Ліна Вікторівна. Моделі та метод обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – інформаційні технології. Східноукраїнський національний університет імені Володимира Даля, Сєверодонецьк, 2021.

Дисертацію присвячено вирішенню актуального науково-прикладного завдання підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об'єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології обробки великих даних.

Об'єктом дослідження є процеси обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів, а предметом досліджень є моделі та метод інформаційної технології обробки та візуалізації великих даних, використовуваних в інформаційно-аналітичних системах моніторингу водних об'єктів.

У роботі виконано аналітичний огляд сучасних методів і технологій обробки великих даних для кращого управління водними ресурсами. Встановлено, що: (1) незважаючи на інтенсивний розвиток технологій інтелектуального зондування та моніторингу, для багатьох водних ресурсів все ще бракує інтегрованих оцінок на основі великих даних, для підтримки динамічної оцінки якості води, моніторингу та управління наглядом; (2) зростаючі вимоги до даних вимагають збільшення потужності та ефективності інструментів та методів їх обробки. З експоненціальним зростанням даних, традиційні алгоритми видобутку та аналізу даних не можуть в повній мірі задовольнити потреби обробки даних, що обумовлює необхідність розробки і використання нових моделей обробки з розумною обчислювальною вартістю; (3) при візуалізації великих даних, сигнали, що надходять до системи моніторингу можуть бути занадто щільні та розмиті, що призводить до проблем з аналізом тривалих записів не дозволяють швидко орієнтуватися в даних, що вимагає удосконалення технологій візуалізації. За результатами аналізу сучасного стану та тенденцій розвитку систем моніторингу водних об'єктів показано необхідність розробки ефективних моделей і методів інформаційних технологій обробки великих даних. Обґрунтовано проведення дисертаційних досліджень в межах вирішення наступних основних завдань: розробка і використання нечітких моделей для реалізації інтегрованих оцінок води на основі великих даних у реальному часі, удосконалення підходів до нечіткої кластеризації на випадок великих даних, удосконалення технологій візуалізації великих даних, розробка та впровадження інформаційної технології та засобів підтримки прийняття рішень в інформаційно-аналітичній системі моніторингу водних об'єктів. У роботі поставлене та вирішене актуальне науково-прикладне завдання підвищення якості роботи інформаційно-аналітичних систем моніторингу водних об'єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології обробки великих даних.

Методологічну основу дисертаційного дослідження становлять методи теоретико-множинного опису, теорії нечітких множин, які використовувались при розробці моделі для аналітичної обробки великих даних; методи нечіткої евристичної кластеризації та інтуїтивістської теорії нечітких множин, міри подібності та метод множника Лагранжа, що використовувались для удосконалення методу обробки великих даних та автоматичного маркування нечітких кластерів, в контексті аналізу якості водойм рибогосподарського призначення; методи Data Mining, MapReduce, OLAP, технології сховищ даних, що використовувались для розробки інформаційної технології; методи візуалізації кореляційних матриць, зокрема хордові діаграми, криві Безьє, технології спрощення полігональних ланцюгів, що використовувались для побудови моделей візуалізації великих даних.

Вперше одержано модель для аналітичної обробки великих даних системи моніторингу водних об'єктів на основі формалізації її атрибутів та інтерпретації

невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об'єктів, для подальшого прийняття рішення;

Удосконалено метод нечіткої кластеризації с-середніх для великих даних, в якому, на відміну від існуючих, виконано узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких множин, що дозволило застосовувати автоматичну розмітку при обробці великих даних і покращити швидкість і конвергенцію алгоритму кластеризації;

Удосконалено технологію візуалізації великих даних за рахунок застосування кореляційних матриць та технології спрощення полігональних ланцюгів, що дозволило проводити динамічну візуалізацію та зменшити час на аналіз даних, отриманих з системи моніторингу, зокрема тривалих записів.

Дістала подальшого розвитку інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації у частині реалізації основних аналітичних функцій.

Усі теоретичні положення дисертації доведено до конкретних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів і вибору варіантів її використання в системі підтримки прийняття рішень. Використання моделей та інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів, зокрема у виробничих процесах рибогосподарських підприємств, дозволило зробити висновок, про їх ефективність в частині підвищення точності інтегральної оцінки якості вод, зменшення часу на прийняття рішень щодо якості вод водойм рибогосподарського призначення та зменшення часу на пошук невідповідностей і ретроспективний аналіз великих даних.

Ключові слова: великі дані, інформаційно-аналітична система, моніторинг, водний об'єкт, підтримка прийняття рішень, аналіз даних.

Lina Barbaruk. Models and method for large data processing in integrated monitoring system of water bodies. – Manuscript copyright.

The dissertation on competition of a scientific degree of the candidate of technical sciences on a specialty 05.13.06 "Information technology". - Volodymyr Dahl East Ukrainian National University, Severodonetsk, 2021.

The thesis is devoted to the actual scientific and applied task of increase of efficiency of integrated monitoring system of water bodies by means of the development and practical application of the ad-hock models and methods of information technology for large data processing. The object of research is the processes of large data processing in integrated monitoring system of water bodies, and the subject of research is the models and method of information technology of processing and visualization of large data used in integrated monitoring system.

The thesis presents an analytical review of modern methods and technologies of big and large data processing for better water resources management. The analysis of the current state and trends in the development of water monitoring systems shows the further need to develop effective models and methods of information technology for large data processing. Therefore, the current scientific and applied task of improving the quality of information and analytical systems for monitoring of water bodies through the development and practical application of models and methods of information technology for large data processing is set and solved. The methodological basis of the research comprises the methods of set-theoretic description, fuzzy set theory used in the model development for analytical processing of large data; fuzzy heuristic clustering, intuitive fuzzy set theory, similarity measures, and Lagrange multiplier method used to improve the data processing and automatic labelling of fuzzy clusters, in the context of quality analysis of fishery reservoirs; Data Mining, MapReduce, OLAP, data warehouse technologies used for information technology development; methods of visualization of correlation matrices, in particular chord diagrams, Bezier curves, technologies of simplification of polygonal chains used to build models of visualization of big data.

The main scientific results are: A new model for analytical large data processing of water monitoring system based on formalization of its attributes and interpretation of uncertainty of water quality assessment in the form of linguistic variables was obtained, which allowed to provide integrated characteristics of fisheries water bodies for further decision making. The method of fuzzy clustering for large data has been improved, which, in contrast to the existing ones, generalizes the procedure of automatic labelling of fuzzy clusters obtained using heuristic algorithms for intuitive fuzzy sets, which allowed the use of automatic mark-up in processing large data and improve convergence of the clustering algorithm. The technology of visualization of large data has been improved by means of correlation matrices and the technology of simplification of polygonal chains, which allowed carrying out dynamic visualization and reducing the time for analysis of data obtained from the monitoring system, including long recordings. The information technology for large data processing in integrated monitoring system of water bodies by its adaptation to tasks of control and management of fisheries enterprises, which provides a semantic basis for integrated automation in terms of implementation of basic analytical functions.

All theoretical provisions of the dissertation are brought to concrete engineering decisions with application of the information and its implementation at the decision support system. This enables to reduce the time to make decisions on the quality of water bodies in fisheries and reduce the time to find discrepancies and retrospective analysis of a large datasets.

Key words: large data, big data, integrated monitoring system, monitoring, water body, decision support, data analysis.

Підписано до друку 19.03.2021 р. Формат 60*84 1/16. Цифровий друк.
Умов. друк. арк. 1,4. Обл.-вид. арк. – 1,36.
Наклад 100 прим. Замовлення № 12, 2021

Віддруковано ФОП Семенова Т.В.
Свідоцтво В02 №112665 видане виконавчим комітетом
Сєверодонецької міської ради Луганської обл. 26.02.2007р.