НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.9

ТЕХНОЛОГІЯ СОЦІАЛІЗАЦІЇ ОСОБИСТОСТЕЙ ЗА СПІЛЬНИМИ ІНТЕРЕСАМИ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА SEO-ТЕХНОЛОГІЙ

Батюк Т. М. – студент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Висоцька В. А. – канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Соціалізація особистостей за спільними інтересами спричинено потребою більшості людей спростити частину життєвих моментів за рахунок зменшення часу на їх реалізацію. З швидкими темпами росту інформації, завантаженості людини в суспільстві та у зв'язку з останніми епідемічними світовими подіями людина стає ізольованою від можливості спілкуватися. А це однією із важливих потреб людської свідомості та самореалізації. Тому є актуальним попитом мати можливість отримувати рекомендований список подібних людей за спільними інтересами як результат інтелектуального пошуку множини релевантних користувачів соціальних мереж через аналіз фото людського обличчя на користувацьких фотографіях (на основі нейронних мереж) і аналіз користувацької інформації (на основі алгоритмів нечіткого пошуку та моделі Noisy Channel).

Мета – розроблення технології для соціалізації особистостей на основі SEO-технології та методу машинного навчання через використання згорткової та сіамської нейронних мереж для ідентифікації користувачів та алгоритмів аналізу тексту для підбору релевантних користувачів майбутнього спілкування.

Метод. При реалізації SEO-технологій обрано алгоритми нечіткого пошуку по словах на основі моделі Noisy Channel з алгоритмами ефективного розподілу текстової інформації. При реалізації машинного навчання розроблено згорткову нейронну мережу для ідентифікації користувачів системи.

Результати. Розроблено інтелектуальну систему соціалізації особистостей за спільними інтересами на основі SEO-технології та методи машинного навчання. Здійснено реалізацію роботи двох нейронних мереж: згорткової та сіамської, що дозволило здійснити пошук людського обличчя, на завантажуваних користувачем фотографіях і порівняти знайдене обличчя з уже наявними в базі даних/Інтернет. Це дає можливість ефективно ідентифікувати справжність користувача та гарантувати, що цього користувача на даний момент нема в базі даних, відповідно він потенційно є реальним. За допомогою алгоритмів нечіткого пошуку, алгоритму Левенштейна та моделі Noisy Channel створено алгоритм аналізу та порівняння користувацької інформації, який для поточного користувача формує список наявних користувачів системи, посортований по спаданню відсоткового співвідношення подібності користувачів та вказує, наскільки інтереси в інших користувачів збігаються з інтересами поточного користувача.

Висновки. Виявлено, що реалізований в системі алгоритм для формування вибірки користувачів є ефективнішою та точнішою приблизно на 25–30% в порівнянні зі звичайним алгоритмом Левенштейна. Також реалізований алгоритм здійснює вибірку приблизно в 10 разів швидше, ніж звичайний алгоритм Левенштейна.

КЛЮЧОВІ СЛОВА: нечіткий пошук, алгоритм Левенштейна, модел Noisy Channel, згорткова нейронна мережа, сіамська нейронна мережа, фотоаналіз обличчя, алгоритм розширення вибірки, алгоритм N-грам.

АБРЕВІАТУРИ

БД – база даних;
ІС – інформаційна система;
ІТ – інформаційна технологія;
МН – машинне навчання;
ПЗ – програмне забезпечення;
SEO – search engine optimization.

НОМЕНКЛАТУРА

S – система соціалізації особистостей;
I – множина вхідних даних;
O – множина вихідних даних;
R – основні правила опрацювання потоку вхідних даних в ІС соціалізації користувача;
U – параметри опрацювання вхідних даних;
N – нейронна мережа;

DOI 10.15588/1607-3274-2022-2-6

α – оператор скачування вхідних даних;

β – оператор опрацювання вхідних даних;

γ – оператор пошуку релевантних користувачів за аналізом профілів та фото;

P – вдосконалена імітаційна модель пошуку релевантних користувачів;

μ – оператор ідентифікації користувача;

χ – оператор формування даних фотогалереї;

ω – оператор формування списку та даних релевантних користувачів;

λ – оператор підтримки соціальних запитів;

i_1 – множина даних ідентифікації (фото, пароль, логін, відбиток пальця, голос);

i_2 – сховище даних фотогалереї обличь конкретної соціальної мережі;

i_3 – різні фото обличчя користувача;

i_4 – конкретний запит користувача;

o_1 – запити з ІС до конкретних релевантних користувачів за вимогою;

o_2 – оновлення для профілю користувача ІС в конкретній соціальній мережі;

o_3 – лайк за вимогою користувача ІС;

r_1 – правила алгоритму взаємодії;

r_2 – правила роботи згорткової нейронної мережі;

r_3 – правила алгоритму нечіткого пошуку;

r_4 – правила алгоритму автозбереження даних;

u_1 – множина рівнів доступу;

u_2 – множина вимог доступу;

u_3 – множина вимог додавання фото;

u_4 – множина вимог формування списку релевантних користувачів;

u_5 – множина вимог підтримки соціалізації;

C_{AU} – контент авторизованого користувача.

ВСТУП

На сьогодення соціалізація особистостей за спільними інтересами є надзвичайно важливим процесом під час ізоляції людей із-за подовженості світової епідемії COVID2019 [1–5]. Паралельно більшість людей завжди намагаються спростити та автоматизувати всі основні життєві процеси, які зазвичай займають багато вільного часу [6–9]. Це ж стосується і процесу соціалізації особистості. МН та SEO-технології на даний момент є надзвичайно важливими в контексті розроблення ІС опрацювання та аналізу великих даних [10–12]. Практично кожна популярна серед великої кількості людей ІС використовує відповідні механізми соціалізації [13–15]. Для ефективної реалізації ІТ соціалізації зазвичай оптимізують існуючі алгоритми і/або створюють нові алгоритмів відповідно до поставлених вимог та розв'язку конкретної задачі [16–21]. Питання МН навчання, соціалізації та SEO-технології є досить популярними і висвітлені в низці статей [22–27].

Об'єктом дослідження є процес соціалізації особистостей, оскільки на сьогодення завдання соціалізації є дуже важливим і всі сучасні соціальні мережі намагаються максимально оптимізувати та автоматизувати соціалізацію різних класів користувачів (за віком, статтю, вподобаннях, хобі тощо) з використанням усіх популярних сучасних ІТ, таких як нейронні мережі та алгоритми аналізу користувацьких текстових повідомлень. Для успішного створення ІС соціалізації особистостей за спільними інтересами найважливішим завданням є визначити конкретну мету соціалізація (наприклад, за якими інтересами/хобі, стилем життя тощо) та відповідно опрацювати/підтримувати процес соціалізації відповідного класу користувачів.

Предметом дослідження метод и та засоби ІТ соціалізації відповідно класу користувачів. Тому здійснюється аналіз/дослідження мети/особливостей користувача ІС, а саме визначення достовірності (справжності існування) конкретного користувача за допомогою пошуку людського обличчя на множині користувацьких фотографіях з використанням нейронних мереж і аналіз користувацької інформації з використанням алгоритмів нечіткого пошуку та моделі Noisy Channel.

Метою дослідження є розроблення ІТ для соціалізації особистостей, яка використовує SEO-технології та методи машинного навчання. Для досягнення мети були поставлені такі завдання:

– удосконалити імітаційну модель для пошуку множини релевантних користувачів в ІС соціалізації особистостей за допомогою пошуку людського обличчя на користувацьких фото з використанням нейронних мереж і аналіз користувацької інформації з використанням алгоритмів нечіткого пошуку та моделі Noisy Channel;

– удосконалити згорткову нейронну мережу, що дозволило ефективно здійснювати пошук людських обличь на фото, та перевіряти наявність вже існуючих людей в БД ІС;

– розробити алгоритму аналізу користувацької інформації та пошуку найбільш релевантних користувачів, відповідно до проаналізованого тексту на основі вже існуючих алгоритмів, таких як алгоритм Левенштейна, алгоритм розширення вибірки, алгоритм N-грам та моделі Noisy Channel;

– розробити та описати ПЗ візуальної пошуку множини релевантних користувачів для соціалізації особистостей за допомогою аналізу фото;

– здійснити аналіз результатів експериментальної апробації запропонованої ІТ пошуку множини релевантних користувачів соціалізації особистостей.

1 ПОСТАНОВКА ПРОБЛЕМИ

Систему соціалізації особистостей *S* подано імітаційною моделлю через кортежем:

$$S = < I, O, R, U, N, \alpha, \beta, \gamma >,$$

де $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$,
$R = \{r_1, r_2, r_3, r_4\}$, $U = \{u_1, u_2, u_3, u_4, u_5\}$.

Основними процесами ІС соціалізації особистості є «Ідентифікація користувача», «Формування даних

DOI 10.15588/1607-3274-2022-2-6

фотогалереї», «Формування списку релевантних користувачів» та «Підтримка соціальних запитів».

Процес ідентифікації користувача ІС соціалізації особистості опишемо суперпозицією:

$$C_{AU} = \mu \circ \beta \circ \alpha,$$
$$C_{AU} = \mu(\beta(\alpha(i_1, i_2, i_4), r_1, u_1), u_2).$$

Процес формування даних фотогалереї користувача ІС соціалізації особистості опишемо суперпозицією: $C_{CU} = \chi \circ \beta \circ \alpha$, тобто

$$C_{CU} = \chi(\beta(\alpha(C_{AU}, i_2, i_3, i_4), r_1, u_3), r_2).$$

Процес формування списку релевантних користувачів опишемо суперпозицією:

$$C_{UL} = \omega \circ \gamma \circ \beta \circ \alpha,$$
$$C_{UL} = \omega(\gamma(\beta(\alpha(C_{CU}, i_2), i_3), u_4), r_3).$$

Процес пітримки соціальних запитів користувача ІС соціалізації особистості опишемо суперпозицією:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha,$$
$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, i_2), i_4), u_5), r_4).$$

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

Під час створення обраної ІС важливим завданням є дослідження та аналіз вже існуючих аналогів для формування множини особливостей, характеристик, переваг та недоліків в існуючих ІС. Під час пошуку аналогів практично не знайдено подібних ІС соціалізації особистостей. Серед знайденого ПЗ виділимо 3 аналоги, які за функціоналом є найбільш наближеними до запропонованої ІС соціалізації, зокрема Tinder, Badoo та Chatous.

Tinder є найбільш популярною та старою ІС соціалізації особистостей. Серед основних переваг є підтримка кросплатформеності як для всіх операційних системах, так і для будь-якого смартфону та чрез браузер. Також перевагою є аналіз фотографії обличчя людини, пошук аналогів обличчя в Tinder та наявність обличчя на фото гарантує те, що ІС користуються лише реальні люди, і при знайомстві, чи під час переписки з користувачем можна бути впевненим, що це реальний користувач, а не шахрай. Ще до переваг можна виділити повну відсутність реклами, що дозволяє не відволікатися і повноцінно використовувати ПЗ. Серед недоліків в першу чергу варто виділити те, що ІС використовує лише базові фільтри пошуку користувачів, а саме вік, стать, та місцезнаходження, що дозволяє лише приблизно звузити область пошуку релевантних користувачів. Також ІС не має ніяких алгоритмів соціалізації та підбору користувачів за спільними інтересами. Ще одним недоліком є обмеження кількості користувачів, яких можна переглянути за день, обмеження відповідно можна зняти за платну підписку.

Badoo є досить популярною новою ІС соціалізації особистостей. Серед основних переваг можна виділити наявність даної програми на всіх платформах. Badoo можна використовувати як програму для робочого столу на всіх операційних системах, як застосунок для смартфону та у браузерній версії. ІС є повністю безкоштовною, відповідно в ній немає обмежень на кількість користувачів, яких можна переглянути за день. Також має безлімітні вподобання, які можна виставляти користувачам та необмежену кількість повідомлень, які можна написати. Серед недоліків варто відзначити величезну кількість реклами, яку неможливо вимкнути, що не дозволяє повноцінно зручно використовувати програму. Badoo не має ніяких алгоритмів соціалізації за спільними інтересами та надає можливість використання лише базових фільтрів пошуку користувачів.

Chatous є найменш популярною ІС соціалізації з усіх розглянутих. Основною перевагою даної ІС є те, що вона повністю безкоштовна для використання, має необмежену кількість вподобань, повідомлень та переглядів користувачів протягом дня. Також в даній програмі повністю відсутня будь яка реклама. Серед переваг можна виділити аналіз фото користувача та підтвердження обличчя користувача на фото, що гарантує те, що ми спілкуємося з реальною людиною. Серед недоліків те, що дана ІС не доступна на всіх платформах, а лише у вигляді браузерної програми, що обмежує кількість можливих користувачів. Chatous має основний недолік, як і в інших аналогах – повністю відсутні алгоритми підбору користувачів по інтересах, що не дозволяє оптимально здійснювати пошук користувачів. Присутні лише базові алгоритми фільтрації основних параметрів користувачів, що є неефективним, так як це означає, що більшість користувачів не будуть відповідати очікуванням пошуку, так як не матимуть спільних інтересів.

З таблиці 1 можна зробити висновок, що важливим пунктом реалізації ІС є відсутність реклами, або її мінімальна кількість. Також перевагою порівняно з Tinder є підтримка необмеженого числа користувачів, їх вподобань та повідомлень, які можна написати протягом дня.

Таблиця 1 – Порівняння аналогів ІС соціалізації

Назва	Tinder	Badoo	Chatous
Відсутність реклами	+	–	+
Необмежені користувачі	–	+	+
Необмежені вподобання	–	+	+
Необмежені повідомлення	–	+	+
Платна версія	+	–	–
Крос платформеність	+	+	–
Аналіз фотографій	+	+	+
Базові фільтри пошуку	+	+	+
Алгоритми соціалізації	–	–	–

DOI 10.15588/1607-3274-2022-2-6

Також перевагами програм Tinder та Badoo є доступність програм на всіх платформах, що значно розширює кількість потенційних користувачів.

3 МАТЕРІАЛИ ТА МЕТОДИ

Одним з найголовніших реалізованих алгоритмів в ІС соціалізації особистостей за спільними інтересами є реалізації нейронної мережі, яка здійснює пошук обличь на фото користувачів та порівнює з вже наявними в поточній БД. Для цього завдання вирішено використовувати глибинні нейронні мережі, основна особливість яких в тому, що вони крім вхідного і вихідного шару також складаються з певної кількості прихованих шарів (рис. 1).

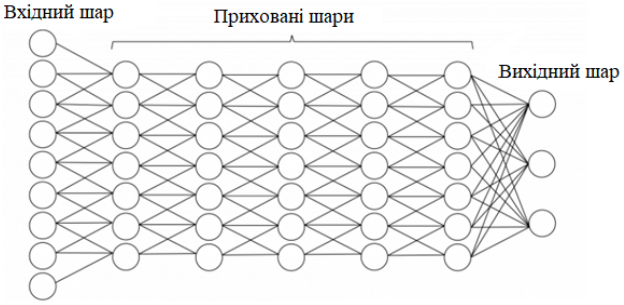

Рисунок 1 – Нейронна мережа з прихованими шарами

Для побудови глибинної нейронної мережі розпізнавання обличь обрано концепцію розподілу шарів на дві повністю окремі підвибірки (згорткові шари та дискретизуючі шари). Основна особливість такого підходу в тому, що кожен нейрон відповідає окремій точці на фіксованому зображенні і також має зв'язок з рецептивним полем (областю зображення, яке подається на вхід). В кожному місці обраного шару можна розподілити певну кількість нейронів. Кожен є унікальним, так як має унікальний набір вхідних ваг, які сформовані за допомогою нейронів в попередньому прямокутному шарі перед поточним шаром опрацювання (рис. 2).

Рисунок 2 – Структура згорткової нейронної мережі

При програмній реалізації даної згорткової мережі перше, що треба зробити, це завантажити фотографію, або декілька фотографій в ІС, зберегти завантажені фотографії за допомогою асинхронної опрацювання в БД, після чого почати опрацювання.

Згорткова нейронна мережа складається з 5 основних шарів: вхідний шар, вихідний шар, та 3 внутрішні приховані шари (P-Net, R-Net та O-net). Кожен шар послідовно здійснює опрацювання завантаженого зображення. Спочатку здійснює опрацювання шар P-Net, який на зображенні виділяє дві квадратні рамки (рис. 3).

На рис. 3 червоне поле відображає ядро 24x24, яке змінилось розміром до початкового зображення. Обчислено ширину та висоту ядра:

$$1500–200 = 300,\ 1800–500 = 300.$$

Отримана ширина та висота – це ширина та висота ядра до початкового розміру. Потім здійснюється множення координатних рамок на 300:

$$0{,}4x300 = 120,\ 0{,}2x300 = 60,$$
$$0{,}9x300 = 270,\ 0{,}7x300 = 210.$$

Додається верхня ліва координата ядра, щоб отримати координати граничного поля:

(200 + 120, 500 + 60) і (200 + 270, 500 + 210)
або (320,560) та (470,710).

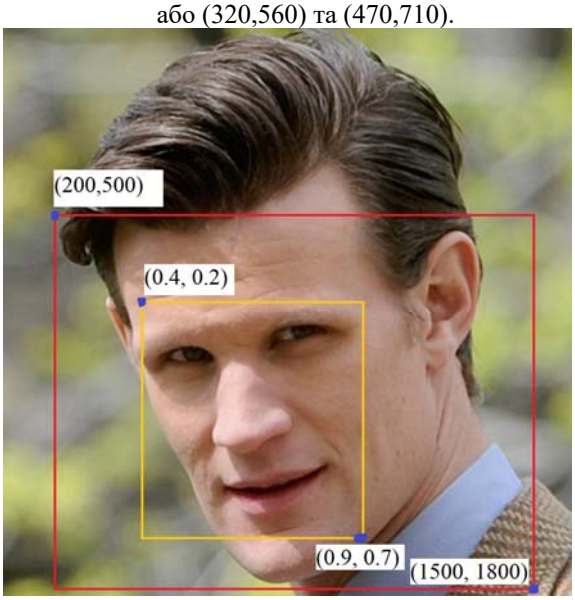

Рисунок 3 – Робота шару P-Net

Отримані дані зберігаються і далі починає працювати шар R-Net. Для кожного обмежувального вікна створюється масив однакового розміру та копіюються значення пікселів у новий масив. Іноді зображення може містити лише частину обличчя, що визирає з боку кадру (рис. 4). У такому випадку шар R-Net може повернути обмежувальний ящик, який частково знаходиться поза кадром.

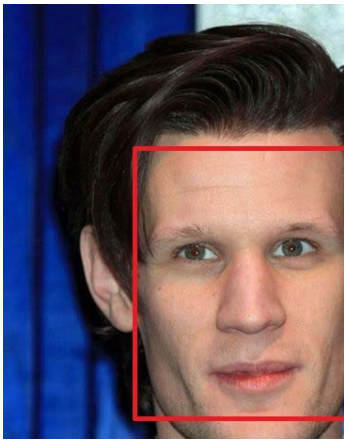

Рисунок 4 – Робота шару R-Net

DOI 10.15588/1607-3274-2022-2-6

Якщо обмежувальне поле не виходить за межі, копіюється частина зображення в обмежувальному полі в новий масив і заповнюється все інше 0. Цей процес заповнення масивів з 0 є padding. Після того, як прокладено масиви обмежувальної коробки, їх розмір змінюється до 24 x 24 пікселів і нормалізується до значень від –1 до 1. В даний час значення пікселів становлять від 0 до 255. Кожне значення пікселя віднімається на половину 255 (127,5) і ділиться на 127,5, після чого зберігається значення між –1 і 1.

Після стандартизації координат переставляються обмежувальні поля на квадрат, який слід передати O-Net. Отримані дані зберігаються і далі починає працювати шар O-Net. Виходи O-Net трохи відрізняються від результатів P-Net та R-Net. O-Net забезпечує 3 виходи: координати обмежувального поля, координати 5 орієнтирів обличчя та рівня довіри кожного поля. Позбавившись від ящиків із нижчим рівнем довіри, здійснюється стандартизація як координатних рамок, так і координат орієнтиру обличчя. Останнім кроком є упаковка всієї інформації у словник з трьома клавішами: «поле», «впевненість» та «ключові точки». «Поле» містить координати обмежувального поля, «впевненість» містить рівень довіри мережі для кожного вікна, а «ключові точки» містять координати кожного орієнтира обличчя (очі, ніс і кінцеві точки рота).

Наступним крок є використання сіамської нейронної мережі, яка відповідає за те, щоб шукати аналогічні обличчя в вже існуючих фото в БД (рис. 5).

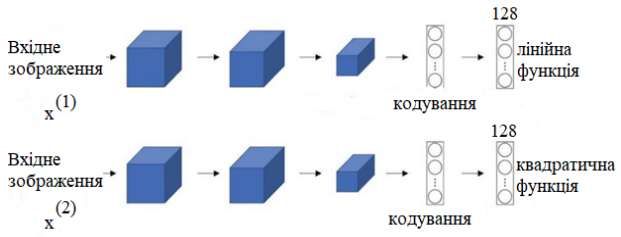

Рисунок 5 – Структура сіамської нейронної мережі

Сіамська нейронна мережа складається з 2 ідентичних шарів, кожен з яких мають однакові точні ваги. Кожен з шарів приймає одну і ту саму картинку в якості вхідних даних. Далі на виході з кожного шару є лінійна та квадратична функції, кожна з яких здійснює порівняння зображення з наявним зображенням всередині БД та формують певний коефіцієнт. Останнім кроком є вирахування різниці між коефіцієнтами та визначення наявності зображення.

4 ЕКСПЕРИМЕНТИ

Головною функцією ІС соціалізації особистостей за спільними інтересами є пошук релевантних користувачів, тому основним завданням є написати оптимізований алгоритм, який максимально автоматизує процес соціалізації користувачів. В даному випадку створений спеціальний алгоритм на основі таких алгоритмів, як алгоритм Левенштейна, розширення вибірки, N-грам та моделі Noisy Channel.

В першу чергу варта виділити вхідні дані за допомогою яких повинна бути сформована кінцева вибірка для користувача. Цими даними є параметри користувача системи, а саме опис користувача, інтереси користувача, та поле, в якому описано кого саме шукає користувач, саме ці частини будуть використовуватися для соціалізації.

З самого початку формується розширена вибірка на основі розформування поданого масиву даних на окремі елементи за допомогою алгоритму розширення вибірки. Цей алгоритм використовується найчастіше у всіх програмах порівняння слів, його особливість в тому, що він відкидає всі найдені результати без співпадінь і відкладає їх для алгоритмів нечіткого пошуку, а сам здійснює пошук по значеннях отриманих елементів. Для кожного слова отриманого з початкової вибірки будується множина значень на основі якої і відбувається реалізація пошуку подальших слів. Цей алгоритм по суті розділяє вибірку на 2 підвибірки, одна з яких буде оброблюватися основними алгоритмами нечіткого пошуку, а інша буде оброблена моделлю Noisy Channel. Алгоритм додатково модифіковано для генерації проміжних варіантів опрацювання тексту з використанням спеціальних правил з відкиданням закінчень слів, щоб уникнути неправильного розподілу на підвибірки через помилку в структуруванні слова з самого початку аналізу тексту. Дана модифікація дозволяє значно оптимізувати роботу алгоритму, так як в цьому випадку можна уникнути використання повних колекцій даних для зберігання тексту, а динамічно подавати дані окремими елементами. Блок-схема алгоритму подана на рис. 6.

Розподіливши дані на 2 підвибірки береться перша підвибірка і для її обробки застосовується алгоритм N-грам. Сам по собі алгоритм є досить старий і використовується вже давно, так як він є досить простий у реалізації, легкий у модифікації відповідно до певних унікальних вимог, та досить швидко працює, тим самим він є найшвидшим алгоритмом нечіткого пошуку серед тих, який використовуються в інформаційній системі соціалізації особистостей за спільними інтересами.

Якщо описувати роботу алгоритму, то він бере сформований масив даних, порівнює його поелементно з порівнюваним масивом даних. Саме порівняння здійснюється за дуже простою формулою, якщо слово 1 співпадає зі словом 2 з врахуванням деяких помилок, то є великий шанс того, що в них буде спільний рядок довжиною N. Блок-схема алгоритму подана на рис. 7.

DOI 10.15588/1607-3274-2022-2-6

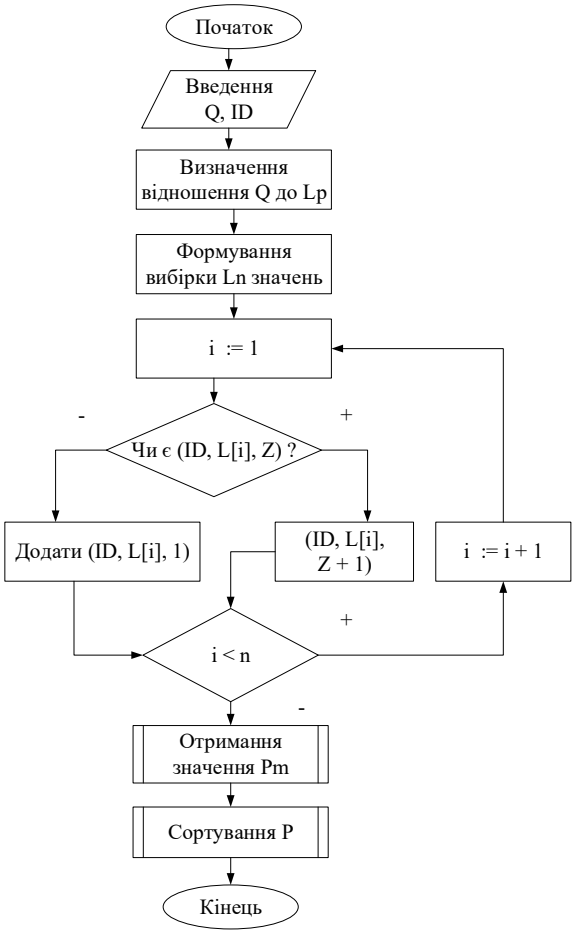

Рисунок 6 – Алгоритм розширення вибірки

Під час індексації потрібне слово розбивається на такі N-грами, щоб середній шанс схожості співпадіння слова був більше 50% і попадає в окремий список слів, найчастіше використовуються триграми, тобто рядки, які складаються з 3 букв. Також оптимізовано пошук і під час обробки даних, дані подаються по декілька елементів, тим самим не створюючи окремий список даних, а просто формуючи кінцевий список.

Також даний алгоритм має певні недоліки, наприклад при створенні всіх триграм рядків є безліч слів, які можуть випасти з вибірки, навіть якщо підходили по параметрам пошуку слова, так як може попастися слово, яке не відповідає поділу і буде помилково не вибране триграмами, тому саме перед цим використано алгоритм розширення вибірки, щоб мінімізувати можливість наявності подібних слів в першій підвибірці.

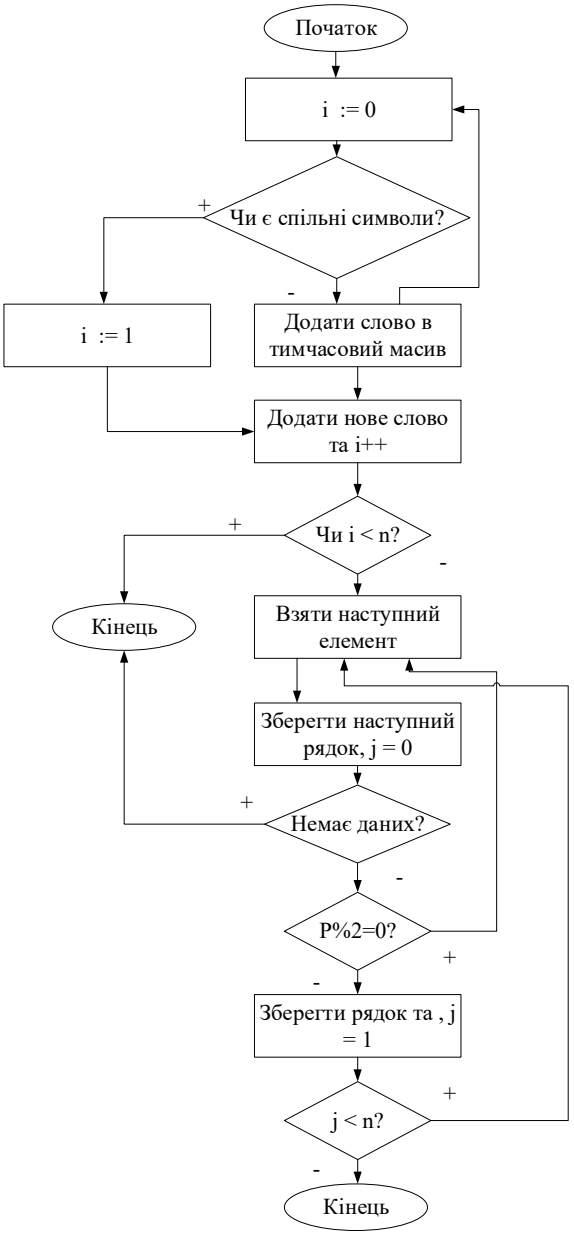

Рисунок 7 – Алгоритм N-грам

Вважається, що більшість таких слів попали саме в другу вибірку, тому відносно неї використовується модель Noisy Channel, її особливість в тому, що вона призначена для опрацювання всіх слів, які скоріше за все були певним чином спотворені, і розуміє під цим створення додаткового словника можливих значень, який і буде ще одним місцем для зберігання та порівняння слів з другої підвибірки, таким чином формується колекція можливих слів, які були спотворені. Сама модель Noisy Channel є дуже повільною у використанні через постійне формування проміжного словника під час обробки тексту, тому для її оптимізації і створено підвибірку слів, які скоріше за все спотворено, щоб не витрачати зайві ресурси. Блок-схема моделі Noisy Channel подана на рис. 8.

DOI 10.15588/1607-3274-2022-2-6

Рисунок 8 – Модель Noisy Channel

Робота моделі Noisy Channel показана на прикладі виправлення слова «actress», суть полягає в підстановці схожих слів за допомогою певних операції обробки слова, при відсутності помилкових літер обрано правильне слово (табл. 2).

Таблиця 2 – Приклад роботи моделі Noisy Channel

Помилка	Можливе виправлення	Правильна літера	Помилкова літера	Тип
acress	actress	t	Немає	Видалення
acress	cress	–	a	Вставка
acress	caress	ca	ac	Транспозиція
acress	access	c	r	Віднімання
acress	across	o	e	Віднімання
acress	acres	–	s	Вставка
acress	acres	–	s	Вставка

Внутрішнім алгоритмом обробки повністю сформованих вибірок слів є алгоритм Левенштейна. Основна суть алгоритму в тому, що він визначає відстань між декількома послідовностями символів.

Таким чином алгоритм на виході визначає певне значення, а саме кількість необхідних замін для того, щоб одне слово співпало з порівнюваним словом у випадку знаходження відмінностей та формує відсоток, який відображає шанс однаковості слів, за допомогою якого можна визначати чи у вибірку попало два аналогічних слова і є співпадіння, чи два слова абсолютно різні і не попадають в кінцеву вибірку. Сам алгоритм оптимізовано за допомогою його побудови з використанням кінцевого автомату, таким чином всі проміжні дані в пам'яті видаляються зі зміною ітерації. Відповідно, спочатку відбувається формування 2 підвибірок слів з вхідних даних користувачів, кожна з двох отриманих підвибірок додатково опрацьовується для збільшення точності пошуку слів, після чого за допомогою алгоритму Левенштейна йде безпосереднє порівняння вибірок слів усіх користувачів системи, кожне слово поточного користувача порівнюється з відповідними параметрами досліджуваного користувача. Блок-схема алгоритму подана на рис. 9.

В ході порівняння формуються за кожне співпадіння формуються бали, за співпадіння інтересів – 3 бали, параметру пошуку – 2 бали, загальної інформації – 1 бал. Перед початком роботи алгоритму Левенштейна формується максимальний бал поточного користувача, так як всі зайві слова вже викинуті і можна вважати, що бал сформований на основі ключових слів. Робота алгоритму Левенштейна на прикладі порівняння слів «elephant» та «relevant» подана в табл. 3.

Після роботи алгоритму при співпадінні кожного слова в користувачів нараховуються бали, які з кожною ітерацію сумуються.

Останнім кроком є порівняння сумарного балу поточного користувача та балів досліджуваного користувача, після чого формується відсоткове співвідношення схожості користувачів. Отримавши готову вибірку користувачів, можна здійснювати взаємодію з користувачами і інформаційна система соціалізації особистостей за спільними інтересами надає для цього всі необхідні функції, які інтуїтивно зрозумілі для звичайного користувача.

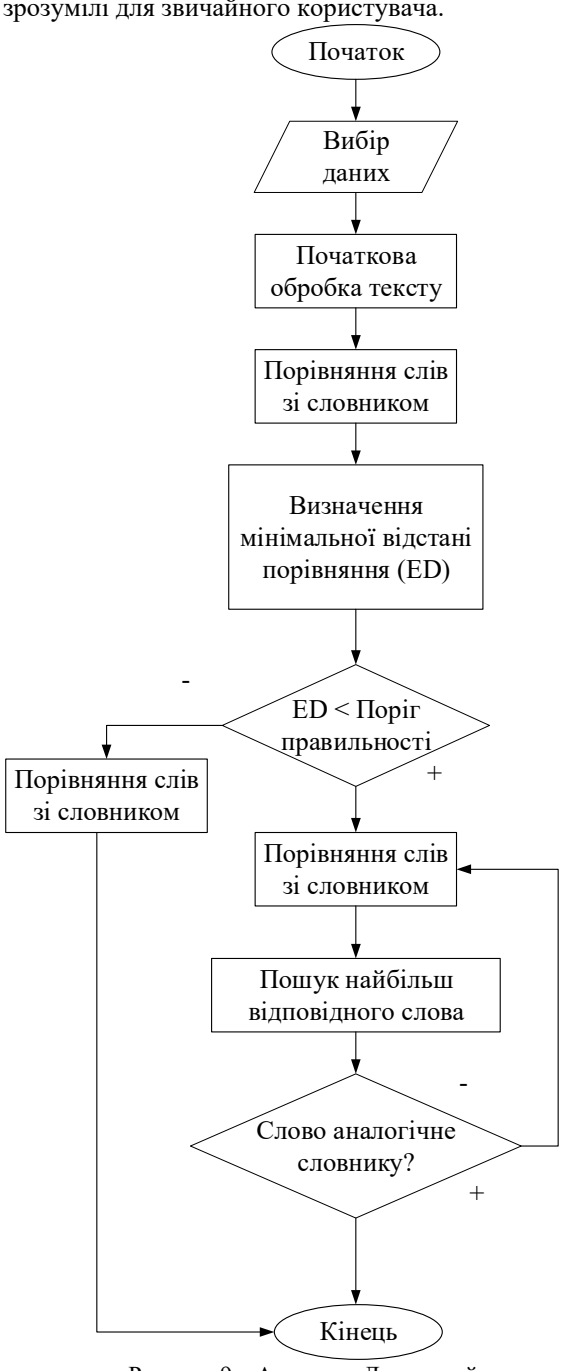

Рисунок 9 – Алгоритм Левенштейна

DOI 10.15588/1607-3274-2022-2-6

Таблиця 3 – Приклад роботи алгоритму Левенштейна

		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

Так метод GetUsers призначений для виведення сформованого списку користувачів системи, спочатку здійснюється пошук поточного користувача, щоб його ж профіль не кидати у вибірку, після чого його id виключається з загальної вибірки.

За допомогою директиви Response.AddPagination() формується відповідь сервера на запит з вибірки користувачів. В відповідь передається поточна сторінка для відображення користувачів, розмір сторінки, кількість елементів та загальна кількість сторінок. Метод повертає сформовану вибірку користувачів відповідно до початкового параметра, відносно якого і формувалася вибірка з усіма сформованими даними сторінки.

Наступним йде метод GetUser, який повертає конкретного користувача для перегляду особистого профілю даного користувача. Метод приймає унікальний ідентифікатор користувача та повертає знайденого користувача у поточній базі даних, а повертає відповідь про те, що сталася помилка і користувача з таким ідентифікатором не існує.

Далі йде метод UpdateUser, який в якості параметрів приймає ідентифікатор користувача, дані про якого потрібно оновити, та об’єкт, який містить всі оновлені дані, надісланий з клієнта програми. В першу чергу перевіряється наявність поточного користувача в базі даних та перевірка, чи користувач авторизований в систему. Далі відбувається звертання до бази даних та асинхронно оновлюється інформація про користувача, у випадку помилки зберігання в базі даних викидається виняток, також код та причина помилки логуються всередині бази даних та зберігаються у відповідний список. Наступним йде метод LikeUser, який призначений для виставлення позначки вподобання даного користувача. Спочатку перевіряється чи авторизований користувач системи, якщо так, то перевіряється чи немає виставленої позначки поточний користувач, якщо має, то викидається повідомлення про помилку і неможливість здійснити цю саму дію ще раз. В іншому випадку, якщо помилок немає, створюється новий об’єкт Like, та додається в базу даних і відправляється запит користувачу, якому цю позначку виставили.

Наступним йде метод CreateMessage для надсилання повідомлень, він приймає 2 основні параметри, це ідентифікатор відправника і об’єкт самого повідомлення. Здійснюється пошук відправника в базі даних, далі в разі успішного пошуку ідентифікатор відправника зберігається і шукається приймач повідомлення за допомогою ідентифікатора в базі даних. В разі успішного пошуку формується саме повідомлення та всі властивості повідомлення, які потрібно зберегти записуються в спеціальну змінну, яка є об’єктом повідомлення для зберігання даних. Спочатку повідомлення зберігається асинхронно в базу даних, далі формується анонімний об’єкт, який за допомогою механізму маршрутизації відправляється в діалог двох користувачів. Всі дані зберігаються в таблицях в реляційній базі даних, всы наявні дані пов’язані між собою за допомогою зв’язків. Схема поточної бази даних подана на рис. 10.

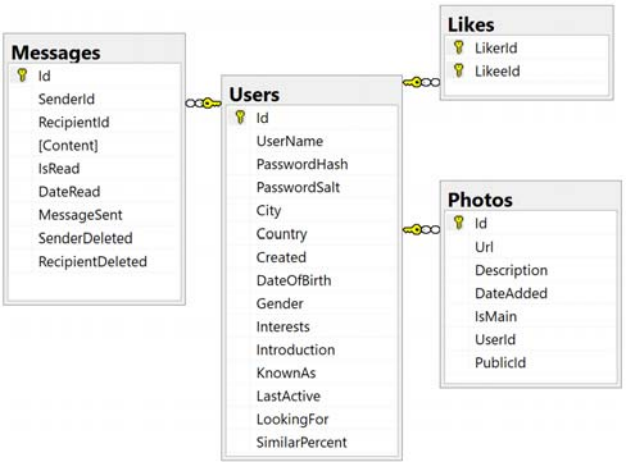

Рисунок 10 – Схема бази даних ІС

В першу чергу варто описати ключову таблицю бази даних – Users, яка представляє поточного користувача системи, вона має такі основні поля: Id – унікальний ідентифікатор користувача, UserName логін користувача, PasswordHash – хеш паролю користувача, PasswordSalt – образ хешу користувача, Gender – стать користувача, DateOfBirth – дата народження, KnownAs – нікнейм користувача всередині системи, Created – дата реєстрації, LastActive – дата останніх відвідин користувача, Introduction – вступна інформація про користувача, LookingFor – основна інформація про те, що саме користувачу потрібно і що він шукає, Interests – опис інтересів користувача системи, City – місто проживання, Country – країна проживання, SimilarPercent – тимчасовий параметр схожості користувача. Також, як видно зі схеми бази даних, таблиця Users пов’язана з іншими таблицями:Photos, Likes та Messages за допомогою зв’язку «один до багатьох», відповідно Users є центральною таблицею.

Далі йде таблиця Photos, в якій є всі основні властивості фотографій користувачів інформаційної системи, а саме Id – унікальний ідентифікатор, Url – посилання на картинку, Description – опис, DateAdded – дата додавання, IsMain – булеве поле, яке визначає аватарку користувача, PublicId – ідентифікатор знаходження картинки в хмарному сервісі зберігання даних та UserId – зовнішній ключ, який означає що

DOI 10.15588/1607-3274-2022-2-6

картинка належить одному з користувачів. Далі таблиця Likes, в якій є ідентифікатори користувачів, які виставляли лайки, а саме LikerId – унікальний ідентифікатор користувача, який поставив лайк та LikeeId – унікальний ідентифікатор користувача, якому поставлено лайк в системі.

Далі йде таблиця Messages, в якій є всі основні властивості повідомлень надісланих користувачами інформаційної системи, а саме Id – унікальний ідентифікатор повідомлення, SenderId – унікальний ідентифікатор відправника, RecipientId – унікальний ідентифікатор отримувача, Content – вміст повідомлення, IsRead – позначка чи прочитане повідомлення, DateRead – дата прочитання, MessageSent – дата відправки, SenderDeleted – чи видалене відправником, RecipientDeleted – чи видалене отримувачем.

5 РЕЗУЛЬТАТИ

Контрольний приклад відображає основні функції і роботу створеної інформаційної системи, на рис. 11 подано головне вікно програми.

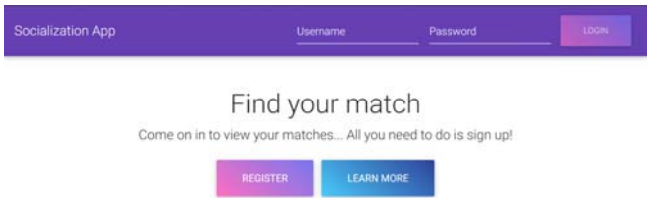

Рисунок 11 – Головне вікно програми

На рис. 12 подано кнопки головного вікна ІС.

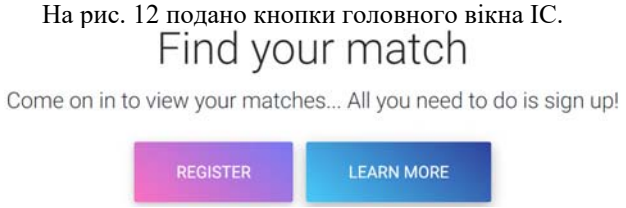

Рисунок 12 – Кнопки головного вікна програми

На рис. 13 подано форму реєстрації користувача.

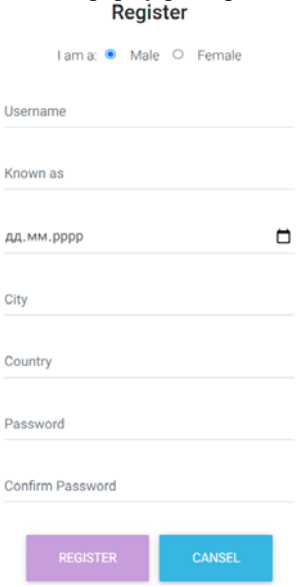

Рисунок 13 – Форма для реєстрації користувача

На рис. 14 подано авторизацію користувача, введення логіну та паролю, на рис. 15 подано повідомлення про успішну авторизацію.

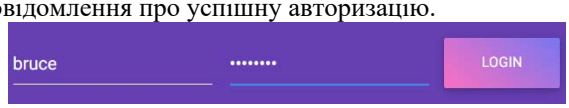

Рисунок 14 – Авторизація користувача

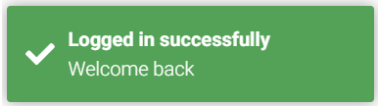

Рисунок 15 – Повідомлення про успішну авторизацію

На рис. 16 подано параметри профілю користувача, на рис. 17 подано заповнений профіль користувача.

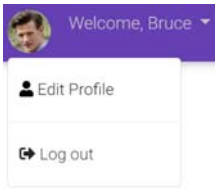

Рисунок 16 – Параметри користувацького профілю

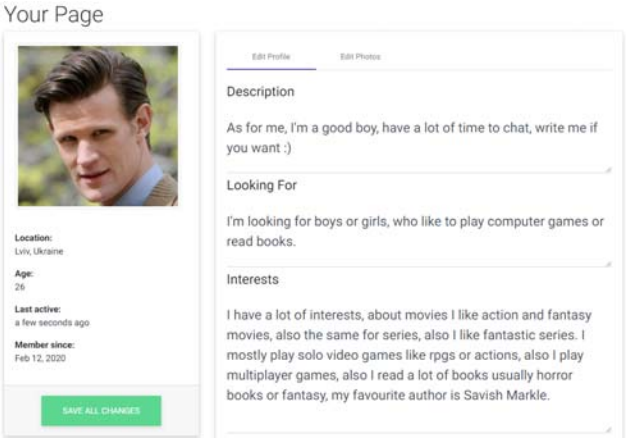

Рисунок 17 – Заповнений профіль користувача

На рис. 18 подано процес завантаження фотографій в систему, можна завантажувати одночасно 1 і більше фотографій перетягнувши їх вручну, або за допомогою провідника.

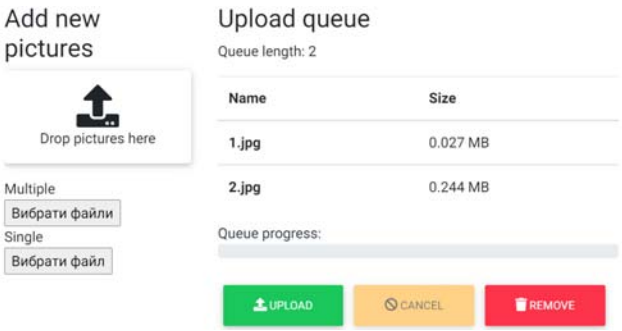

Рисунок 18 – Завантаження фотографій

DOI 10.15588/1607-3274-2022-2-6

На рис. 19 подано завантажені фотографії користувача, можна видалити всі фотографії, крім поточної головної фотографії та нейронні мережі обробили всі фотографії, і ті, на яких не знайдено лиця недоступні для виставлення основними фотографіями користувача.

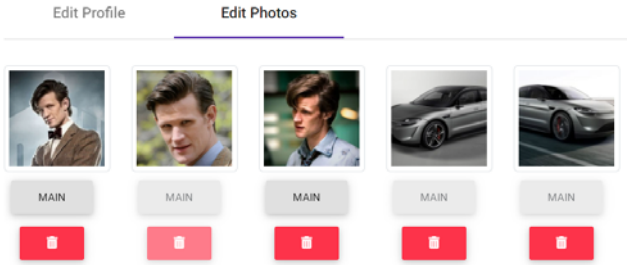

Рисунок 19 – Завантажені фотографії користувача

На рис. 20 подано сформований список користувачів за допомогою алгоритмів обробки тексту та посортований по спаданню відсоткового співвідношення схожості користувачів.

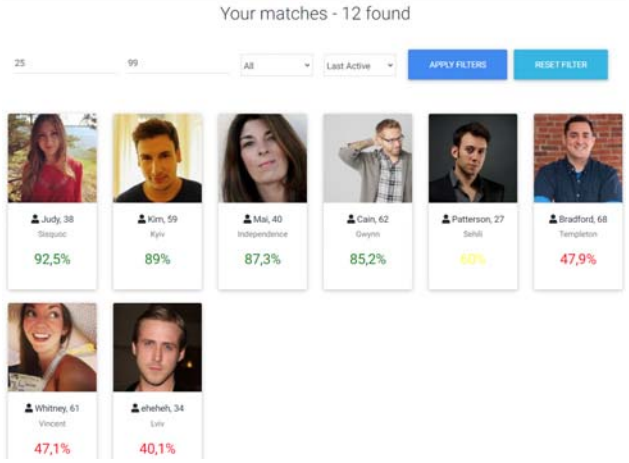

Рисунок 20 – Сформований список користувачів

На рис. 21 подано використання фільтрів для пошуку в вже сформованому списку.

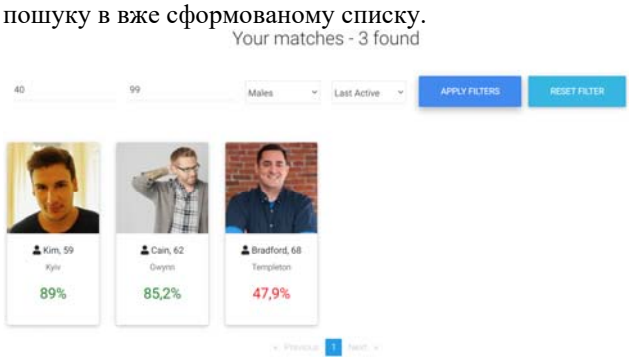

Рисунок 21 – Фільтрація списку

На рис. 22 подано вибір профілю користувача, видно можливість переглянути профіль користувача, поставити лайк і написати приватне повідомлення.

На рис. 23–24 подано перегляд вкладки з інформацією про вподобання користувачів, які вибрали нас, і яких вибрали ми.

На рис. 25–27 подано основну інформацію профілю вибраного користувача, вкладку з інтересами користувача, та вкладку з усіма фотографіями користувача.

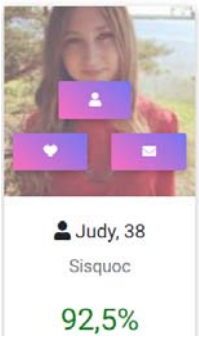

Рисунок 22 – Вибір користувача

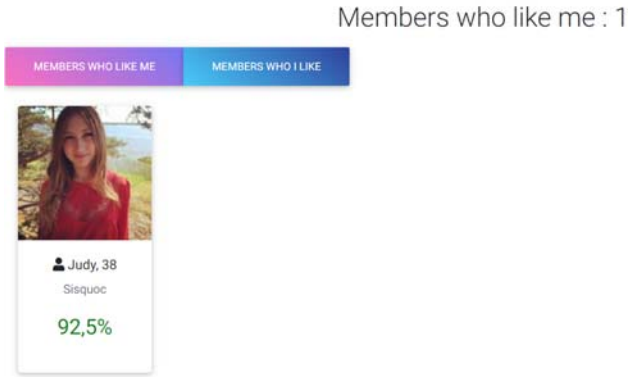

Рисунок 23 – Користувачі, які вибрали нас

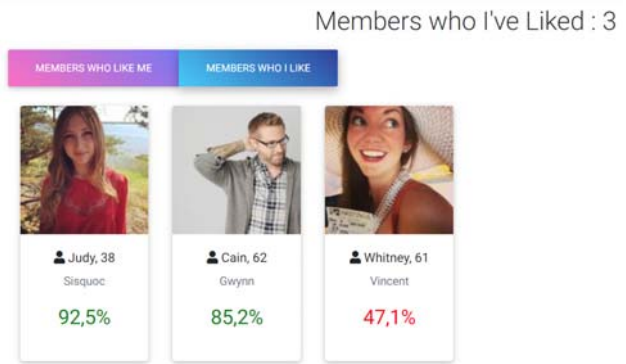

Рисунок 24 – Користувачі, яких вибрали ми

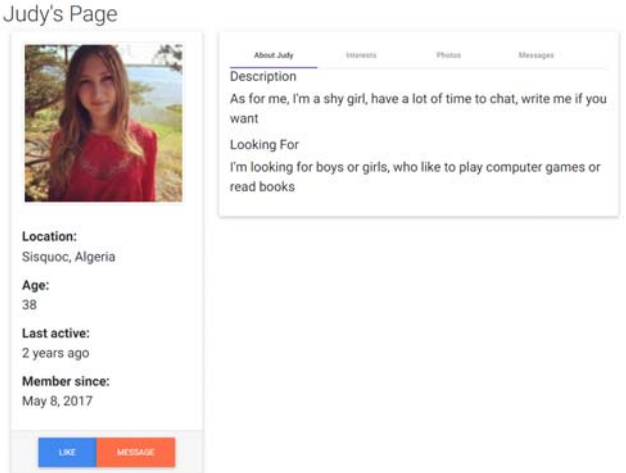

Рисунок 25 – Основна інформація про користувача

DOI 10.15588/1607-3274-2022-2-6

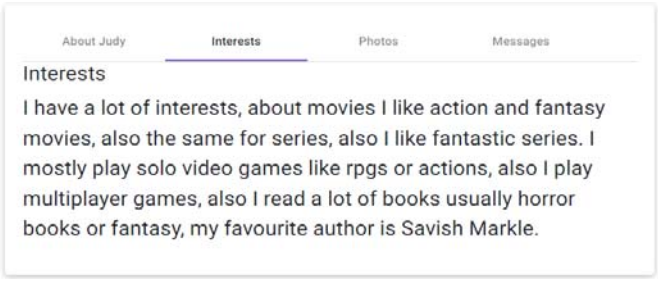

Рисунок 26 – Інтереси користувача

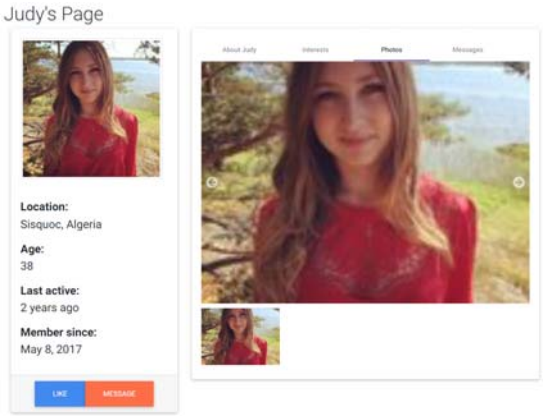

Рисунок 27 – Фотографії користувача

На рис. 28 подано вкладку з приватною перепискою з користувачем. В переписці подано ніки користувачів, фотографії, час надіслання і час прочитання повідомлень. На рис. 29–30 подано сторінку з інформацією про всі повідомлення, подано непрочитані, отримані та надіслані, можна здійснювати керування повідомленнями, а саме перегляд вибраного повідомлення за допомогою переходу в діалог з користувачем, або видалення вибраного свого повідомлення для всіх, або чужого повідомлення лише для себе.

На рис. 31–32 подано вхід з профілю іншого користувача, якого обрано першим користувачем системи та перегляд списку користувачів, які нас обрали, що дозволяє почати приватну переписку між двома користувачами. Що вибрали один одного.

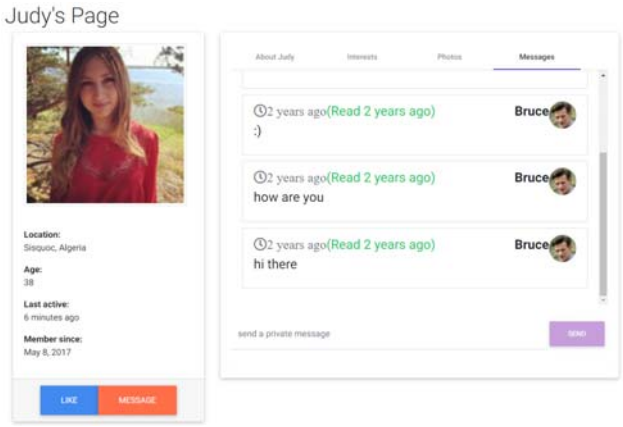

Рисунок 28 – Приватна переписка з користувачем

Рисунок 29 – Отримані повідомлення

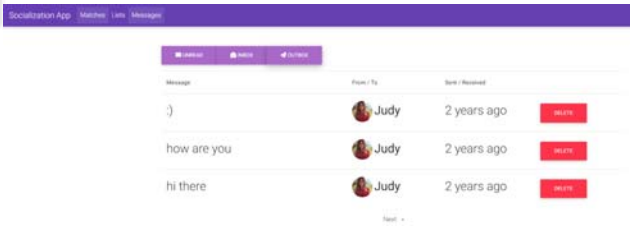

Рисунок 30 – Надіслані повідомлення

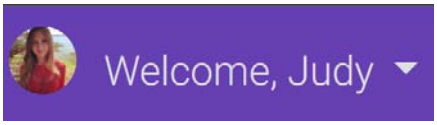

Рисунок 31 – Вхід іншого користувача в систему

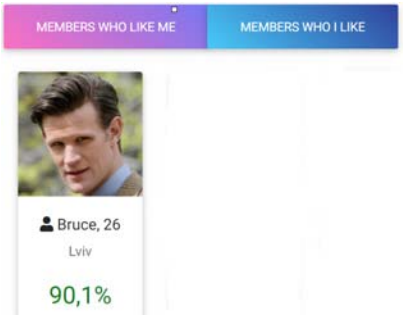

Рисунок 32 – Користувачі, що вибрали поточного користувача

На рис. 33 подано приватну переписку з початковим користувачем, від лиця вибраного користувача системи.

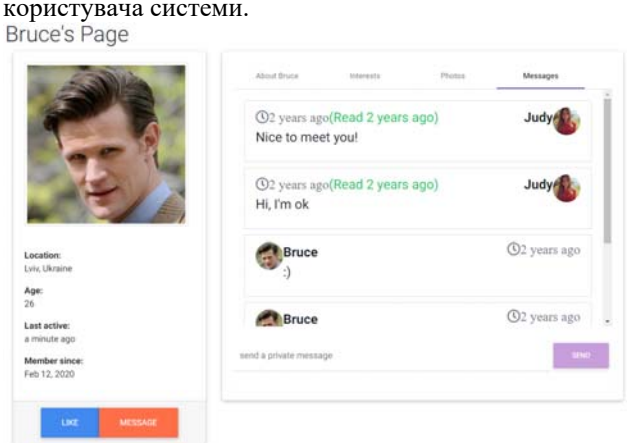

Рисунок 33 – Приватна переписка користувачів

6 ОБГОВОРЕННЯ

Отримавши реалізовану ІС здійснено статистичний аналіз 2 параметрів системи, а саме порівняння швидкості виконання формування вибірки

DOI 10.15588/1607-3274-2022-2-6

користувачів та точність отриманого відсоткового співвідношення. Порівнювались відповідно реалізована в інформаційній системі комбінація алгоритмів Левенштейна, N-грам, розширення вибірки, моделі Noisy Channel та звичайного алгоритму Левенштейна, який найчастіше використовують в схожих системах соціалізації особистостей. Спочатку проаналізовано ефективність формування вибірки користувачів, обрано 12 користувачів системи і кожному формувалась вибірка з використанням комбінації алгоритмів, та з використанням звичайного алгоритму Левенштейна, отримана діаграма подана на рис. 34, з якої можна зробити висновок, що реалізована в системі комбінація алгоритмів є ефективнішою та точнішою приблизно на 25–30% в порівнянні зі звичайним алгоритмом Левенштейна.

Рисунок 34 – Відсоткове співвідношення між користувачами

Далі проаналізовано швидкість формування вибірки користувачів, знову обрано 12 користувачів системи і кожному формувалась вибірка з використанням комбінації алгоритмів, та з використанням звичайного алгоритму Левенштейна, отримана діаграма подана на рис. 35, з якої можна зробити висновок, що реалізована в системі комбінація алгоритмів здійснює вибірку приблизно в 10 разів швидше, ніж звичайний алгоритм Левенштейна.

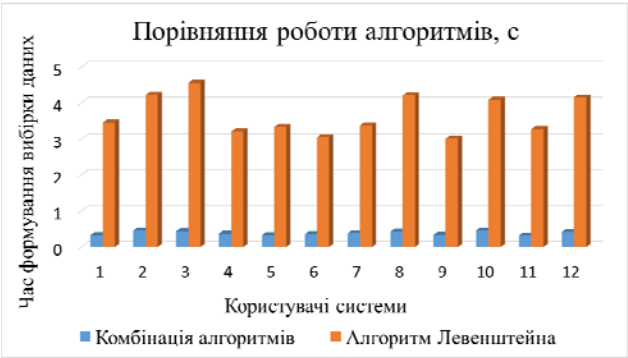

Рисунок 35 – Час формування вибірки даних

ВИСНОВКИ

В наш час соціалізація особистостей за спільними інтересами є надзвичайно важливим процесом, так як більшість людей намагаються спростити та автоматизувати всі основні життєві процеси, які зазвичай займають багато вільного часу, те саме стосується і процесу соціалізації, створена інформаційна система в цьому плані грає важливу роль. ІС розроблена з використанням алгоритмів нечіткого пошуку по словах з використанням моделі Noisy Channel з алгоритмами ефективного розподілу текстової інформації, та з використанням згорткової нейронної мережі для ідентифікації користувачів системи, так як на даний момент немає такої системи, яка б здійснювала аналіз вказаної про користувача інформації та формувала б список найбільш релевантних користувачів.

Створення ІС соціалізації особистостей є актуальним завданням, так як в сучасному світі люди намагаються оптимізувати всі життєві процеси для економії часу. Користувачі при пошуку тих чи інших програм в першу чергу вибирають ті, які економлять час, оптимізують роботу і достатньо автоматизовані, щоб виконувати більшість дій замість користувача системи. Дана інформаційна система об'єднує в собі зразу два важливих завдання: соціалізацію користувачів та максимально оптимізує і автоматизує сам процес соціалізації.

Якщо здійснити економічну оцінку, то розробка даного ПЗ є вигідною, оскільки створювана ІС пропонує унікальні можливості в плані пошуку та аналізу користувачів, відповідно крупні компанії з розробки програмної продукції будуть зацікавлені в покупці даного програмного продукту. Якщо провести маркетингову оцінку, то ІС даного типу є актуальною серед користувачів, оскільки в першу чергу вона надає можливості, які відсутні в схожих за функціоналом програмних продуктах, є легкою та зручною у використанні і пропонує повністю новий підхід до процесу пошуку, аналізу та соціалізації користувачів всередині системи.

Найважливішим кроком була практична реалізація інформаційної системи соціалізації особистостей за спільними інтересами, в першу чергу здійснено написання простого та надійного функціоналу реєстрації та подальшої авторизації користувача системи з використанням методів Identity та JWT токенів, що дозволило надійно зберігати паролі користувачів в базі даних та оптимізовано створювати сесію та надавати весь необхідний функціонал під час роботи користувача в системі. Далі здійснено реалізацію роботи двох нейронних мереж: згорткової та сіамської, що дозволило здійснити пошук людського лиця, на фотографіях що завантажує користувач і порівняти знайдене лице з уже наявними в базі даних лицями, що дає можливість ефективно ідентифікувати справжність користувача та гарантувати, що цього користувача на даний момент нема базі даних, відповідно він є реальним.

За допомогою алгоритмів нечіткого пошуку, алгоритму Левенштейна та моделі Noisy Channel створено алгоритм аналізу та порівняння користувацької інформації, який для поточного користувача формує список наявних користувачів

DOI 10.15588/1607-3274-2022-2-6

системи, посортований по спаданню відсоткового співвідношення схожості користувачів та вказує, наскільки інтереси в інших користувачів збігаються з інтересами поточного користувача.

В ході виконання роботи в більшості були досягнуті основні цілі відповідно до поставленої мети, існує ще багато речей, які можна удосконалити всередині створеної інформаційної системи та удосконалити, але той функціонал, який є доступним всередині інформаційної системи на даний момент відповідає поставленій меті створення інформаційної системи і є реалізована можливість усім користувачам системи пройти реєстрацію, вказати про себе всі необхідні дані, завантажити фотографії зі своїм лицем і після цього здійснювати ефективну соціалізацію, а саме переглядати користувачів зі сформованого списку, ставити відмітку про вподобання користувача та здійснювати приватну переписку всередині системи з вибраним користувачем.

До наукової новизни одержаних результатів варто віднести розроблення нового алгоритму аналізу користувацької інформації та пошуку найбільш релевантних користувачів ІС відповідно до проаналізованого тексту повідомлень профілю на основі вже існуючих алгоритмів Левенштейна, розширення вибірки, N-грам та моделі Noisy Channel. Для створення динамічної ІС соціалізації використано шаблон асинхронного програмування. Удосконалено згорткову нейронну мережу, що дозволило ефективно здійснювати пошук людських обличь на фото та перевіряти наявність вже існуючих людей в БД ІС.

Практичне значення створеної ІС соціалізації особистостей є дуже важливим, оскільки вона виконує важливу функцію соціалізації особистостей згідно потреб сучасних користувачів соціальних мереж. Також ІС має важливе значення в плані інновацій, так як на даний момент не існує аналогічних систем для соціалізації особистостей, які б використовували розроблені алгоритми. Система дозволить ефективно та швидко здійснювати підбір, аналіз, опрацювання текстових даних та формування кінцевого результату. В системі використовуються SEO-технології для ефективного та якісного інтелектуального пошуку та опрацювання відповідних даних за потребою конкретного користувача. Нейронна мережа дозволяє ефективно здійснювати ідентифікацію користувача по його фото. Загалом використовувані алгоритми дозволяють створити зручну ІС соціалізації з використанням необхідних для цього алгоритмів.

Варто зазначити важливість оптимізації наявної в ІС, в першу чергу це повна асинхронність системи, що дозволить уникнути всіх довгих очікувань та важких в плані опрацювання та аналізу запитів, система дозволяє ефективно та динамічно працювати з різними обсягами великих даних, здійснювати їх аналіз, опрацювання та формування нових даних необхідних користувачам ІС. Також використовується хмарний сервіс, який дозволить здійснити розподіл даних, відповідно можна буде зберігати всі найбільш важкі дані в хмарному середовищі і з використанням простого програмного інтерфейсу ІС за допомогою запитів здійснювати завантаження всіх необхідних даних. Таким чином, можна стверджувати, що створення даної ІС є важливим як і в соціальному плані, так і в плані реалізації всіх алгоритмів, які забезпечують необхідний функціонал ІС.

DOI 10.15588/1607-3274-2022-2-6

ПОДЯКИ

Роботу виконано в рамках держбюджетної теми «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (ID:839 2017-05-15 09:20:01 (2459-315)). Дослідження провадилось в межах спільних наукових досліджень кафедри інформаційних систем та мереж НУ «Львівська політехніка» на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ НУ «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Chu S. C. Using a consumer socialization framework to understand electronic word-of-mouth (eWOM) group membership among brand followers on Twitter / S. C. Chu, Y. Sung // Electronic Commerce Research and Applications. – 2015. – Vol. 14(4). – P. 251–260. DOI: 10.1016/j.elerap.2015.04.002
2. De-Gregorio F. Understanding attitudes toward and behaviors in response to product placement: A Consumer Socialization Framework / F. De-Gregorio, Y. Sung // Journal of Advertising. – 2010. – Vol. 39(1). – P. 83–96. DOI: 10.2753/JOA0091-3367390106
3. Peer-based social media features in behavior change interventions: Systematic review / [S. M. R. A. Elaheebocus, M. Weal, L. Morrison, L. Yardley] // Journal of Medical Internet Research. – 2018. – Vol. 20(2). – P. 1–20. 10.2196/jmir.8342
4. Erkan I. The influence of e-WOM in social media on consumers' purchase intentions: An extended approach to information adoption / I. Erkan // Computers in Human Behavior. – 2016. – Vol. 4. – P. 47–55. DOI: 10.1016/j.chb.2016.03.003
5. Ferrara E. Online popularity and topical interests through the lens of Instagram / E. Ferrara, R. Interdonato, A. Tagarelli // Hypertext and Social Media. – 2014. – Vol. 2. – P. 24–23. DOI: 10.1145/2631775.2631808
6. Gao L. Online consumer behavior and its relationship to website atmospheric induced flow: Insights into online travel agencies in China / L. Gao, X. Bai // Journal of Retailing and Consumer Services. – 2014. – Vol. 21(4). – P. 653–655. DOI: 10.1016/j.jretconser.2014.01.001
7. Geurin-Eagleman A. N. Communicating via photographs: A gendered analysis of Olympic athletes' visual self – presentation on Instagram / A. N. Geurin-Eagleman,

L. M. Burch // Sport Management Review. – 2016. – Vol. 19(2). – P. 133–145. DOI: 10.1016/j.smr.2015.03.002
8. From McDonalds fail to Dominos sucks: An analysis of Instagram images about the 10 largest fast food companies / [J. D. Guidry, M. Messner, Y. Jin, V. Medina-Messner] // Corporate Communications: An International Journal. – 2015. – Vol. 20(3). – P. 344–359. DOI: 10.1108/CCIJ-04-2014-0027
9. Hanna R. We're all connected: The power of the social media ecosystem / R. Hanna, A. Rohm, V. L. Crittenden // Business Horizons. – 2011. – Vol. 54(3). – P. 265–273. DOI: 10.1016/j.bushor.2011.01.007
10. Salganik M. J. Bit by bit: Social Research in the Digital Age / M. J. Salganik. – Princeton : Princeton University Press, 2019. – 448 p.
11. The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees / [S. Hudson, M. Roth, T. J. Madden, R. Hudson] // Tourism Management. – 2015. – Vol. 47. – P. 68–76. DOI: 10.1016/j.tourman.2014.09.001
12. Managing brand presence through social media: The case of UK football clubs / [M. Jeff, R. Jennifer, J. Catherine, P. Elke] // Internet Research. – 2014. – Vol. 24(2). – P. 181–204. DOI: 10.1108/IntR-08-2012-0154
13. Kim E. Brand followers' retweeting behaviour on Twitter: How brand relationship influence brand electronic word-of-mouth / E. Kim, Y. Sung, H. Kang // Computers in Human Behavior. – 2014. – Vol. 37. – P. 18–25. DOI: 10.1016/j.chb.2014.04.020
14. Kudeshia C. Spreading love through fan page liking: A perspective on small scale entrepreneurs / C. Kudeshia, P. Sikdar, A. Mittal // Computers in Human Behavior. – 2016. – Vol. 54). – P. 257–270. DOI: 10.1016/j.chb.2015.08.003
15. Lueg J. E. Interpersonal communication in the consumer socialization process: Scale development and validation / J. E. Lueg, R. Z. Finney // Journal of Marketing Theory and Practice. – 2007. – Vol. 15(1). – P. 25–39. DOI: 10.2753/MTP1069-6679150102
16. Mousavijad M. The effect of socialization factors on decision making of teenagers consumers in schools / M. Mousavijad, S. Payvandi // Journal of School Administration. – 2017. – Vol. 5(1). – P. 217–234.
17. Schnell R. Enhancing Surveys with Objective Measurements and Observer Ratings / R. Schnell // Improving Survey Methods. – London : Routledge, 2014. – P. 310–324.
18. Parry M. E. The effect of personal and virtual word-of-mouth on technology acceptance / M. E. Parry, T. Kawakami, K. Kishiya // Journal of Product Innovation Management. – 2012. – Vol. 29(6). – P. 952–966. DOI: 10.1111/j.1540-5885.2012.00972.x
19. Quan-Haase A. Introduction to the Handbook of Social Media Research Methods: Goals, Challenges and Innovations / A. Quan-Haase, L. Sloan // The Sage Handbook of Social Media Research Methods. – 2017. – Vol. 1. – P. 1–9.
20. Murphy S. T. Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures / S. T. Murphy, R. B. Zajonc // Journal of Personality and Social Psychology. – 1993. – Vol. 64(5). – P. 723–739. DOI: 10.1037/0022-3514.64.5.723
21. Ngoma M. Word of mouth communication: A mediator of relationship marketing and customer loyalty / M. Ngoma, P. D. Ntale // Cogent Business and Management. – 2019. – Vol. 6(1). – A. 1580123. DOI: 10.1080/23311975.2019.1580123
22. Analyses of word-of-mouth communication and its effect on students' university preferences / [A. Ozdemir, B. Tozlu, E. Şen, A. Ateşoğlu] // Procedia – Social and Behavioral Sciences. – 2016. – Vol. 235. – P. 22–35. DOI: 10.1016/j.sbspro.2016.11.022
23. Park J. Style in the age of Instagram: Predicting success within the fashion industry using social media / J. Park, G. L. Ciampaglia, F. Ferrara // Computer-Supported Cooperative Work & Social Computing : 19th ACM Conference, San Francisco, CA, USA, February 27 – March 2, 2016 : proceedings. – ACM : Permissions@acm.org, 2019. – P. 64–73. DOI: 10.1145/2818048.2820065
24. The Intelligent System Development for Psychological Analysis of the Person's Condition / [O. Oborska, V. Andrunyk, L. Chyrun et al.] // Computational Linguistics and Intelligent Systems (COLINS 2021) : 5th International Conference, Lviv, 22–23 April 2021 : CEUR workshop proceedings. – Aachen: CEUR-WS.org, 2021. – Vol. 2870. – P. 1390–1419.
25. Ranjbaran B. A survey for identification of major factors influencing customers attitude toward machine made carpet brands / B. Ranjbaran, M. Jamshidian, Z. Dehghan // Journal of Commercial Strategies. – 2020. – Vol. 5(23). – P. 109–118.
26. Batiuk T. Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning / T. Batiuk, V. Vysotska, V. Lytvyn // Computational Linguistics and Intelligent Systems (COLINS 2020) : 4th International Conference, Lviv, 23–24 April 2020 : CEUR workshop proceedings. – Aachen: CEUR-WS.org, 2020. – Vol. 2604. – P. 1237–1250.
27. Vysotska V. Information Technology for Internet Resources Promotion in Search Systems Based on Content Analysis of Web-Page Keywords / V. Vysotska // Radio Electronics, Computer Science, Control. – 2021. – No 3. – P. 133–151. DOI: 10.15588/1607-3274-2021-3-12

Стаття надійшла до редакції 19.12.2021.
Після доробки 17.01.2022.

УДК 004.9

ТЕХНОЛОГИЯ СОЦИАЛИЗАЦИИ ЛИЧНОСТЕЙ ЗА ОБЩИМИ ИНТЕРЕСАМИ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И SEO-ТЕХНОЛОГИЙ

Батюк Т. М. – студент кафедры «Информационные системы и сети», Национальный университет «Львовская политехника», Украина.

Высоцкая В. А. – канд. техн. наук, доцент, доцент кафедры «Информационные системы и сети», Национальный университет «Львовская политехника», Украина.

АННОТАЦИЯ

Актуальность. Социализация личностей по общим интересам вызвана потребностью большинства людей упростить часть жизненных моментов за счет уменьшения времени их реализации. С быстрыми темпами роста информации,

DOI 10.15588/1607-3274-2022-2-6

загруженности человека в обществе и в связи с последними эпидемическими событиями человек становится изолированным от возможности общаться. А это одна из важных потребностей человеческого сознания и самореализации. Поэтому является актуальным спросом возможность получать рекомендованный список подобных людей по общим интересам как результат интеллектуального поиска множества релевантных пользователей социальных сетей через анализ фото человеческого лица на пользовательских фотографиях (на основе нейронных сетей) и анализ пользовательской информации (на основе алгоритмов нечеткого поиска и модели Noisy Channel).

Целью исследования является разработка технологии для социализации личностей на основе SEO технологии и метода машинного обучения через использование сверточной и сиамской нейронных сетей для идентификации пользователей и алгоритмов анализа текста для подбора релевантных пользователей будущего общения.

Метод. При реализации SEO-технологий выбраны алгоритмы нечеткого поиска по словам на основе модели Noisy Channel с алгоритмами эффективного распределения текстовой информации. При реализации машинного обучения разработана сверточная нейронная сеть для идентификации пользователей системы.

Результаты. Разработана интеллектуальная система социализации личностей по общим интересам на основе SEO-технологии и методы машинного обучения. Осуществлена реализация работы двух нейронных сетей: сверточной и сиамской, что позволило осуществить поиск человеческого лица, на загружаемых пользователем фотографиях и сравнить найденное лицо с уже имеющимися в базе данных/Интернет. Это позволяет эффективно идентифицировать подлинность пользователя и гарантировать, что этого пользователя на данный момент нет в базе данных, соответственно он потенциально реальный. С помощью алгоритмов нечеткого поиска, алгоритма Левенштейна и модели Noisy Channel создан алгоритм анализа и сравнения пользовательской информации, который для текущего пользователя формирует список имеющихся пользователей системы, рассортированный по убыванию процентного соотношения сходства пользователей и указывает, насколько интересы других пользователей совпадают с интересами текущего пользователя .

Выводы. Выявлено, что реализуемый в системе алгоритм для формирования выборки пользователей является более эффективным и точным примерно на 25–30% по сравнению с обычным алгоритмом Левенштейна. Также реализуемый алгоритм осуществляет выборку примерно в 10 раз быстрее, чем обычный алгоритм Левенштейна.

КЛЮЧЕВЫЕ СЛОВА: нечеткий поиск, алгоритм Левенштейна, модель Noisy Channel, сверточная нейронная сеть, сиамская нейронная сеть, фотоанализ лица, алгоритм расширения выборки, алгоритм N-грамм.

UDC 004.9

TECHNOLOGY FOR PERSONALITIES SOCIALIZATION BY COMMON INTERESTS BASED ON MACHINE LEARNING METHODS AND SEO-TECHNOLOGIES

Batiuk T. – Student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.
Vysotska V. – PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. The socialization of individuals with common interests is caused by the need of most people to simplify some of the moments of life by reducing the time for their implementation. With the rapid growth of information, the human workload in society and the recent epidemics of the world, people are becoming isolated from the opportunity to communicate. And this is one of the important needs of human consciousness and self-realization. Therefore, there is an urgent need to be able to obtain a recommended list of similar people of common interest as a result of intelligent search of many relevant users of social networks through analysis of human faces in user photos (based on neural networks) and analysis of user information based on fuzzy search algorithms and Noisy model. Channel).

Objective of the study is to develop technology for socialization of individuals based on SEO-technology and machine learning through the use of convolutional and Siamese neural networks to identify users and text analysis algorithms to select relevant users of future communication.

Method. In the implementation of SEO-technologies selected fuzzy word search algorithms based on the Noisy Channel model with algorithms for efficient distribution of textual information. During the implementation of machine learning, a convolutional neural network was developed to identify users of the system.

Results. An intelligent system of socialization of individuals by common interests based on SEO-technology and machine learning methods has been developed. The work of two neural networks was implemented: convolutional and Siamese, which allowed to search for a human face in photos uploaded by the user and compare the found face with those already available in the database / Internet. This makes it possible to effectively identify the authenticity of the user and ensure that this user is not currently in the database, so it is potentially real. Using fuzzy search algorithms, Levenstein's algorithm and the Noisy Channel model, an algorithm for analyzing and comparing user information was created, which for the current user forms a list of available users of the system, sorted by descending percentage of similarity and indicates how other users' interests coincide.

Conclusions. It was found that the algorithm implemented in the system for forming a sample of users is more efficient and accurate by about 25–30% compared to the usual Levenstein algorithm. Also, the implemented algorithm performs sampling approximately 10 times faster than the usual Levenstein algorithm.

KEYWORDS: fuzzy search, Levenstein algorithm, Noisy Channel model, convolutional neural network, Siamese neural network, facial photoanalysis, sample expansion algorithm, N-gram algorithm.

DOI 10.15588/1607-3274-2022-2-6

REFERENCES

1. Chu S. C., Sung Y. Using a consumer socialization framework to understand electronic word-of-mouth (eWOM) group membership among brand followers on Twitter, *Electronic Commerce Research and Applications,* 2015, Vol. 14(4), pp. 251–260. DOI: 10.1016/j.elerap.2015.04.002
2. De-Gregorio F., Sung Y. Understanding attitudes toward and behaviors in response to product placement: A Consumer Socialization Framework, *Journal of Advertising*, 2010, Vol. 39(1), pp. 83–96. DOI: 10.2753/JOA0091-3367390106
3. Elaheebocus S. M. R. A., Weal M., Morrison L., Yardley L. Peer-based social media features in behavior change interventions: Systematic review, *Journal of Medical Internet Research,* 2018, Vol. 20(2), pp. 1–20. 10.2196/jmir.8342
4. Erkan I. The influence of e-WOM in social media on consumers' purchase intentions: An extended approach to information adoption, *Computers in Human Behavior,* 2016. Vol. 4, pp. 47–55. DOI: 10.1016/j.chb.2016.03.003
5. Ferrara E., Interdonato R., Tagarelli A. Online popularity and topical interests through the lens of Instagram, *Hypertext and Social Media*, 2014, Vol. 2, pp. 24–23. DOI: 10.1145/2631775.2631808
6. Gao L., Bai X. Online consumer behavior and its relationship to website atmospheric induced flow: Insights into online travel agencies in China, *Journal of Retailing and Consumer Services,* 2014, Vol. 21(4), pp. 653–655. DOI: 10.1016/j.jretconser.2014.01.001
7. Geurin-Eagleman A. N., Burch L. M. Communicating via photographs: A gendered analysis of Olympic athletes' visual self-presentation on Instagram, Sport Management Review, 2016, Vol. 19(2), pp. 133–145. DOI: 10.1016/j.smr.2015.03.002
8. Guidry J. D., Messner M., Jin Y., Medina-Messner V. From McDonalds fail to Dominos sucks: An analysis of Instagram images about the 10 largest fast food companies, *Corporate Communications: An International Journal*, 2015, Vol. 20(3), pp. 344–359. DOI: 10.1108/CCIJ-04-2014-0027
9. Hanna R., Rohm A., Crittenden V. L. We're all connected: The power of the social media ecosystem, *Business Horizons*, 2011, Vol. 54(3), pp. 265–273. DOI: 10.1016/j.bushor.2011.01.007
10. Salganik M. J. Bit by bit: Social Research in the Digital Age. Princeton, Princeton University Press, 2019, 448 p.
11. Hudson S., Roth M., Madden T. J., Hudson R. The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees, *Tourism Management*, 2015, Vol. 47, pp. 68–76. DOI: 10.1016/j.tourman.2014.09.001
12. Jeff M., Jennifer R., Catherine J., Elke P. Managing brand presence through social media: The case of UK football clubs, *Internet Research*, 2014, Vol. 24(2), pp. 181–204. DOI: 10.1108/IntR-08-2012-0154
13. Kim E., Sung Y., Kang H. Brand followers' retweeting behaviour on Twitter: How brand relationship influence brand electronic word-of-mouth, *Computers in Human Behavior*, 2014, Vol. 37, pp. 18–25. DOI: 10.1016/j.chb.2014.04.020
14. Kudeshia C., Sikdar P., Mittal A. Spreading love through fan page liking: A perspective on small scale entrepreneurs, *Computers in Human Behavior,* 2016, Vol. 54, pp. 257–270. DOI: 10.1016/j.chb.2015.08.003
15. Lueg J. E., Finney R. Z. Interpersonal communication in the consumer socialization process: Scale development and validation, *Journal of Marketing Theory and Practice*, 2007, Vol. 15(1), pp. 25–39. DOI: 10.2753/MTP1069-6679150102
16. Mousavijad M., Payvandi S. The effect of socialization factors on decision making of teenagers consumers in schools, *Journal of School Administration,* 2017, Vol. 5(1), P. 217–234.
17. Schnell R. Enhancing Surveys with Objective Measurements and Observer Ratings, *Improving Survey Methods*. London, Routledge, 2014, pp. 310–324.
18. Parry M. E., Kawakami T., Kishiya K. The effect of personal and virtual word-of-mouth on technology acceptance, *Journal of Product Innovation Management,* 2012, Vol. 29(6), pp. 952–966. DOI: 10.1111/j.1540-5885.2012.00972.x
19. Quan-Haase A., Sloan L. Introduction to the Handbook of Social Media Research Methods: Goals, Challenges and Innovations, *The Sage Handbook of Social Media Research Methods*, 2017, Vol. 1, pp. 1–9.
20. Murphy S. T., Zajonc R. B. Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures, *Journal of Personality and Social Psychology,* 1993,Vol. 64(5), pp. 723–739. DOI: 10.1037/0022-3514.64.5.723
21. Ngoma M., Ntale P. D. Word of mouth communication: A mediator of relationship marketing and customer loyalty, *Cogent Business and Management*, 2019, Vol. 6(1), A. 1580123. DOI: 10.1080/23311975.2019.1580123
22. Ozdemir A., Tozlu B., Şen E., Ateşoğlu A. Analyses of word-of-mouth communication and its effect on students' university preferences, *Procedia – Social and Behavioral Sciences,* 2016, Vol. 235, pp. 22–35. DOI: 10.1016/j.sbspro.2016.11.022
23. Park J., Ciampaglia G. L., Ferrara F. Style in the age of Instagram: Predicting success within the fashion industry using social media, *Computer-Supported Cooperative Work & Social Computing, 19th ACM Conference, San Francisco, CA, USA, February 27 – March 2, 2016, proceedings*. ACM, Permissions@acm.org, 2019, pp. 64–73. DOI: 10.1145/2818048.2820065
24. Oborska O., Andrunyk V., Chyrun L., Hasko R., Vysotskyi A., Mushasta S., Petruchenko O., Shakleina I. The Intelligent System Development for Psychological Analysis of the Person's Condition, *Computational Linguistics and Intelligent Systems (COLINS 2021), 5th International Conference, Lviv, 22–23 April 2021, CEUR workshop proceedings.* Aachen, CEUR-WS.org, 2021, Vol. 2870, pp. 1390–1419.
25. Ranjbaran B., Jamshidian M., Dehghan Z. A survey for identification of major factors influencing customers attitude toward machine made carpet brands, *Journal of Commercial Strategies*, 2020, Vol. 5(23), pp. 109–118.
26. Batiuk T., Vysotska V., Lytvyn V. Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning, *Computational Linguistics and Intelligent Systems (COLINS 2020), 4th International Conference, Lviv, 23–24 April 2020, CEUR workshop proceedings*. Aachen, CEUR-WS.org, 2020, Vol. 2604, pp. 1237–1250.
27. Vysotska V. Information Technology for Internet Resources Promotion in Search Systems Based on Content Analysis of Web-Page Keywords, *Radio Electronics, Computer Science, Control,* 2021, No 3, pp. 133–151. DOI: 10.15588/1607-3274-2021-3-12

DOI 10.15588/1607-3274-2022-2-6

UDC 004.896

NEURAL NETWORK DIAGNOSTICS OF AIRCRAFT PARTS BASED ON THE RESULTS OF OPERATIONAL PROCESSES

Leoshchenko S. D. – Post-graduate student of the Department of Software Tools, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Pukhalska H. V. – PhD, Associate Professor, Associate Professor of the Department of Machinery Engineering Technology, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Subbotin S. A. – Dr. Sc., Professor, Head of the Department of Software Tools, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Oliinyk A. O. – Dr. Sc., Professor, Professor of the Department of Software Tools, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Gofman Ye. O. – PhD, Senior Researcher of the Research Unit, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

ABSTRACT

Context. The problem of synthesis of an optimal neural network model for diagnostics of aircraft parts after operational processes is considered. The object of the study is the process of synthesis of neural network diagnostic models for aircraft parts based on the results of operational processes

Objective is to synthesize neural network diagnostic models of aircraft parts after operational processes with a high level of accuracy.

Method. It is proposed to research the use of two approaches to the synthesis of neural network diagnostic models. So, using a system of indicators, the topology of the neural network is calculated, which will be trained using the method of Backpropagation method in the future. The second approach is based on the use of a neuroevolutionary approach, which allows for a complete synthesis of the neural network, dynamically modifying the topology of the solution in addition to the parameters. the final decisions are compared in the accuracy of work on the training and test data set. This approach will allow to determine the possibility and correctness of using neuroevolutionary methods for the synthesis of diagnostic models.

Results. Neuromodels for diagnostics of aircraft parts based on the results of operational processes have been obtained. The obtained results of comparing the methods used for synthesis made it possible to form recommendations for the implementation of neuroevolutionary methods in the synthesis of diagnostic neuromodels.

Conclusions. The results obtained during the experiments confirmed the operability of the mathematical software used and allowed us to form recommendations for further use of the considered methods in practice in order to synthesize diagnostic neuromodels. The prospects for further research may consist in expanding the input data sets in order to synthesize and study more complex topologies of neural network models.

KEYWORDS: diagnostics, aviation parts, synthesis, training, neuroevolution, data sampling, operational processes.

ABBREVIATIONS

ANN is an artificial neural net;
CPU is central processing unit;
FDI is fault detection and identification method;
MGA is modification genetic algorithm;
SSD is solid-state drive;
UAE is United Arab Emirates.

NOMENCLATURE

ε is discrepancy, which is the difference between the real and calculated outputs;
δ is error of the neurons in the output layer;
θ is error of the neurons in the hidden layer;
μ is the learning rate;
$\sigma_{0,2}$ is yield strength;
$\sigma_в$ is tensile strength;
$comp$ is separate component of system;
$char$ is separate characteristic of component;
DV_q is vector of desired outputs;
f_a is activation function;
FB is recurrent connections in ANN;
k is number of components at system;
l is number of neurons at the network input;
m is number of dependent (categorical) features of sample instances;
N_i is multiple neurons at the network input;
N_{i_l} is neuron at the network input;
N_o is multiple neurons at the network output;
N_{o_p} is neuron at the network output;
N_h is a multiple neurons of the hidden network layer;
N_{h_r} is hidden network layer neuron;
NN is a neural network;
NN_{struct} is structure of neural network;
p is number of neurons at the network output;
r is number of neurons in the hidden network layer;
$status$ is main defect characteristic of component;
$Sample$ is data set;
$System$ is general information about aviation system;
x_n is independent attribute of the sample instance;
X is input vector of perceptron;
$w_{i_l h_r}$ is coefficient of the input and hidden layers;

DOI 10.15588/1607-3274-2022-2-7

$w_{h_r o_p}$ is coefficient of the hidden and output layers;

W^{ih} is matrix of coefficients of the input and hidden layers;

W^{ho} is matrix of coefficient of the hidden and output layers;

y_m is value of the dependent variable (attribute) of the sample instance;

Y is output vector of perceptron.

Y_m is the vector of parameters calculated using the analytical model;

Y_{real} is the vector of the output parameters; of the engine obtained by measuring using sensors.

INTRODUCTION

At present, when the plane crashes have become a global problem, the problem of early detection of malfunctions of aircraft parts and systems has become particularly relevant [1, 2]. Traditionally, the process of diagnosing malfunctions of aviation systems is carried out using analytical models based on physical patterns, as well as by statistical processing of flight monitoring data. Specialists dealing with this problem install sensors that measure the parameters of aircraft engines during flights [1–3]. The flight monitoring data file usually contains parameters such as [1–3]:

– flight number;
– flight date;
– total engine operating time;
– temperature and air pressure at the engine inlet;
– temperature and gas pressure behind the turbine;
– temperature of the blades;
– oil level and temperature in the oil block;
– Mach number, etc.

The number of flight parameters can reach hundreds or more units.

After performing a certain number of flights, the engine, blades, transmission and other parts are removed from the aircraft and subjected to bench disassembly, during which a number of defects are identified and eliminated [1–4].

The task of the diagnostic engineer is to use flight monitoring data to identify system defects before they fail or before preventive disassembly. As already noted, traditionally this problem is solved by applying techniques based on physical laws: each defect causes certain deviations of certain parameters of work, physical characteristics, etc., therefore, analyzing their nature of change, it is possible to make assumptions about the appearance of defects that cause these changes. It is clear that due to the significant amounts of information and the complexity of the existing relationships between defects and measured parameters, the task of analyzing flight monitoring data and detecting defects is far from trivial and in many cases is not solved reliably and qualitatively enough [4].

The main directions determining the improvement of the quality of information technologies for diagnosing the technical condition of aviation should be considered the intellectualization of information processing processes involving data mining methods [2]. Such an approach is capable of improving the quality of recognition of the technical condition under the action of the above defined (measurable) and uncertain factors, as well as the integration of information processes (distributed local databases and knowledge into a global database and knowledge) [1, 3].

Data analyzing methods represent a new direction that complements and develops classical statistical research methods, often referred to in domestic and foreign literature as Data Mining and knowledge discovery. Data Mining uses modern intelligent technologies, including neural networks, fuzzy logic, expert systems. These technologies are used in this work to solve a wide range of problems of diagnostics of the technical condition of complex technical systems and their components [4].

In this article, it is proposed to solve this problem using a neural network basis, for example, using a multilayer perceptron with sigmoid activation functions. First of all, it should be noted that in the input vector X of the perceptron, places should be provided for all monitoring parameters, the values of which are affected by the appearance of detected defects. Possible defects of the aircraft engine can be encoded in the output vector Y, for example, using zeros and ones. Vectors of desired outputs are compiled DV_q based on the results of bench disassembly of engines [1–4].

The object of study is the process of using a model based on a neural network to diagnose aircraft parts after operational processes with high accuracy.

To test and investigate various approaches to the synthesis of neural network diagnostic models.

The subject of the study is a neural network model for the diagnosis of aircraft parts after operational processes, characterized by high accuracy.

Using information about operational processes and fixed traces of these processes to synthesize neural network diagnostic models.

The purpose of the work is to build and study a diagnostic neuromodel for aircraft parts after operational processes.

1 PROBLEM STATEMENT

The task of diagnosing aircraft parts based on the results of operational processes can be presented as a diagnostic task where it is necessary to determine whether the part is serviceable or not.

Thus, let's imagine an aviation system as a set of individual components $System = \{comp_1, comp_2, comp_3, ..., comp_k\}$, where k is the number of components (parts) of the system. Each component has a number of characteristics of features that can be measured with bench measurements or in real time using specialized sensors $comp = \{char_1, char_2, char_3, ..., char_l; status\}$, and in addition to physical (or chemical) characteristics, the component also has a characteristic of its defect: $status$. Thus, to train a diagnostic neuromodel NN, a sample is ob-

DOI 10.15588/1607-3274-2022-2-7

tained: $Sample=\langle X,Y\rangle$, where X the set of input features consists of the characteristics of the part $X=\{x_1=char_1, x_2=char_2, x_3=char_3,...,x_l=char_l\}$, and the set of output features consists of the characteristics of the defect of the part $Y=\{y=status\}$.

Then the diagnostic neuromodel of an aircraft part can be represented as an ANN: NN, consisting of structural elements and a set of parameters $NN=(struct, param)$. The structure of such a neuromodel is determined by sets of computational nodes – neurons and connections between them: $struct=\{N,c\}, N=\{N_i,N_h,N_o\}, c=\{c\}$. In turn, the aggregates of many neurons are divided into subsets by layers: $N_i=\{N_{i_1},N_{i_2},...,N_{i_l}\}, l=1,2,...,|N_i|$ neurons of the input layer, $N_o=\{N_{o_1},N_{o_2},...,N_{o_p}\}, p=1,2,...,|N_o|$ output and hidden $N_h=\{N_{h_1},N_{h_2},...,N_{h_r}\}, r=1,2,...,|N_h|$. It should be noted that the neurons of the input layer take values from the set of input features X, so their number is equal. The subset of links consists of the links themselves and their weighting coefficients: $c=\{c_1,c_2,..c_k\}, k=1,2,...,|c|$, $w=\{w_k\}$.

Accordingly, the task can be presented as a synthesis of ANN with optimal structure and accuracy $NN=(struct, param)$, based on a sample of initial, experimental data about the object under study $Sample=\langle X,Y\rangle$. For further automation of the process of diagnostics of aircraft parts based on the results of operation, as a particular classification task.

2 REVIEW OF THE LITERATURE

The analysis of works in the field of automation of the process of diagnosing the condition of aircraft parts based on analytical models, including ANNs [1–4], demonstrates that today such work is being carried out extremely actively. However, it is worth noting that a number of such works are poorly covered due to a number of factors: secrecy, military or corporate secrecy, narrow specialization of the tasks being solved. A number of works do not cover engineering solutions or give only general theoretical and practical recommendations for solving such problems.

The use of the FDI method is recognized as a common approach in similar tasks [5–8]. This methodology for solving problems of automation of diagnostics of the technical condition of aircraft parts is based on the principle of comparing the measurement results of physical (or chemical) parameters of a real part (system) with the calculated parameters calculated on the basis of a mathematical model [5–8].

Fig. 1 shows the general scheme of using the FDI method to automate the task of automatic technical diagnostics. So in the diagram, where X is the vector of control actions; Y_m is the vector of parameters calculated using the analytical model of the part (system); Y_{real} is the vector of the output parameters of the engine obtained by measuring using sensors; $\varepsilon=Y_{real}-Y_m$ the discrepancy, which is the difference between the vectors Y_m and Y_{real}.

As a category of work, he suggests using ANNs as an analytical model of a technical part (or system) [9, 10]. The range of tasks solved using such a model within the framework of the FDI method is quite wide: from the tasks of monitoring and diagnosing the technical condition to debugging parameters [9, 10].

The main stages of the engineering methodology for building an INS model include [9, 10]:

1) preliminary data analysis at the stage of setting the task and choosing the neural network architecture;
2) data transformation (preprocessing) to build a more efficient network setup procedure;
3) the choice of neural network architecture;
4) selection of the neural network structure;
5) selection of the learning algorithm;
6) neural network training and testing;
7) analysis of the accuracy of the neural network solution;
8) making a decision based on the results obtained.

The analysis of the published works devoted to the use of ANNs for diagnosing the parameters of aircraft parts shows that in these works the main trends and characteristic features of solving the problems of diagnostics of parts based on ANN are highlighted. At the same time, they are devoted, as a rule, to solving particular problems, for example [11–14]:

– diagnosing the condition of the turbine blades of a gas turbine engine;

– formation of a space of diagnostic signs of the state of a gas turbine engine for the construction of a neural network classifier;

– indirect measurement of the temperature of gases behind the combustion chamber based on the ANN to diagnose the thermal condition of the engine.

They do not contain instructions on the choice of architecture, structure and methods of ANN training; there is no engineering methodology for designing such networks in relation to the tasks of diagnosing the technical condition of aircraft engines. Neural network methods for solving problems of diagnostics of aircraft parts are investigated below in order to identify the main patterns of their use and develop appropriate methods and techniques for the implementation of diagnostics of technical condition based on ANN [9, 10].

3 MATERIALS AND METHODS

In general, an ANN is a mathematical model, as well as its software implementation (or imitation), working on the principle of the human brain: it runs input data through a system of neurons: computing nodes interacting with each other, after which it outputs a certain result of calculations based on this interaction [15–19]. Also, in more complex architectures of such models, previous experience and mistakes of past launches play an important role in decision-making. This behavior of the model

DOI 10.15588/1607-3274-2022-2-7

leads to the thesis about a certain level of self-learning of the ANNs as an artificial intelligence system [15–19].

Today, ANNs solve a wide range of tasks: from digital image processing to forecasting financial processes. Accordingly, in some tasks, models based on ANNs can replace experts: in medicine the doctors, in technical tasks the operators, etc. [15–19].

The main advantage of ANNs is their ability to study patterns in training data and how best to associate it with the target variable that needs to be determined (or forecasted). From an analytical point of view, ANNs are capable of recreating any function and have proven themselves as a universal approximation device, that is, the emulation of some objects into other, more simplified ones [15–19].

The Multilayer Perceptron is one of the simplest ANN models that emulates a primitive model of the biological brain within the framework of machine learning and can be used to solve complex computational tasks such as classification or prediction. To put it simply, it can be noted that a perceptron is a model of a single neuron, which was the predecessor of larger and more complex ANNs capable of more accurately emulating brain function and using more natural approaches to model learning [15–19].

The basic computing nodes (perceptron blocks) are artificial neurons, simple computing blocks that have weighted input signals and generate an output signal using the activation function. The parameter of the weighting coefficient of such a neuron is similar to the coefficients used in the equation from the theory of linear regression [15–19]. Similar to linear regression, each neuron also has bias, which can be considered as an input weight, by default equal to one. For example, a neuron may have two input data sources, in which case three weights are required: one for each input source and one for the weights. Weights are often initialized with small random values, but more complex initialization schemes can be used for more complex ANNs topologies [15–19]. As in linear regression, large weights indicate an increased complexity of the model. It is desirable that the weights in the network are small, then regularization methods are applicable. The weighted input data is summed up and transmitted via an activation function, sometimes called a transfer function. This is a simple display of the summed weighted input and output of a neuron. The function determines the threshold at which the neuron is activated and the strength of the output signal. Nonlinear activation functions are traditionally used. This allows the network to combine input data in a more complex way and, in turn, expand the capabilities of the functions that they can model [15–19].

Neurons are organized into a network. A number of neurons are called a layer, and one network can consist of several layers. The architecture of neurons in a network is often referred to as network topology. The initial input layer, which accepts input data from a dataset, is called visible because it is an open part of the network [15–19]. The layers after the input are called hidden because they are not directly exposed. The simplest network structure is to have a single neuron in the hidden layer that directly outputs the value [15–19]. With the availability of computing power and efficient software libraries, it is possible to build neural networks of deep learning, which means a lot of hidden layers. The last hidden layer is called the output layer, and it is responsible for the output of values or their vector in the appropriate format. After setting up, the neural network needs to be trained on your dataset [15–19].

One of the most common methods of ANNs training is the Backpropagation method [20, 21]. Having a simple perceptron, as in Fig. 2, it is noted the input layer, where data is received by one hidden layer, and the output layer [20, 21]. The input layer contains the number of neurons corresponding to the input data of the neuron, one of which is called the displacement neuron. The displacement neuron always contains the same value, for example, one and is designed to supply a constant displacement to all subsequent neurons with which it is connected, it can be disabled by setting it to 0. Next comes a hidden layer consisting of a given number of neurons, again one of which is a displacement neuron. Note that it is connected only to the subsequent output layer, no connections are received from the input layer, since it does not change its state. The result of the network is calculated on the output layer, in which the number of neurons is determined beforehand. As a rule, this number depends on the number of target variables [20, 21].

Each subsequent layer is connected to the previous layer by links with certain weight coefficients. There may be several hidden layers in the network. The network is called a direct distribution network because the first layer is connected to the second, the second to the third, and so on, and there are no feedbacks, for example, from the output layer to the input. Networks with feedbacks are called recurrent networks and are more complex and resource-intensive in operation [20, 21].

For convenience, in all cases, the displacement neuron number is assumed to be zero. The coupling coefficients of the input and hidden layers can be denoted as $w_{i_l h_r}$, and the matrix of these coefficients is denoted by W^{ih}. Thus, $w_{i_0 h_1}$ it determines the connection of the input layer displacement neuron with the first neuron of the hidden layer, and $w_{i_3 h_2}$ sets the connection of the third neuron of the input layer with the second neuron of the hidden layer [20, 21].

The coupling coefficients of the hidden and output layers can be denoted as $w_{h_r o_p}$, and the matrix of these coefficients will be called W^{ho} large.

DOI 10.15588/1607-3274-2022-2-7

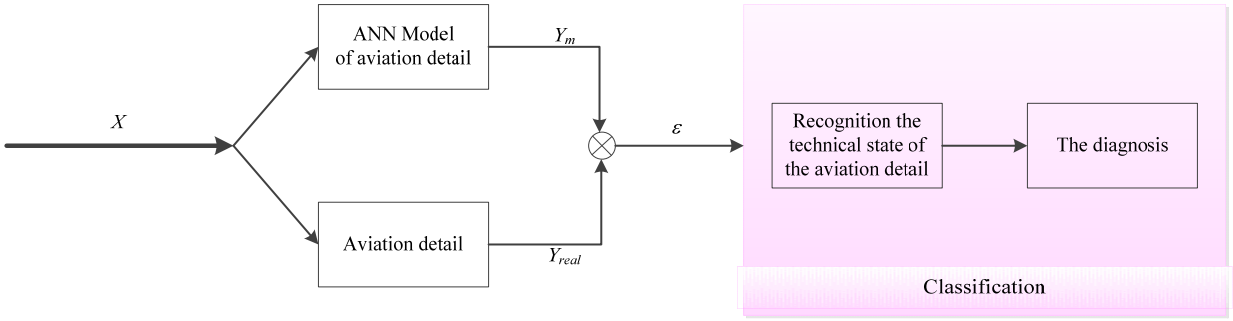

Figure 1 – Implementation of FDI method

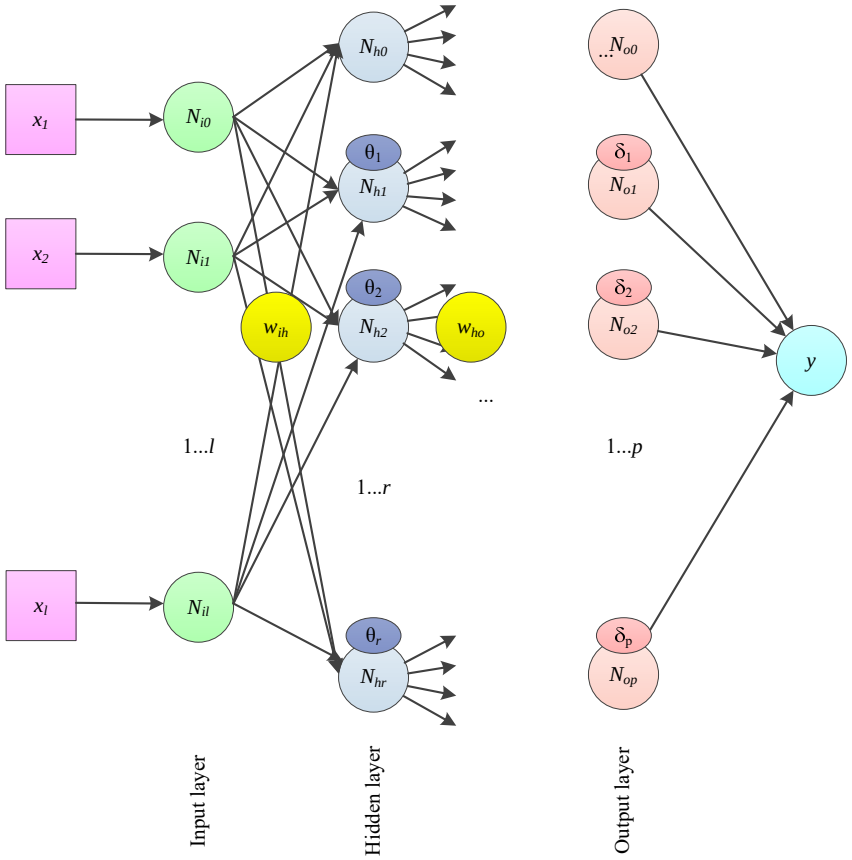

Figure 2 – General implementation of Backpropagation method

$$W^{ih} = \begin{pmatrix} w_{i_0h_1} & w_{i_0h_2} & w_{i_0h_3} & \dots & w_{i_0h} \\ w_{i_1h_1} & w_{i_1h_2} & w_{i_1h_3} & \dots & w_{i_1h} \\ w_{i_2h_1} & w_{i_2h_2} & w_{i_2h_3} & \dots & w_{i_2h} \\ \dots & \dots & \dots & \dots & \dots \\ w_{ih_1} & w_{ih_2} & w_{ih_3} & \dots & w_{ih} \end{pmatrix}, \quad (1)$$

$$W^{ho} = \begin{pmatrix} w_{h_0o_1} & w_{h_0o_2} & \dots & w_{h_0o} \\ w_{h_1o_1} & w_{h_1o_2} & \dots & w_{h_1o} \\ \dots & \dots & \dots & \dots \\ w_{ho_1} & w_{ho_2} & \dots & w_{ho} \end{pmatrix}. \quad (2)$$

The values of the input layer can be represented by a vector $X = N_i$. The values of the neurons of the hidden layer N_{h_r}, they make up a vector N_h, and the output values N_{o_p} are a vector N_o. The information is processed sequentially, first the values of the hidden layer N_h are calculated, then the values of the output layer N_o [20, 21].

The formula for calculating the values of the hidden layer is indicated by a number:

$$N_{h_r} = f_a\left(\sum_{i=0...l} N_{i_l} \cdot w_{ih}\right). \quad (3)$$

DOI 10.15588/1607-3274-2022-2-7

Each neuron calculates a combined input consisting of the sum of the products of the input value by the corresponding weight, and then the result is run through the activation function of this neuron f_a.

By analogy with the previous layer, the formula (4) is compiled to calculate the output values. The combined input is the sum of the products of the values of the intermediate layer by the values N_{h_r} of the weights $w_{h_r o_p}$. The result is fed to the activation function [20, 21].

$$N_{o_p} = f_a\left(\sum_{h=0...r} N_{h_r} \cdot w_{h_r o_p}\right). \quad (4)$$

The essence of the method is that when submitting a training set of examples, the result of the network is compared with the target value, errors in the output layer are determined as δ (Fig. 2), and then these errors are propagated in the opposite direction and the errors of the neurons of the hidden layers are calculated as θ, and at the last step, the values of all weights are adjusted based on the values errors found [20, 21].

It is necessary to use the general (5), in which the difference between the target N_{o_p} and real values y_p is multiplied with the value of the derivative of the activation function:

$$\delta_p = \left(y_p - N_{o_p}\right) \cdot f_a'\left(net_p\right). \quad (5)$$

Next, we will find the errors of the neurons of the hidden layer: θ. The error for the displacement neuron is not calculated:

$$\theta_r = f_a'(net_r) \sum_{p=1...p} \delta_p \cdot w_{ho}. \quad (6)$$

Thus, the error of a hidden layer neuron is a combination of the errors of all the neurons that it affects. The larger the connection $w_{h_r o_p}$, the more the error of the output layer δ affects the error of the neuron of the hidden layer. Thus, the error is propagated backwards from the network output to its hidden layers [20, 21].

The final stage is the adjustment of the weights of the arrays W^{ih} and W^{ho} [20, 21]:

$$w_{ih}' = w_{ih} + \Delta w_{ih} = w_{ih} + \mu N_{i_l} \theta_r, \quad (7)$$

$$w_{ho}' = w_{ho} + \Delta w_{ho} = w_{ho} + \mu N_{h_r} \delta_p, \quad (8)$$

where μ is the learning rate, which is set in the range $[0.1;0.4]$.

However, analyzing the above method, it can be concluded that in general, the training of the model based on the ANN is reduced to iterative iteration of trial and error, since the Backpropagation method does not involve the selection and fine-tuning of the architecture, but works with the already selected topology. Moreover, a number of papers note problems in the areas of local optima. Therefore, since the 2010s, more and more attention has been paid to neuroevolutionary methods of ANN synthesis [22, 23]. Such approaches existed before, but it was with the growth of computational capabilities that they began to show better results in comparison with gradient learning methods [22, 23].

The neuroevolutionary approach to the synthesis of INS uses evolutionary methods to create an ANN: the selection of its parameters, topology and rules. Neuroevolution is usually used as part of the reinforcement learning paradigm, and it can be contrasted with traditional deep learning methods that use gradient descent in a neural network with a fixed topology. Due to more flexible settings of synthesis parameters, the process allows fine-tuning and selecting the ANNs architecture for each task, avoiding the problem of retraining [22, 23].

Of course, this approach involves the use of large computing and time resources. So the synthesis process begins with the installation of metaparameters and the synthesis framework: the accuracy of the ANN, the number of epochs, the learning rate and topological complexity. The complexity can be set by limiting the number of hidden layers and neurons in them, the presence of feedbacks in the neurons of the hidden layer, etc. [22, 23].

As a neuroevolutionary method, consider a MGA. So, at the beginning, restrictions are set on the structural complexity of the final solution: the presence of feedbacks ($FB = 0 \,||\, FB = 1$), the number and depth of hidden layers ($|N_h|$) and stopping criteria. After that, a population is generated from simple ANN, and their genetic information is encoded based on interneuronal connections. Further, relying on the mechanisms of selective pressure and smart crossover, the main stages of GA are performed: crossing, mutation of a new generation and selection of individuals into the parent pool [23]. In general, the method can be represented schematically as in Fig. 3.

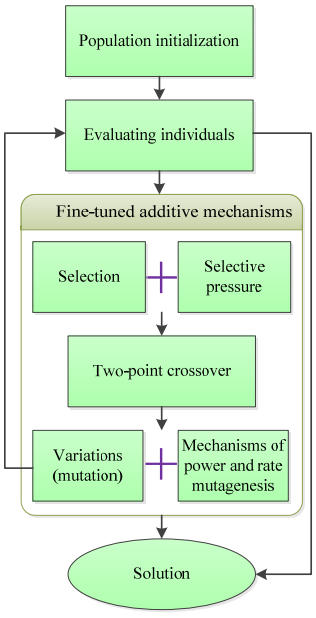

Figure 3 – General scheme of MGA method

DOI 10.15588/1607-3274-2022-2-7

4 EXPERIMENTS

The blades of the first stage of the compressor of the Klimov TV3-117 engine, having operational damage to the feather of the blades of the engines, were selected as the object of research [24]. In the studies, engines that were in operation in different countries were observed, respectively, the physical characteristics of the operational processes differed. From this it can be concluded that aircraft parts have different operating time and, accordingly, different degrees of damage to the blades. The engines were operated and observed in the following countries and enterprises: Yemen, India, UAE, Peru, Cyprus, Utair (Tyumen), Algeria, Spain [24].

Table 1 shows an example of sampling input data.

The table shows that x_1 is the average temperature in the region where the operational process took place; x_2 and x_3 are the values of the chord, in sections A2–A2 and A8–A8; x_4 is HB, the hardness of the initial blade, HRC; x_5 is $б_{0,2}$, yield strength, MPa; x_6 is $б_в$ tensile strength, MPa; x_7 is the frequency of natural vibrations of the blades, Hz.

y_1: T_1 total operating time; y_2: T2 operating time up to first repair.

For the experiments, a workstation with the following characteristics was used: Intel Core i5-8250U CPU (1.60–3.40 GHz (Intel Turbo Boost 2.0), 4 cores and 8 threads), 16 Gb RAM (dual-channel mode), SK hynix SC308 128 GB SSD (M.2), the Java programming language.

5 RESULTS

Table 2 shows the selected information features with their weight coefficients.

Table 3 shows a comparison of the results of the two methods. So the work of the methods was compared according to the following parameters:

– work time: time spent on the synthesis of ANN;

– accuracy of work on the training sample: accuracy of the model during training;

– accuracy of work on the test sample: accuracy of the model during testing.

Tables 4 and 5 shows the neural network models obtained.

Table 1 – Example of fragment from data set

Blade number	Average temperature	Blade A2-A2	Blade A8-A8	HRC	$б_{0.2}$	$б_в$	Self-frequency of natural vibrations	T_1	T_2
Index	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y_1	y_2
India-1	24.6	26.7	28.2	38	970	1180	603.2	1652	724
India-2	24.6	26.55	28.22	38	970	1180	617.1	1652	724
India-3	24.6	26.73	28.1	38	970	1180	631.8	1652	724
India-4	24.6	26.75	28.09	38	970	1180	623.9	1652	724
India-5	24.6	26.59	28.12	38	970	1180	634.9	1652	724
India-6	24.6	26.56	28.22	38	970	1180	624	1652	724
India-7	24.6	26.6	28.13	38	970	1180	629.9	1652	724
India-8	24.6	26.53	28	38	970	1180	637.2	1652	724
India-9	24.6	26.83	28.2	38	970	1180	615	1652	724
India-10	24.6	26.3	28.28	38	970	1180	625.4	1652	724
…	…	…	…	…	…	…	…	…	…
Yemen-20	20.5	26.63	28	990	451	32	950	1100	627.4

Table 2 – Results of feature selection

	y_2	y_3
x_1	0.2153	–0.1901
x_2	–0.0323	–0.0030
x_3	–0.3629	–0.0191
x_6	0.7211	0.4971
x_7	0.0844	0.0336

Table 3 – Results of using different methods for neuromodels synthesis

Target variable	Method	The synthesis time. s	Accuracy of work on the training sample	Accuracy of work on the test sample
y_1	Backpropagation	15.3726	0.0002	0.00025
	MGA	64.2397	0.00017	0.00024
y_2	Backpropagation	15.2863	0.0001	0.0001
	MGA	52.6493	0.00014	0.0001

DOI 10.15588/1607-3274-2022-2-7

Table 4 – Coefficients matrices of resulting neuromodels for y_1

	Number of layer	Number of neuron at layer	Number of input of neuron			
			0	1	2	3
Backpropagation y_1	1	1	–40.2890	64.2278	–1.2651	152.6108
		2	–8.7750	–7.2652	0.0004	–104.6105
		3	10.5814	11.9088	0.0000	92.3337
		4	22.2327	–9.9015	0.2474	130.9445
		5	–37.1694	–34.9135	0.0009	57.2398
	2	1	17.5119	–10.0749	0.0060	–42.5008
	Number of layer	Number of neuron at layer	Number of input of neuron			
			0	1	2	3
MGA y_1	1	1	–10.3678	3.9329	0.1771	44.4167
		2	–42.0171	0.0407	0.0058	0.3170
		3	–79.2515	0.2169	0.1018	–85.1717
		4	20.4838	0.5891	–0.0395	10.0486
		5	21.3511	0.1410	0.0512	9.8286
	2	1	37.8691	–3.0681	0.8152	41.7886

Table 5 – Coefficients matrices of resulting neuromodels y_2

	Number of layer	Number of neuron at layer	Number of input of neuron			
			0	1	2	3
Backpropagation y_2	1	1	–2.3766	–2.7395	–2.7277	5.7025
		2	1.8790	0.0000	1.7174	0.0000
		3	–15.9818	0.0000	–15.2190	0.0000
		4	–7.6228	1.7200	–7.7104	0.8612
		5	37.7405	0.0000	36.0519	0.0000
	2	1	–7.1881	1.8330	–7.0854	4.2028
	Number of layer	Number of neuron at layer	Number of input of neuron			
			0	1	2	3
MGA y_2	1	1	13.5323	–9.2208	5.8346	–15.7445
		2	–0.6086	–0.6000	0.4055	0.4643
		3	–4.1636	–4.1129	2.0460	–106.7364
		4	7.7830	–2.6386	4.4783	–13.5350
		5	8.0768	–2.3306	2.9106	–13.5143
	2	1	–2.3971	–4.4495	6.9384	–3.2805

6 DISCUSSION

For the operating time in both cases, the combination of informatively important features is the same. And in both cases, the frequency of natural vibrations of the blades is an important sign.

When initializing the synthesis process using MGA, restrictions were set on the absence of feedbacks and excessive growth of hidden layers. Based on the assessment of the complexity of the task, the optimal number of neurons in the hidden layer was chosen 4 [25]. During neuroevolutionary synthesis, this number of neurons was confirmed.

Comparing the operating time, it can be noted that the MGA method worked much slower, this is due to the fact that the method worked in single-threaded mode and completely synthesized a new network architecture, operating with a population of non-network models. From this we can conclude that in simple tasks, neuroevolutionary methods may need increased time resources. At the same time, the higher accuracy of the synthesized solution (which has been confirmed experimentally) may not fully justify such time expenditures.

However, in complex tasks, when the process of input data preprocessing is not possible or is largely difficult and the accuracy of the model is extremely important, neuroevolutionary methods can show great efficiency. This is due to the lower dependence of the operation of such methods on the noise of the input data, as well as the proportionally increasing time spent on training complex topologies using iterative methods.

CONCLUSIONS

The urgent scientific and applied problem of synthesis of an optimal neural network model for diagnostics of aircraft parts after operational processes has been solved.

The scientific novelty lies in the fact that it is proposed to use different methods for the synthesis of neuromodels. Thus, the Backpropagation method was used to train a predefined ANN structure based on an assessment of the complexity of the simulated task. The MGA method was also used for neuroevolutionary synthesis of the model. As a result, both methods presented similar perceptron topologies with the same structures.

The practical significance lies in the fact that the rationality of approaches to the synthesis of neuromodels has been investigated. So for y_1 and y_2, MGA worked slower by 23.93% and 29.03%, respectively. At the same time, the accuracy of the resulting models differed by 0.3–0.1 percentage points. From this we can form a recommendation: for such tasks, the use of neuroevolutionary methods may not be justified precisely in the case of the time resources spent. However, for more complex tasks, where accuracy is more important, the neuroevolutionary approach will be preferable.

Prospects for further research are to expand the dataset of input characteristics of aircraft parts to use

DOI 10.15588/1607-3274-2022-2-7

complex ANN topologies and monitor the accuracy of their operation.

ACKNOWLEDGEMENTS

The work was carried out with the support of the state budget research projects of the state budget of the National University "Zaporozhzhia Polytechnic" "Development of methods and tools for analysis and prediction of dynamic behavior of nonlinear objects" (state registration number 0121U107499) and "Intelligent methods and tools for diagnosing and predicting the state of complex objects" (state registration number 0122U000972).

REFERENCES

1. Smith D. J. Reliability, Maintainability and Risk: Practical Methods for Engineers, Oxford, Butterworth-Heinemann, 2021, 516 p.
2. Høyer C. B., Nielsen T. S., Nagel L. L., Uhrenholt L., Boel L.W.T. Investigation of a fatal airplane crash: autopsy, computed tomography, and injury pattern analysis used to determine who was steering the plane at the time of the accident. A case report, *Forensic Science, Medicine and Pathology*, 2012, Vol. 8(2), P. 179–188. DOI: 10.1007/s12024-011-9239-4
3. Maltry G.W. Airplane Crash Analysis [Electronic resource]. Access mode: https://www.edtengineers.com/blog-post/airplane-crash-analysis
4. Yunusov S., Labendik V., Guseynov S. Monitoring and Diagnostics of Aircraft Gas Turbine Engines: Improvement of Models and Methods for Diagnosis of Gas Path of Gas Turbine Engines, Chisinau: LAP LAMBERT Academic Publishing, 2014, 204 p.
5. Johri P., Anand A., Vain J., Singh J., Quasim M.T. System Assurances: Modeling and Management (Emerging Methodologies and Applications in Modelling, Identification and Control), Cambridge: Academic Press, 2022, 614 p.
6. Sun W., Paiva A.R.C., Xu P., Sundaram A., Braatz R.D. Fault Detection and Identification using Bayesian Recurrent Neural Networks, *Computers & Chemical Engineering*, 2019, Vol. 141, pp. 1–43. DOI: 10.1016/j.compchemeng.2020.106991
7. Nguyen H.V., Golinval J.-C. Fault detection based on Kernel Principal Component Analysis, *Engineering Structures*, 2010, Vol. 32(11), pp. 3683–3691. DOI: 10.1016/j.engstruct.2010.08.012
8. Aldrich C., Auret L. Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods, Berlin: Springer, 2013, 396 p.
9. Adouni A., Chariag D., Diallo D., Hamed M.B., Sbita L. FDI based on Artificial Neural Network for Low-Voltage-Ride-Through in DFIG-based Wind Turbine, *ISA Transactions*, 2016, Vol. 64, pp. 353–364. DOI: 10.1016/j.isatra.2016.05.009
10. Plikynas D., Akbar Y. H. Neural Network Approaches to Estimating FDI Flows: Evidence from Central and Eastern Europe, *Eastern European Economics*, Vol. 44, No. 3, 2006, pp. 29–59.
11. Babenko O., Pribora T. Analysis of the results of the study of the frequencies and forms of natural vibrations of the working blade of the 1st stage of the SLP, Bulletin of Engine Building, 2018, Vol. 2, pp. 91–98. [In Russian]
12. Dvirnyk Ya., Pavlenko D. The influence of dust erosion on the gas dynamic characteristics of the axial compressor of the GTE Vestnik dvigatelstroeniya, Bulletin of Engine Building, 2017, Vol. 1, pp. 56–66. [In Russian]
13. Yefanov V., Prokopenko O., Ovchinnikov O., Vnukov U. Erosion resistance of compressor blades of helicopter gas turbine engines with various types of coatings, Bulletin of Engine Building, 2017, Vol. 1, pp. 120–123. [In Russian]
14. Dvirnyk Ya., Pavlenko D. Patterns of wear of the compressor blades of helicopter engines operating in a dusty atmosphere, Bulletin of Engine Building, 2016, Vol. 1, pp. 42–51. [In Russian]
15. Huang H. Statistical Mechanics of Neural Networks, Berlin: Springer, 2022, 314 p.
16. Ekman M. Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow, Boston, Addison-Wesley Professional, 2021, 752 p.
17. Aggarwal C. C. Neural Networks and Deep Learning: A Textbook, Berlin, Springer, 2018, 520 p.
18. Wu L., Cui P., Pei J., Zhao L. Graph Neural Networks: Foundations, Frontiers, and Applications, Berlin, Springer, 2022, 752 p.
19. Kneusel R.T. Math for Deep Learning: What You Need to Know to Understand Neural Networks, San Francisco, No Starch Press, 2021, 344 p.
20. Chapmann J. Neural Networks: Introduction to Artificial Neurons, Backpropagation Algorithms and Multilayer Feedforward Networks (Advanced Data Analytics), Scotts Valley, CreateSpace Independent Publishing Platform, 2017, 108 p.
21. Wadi H. Learn From Scratch Backpropagation Neural Networks using Python GUI & MariaDB, Chicago, Independently published, 2021, 590 p.
22. Milani A., Carpi A., Poggioni V. Evolutionary Algorithms in Intelligent Systems, Basel: Mdpi AG, 2020, 144 p.
23. Leoshchenko S., Oliinyk A., Subbotin S., Lytvyn V., Shkarupylo V. Modification and parallelization of genetic algorithm for synthesis of artificial neural networks, *Radio Electronics, Computer Science, Control*, 2019, No. 4, pp. 68–82. DOI: 10.15588/1607-3274-2019-4-7
24. Subbotin S., Pukhalska H., Leoshchenko S., Oliinyk A., Gofman Ye. Neuromodeling of operational processes, *Radio electronics, computer science, control*, 2022, No. 1, pp. 120–129.
25. Leoshchenko S., Subbotin S., Oliinyk A., Narivs'kiy O. Implementation of the indicator system in modeling complex technical systems, *Radio electronics, computer science, control*, 2021, No. 1, pp. 117–127. DOI: 10.15588/1607-3274-2021-1-12.

Received 15.03.2022.
Accepted 22.03.2022.

DOI 10.15588/1607-3274-2022-2-7

УДК 004.896

НЕЙРОМЕРЕЖЕВЕ ДІАГНОСТУВАННЯ АВІАЦІЙНИХ ДЕТАЛЕЙ ЗА РЕЗУЛЬТАТАМИ ЕКСПЛУАТАЦІЙНИХ ПРОЦЕСІВ

Леощенко С. Д. – аспірант кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя Україна.

Пухальська Г.В. – канд. техн. наук, доцент кафедри технологія машинобудування Національного університету «Запорізька політехніка», Запоріжжя, Україна.

Субботін С. О. – д-р техн. наук, професор, завідувач кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя, Україна.

Олійник А. О. – д-р техн. наук, професор, професор кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя, Україна.

Гофман Є. О. – канд. техн. наук, старший науковий співробітник науково-дослідної частини Національного університету «Запорізька політехніка», Запоріжжя, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто завдання синтезу оптимальної нейромережевої моделі для діагностики авіаційних деталей після експлуатаційних процесів. Об'єктом дослідження є процес синтезу нейромережевих діагностичних моделей для авіаційних деталей за результатами експлуатаційних процесів.

Мета роботи полягає в синтезі нейромережевих діагностичних моделей авіаційних деталей після експлуатаційних процесів з високим рівнем точності.

Метод. Запропоновано дослідити використання двох підходів до синтезу нейромережевих діагностичних моделей. Так використовуючи систему індикаторів, обчислюється топологія нейронної мережі, яка в подальшому буде навчена з використанням методу зворотного поширення помилки. Другий же підхід грунтується на використанні нейроеволюційного підходу, який дозволяє зробити повний синтез нейронної мережі, динамічно модифікуючи крім параметрів і топологію рішення. підсумкові рішення порівнюються в точності роботи на навчальному і тестовому наборі даних. Такий підхід дозволить визначити можливість і коректність використання нейроеволюційних методів для синтезу діагностичних моделей.

Результати. Отримано нейромоделі для діагностики авіаційних деталей за результатами експлуатаційних процесів. Отримані результати порівняння використовуваних для синтезу методів дозволили сформувати рекомендації для імплементації нейроеволюційних методів в процеси синтезу діагностичних нейромоделей.

Висновок. Отримані в ході експериментів результати підтвердили працездатність використовуваного математичного забезпечення і дозволили сформувати рекомендації для подальшого використання розглянутих методів на практиці з метою синтезу діагностичних нейромоделей. Перспективи подальших досліджень можуть полягати в розширенні вхідних наборів даних з метою синтезу і дослідження більш складних топологій нейромережевих моделей.

КЛЮЧОВІ СЛОВА: діагностування, авіаційні деталі, синтез, навчання, нейроеволюція, вибірка даних, експлуатаційні процеси.

УДК 004.896

НЕЙРОСЕТЕВОЕ ДИАГНОСТИРОВАНИЕ АВИАЦИОННЫХ ДЕТАЛЕЙ ПО РЕЗУЛЬТАТАМ ЭКСПЛУАТАЦИОННЫХ ПРОЦЕССОВ

Леощенко С. Д. – аспирант кафедры программных средств Национального университета «Запорожская политехника», Запорожье Украина.

Пухальская Г.В. – канд. техн. наук, доцент кафедры технология машиностроения Национального университета «Запорожская политехника», Запорожье Украина.

Субботин С. А. – д-р техн. наук, профессор, заведующий кафедрой программных средств Национального университета «Запорожская политехника», Запорожье Украина.

Олейник А. А. – д-р техн. наук, профессор, профессор кафедры программных средств Национального университета «Запорожская политехника», Запорожье Украина.

Гофман Е. А. – старший научный сотрудник научно-исследовательской части Национального университета «Запорожская политехника», Запорожье Украина.

АННОТАЦИЯ

Актуальность. Рассмотрена задача синтеза оптимальной нейросетевой модели для диагностики авиационных деталей после эксплуатационных процессов. Объектом исследования является процесс синтеза нейросетевых диагностических моделей для авиационных деталей по результатам эксплуатационных процессов.

Цель работы заключается в синтезе нейросетевых диагностических моделей авиационных деталей после эксплуатационных процессов с высоким уровнем точности.

Метод. Предложено исследовать использование двух подходов к синтезу нейросетевых диагностических моделей. Так используя систему индикаторов, вычисляется топология нейронной сети, которая в дальнейшем будет обучена с использованием метода обратного распространения ошибки. Второй же подход основывается на использовании нейроэволюционного подхода, который позволяет произвести полный синтез нейронной сети, динамично модифицируя помимо параметров и топологию решения. итоговые решения сравниваются в точности работы на обучающем и тестовом наборе данных. Такой

DOI 10.15588/1607-3274-2022-2-7

подход позволит определить возможность и корректность использования нейроэволюционных методов для синтеза диагностических моделей.

Результаты. Получены нейромодели для диагностики авиационных деталей по результатам эксплуатационных процессов. Полученные результаты сравнения используемых для синтеза методов позволили сформировать рекомендации для имплементации нейроэволюционных методов в процессы синтеза диагностических нейромоделей.

Выводы. Полученные в ходе экспериментов результаты подтвердили работоспособность используемого математического обеспечения и позволили сформировать рекомендации для дальнейшего использования рассматриваемых методов на практике с целью синтеза диагностических нейромоделей. Перспективы дальнейших исследований могут заключаться в расширении входных наборов данных с целью синтеза и исследования более сложных топологий нейросетевых моделей.

КЛЮЧЕВЫЕ СЛОВА: диагностирование, авиационные детали, синтез, обучение, нейроэволюция, выборка данных, эксплуатационные процессы.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Smith D.J . Reliability, Maintainability and Risk : Practical Methods for Engineers / D. J. Smith. – Oxford : Butterworth-Heinemann, 2021. – 516 p.
2. Investigation of a fatal airplane crash: autopsy, computed tomography, and injury pattern analysis used to determine who was steering the plane at the time of the accident. A case report / [C. B. Høyer, T. S. Nielsen, L. L. Nagel et al.] // Forensic Science, Medicine and Pathology. – 2012. – Vol. 8(2). – P. 179–188. DOI: 10.1007/s12024-011-9239-4
3. Maltry G.W. Airplane Crash Analysis [Electronic resource]. Access mode: https://www.edtengineers.com/blog-post/airplane-crash-analysis
4. Yunusov S. Monitoring and Diagnostics of Aircraft Gas Turbine Engines: Improvement of Models and Methods for Diagnosis of Gas Path of Gas Turbine Engines / S. Yunusov, V. Labendik, S. Guseynov. – Chisinau : LAP LAMBERT Academic Publishing, 2014. – 204 p.
5. System Assurances: Modeling and Management (Emerging Methodologies and Applications in Modelling, Identification and Control) / [P. Johri, A. Anand, J. Vain et al.]. – Cambridge : Academic Press, 2022. – 614 p.
6. Fault Detection and Identification using Bayesian Recurrent Neural Networks / [W. Sun, A.R.C. Paiva, P. Xu et al.] // Computers & Chemical Engineering. – 2019. – Vol. 141. – P. 1–43. DOI: 10.1016/j.compchemeng.2020.106991
7. Nguyen H. V. Fault detection based on Kernel Principal Component Analysis / H. V. Nguyen, J.-C. Golinval // Engineering Structures. – 2010. – Vol. 32(11). – P. 3683–3691. DOI: 10.1016/j.engstruct.2010.08.012
8. Aldrich C. Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods / C. Aldrich, L. Auret. – Berlin : Springer, 2013. – 396 p.
9. FDI based on Artificial Neural Network for Low-Voltage-Ride-Through in DFIG-based Wind Turbine / [A. Adouni, D. Chariag, D. Diallo et al.] // ISA Transactions. – 2016. – Vol. 64. – P. 353–364. DOI: 10.1016/j.isatra.2016.05.009
10. Plikynas D. Neural Network Approaches to Estimating FDI Flows: Evidence from Central and Eastern Europe / D. Plikynas, Y. H. Akbar // Eastern European Economics. – 2006. – Vol. 44, No. 3. –P. 29–59.
11. Бабенко О. Н. Анализ результатов исследования частот и форм собственных колебаний рабочей лопатки 1 ступени КНД / О. Н. Бабенко, Т. И. Прибора // Вестник двигателестроения. – 2018. – № 2. – С. 91–98.
12. Двирник Я.В. Влияние пылевой эрозии на газодинамические характеристики осевого компрессора ГТД / Я. В. Двирник, Д. В. Павленко // Вестник двигателестроения. – 2017. – № 1. – С. 56–66.
13. Эрозионная стойкость лопаток компрессора вертолетных газотурбинных двигателей с различными типами покрытий / [В. С. Ефанов, А. Н. Прокопенко, А. В. Овчинников и др.] // Вестник двигателестроения. – 2017 – № 1. – С. 120–123.
14. Павленко Д. В. Закономерности изнашивания рабочих лопаток компрессора вертолетных двигателей, эксплуатирующихся в условиях запыленной атмосферы / Д. В. Павленко, Я. В. Двирник // Вестник двигателестроения. – 2016. – № 1. – С. 42–51.
15. Huang H. Statistical Mechanics of Neural Networks / H. Huang. – Berlin : Springer, 2022. – 314 p.
16. Ekman M. Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow / M. Ekman. – Boston : Addison-Wesley Professional, 2021. – 752 p.
17. Aggarwal C. C. Neural Networks and Deep Learning: A Textbook / C. C. Aggarwal. – Berlin : Springer, 2018. – 520 p.
18. Graph Neural Networks: Foundations, Frontiers, and Applications / [L. Wu, P. Cui, J. Pei, L. Zhao]. – Berlin : Springer, 2022. – 752 p.
19. Kneusel R. T. Math for Deep Learning: What You Need to Know to Understand Neural Networks / R. T. Kneusel. – San Francisco : No Starch Press, 2021. – 344 p.
20. Chapmann J. Neural Networks: Introduction to Artificial Neurons, Backpropagation Algorithms and Multilayer Feedforward Networks (Advanced Data Analytics) / J. Chapmann. – Scotts Valley : CreateSpace Independent Publishing Platform, 2017. – 108 p.
21. Wadi H. Learn From Scratch Backpropagation Neural Networks using Python GUI & MariaDB / H. Wadi. – Chicago : Independently published, 2021. – 590 p.
22. Evolutionary Algorithms in Intelligent Systems / [A. Milani, A. Carpi, V. Poggioni]. – Basel : Mdpi AG, 2020. – 144 p.
23. Modification and parallelization of genetic algorithm for synthesis of artificial neural networks / [S. D. Leoshchenko, A. O. Oliinyk, S. A. Subbotin at al.] // Radio Electronics, Computer Science, Control. – 2019. – № 4. – P. 68–82. DOI: 10.15588/1607-3274-2019-4-7
24. Neuromodeling of operational processes / [S. A. Subbotin, H. V. Pukhalska, S. D. Leoshchenko et al.] // Radio electronics, computer science, control. – 2022. – № 1. – P. 120–129.
25. Implementation of the indicator system in modeling complex technical systems / [S. D. Leoshchenko, S. A. Subbotin, A. O. Oliinyk, O. E. Narivs'kiy] // Radio electronics, computer science, control. – 2021. – № 1. – P. 117–127. DOI: 10.15588/1607-3274-2021-1-12.

DOI 10.15588/1607-3274-2022-2-7

UDC 303.732.4 + 004.67 + 004.8

AUTOMATIC CLASSIFICATION OF PAINTINGS BY YEAR OF CREATION

Martynenko A. A. – Senior Lecturer of Department of Software Engineering, Dnipro University of Technology, Dnipro, Ukraine.

Tevyashev A. D. – Dr. Sc., Professor, Head of Department of Applied Mathematics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Kulishova N. E. – PhD, Assistant Professor, Professor of Department of Media Systems and Technologies, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Moroz B. I. – Dr. Sc., Professor, Corresponding Member of the Academy of Applied Electronics, Professor of the Software Engineering Department, Dnipro University of Technology, Dnipro, Ukraine.

Sergienko A. S. – Student of Department of Applied Mathematics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. The problem of automatic verification of the legitimacy of the export of works of art is considered.

Objective. A method is proposed for automatically determining the age of a painting from a digital photograph using a classification that is performed by an intelligent decision-making system.

Method. It is proposed to use the attribute of picture year of creation as the main criterion for making a decision during the customs check of exports legitimacy. Instead of a long and expensive museum examination, photographing works of art in customs conditions and processing photos using a set of descriptors is used. The set of descriptors is proposed, include local binary patterns, their color modification, Haralik's texture features, the first four moments, Tamura's texturt features, SIFT descriptor. The data obtained as a result of descriptors action give the values of several dozen private attributes. They form data vectors, which are then concatenated into a generalized object description vector. In the feature space thus created, automatic classification by weighted k-nearest neighbors is performed. The proposed algorithm calculates the distance between objects in a multidimensional space of attribute values and assigns new objects to already formed classes. The criterion for creating classes is the age of the painting from the existing database. As a measure of the objects proximity, it is proposed to use the Euclid and Minkowski metrics. The calculation of weights for the proposed classification algorithm is performed by the Fisher method.

Results. The effectiveness of the proposed method was investigated in the course of experiments with an image database containing photos of paintings by world, European and Ukrainian artists. Algorithm configuration parameters that provide high classification accuracy are found.

Conclusions. The performed experiments have shown the effectiveness of the selected descriptors for the formation of vector descriptions of images of paintings. The greatest accuracy is provided by descriptor merging, which reveals significant differences in the structural properties of images. This approach to the description of objects in combination with the proposed classification algorithm and the chosen main criterion ensures high accuracy of the obtained solutions. The direction of further research may include the use of convolutional neural networks to improve the accuracy of classification under the condition of a static database.

KEYWORDS: intelligent decision-making system, automatic classification, k-nearest neighbors, image descriptors, feature vector, customs examination, paintings.

ABBREVIATIONS

SOM are self-organizing maps;

SVM are support vector machines;

k-NN is a *k*-nearest neighbors method;

CNN is a convolutional neural network;

LBP are local binary patterns;

SIFT is a scale-invariant feature transform descriptor;

RGB LBP are local binary patterns in RGB color space;

Color LBP are local binary patterns found in color channels.

NOMENCLATURE

g is a pixel brightness;

$s(\bullet)$ is a the Heaviside step function;

P is a size of pixel neighborhood;

g_c is a value of brightness of neighborhood central pixel;

g_p is a brightness value of p-th pixel of neighborhood with the size P;

μ is a pixel brightness average value;

σ^2 is a pixel brightness deviation from the average;

μ_n is a n-th order central moment of random pixel brightness distribution;

μ_3 is a 3-th order central moment or asymmetry of random pixel brightness distribution;

μ_4 is a 4-th order central moment of random pixel brightness distribution;

I is an image;

$P(i, j)$ is a contingency matrix;

i, j are pixel coordinates;

C_H is an image contrast;

DOI 10.15588/1607-3274-2022-2-8

$Corr_H$ is an image correlation;

$Entropy_H$ is an image entropy;

$Energy_H$ is an image energy;

M_B is an image palette redundancy;

$H_{\max}$ is a maximum image entropy;

H_{RGB} is an entropy, calculated for individual *R*, *G*, *B* channels;

A_k is an average value of pixel brightness in neighborhood;

E_k is a texture roughness;

C_k is a texture contrast;

α_4 is a kurtosis;

$H_{dir}(a)$ is a quantized edge directions;

D_k is a histogram of quantized edge directions;

n_{peaks} is a histogram peaks number;

a_p is a peak angular direction;

r is a coefficient that depends on quantization levels of angles;

L_k is a linear similarity;

R_k is a texture regularity;

$\sigma_{coarseness}$ is a standard deviation of texture coarseness;

$\sigma_{contrast}$ is a standard deviation of texture contrast;

$\sigma_{directionality}$ is a standard deviation of texture directionality;

$\sigma_{linelikeness}$ is a standard deviation of texture linelikeliness;

$d(x_i, x_j)$ is a measure of similarity between objects, equal to metric distance between data points;

x_i, x_j are objects to be compared;

f_i is an attribute of object matching;

$\mathbf{c}_i$ is an attributes value;

d_E is an Euclidean metric;

d_M is a Minkowski metric.

INTRODUCTION

Painting has long ceased to be art for the elite – reproductions of paintings can be found on items of clothing, bags, in the form of curtains, as graffiti on the walls of buildings. Such popularization undoubtedly leads to the fact that the originals of paintings are constantly growing in value and have long since turned from objects of art into an accumulating value means. This raises many problems for customs services – export of valuable paintings undermines the economic security of the state.

Verifying the authenticity and value of art objects when crossing state borders is an important, urgent and difficult task. The procedure for exporting cultural property during customs control for examination, organization of expertise and other aspects are regulated by the 1970 UNESCO Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property, 1995 UNIDROIT Convention on Stolen or Illegally Exported Cultural Objects, 1954 Europe Cultural Convention [1–3]. In particular, according to the approved procedure, the export of cultural property is possible only if it is confirmed by certificate for the right to export, issued by the Department for the movement of cultural property of the Department of Museum Affairs and Cultural Property under the Ministry of Culture of Ukraine**.** The paintings authenticity takes place during expertise carried out by qualified historians and art critics for a fairly long time. However, it is not uncommon for malefactors to deliberately hide true value of paintings for export, passing them off as much less valuable and therefore do not require a certificate for the right to export. Then the customs service is faced with the need to quickly and accurately assess whether the picture being transported can be classified as a cultural property or not. According to regulatory documents, antiques are items over 100 years old. That is, an operational customs check when exporting paintings abroad is reduced to determining the painting age. The most reliable techniques for this use X-ray fluorescence analysis, infrared and ultraviolet spectroscopy and other methods of analysis. Unfortunately, all of them are now absent in customs arsenal, as well as specialists of corresponding qualifications. At the same time, organizing photographing a picture using a digital camera is a solution that is affordable both in cost and in terms of technical capabilities. An intelligent decision-making system [4, 5] provides painting automatic identification by painting photo and establishing its authenticity and value. Obviously, for operational customs control during the paintings export, it is enough to estimate the painting age and, based on this information, make a decision on export possibility or impossibility.

The object of study is a decision-making process for permission to export paintings during a customs check, which is implemented in an intelligent decision-making system.

The subject of study are methods for automatic classification of paintings images based on a generalized description of their properties with the year of creation as a key attribute.

The main purpose of the work is automatically determining the age of a painting from a digital photograph during classification performed by an intelligent decision-making system.

1 PROBLEM STATEMENT

Suppose given a set of images $X = \{x_1, x_2, ..., x_N\}$, N – number of them. Every image could be described by several characteristics f_i, $i = \overline{1,m}$, their values $\mathbf{c}_i = (c_{f_i 1}, ..., c_{f_i k}, ..., c_{f_i n})$, $k = \overline{1,n}$ are results of some

DOI 10.15588/1607-3274-2022-2-8

image processing descriptors. One image characteristic – a key attribute $Y_l, l = 1,...N$ – is known an advance. According to attribute Y set X is marked by some class labels $Z = \{z_1, z_2, ..., z_M\}$. The same label assigns to images x_i, x_j, that have a distinction $d(x_i, x_j) \le \varepsilon$ where ε – similarity level.

The mathematical problem is to find a classification rule $g(X,Y): X \to Z$, that for a given image x_{N+1} minimizes a distinction measure $d(x_{N+1}, x_j \mid \forall x_j : g(x_{N+1}, y_{N+1}) = g(x_j, y_j),\ j = 1,...N)$.

2 REVIEW OF THE LITERATURE

With development of computing power, which made it possible to process digital photos in high resolution to analyze of high-dimensional data, scientists and engineers began to solve the problem of automatically classifying works of art by photo. These studies use a machine learning approach and are ongoing [6–10]. For pictures automatic classification the most widely used methods of self-organizing maps (SOM) [11], support vector machines (SVM) [12], *k*-nearest neighbors (*k*-NN) [13–16]. Their undoubted advantages are high classification accuracy, ability to fast updating of training datasets, a high learning rate.

Convolutional neural networks (CNN), which have proven their high efficiency in a wide range of tasks related to processing of images of various kinds, are in serious competition for mentioned techniques. CNNs provide higher accuracy compared to other machine learning methods and are fast. Many works demonstrate that application of CNN to automatic classification of picture photos gives positive results [17–21]. However, such networks have a number of disadvantages that limit their use for expeditious customs inspection of paintings for their export possibility. These disadvantages include need for hundreds of thousands or millions of objects for convolutional networks training; training duration, which will increase significantly if it is necessary to update the set of training samples. These factors make convolutional networks computationally and financially costly.

In overwhelming majority of works, automatic classification of paintings is carried out according to several main criteria: by the artist name; by the artistic style or genre to which work can be attributed. This is necessary when identifying and confirming paintings authenticity. Prompt check of painting items value at customs lead to the need to classify paintings by age. There are not so many such works, since dozens of genres can be represented in painting at the same time. Nevertheless, researchers do not abandon this attribute, successfully including it in paintings automatic categorization systems [17].

In this paper, it is proposed to use weighted *k*-nearest neighbors algorithm, which has successfully proven itself in solving complex problems of image classification and allows to successfully updating the dataset for training, and use the picture age as the main classification criterion.

3 MATERIALS AND METHODS

The classification accuracy depends on choice of attributes characterizing objects. When working with images, such algorithms and descriptors as Local binary patterns (LBP) [22] and their color modifications; Haralik's texture features [23]; SIFT descriptor [24] have successfully proven themselves.

Local binary patterns (LBP) [22] – descriptors describing properties of neighborhoods of a given pixel in the image:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p, \quad (1)$$

where $s(\bullet)$ – the Heaviside step function *step* (*x*), which returns 0 if $x \ge 0$, and 1 otherwise.

Based on results of comparing brightness g_c and g_p, a histogram for each pixel is built. These histograms are normalized, compressed and combined into a single data vector. The method turned out to be extremely effective, especially in the tasks of separating object from background. One of its modifications – RGB-LBP – allows color images processing. In this case, local binary pattern is calculated in RGB space for each color component separately, and then descriptions are combined. In general, LBPs can be defined in any color space, such modifications are known as Color LBPs.

In analysis practice, an image is often considered as a random process characterized by a certain law of distribution of pixel brightness g as a random variable. The main parameters describe this random variable are the mathematical expectation, variance, and central moments of brightness distribution.

The mathematical expectation in the case of images with a finite number of pixels P is represented by its approximation – the average value:

$$\mu = \lim_{p \to \infty} \frac{1}{P} \sum_{p=1}^{P} g_p. \quad (2)$$

Dispersion allows estimating degree of pixel brightness possible values deviation from the average:

$$\sigma^2 = \sum_{p=1}^{P} (g_p - \mu)^2. \quad (3)$$

The central moment of the *n*-th order of random variable distribution in the general case for an image can be estimated using the relation:

DOI 10.15588/1607-3274-2022-2-8

$$\mu_n = \sum_{p=1}^{P} \left(g_p - \mu \right)^n . \quad (4)$$

The first central moment is equal to zero, the second is equal to the variance. The third central moment μ_3 is asymmetry. It demonstrates the asymmetry of the probability density function about the mean. The fourth central moment μ_4 characterizes the sharpness of the probability density function top. Thus, combination of these four indicators – pixel brightness average value, variance, third and fourth central moments – exhaustively describes properties of distribution function of pixel brightness for a specific image. Therefore, they are called "first four moments" and are often used in image analysis problems solving.

Haralik's texture features [23] also describe brightness values statistical properties and are calculated based on the contingency matrix:

$$P(i,j) = \frac{\neq \left[(g_1, g_2) \in I \,|\, (g_1 = i) \wedge (g_2 = j) \right]}{\neq I} . \quad (5)$$

where g_1, g_2 – pixels belonging to the image I. Then contrast is found in accordance with the expression

$$C_H(x,y) = \sum_{i,j} (i-j)^2 P(i,j) ; \quad (6)$$

the correlation is calculated as

$$Corr_H(x,y) = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j) P(i,j)}{\sigma_i \sigma_j} , \quad (7)$$

entropy:

$$Entropy_H(x,y) = \sum_i \sum_j P(i,j) \log_2 P(i,j) , \quad (8)$$

energy:

$$Energy_H(x,y) = \sum_i \sum_j P(i,j)^2 . \quad (9)$$

The important information about images is clearly related to color data. To generalize them, such an indicator as palette redundancy is used [25]:

$$M_B = \frac{H_{\max} - H_{RGB}}{H_{\max}} . \quad (10)$$

where $H_{\max}$ is maximum image entropy, which for 8-bit color coding is 8 * 3 = 24; H_{RGB} – entropy, calculated by (8), for individual *R*, *G*, *B* channels.

Each image can be viewed as a texture formed by a collection of some repeating and non-repeating elements. The well-known Tamura features effectively describe the texture properties. They include roughness, contrast, directionality, linearity, roughness, and regularity.

The texture roughness characterizes dimensions of main details that form the image. Its estimate is based on calculation of average values within pixels neighborhood:

$$A_k(x,y) = \sum_P \frac{g(i,j)}{2^{2P}} , \quad (11)$$

where $g(i,j)$ – brightness of pixel with coordinates i, j; P is the size of the neighborhood; the texture roughness is then

$$E_k(x,y) = A_k(x,y) - A_k(x',y), x' \neq x . \quad (12)$$

The texture contrast is estimated based on the fourth moment μ_4 relative to mathematical expectation and variance σ^2 within the neighborhood:

$$C_k(x,y) = \frac{\sigma}{(\alpha_4)^{0.25}} . \quad (13)$$

where $\alpha_4 = \frac{\mu_4}{\sigma^4}$ – kurtosis.

The texture directivity is estimated based on a histogram of quantized edge directions $H_{dir}(a)$:

$$D_k(x,y) = 1 - r n_{peaks} \sum_p \sum_{a \in w_p} \left(a - a_p \right)^2 H_{dir}(a) , \quad (14)$$

where n_{peaks} – histogram peaks number; a_p – peak angular direction; r – coefficient that depends on quantization levels of angles a_p; $a_p = \arctan \frac{\Delta x}{\Delta y}$ are calculated with Prewitt contour detector.

Linear similarities $L_k(x,y)$ are evaluated as average coincidence of edge directions that coincide in pairs of pixels separated by a distance along edge direction in each pixel.

Texture regularity is a generalized feature defined as

$$R_k(x,y) = 1 - r\left(\sigma_{coarseness} + \sigma_{contrast} + \sigma_{directionality} + \sigma_{linelikeness} \right) , \quad (15)$$

where $\sigma_{coarseness}, \sigma_{contrast}, \sigma_{directionality}, \sigma_{linelikeness}$ are standard deviations for each feature.

Roughness summarizes the contrast and roughness of texture as follows:

DOI 10.15588/1607-3274-2022-2-8

$$Roughness_k(x,y) = E_k(x,y) + C_k(x,y). \quad (16)$$

The well-known SIFT descriptor [24] collects information about the statistics of local directions of pixel brightness gradient. It is stable to shifts, rotations and scale transformations. In problems of image classification, these properties of descriptor turn out to be indispensable, since they allow comparing objects regardless of differences in size, orientation and location in image.

Vectors obtained as a result of separate descriptors using are combined into one common vector for describing the object.

The simplest metric classification method determines the similarity between data points using the chosen similarity measure, and based on this information, assigns new data points to one or another existing class. The algorithm refers to supervised learning methods, is distinguished by its implementation simplicity and rather high performance if data attributes number is small, and classification objects number does not exceed 10^3. The Euclidean distance is used as a measure of similarity:

$$d_E(x_i,x_j) = \sqrt{\left(c_{f_i1} - c_{f_j1}\right)^2 + ... + \left(c_{f_in} - c_{f_jn}\right)^2} = \left\|\mathbf{c}_i - \mathbf{c}_j\right\|, \quad (17)$$

where $d(x_i,x_j)$ – measure of similarity between objects, equal to metric distance between data points, x_i, x_j are the objects to be compared, f_i, $i = \overline{1,m}$ are the attributes of object matching, $\mathbf{c}_i = \left(c_{f_i1},...,c_{f_ik},...,c_{f_in}\right), k = \overline{1,n}$ are the attributes values.

Minkowski metric

$$d_M(x_i,x_j) = \sqrt[1/p]{\left(c_{f_i1} - c_{f_j1}\right)^p + ... + \left(c_{f_in} - c_{f_jn}\right)^p} = \left\|\mathbf{c}_i - \mathbf{c}_j\right\|^p \quad (18)$$

also extremely useful in image classification tasks.

In this paper, we consider a system for which the features number is large enough. A weakness of weighted *k*-nearest neighbors method is that when you add up a large number of dissimilarities between data points, the sums can be approximately equal. Because of this, classification objects become poorly distinguishable in selected feature space. To make features more distinguishable, use weights assigned to attributes or data points. The simplest solution is to assign the weight value to reciprocal of distance between the points.

In a multidimensional data space, nearest neighbor search can be performed in different ways also. Known modifications suggest dividing the space by hyperplanes, as in k-d tree algorithm [26]. The modification provides high algorithm performance if number of attributes does not exceed 20.

For problems with a large number of dimensions of data space, so-called BallTree algorithm is used [27]. In this case, space is divided into hyperspheres with centers at data points. The known distance between current data point and the centroid of hypersphere allows defining boundaries of distances to all points within hypersphere. This approach reduces time needed to find the nearest neighbor and is most effective for highly structured data, even with very large space dimensions.

4 EXPERIMENTS

To research the approach effectiveness, a set of images of paintings by 50 famous artists who lived at different times, from 15th century to mid-20th century, was used [28]. The set objects are characterized by such characteristics as artist name, years of life, genres, nationality, biographical facts. For artists who have searched for style in their work, there is information about several genres related to the same period of life. Undoubtedly, each of characteristics of picture description can act as a target attribute in classification. In this work, the attribute of artist's lifetime is chosen as the target feature. The total number of images in dataset was 1169. The number of works for studied artists is shown in Table 1.

5 RESULTS

In the first part of experiment, it was studied the influence of descriptor choice on data classification accuracy. Descriptors LBP (1), Color LBP; the first four moments, calculated by (2)–(4); Haralik parameters calculated by (5)–(10); Tamura texture features, estimated by (11)–(16); SIFT descriptor. Examples of original images processing using selected descriptors are shown in Fig. 1, 2.

Applying descriptors to photos of pictures from a dataset gives feature vectors of different dimensions. For example, color LBP gives a feature vector with dimensions 512×1, SIFT descriptor – a feature vector with dimensions 788×128. To solve the classification problem, all feature vectors must be converted into columns, so the final size of the feature vector for SIFT was 100864×1. A generalized feature vector is formed from such vectors by concatenation. The results of classification using single feature vectors for each descriptor and a generalized vector are presented in Table 2.

Table 1 – The number of paintings images included in studied dataset, depending on artist name

Artist name	Modigliani	Kandinsky	C. Monet	Rivera	Magritte	Dali	Klimt
Number of paintings images by author	193	87	73	70	194	138	117

DOI 10.15588/1607-3274-2022-2-8

6 DISCUSSION

The examples (Fig. 1, 2) show that the proposed descriptors unambiguously and fully reflect the structural, morphological and color paintings properties. LBP, color LBP and the first 4 moments reveal the contours and fine details of images, the Haralik features segment the images according to textural characteristics with high accuracy, and the SIFT descriptor finds the feature points in contours.

During the experiments, the hypothesis that one descriptor is enough to accurately classify the picture according to the creation time based on received information was tested.

This hypothesis is explained by the need to provide high algorithm performance simultaneously with high accuracy. The longer image attributes vector, the lower the classifier speed.

The data in Table 2 shows that the use of only one descriptor provides a low quality classifier: solution accuracy varies from 62% to almost 73%. Combining the data vectors received from several descriptors into one vector made it possible to increase the accuracy of the solution by almost 10%, bringing it to 82.71%. Since during the customs check it is not expected that new images will arrive in the stream at a speed comparable to video, we can conclude that the use of generalized description vectors turned out to be a very effective solution.

In the second part of the experiment, it was necessary to find the settings of the classifier that implements the *k*-nearest neighbors method. We checked such settings as a search tree creating algorithm (k-d tree, Balltree), metric (Euclidean, Minkowski), method for calculating weights (inverse to distance, Fisher score), number of neighboring points when deciding whether to belong to a class and size leaves in the search tree.

For the considered problem, attributes properties were such that the Euclidean metric provided sufficient distinguishability of data points. The highest accuracy of the classifier is provided by settings that are given in Table 3.

The artistic manner of each artist influences the classification result. For several of most famous masters in the third part of the experiment, categorization by age was performed. The results are shown in Table 4.

The listed artists worked in the late 19th and early 20th centuries, but their artistic style varies greatly. Despite the dissimilarity of style, the proposed algorithm carried out the classification by the paintings creation time for these masters with high accuracy – 80–82%.

CONCLUSIONS

The paper considers the problem of automatic paintings classification by age. The authors propose a solution in the form of a classifier, which action is based on a weighted k-nearest neighbors algorithm.

To ensure high accuracy in the work, a set of attributes is proposed, including color, texture, statistical and other characteristics of images. Attribute values are generated from paintings photos in an intelligent decision-making system, which then classifies the painting by age.

To calculate the weights during k-nearest neighbors algorithm implementation, it is proposed to use Fisher score; to calculate the similarity measure, the authors propose to apply the Euclidean metric.

As a dataset for experimental research, it was proposed to use a set that includes works of famous world, European and Ukrainian artists, as well as metadata with artists' biographies, life period, and paintings genres description.

The scientific novelty of the work consists in the formation of a set of descriptors for paintings photos processing, which provides an accurate categorization of paintings by the time of creation.

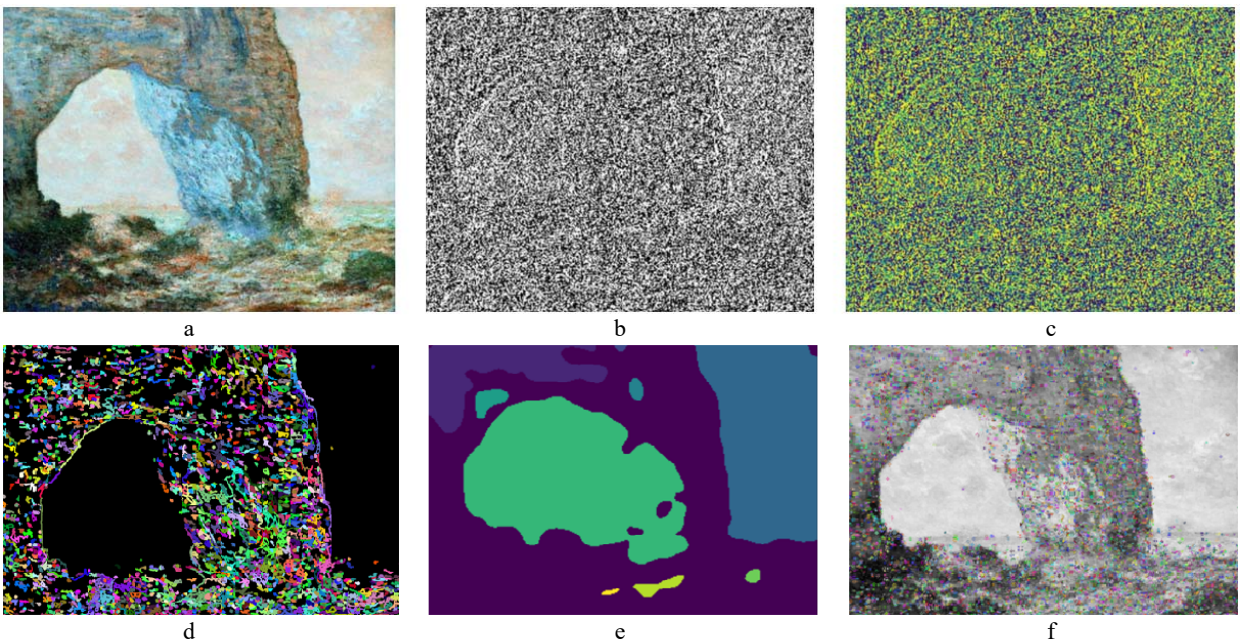

Figure 1 – The results of features vectors calculating:
a – the original image of C. Monet "La Manneport" (1883); image processed using descriptors: b – LBP; c – color LBP; d – the first four moments; e – Haralik texture features; f) SIFT

DOI 10.15588/1607-3274-2022-2-8

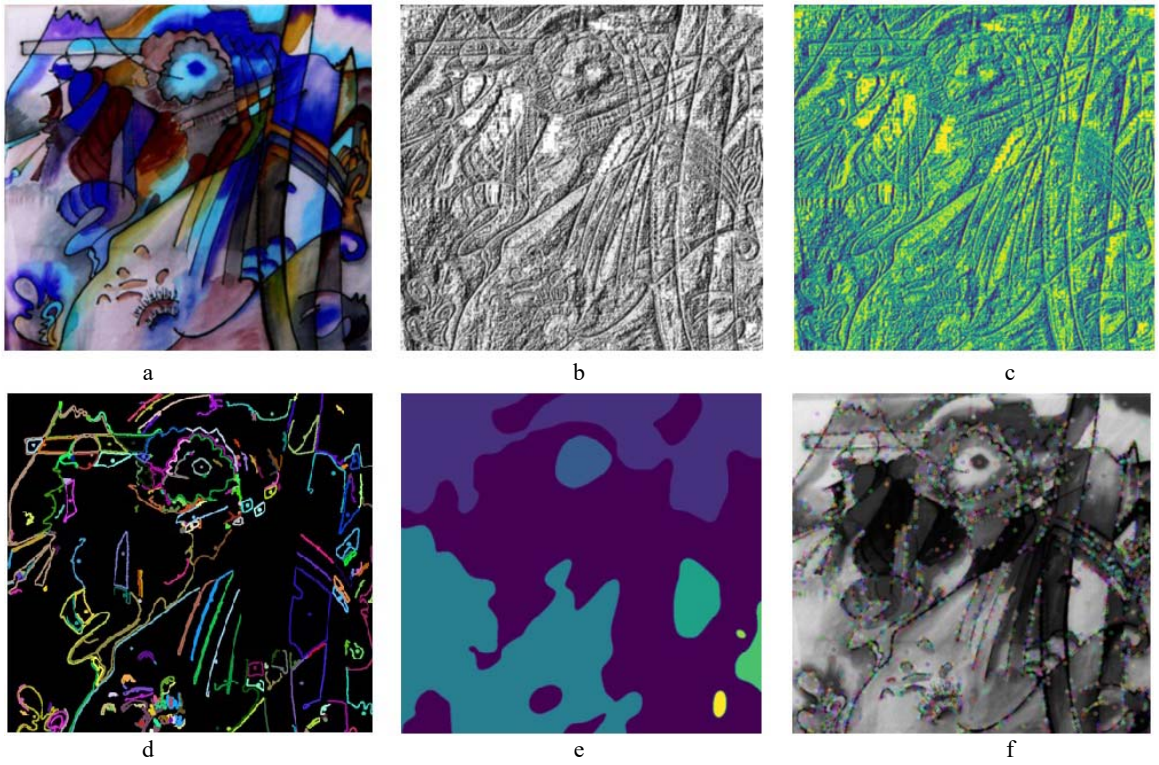

Figure 2 – The results of features vectors calculating:
a – the original image of V. Kandinsky "Composition VIII" (1923); image processed using descriptors: b – LBP; c – color LBP; d – the first four moments; e – Haralik texture features; f – SIFT

Table 2 – Accuracy of picture images classification depending on descriptor used to form the feature vector

Descriptor name	LBP	Color LBP	First 4 moments	Haralik texture features	Palette redundancy	SIFT descriptor	Tamura texture features	Generalized vector
Accuracy	70.92%	66.18%	62.81%	71.80%	63.89%	66.55%	72.92%	82.71%

Table 3 – Configuration parameters of classification algorithm

Algorithm configuration parameter name	Algorithm for constructing a search tree	Metric	Technique for data point weights calculating	Best number of neighboring points in a neighborhood	Best leaf size of a search tree
Parameter value	BallTree	Euclidean	Weight value is reciprocal of distance between the points	11	1

Table 4 – The results of paintings classification by artists by year of creation

Name	Amedeo Modigliani	Vasiliy Kandinskiy	Diego Rivera	Claude Monet	Rene Magritte	Salvador Dali	Gustav Klimt	Kazimir Malevich	Mikhail Vrubel
Accuracy	82.55%	80.48%	80.02%	82.29%	81.62%	81.92%	81.80%	80.71%	82.18%

The practical significance of the results is reducing the time and cost of customs verification of the paintings export legality.

Prospect for further research is related to further improving the accuracy of categorization by modifying the *k*-nearest neighbors algorithm.

ACKNOWLEDGEMENTS

The work was performed at the Department of Applied Mathematics, Kharkiv National University of Radio Electronics and the Department of Software Engineering, Dnipro University of Technology within the framework of scientific research conducted by the departments.

DOI 10.15588/1607-3274-2022-2-8

REFERENCES

1. Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property 1970 [Electronic resource]. Access mode: http://portal.unesco.org/en/ev.php-URL_ID=13039&URL_DO=DO_TOPIC&URL_SECTION=201.html
2. UNIDROIT Convention on Stolen or Illegally Exported Cultural Objects [Electronic resource]. Access mode: https://www.unidroit.org/instruments/cultural-property/1995-convention/
3. European Cultural Convention (ETS No. 018) [Electronic resource]. Access mode: https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=018
4. Martynenko A., Moroz V., Hulina I. An intelligent decision support system for cultural property identification, *Computer-Integrated Technologies: Education, Science, Production*, 2020, Vol. 39, pp. 78–82. URL: http://cit-journal.com.ua/index.php/cit/article/view/126
5. Martynenko A.A., Tevyashev A. D., Kulishova N. Ye., Moroz B. I. System analysis of the problem of establishing the authenticity and authority of painting works, *System Research and Information Technologies,* 2022, No. 1, pp. 1–16.
6. Fiorucci M. Khoroshiltseva M., Pontil M., Traviglia A., Del Bue A., James S. Machine learning for cultural heritage: A survey, *Pattern Recognition Letters*, 2020, Vol. 133, pp. 102–108.
7. Belhi A., Bouras A., Al-Ali A. Kh., Sadka A. H., Eds. Data Analytics for Cultural Heritage: Current Trends and Concepts. – Springer, 2021, 279 p.
8. Shamir L., Macura T., Orlov N., Eckley D. M., Goldberg I. G. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art, *ACM Transactions on Applied Perception (TAP),* 2010, Vol. 7(2), pp. 1–17.
9. Falomir Z., Museros L., Sanz I., Gonzalez-Abril L. Categorizing paintings in art styles based on qualitative color descriptors, quantitative global features and machine learning (qartlearn), *Expert Systems with Applications*, 2018, Vol. 97, pp. 83–94.
10. Sabatelli M., Kestemont M., Daelemans W., Geurts P. Deep transfer learning for art classification problems, *Proceedings of the European Conference on Computer Vision (ECCV),* 2018. https://doi.org/10.1007/978-3-030-11012-3_48.
11. Lee S.-G., Cha E.-Y. Style classification and visualization of art painting's genre using self-organizing maps, *Human-centric Computing and Information Sciences*, December 2016, Vol. 6(1). DOI: 10.1186/s13673-016-0063-4
12. Falomir Z., Museros L., Sanz I., Gonzalez-Abril L. Categorizing Paintings in Art Styles Based on Qualitative Color Descriptors, Quantitative Global Features and Machine Learning (QArt-Learn), *Expert Systems with Applications,* 2018, Vol. 97, pp. 83–94. doi.org/10.1016/j.eswa.2017.11.056
13. Jankovic´ R. Machine Learning Models for Cultural Heritage Image Classification: Comparison Based on Attribute Selection, *Information,* 2020, Vol. 11, 12. DOI: 10.3390/info11010012
14. Ye Z., Su L. The use of data mining and artificial intelligence technology in art colors and graph and images of computer vision under 6G internet of things communication, *International Journal of System Assurance Engineering and Management, The Society for Reliability, Engineering Quality and Operations Management (SREQOM).* India, and Division of Operation and Maintenance, Lulea University of Technology, Sweden, August, 2021. Vol. 12(4), pp. 689–695. DOI: 10.1007/s13198-021-01063-5
15. Wu Y. Application of Improved Boosting Algorithm for Art Image Classification, *Scientific Programming*, 2021, Vol. 2021. doi.org/10.1155/2021/3480414
16. Hamilton M., Fu S., Lu M., Bui J., Bopp D., Chen Zh., Tran F., Wang M., Rogers M., Zhang L., Hoder C., Freeman W. T. MosAIc: Finding Artistic Connections across Culture with Conditional Image Retrieval, *Proceedings of Machine Learning Research,* 2021, Vol. 1, pp. 1–23. arXiv:2007.07177v3
17. Pancaroglu D. Eds.: D.C. Wyld et al. Artist, style and year classification using face recognition and clustering with convolutional neural networks, *COMIT, SIPO, AISCA, MLIOB, BDHI,* 2020, CS & IT, CSCP 2020, pp. 41–54. DOI: 10.5121/csit.2020.101604
18. Banar N., Sabatelli M., Geurts P., Daelemans W., Kestemont M. Transfer Learning with Style Transfer between the Photorealistic and Artistic Domain, *Electronic Imaging, Computer Vision and Image Analysis of Art*, 2021, Vol. 41, pp. 1–9. doi.org/10.2352/ISSN.2470-1173.2021.14.CVAA-041
19. Cetinic E., Lipic T., Grgic S. Fine-tuning Convolutional Neural Networks for fine art classification, *Expert Systems with Applications*, 2018, Vol. 114, pp. 107–118. doi.org/10.1016/j.eswa.2018.07.026
20. Chen T., Yang J. A Novel Multi-Feature Fusion Method in Merging Information of Heterogenous-View Data for Oil Painting Image Feature Extraction and Recognition, *Front. Neurorobot.,* 2021, Vol. 12. doi.org/10.3389/fnbot.2021.709043
21. Zhao W., Zhou D., Qiu X., Jiang W. Compare the performance of the models in art classification, *PLoS ONE*, 2021, Vol. 16(3). doi.org/10.1371/journal.pone.0248414
22. Guo Z., Zhang L., Zhang D. A completed modeling of local binary pattern operator for texture classification, *IEEE Trans. on Image Processing,* 2010, Vol. 19, Issue 6, pp. 1657–1663.
23. Haralick R. M., Shanmugam K., Dinstein I. Textural Features for Image Classification, *IEEE Transactions on Systems, Man, and Cybernetics*, 1973, Vol. 3, pp. 610–621. doi:10.1109/TSMC.1973.4309314
24. Otero I. R., Delbracio M. Anatomy of the SIFT method, *Image Processing On Line*, 2014, Vol. 4, pp. 370–396.
25. Wang D., Chen Y. Calculation and Application of Xin'an Painting School Art Style Model, *ICAITA 2020,* 2020, Vol. 1651, 012033. DOI: 10.1088/1742-6596/1651/1/012033
26. Bentley J. L. Multidimensional binary search trees used for associative searching, *Communications of the ACM,* 1975, Vol. 18 (9), pp. 509–517. DOI: 10.1145/361002.361007. S2CID 13091446.
27. Omohundro S. M. Five balltree construction algorithms, *International Computer Science Institute Technical Report,* 1989.
28. Best Artworks of All Time [Electronic resource]. – Access mode: https://www.kaggle.com/ikarus777/best-artworks-of-all-time/tasks

Received 24.01.2022.
Accepted 17.02.2022.

УДК 303.732.4 + 004.67 + 004.8

АВТОМАТИЧНА КЛАСИФІКАЦІЯ КАРТИН ЗА РОКОМ СТВОРЕННЯ

Мартиненко А. А. – старший викладач кафедри програмного забезпечення комп'ютерних систем Національного технічного університету «Дніпровська політехніка», Дніпро, Україна.

Тевяшев А. Д. – д-р техн. наук, професор, завідувач кафедри прикладної математики Харківського національного університету радіоелектроніки, м. Харків, Україна.

Кулішова Н. Є. – канд. техн. наук, доцент, професор кафедри медіасистем і технологій Харківського національного університету радіоелектроніки, Харків, Україна.

DOI 10.15588/1607-3274-2022-2-8

Мороз Б. І. – д-р техн. наук, професор, член-кореспондент Академії прикладної електроніки, професор кафедри програмного забезпечення комп'ютерних систем Національного технічного університету «Дніпровська політехніка», м. Дніпро, Україна.

Сергієнко О. С. – студентка кафедри прикладної математики Харківського національного університету радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Розглядається завдання автоматичної перевірки легітимності експорту творів живопису.

Мета. Запропоновано метод автоматичного визначення віку картини з цифрової фотографії за допомогою класифікації, яку виконує інтелектуальна система прийняття рішень.

Метод. Пропонується використовувати атрибут року створення картини як головний критерій для прийняття рішення під час митної перевірки легітимності експорту. Замість тривалої та дорогої музейної експертизи застосовується фотографування творів живопису в умовах митниці та обробка фото за допомогою набору дескрипторів. До набору дескрипторів пропонується включити локальні бінарні патерни, їх колірну модифікацію, текстурні ознаки Хараліка, перші чотири моменти, текстурні ознаки Тамури, SIFT дескриптор. Дані, отримані внаслідок дії дескрипторів, утворюють значення кількох десятків окремих атрибутів. Вони формують вектори даних, які потім конкатенуються в узагальнений опис вектора-об'єкта. У просторі ознак, створеному таким чином, виконується автоматична класифікація методом зважених *k*-найближчих сусідів. Пропонований алгоритм розраховує відстань між об'єктами в багатовимірному просторі значень атрибутів, і відносить нові об'єкти до сформованих класів. Критерієм для створення класів є вік картини із існуючої бази даних. Як міру близькості об'єктів пропонується використовувати метрики Евкліда та Мінковського. Розрахунок вагів для алгоритму класифікації запропоновано виконувати методом Фішера.

Результати. Ефективність запропонованого методу була досліджена під час експериментів із базою зображень, що містить фото картин світових, європейських та українських художників. Знайдено параметри конфігурації алгоритму, що забезпечують високу точність класифікації.

Висновки. Проведені експерименти показали ефективність вибраних дескрипторів формування векторів-описів зображень картин. Найбільшу точність забезпечує поєднання дескрипторів, яке виявляє суттєві відмінності у структурних властивостях зображень. Такий підхід до опису об'єктів у поєднанні із запропонованим алгоритмом класифікації та обраним головним критерієм забезпечує високу точність отриманих рішень. Напрямок подальших досліджень може включати використання згорткових нейронних мереж для підвищення точності класифікації за умови статичності бази даних.

КЛЮЧОВІ СЛОВА: інтелектуальна система прийняття рішень, автоматична класифікація, *k*-найближчих сусідів, дескриптори зображень, вектор ознак, митна експертиза, твори живопису.

УДК 303.732.4 + 004.67 + 004.8

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ КАРТИН ПО ГОДУ СОЗДАНИЯ

Мартыненко А. А. – старший преподаватель кафедры программного обеспечения компьютерных систем Национального технического университета «Днепровская политехника», Днепр, Украина.

Тевяшев А. Д. – д-р техн. наук, профессор, заведующий кафедрой прикладной математики Харьковского национального университета радиоэлектроники, Харьков, Украина.

Кулишова Н. Е. – канд. техн. наук, доцент, профессор кафедры медиасистем и технологий Харьковского национального университета радиоэлектроники, Харьков, Украина.

Мороз Б. И. – д-р техн. наук, профессор, член-корреспондент Академии прикладной электроники, профессор кафедры программного обеспечения компьютерных систем Национального технического университета «Днепровская политехника», Днепр, Украина.

Сергиенко А. С. – студентка кафедры прикладной математики Харьковского национального университета радиоэлектроники, Харьков, Украина.

АННОТАЦИЯ

Актуальность. Рассматривается задача автоматической проверки легитимности экспорта произведений живописи.

Цель. Предложен метод автоматического определения возраста картины по цифровой фотографии с помощью классификации, которую выполняет интеллектуальная система принятия решений.

Метод. Предлагается использовать атрибут года создания картины в качестве главного критерия для принятия решения в ходе таможенной проверки легитимности экспорта. Вместо длительной и дорогостоящей музейной экспертизы применяется фотографирование произведений живописи в условиях таможни и обработка фото с помощью набора дескрипторов. В набор дескрипторов предлагается включить локальные бинарные паттерны, их цветовую модификацию, текстурные признаки Харалика, первые четыре момента, текстурные признаки Тамуры, SIFT дескриптор. Данные, полученные в результате действия дескрипторов, образуют значения нескольких десятков частных атрибутов. Они формируют векторы данных, которые затем конкатенируются в обобщенный вектор-описание объекта. В пространстве признаков, созданном таким образом, выполняется автоматическая классификация методом взвешенных *k*-ближайших соседей. Предлагаемый алгоритм рассчитывает расстояние между объектами в многомерном пространстве значений атрибутов, и относит новые объекты к уже сформированным классам. Критерием для создания классов является возраст картины из существующей базы данных. В качестве меры близости объектов предлагается использовать метрики Евклида и Минковского. Расчет весов для алгоритма классификации предложено выполнять методом Фишера.

Результаты. Эффективность предложенного метода была исследована в ходе экспериментов с базой изображений, содержащей фото картин мировых, европейских и украинских художников. Найдены параметры конфигурации алгоритма, которые обеспечивают высокую точность классификации.

DOI 10.15588/1607-3274-2022-2-8

Выводы. Проведенные эксперименты показали эффективность выбранных дескрипторов для формирования векторов-описаний изображений картин. Наибольшую точность обеспечивает объединение дескрипторов, которое обнаруживает существенные различия в структурных свойствах изображений. Такой подход к описанию объектов в сочетании с предложенным алгоритмом классификации и выбранным главным критерием обеспечивает высокую точность полученных решений. Направление дальнейших исследований может включать использование сверточных нейронных сетей для повышения точности классификации при условии статичности базы данных.

КЛЮЧЕВЫЕ СЛОВА: интеллектуальная система принятия решений, автоматическая классификация, *k*-ближайших соседей, дескрипторы изображений, вектор признаков, таможенная экспертиза, произведения живописи.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Convention on the Means of Prohibiting and Preventing the Illicit Import, Export and Transfer of Ownership of Cultural Property 1970 [Electronic resource]. – Access mode: http://portal.unesco.org/en/ev.php-URL_ID=13039&URL_DO=DO_TOPIC&URL_SECTION=201.html
2. UNIDROIT Convention on Stolen or Illegally Exported Cultural Objects [Electronic resource]. – Access mode: https://www.unidroit.org/instruments/cultural-property/1995-convention/
3. European Cultural Convention (ETS No. 018) [Electronic resource]. – Access mode: https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=018
4. Martynenko A. An intelligent decision support system for cultural property identification / A. Martynenko, V. Moroz, I. Hulina // Computer-Integrated Technologies: Education, Science, Production. – 2020. – Vol. 39. – P. 78–82. URL: http://cit-journal.com.ua/index.php/cit/article/view/126
5. System analysis of the problem of establishing the authenticity and authority of painting works / [A. A. Martynenko, A. D. Tevyashev, N. Ye. Kulishova, B. I. Moroz] // System Research and Information Technologies. – 2022. – № 1. – P. 1–16.
6. Fiorucci M. Machine learning for cultural heritage: A survey / [M. Fiorucci, M. Khoroshiltseva, M. Pontil et al.] // Pattern Recognition Letters. – 2020. – Vol. 133. – P. 102–108.
7. Data Analytics for Cultural Heritage: Current Trends and Concepts / Eds.: A. Belhi, A. Bouras, A. Kh. Al-Ali, A. H. Sadka. – Springer, 2021. – 279 p.
8. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art / [L. Shamir, T. Macura, N. Orlov et al.] // ACM Transactions on Applied Perception (TAP). – 2010. – Vol. 7(2). – P. 1–17.
9. Categorizing paintings in art styles based on qualitative color descriptors, quantitative global features and machine learning (qartlearn) / [Z. Falomir, L. Museros, I. Sanz, L. Gonzalez-Abril] // Expert Systems with Applications. – 2018. – Vol. 97. – P. 83–94.
10. Deep transfer learning for art classification problems / [M. Sabatelli, M. Kestemont, W. Daelemans, P. Geurts] // Proceedings of the European Conference on Computer Vision (ECCV), 2018. https://doi.org/10.1007/978-3-030-11012-3_48.
11. Lee S.-G. Style classification and visualization of art painting's genre using self-organizing maps / S.-G. Lee, E.-Y. Cha // Human-centric Computing and Information Sciences, December. – 2016. – Vol. 6(1). DOI: 10.1186/s13673-016-0063-4
12. Categorizing Paintings in Art Styles Based on Qualitative Color Descriptors, Quantitative Global Features and Machine Learning (QArt-Learn) / [Z. Falomir, L. Museros, I. Sanz, L. Gonzalez-Abril] // Expert Systems with Applications. – 2018. – Vol. 97. – P. 83–94. doi.org/10.1016/j.eswa.2017.11.056
13. Jankovic´ R. Machine Learning Models for Cultural Heritage Image Classification: Comparison Based on Attribute Selection / R. Jankovic´ // Information. – 2020. – Vol. 11, 12. DOI: 10.3390/info11010012
14. Ye Z. The use of data mining and artificial intelligence technology in art colors and graph and images of computer vision under 6G internet of things communication / Z. Ye, L. Su // International Journal of System Assurance Engineering and Management, The Society for Reliability, Engineering Quality and Operations Management (SREQOM), India, and Division of Operation and Maintenance, Lulea University of Technology, Sweden, August. – 2021. – Vol. 12(4). – P. 689–695. DOI: 10.1007/s13198-021-01063-5
15. Wu Y. Application of Improved Boosting Algorithm for Art Image Classification/ Y. Wu // Scientific Programming, 2021. – Vol. 2021. doi.org/10.1155/2021/3480414
16. Hamilton M. MosAIc: Finding Artistic Connections across Culture with Conditional Image Retrieval / [M. Hamilton, S. Fu, M. Lu et al.] // Proceedings of Machine Learning Research. – 2021. – Vol. 1. – P. 1–23. arXiv:2007.07177v3
17. Pancaroglu D. Artist, style and year classification using face recognition and clustering with convolutional neural networks / D. Pancaroglu // Eds.: D. C. Wyld et al. COMIT, SIPO, AISCA, MLIOB, BDHI – 2020, CS & IT – CSCP 2020. – P. 41–54. DOI: 10.5121/csit.2020.101604
18. Transfer Learning with Style Transfer between the Photorealistic and Artistic Domain / [N. Banar, M. Sabatelli, P. Geurts et al] // Electronic Imaging, Computer Vision and Image Analysis of Art. – 2021. – Vol. 41. – P. 1–9. doi.org/10.2352/ISSN.2470-1173.2021.14.CVAA-041
19. Cetinic E. Fine-tuning Convolutional Neural Networks for fine art classification / E. Cetinic, T. Lipic, S. Grgic // Expert Systems with Applications. – 2018. – Vol. 114. – P. 107–118. doi.org/10.1016/j.eswa.2018.07.026
20. Chen T. A Novel Multi-Feature Fusion Method in Merging Information of Heterogenous-View Data for Oil Painting Image Feature Extraction and Recognition / T. Chen, J. Yang // Front. Neurorobot. – 2021. – Vol. 12. doi.org/10.3389/fnbot.2021.709043
21. Compare the performance of the models in art classification / [W. Zhao, D. Zhou, X. Qiu, W. Jiang] // PLoS ONE. – 2021. – Vol. 16(3). doi.org/10.1371/journal.pone.0248414
22. Guo Z. A completed modeling of local binary pattern operator for texture classification / Z. Guo, L. Zhang, D. Zhang // IEEE Trans. on Image Processing. – 2010. – Vol. 19, Issue 6. – P. 1657–1663.
23. Haralick, R. M. Textural Features for Image Classification / R. M. Haralick, K. Shanmugam, I. Dinstein // IEEE Transactions on Systems, Man, and Cybernetics. – 1973. – Vol. 3. – P. 610–621. DOI: 10.1109/TSMC.1973.4309314
24. Otero I. R. Anatomy of the SIFT method / I. R. Otero, M. Delbracio // Image Processing On Line. – 2014. – Vol. 4. – P. 370–396.
25. Wang D. Calculation and Application of Xin'an Painting School Art Style Model / D. Wang, Y. Chen // ICAITA 2020. – 2020. – Vol. 1651. – 012033. DOI: 10.1088/1742-6596/1651/1/012033
26. Bentley J. L. Multidimensional binary search trees used for associative searching / J. L. Bentley // Communications of the ACM. – 1975. – Vol. 18 (9). – P. 509–517. doi:10.1145/361002.361007. S2CID 13091446.
27. Omohundro S. M. Five balltree construction algorithms / S. M. Omohundro // International Computer Science Institute Technical Report, 1989.
28. Best Artworks of All Time [Electronic resource]. – Access mode: https://www.kaggle.com/ikarus777/best-artworks-of-all-time/tasks

DOI 10.15588/1607-3274-2022-2-8

e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2022. № 2
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2022. № 2

UDC 004.35

PROTOTYPING SMART HOME FOR IMMOBILIZED PEOPLE: EEG/MQTT-BASED BRAIN-TO-THING COMMUNICATION

Zubov D. A. – Dr. Sc., Associate Professor of the Department of Computer Science, University of Central Asia, Bishkek, Kyrgyzstan.

Qureshi M. S. – Dr. Sc., Assistant Professor of the Department of Computer Science, University of Central Asia, Bishkek, Kyrgyzstan.

Köse U. – Dr. Sc., Associate Professor the Department of Computer Engineering, Süleyman Demirel University, Isparta, Turkey.

Kupin A. I. – Dr. Sc., Professor, Head of the Department of Computer Systems and Networks, Kryvyi Rih National University, Kryvyi Rih, Ukraine.

ABSTRACT

Context. Immobilized people face additional barriers in almost all areas of life, including simple operations like turning the light on/off and controlling the air conditioner. The object of the study was to develop the brain-to-thin communication of affordable price to control the smart home appliances by immobilized people from neck to toes.

Objective. The goal of the work is to manage smart home appliances via brain-to-thing communication with EEG non-invasive electrodes, edge IoT devices, and MQTT protocol if the brain and eye control of the disabled work normally.

Method. A non-invasive Sichiray TGAM brainwave EEG sensor kit captures signals and then transmit them via Bluetooth to the HC-05 module connected to the Arduino Mega microcontroller. Information about edge IoT devices is presented to the disabled on the LCD 1602 display wired to the same Arduino Mega. The disabled person chooses the option shown on display via the double blink that is detected if the quality of signal equals zero and low/mid gamma waves are less than ten in three consecutive Bluetooth packets. Control commands are sent from Arduino Mega (MQTT publisher) to the edge IoT devices (MQTT subscribers) that analyze them and start a specific operation like opening a door and turning the alarm on/off.

Results. Five females and five males of different ages from 8 to 59 years old examined the control of smart home appliances with the Sichiray TGAM brainwave sensor kit. Everyone successfully handled the Sichiray headset and showed satisfaction with the brain-to-thing system.

Conclusions. In this work, a smart home concept for immobilized people was developed using the brain-to-thing approach and the MQTT communication between the MQTT publisher, Sichiray TGAM brainwave EEG sensor kit connected via Bluetooth to the Arduino Mega microcontroller, and edge IoT devices total priced at USD 150. The most likely prospect of the presented work is to produce the sample that is ready to market.

KEYWORDS: brain-to-thing, immobilized people, EEG sensor, MQTT.

ABBREVIATIONS

AC is an air conditioner;
AI is artificial intelligence;
AMQP is an advanced message queuing protocol;
ASIC is an application specific integrated circuit;
BCI is a brain-computer interface;
BTC is a brain-to-thing communication;
CoAP is a constrained application protocol;
DDS is a data distribution service;
EEG is an electroencephalogram;
IIC is an inter-integrated circuit;
IoT is an Internet of Things;
LCD is a liquid-crystal display;
MAC is media access control;
MDP is a Markov decision process;
MQTT is a message queueing telemetry transport;
TGAM is ThinkGear ASIC module;
WBCI is a wireless brain-computer interface;
XMPP is an extensible messaging and presence protocol.

NOMENCLATURE

γ_L is a value of low gamma wave;
γ_M is a value of mid gamma wave;
$\max_{\gamma L}$ is a maximum value of low gamma waves;
$\max_{\gamma M}$ is a maximum value of mid gamma waves;
N is the number of IoT devices;
t is a current period of time.

INTRODUCTION

Disabled people without physical movement face additional barriers in almost all areas of life, including simple operations like turning the light on/off and controlling the AC. During the COVID-19 pandemic, they are disproportionately affected [1, 2] since regular transportation of medical and other support staff has got broken in some areas. For instance, public transport is suspended, and passengers may disembark from regional buses and trains but may not board in red zones under the Ukrainian adaptive quarantine regime [3]. Because of that and other factors, people without physical movement spend some time alone. The problem is getting more complicated for the disabled who are physically immobilized from neck to toes.

This paper presents a thought-managed smart home system based on BTC [4, 5] for people physically immobilized from neck to toes whose brain and eye control work normally. A non-invasive WBCI Sichiray TGAM brainwave EEG sensor kit is proposed to capture EEG signals. The EEG signals are transmitted via Bluetooth to the HC-05 module connected to the Arduino Mega board. The information about edge IoT devices [4–19] is pre-

DOI 10.15588/1607-3274-2022-2-9

sented to the disabled on the LCD 1602 display with an IIC adapter wired to the same Arduino Mega microcontroller. The double blink is used to detect an event where the disabled person chooses the option shown on display. Control commands are sent from Arduino Mega (MQTT publisher) to the edge IoT devices (MQTT subscribers) that analyze them and start a specific operation like opening a door and turning the light on/off. The operation's confirmation is shown on the LCD 1602 display. To support the personalized usage of the BTC system with an AI touch, a combination of MDP [20] and Q-Learning (Q–L) [21] is proposed to identify the personalized suggestion flow directly. The BTC system has been examined by five males and five females of different ages between 8 and 59 years. Everyone successfully handled the BCI headset and showed satisfaction with the BTC system.

The object of study is the brain-to-thin communication that manages smart home appliances by immobilized people.

Three types of BCIs, active, reactive, and passive, describe possible communication pathways between the human brain and external devices [14]. Over twenty companies produce EEG devices that can derive the brain signals in several non-overlapping frequency bands, e.g., Alpha, Beta, Delta, Gamma, and Theta. The Sichiray TGAM brainwave EEG sensor kit with a separate Bluetooth board [15, 16] has an advantage over alternatives because of its low price of USD 50.

The subject of study is the non-invasive Sichiray TGAM brainwave EEG sensor kit that captures signals transmitted via Bluetooth to the HC-05 module connected to the Arduino Mega board. In this soft-/hardware complex, the information about edge IoT devices is presented to the disabled person on the LCD 1602 display wired to the same Arduino Mega microcontroller.

Analysis of the concepts and soft-/hardware solutions presented in [4–16] shows that none of them implement WBCI to control smart home appliances using IoT data protocol(s) along with budget EEG-based brain kit(s). The perspective approach is the WBCI based on the low-cost brainwave EEG sensor such as Sichiray TGAM and lightweight publish/subscribe IoT data protocol such as MQTT.

The purpose of the work is to develop the brain-to-thin communication of affordable price to control the smart home appliances by immobilized people.

1 PROBLEM STATEMENT

Over a billion people live with some form of disability around the world [1] and this number is dramatically increasing nowadays because of the COVID-19 and other pandemics. Military conflicts and world economic stagnation reduce the financial support of social activities, including assistive devices and support staff for immobilized people. None of the existing devices is of the price at USD 150 which is the aim of this work.

2 REVIEW OF THE LITERATURE

Many papers released were related to the case when the brain and eye control of immobilized people work normally and can be employed to manage the smart home. The BCI with EEG non-invasive electrodes and a headset installation is broadly applied to manage smart home appliances using IoT techniques [4–16]. It is also important to personalize smart home interactions to help disabled people improve their daily life routines through solutions based on AI. For instance, the menu options can be shown in a different order in accord with previous selections, i.e., rates/ranks of the options, of the disabled. Here, a good solution may be giving suggestions to save time while interacting with the smart home environment.

Edge IoT devices exchange data via IoT protocols such as MQTT, CoAP, XMPP, DDS, AMQP [17] using WiFi, Bluetooth, and Ethernet networking technologies. On a short distance of up to 10 m, most IoT hardware employs Bluetooth classes 2 or 3 to be connected directly [18]. WiFi and Ethernet connections are mainly applied for longer distances. MQTT IoT software can be executed on thin clients like Arduino Uno since it takes approximately 10 KB of random-access memory. MQTT brokers reliably work with 100,000 publishers and 100 subscribers [19] that satisfies the smart and intelligent home network requirements.

3 MATERIALS AND METHODS

In this research work, smart home appliances are wirelessly controlled via the NodeMCU ESP8266 ESP-12 modules [22] with 3.3 V relays [23] (5 V relays can be used as well), MQTT subscribers, that receive control commands through the MQTT IoT data protocol from Arduino Mega microcontroller, the MQTT publisher. Ethernet shield connects Arduino Mega to the smart home network, as well as the MQTT broker is started on the Lenovo G510 laptop with Windows 10 in this project. LCD 1602 display presents information about edge IoT devices (e.g., light, door, and AC), it is wired to the Arduino Mega board. Sichiray TGAM brainwave EEG sensor kit captures EEG signals and then transmits them to the HC-05 Bluetooth module [16] connected to the Arduino Mega. C programs were developed in Arduino IDE for Arduino Mega and NodeMCU ESP8266 ESP-12 boards. The various parts involved in the design of the BTC system for controlling smart home devices are shown in Fig. 1.

During the EEG recording, the eye blinking can generate bioelectric signals with the amplitude ten times larger/smaller than some previous values [24]. Hence, double blinking is used to select an option shown on the LCD 1602 display in this project. It is detected if the following dependencies appear in three Bluetooth packets [16] in a row:

1. Quality of signal equals 0.
2. $Y_L[t] < \max_{Y_L}$, $Y_M[t] < \max_{Y_M}$, $\max_{Y_L} = 10$, $\max_{Y_M} = 10$. Maximum values 10 for $\max_{\gamma L}$ and $\max_{\gamma M}$ are proposed for the experienced users in this project. Gamma waves are applied to detect the double blink event in this project since they are the only type of brain waves that affect the entire brain [25, 26].

DOI 10.15588/1607-3274-2022-2-9

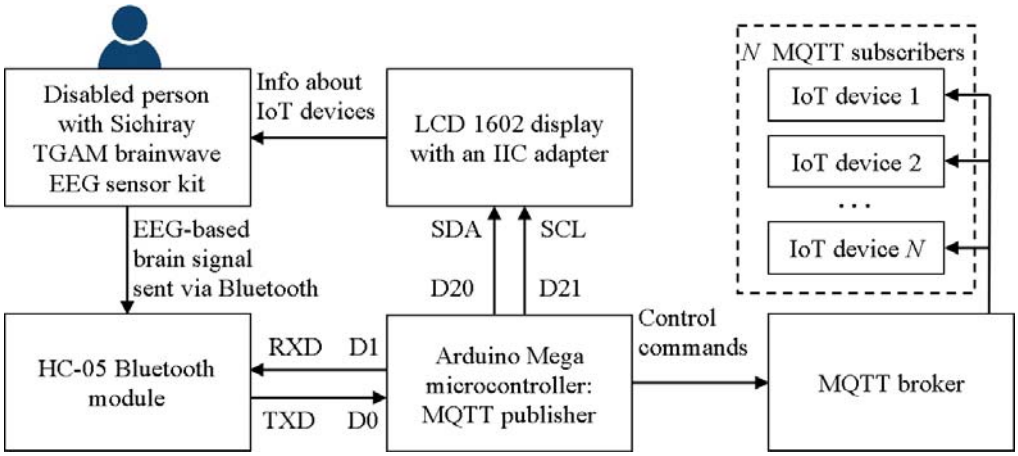

Figure 1 – Block diagram of the brain-controlled smart home with Sichiray TGAM brainwave EEG sensor kit and MQTT IoT data protocol

Then, three Bluetooth packets are skipped since two above-stated dependencies may be repeated, and hence the duplication of the double blink can be detected again that is wrong.

Boards with static IP addresses are connected via the router (two routers, TP-Link TL-WR940N and TL-MR3020, were successfully tested). Additional MAC filtering can optionally be applied to improve the security of the BTC system.

Like a similar WBCI system [6], the proposed BTC project employs a single dry EEG sensor to prevent the inconvenience of wearing many EEG electrodes, making the device completely wearable. The low power consumption with 3 V and 5 V supplies and recharging ability adds to its flexibility. Therefore, the proposed BTC is a comprehensive system in terms of functionality and design which provides independence and ease of its use for disabled people without physical movement.

For the long-term exploitation, the BTC system is proposed to employ MDP and Q-L techniques to run reinforcement learning for detecting optimum rewards of actions so it will personalize the order of menu options in future use cases.

4 EXPERIMENTS

The outcomes of the proposed BTC system for controlling a smart home for immobilized people are achieved using two parts, headset and edge wireless IoT devices.

In the headset part, the information on IoT devices is displayed and brain signals are sent to the HC-05 Bluetooth module connected to the Arduino Mega (MQTT publisher).

Sichiray TGAM brainwave EEG sensor kit and an example of its installation on the head are shown in Fig. 2. It safely measures and transfers signals such as human attention and meditation, Alpha, Beta, Delta, Gamma, and Theta waves via the Bluetooth slave module. The kit consists of an EEG electrode, two ear clips with ground electrodes, TGAM and Bluetooth modules, and a 2xAAA battery holder with a switch. An additional bandage is needed to rest the EEG electrode on the forehead above the eyes. The baud rate of 57600 bits/sec is used in this project [16].

The receiving part consists of the Arduino Mega board wired with the HC-05 Bluetooth master module (see Fig. 3), where HC-05 TXD and RXD pins are wired with Arduino Mega RX(D0) and TX(D1) pins, respectively. These pins must be disconnected during the Arduino sketch uploading into the Arduino Mega board. In this project, the logic level converter [16] is not used.

Arduino Brain Library is the most known way to parse data from Neurosky-based EEG headsets [27] nowadays. The standard test, the Arduino sketch BrainSerialTest.ino, shows only errors in the data received by the HC-05 module [16]; hence, the solution is the analysis of the Bluetooth packets sent by the Sichiray Bluetooth module.

Information about IoT devices is presented to the disabled on the LCD 1602 display with an IIC adapter wired via the SDA and SCL pins to the Arduino Mega microcontroller (D20 and D21 pins, respectively). In this project, LCD 1602 module shows the IoT device's name and state in the first line and the brain signals in the second one, a maximum of 16 characters in a line. The disabled person must double blink to choose a specific option. Options are changed every 10 sec. Examples are as follows:

– "Alarm": the option "Alarm" is shown to the disabled person;

– "Alarm ON": the option "Alarm ON" is shown to the disabled person;

– "Alarm OFF": the option "Alarm OFF" is shown to the disabled person;

– "Alarm ON +++++++": this text is the confirmation of the option "Alarm ON" selected by the disabled person;

– "Alarm OFF ------": this text is the confirmation of the option "Alarm OFF" selected by the disabled person.

DOI 10.15588/1607-3274-2022-2-9

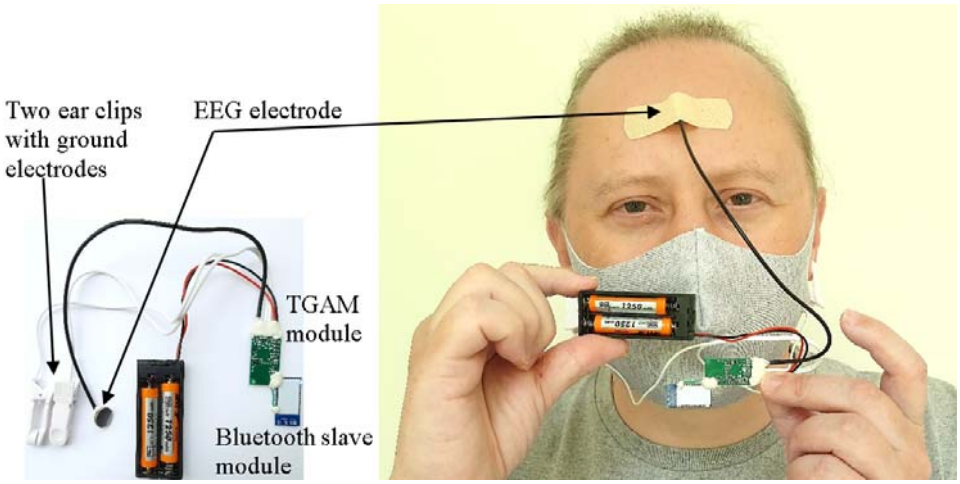

Figure 2 – Sichiray TGAM brainwave EEG sensor kit (on the left) and an example of its installation on the head (on the right)

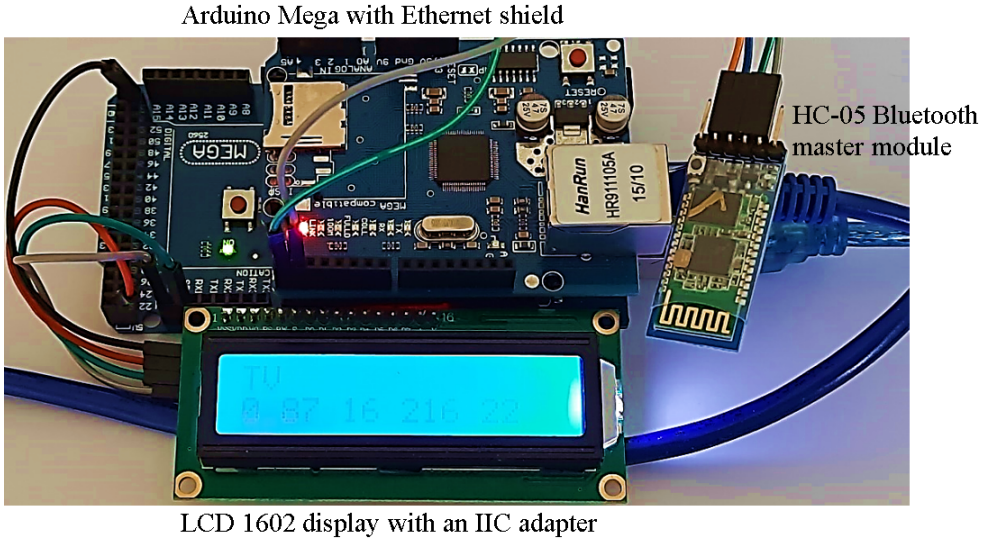

Figure 3 – Arduino Mega microcontroller with HC-05 Bluetooth module, LCD 1602 display, and Ethernet shield

The quality of signal and ten brain signals sent by the Sichiray TGAM brainwave EEG sensor kit are as follows: attention, meditation, delta wave, theta wave, low alpha wave, high alpha wave, low beta wave, high beta wave, low gamma wave, mid gamma wave. However, quality of the signal, attention, meditation, low gamma wave, and mid gamma wave are only shown in the second line of the LCD 1602 display since the first parameter is a selector that excludes Bluetooth packets with poor quality of signal, second and third parameters are the integrative ones because they are calculated using different brain waves, last two parameters are used to detect the double blink event in this project.

To manage the edge IoT devices via the MQTT IoT data protocol, the following control commands are applied:

– Alarm (static IP address 192.168.0.105): ON (control command “101”), OFF (control command “102”);

– Light (static IP address 192.168.0.106): ON (control command “201”), OFF (control command ”202”);

– AC (static IP address 192.168.0.102): ON (control command “301”), OFF (control command “302”);

– TV (television; static IP address 192.168.0.103): ON (control command “401”), OFF (control command “402”);

– Door (static IP address 192.168.0.104): Open (control command “501”), Close (control command “502”).

In this project, the MQTT broker and the MQTT publisher are assigned static IP addresses 192.168.0.100 and 192.168.0.101, respectively.

A flow chart of the flow of events in the BTC smart home for disabled people without physical movement is shown in Fig. 4.

DOI 10.15588/1607-3274-2022-2-9

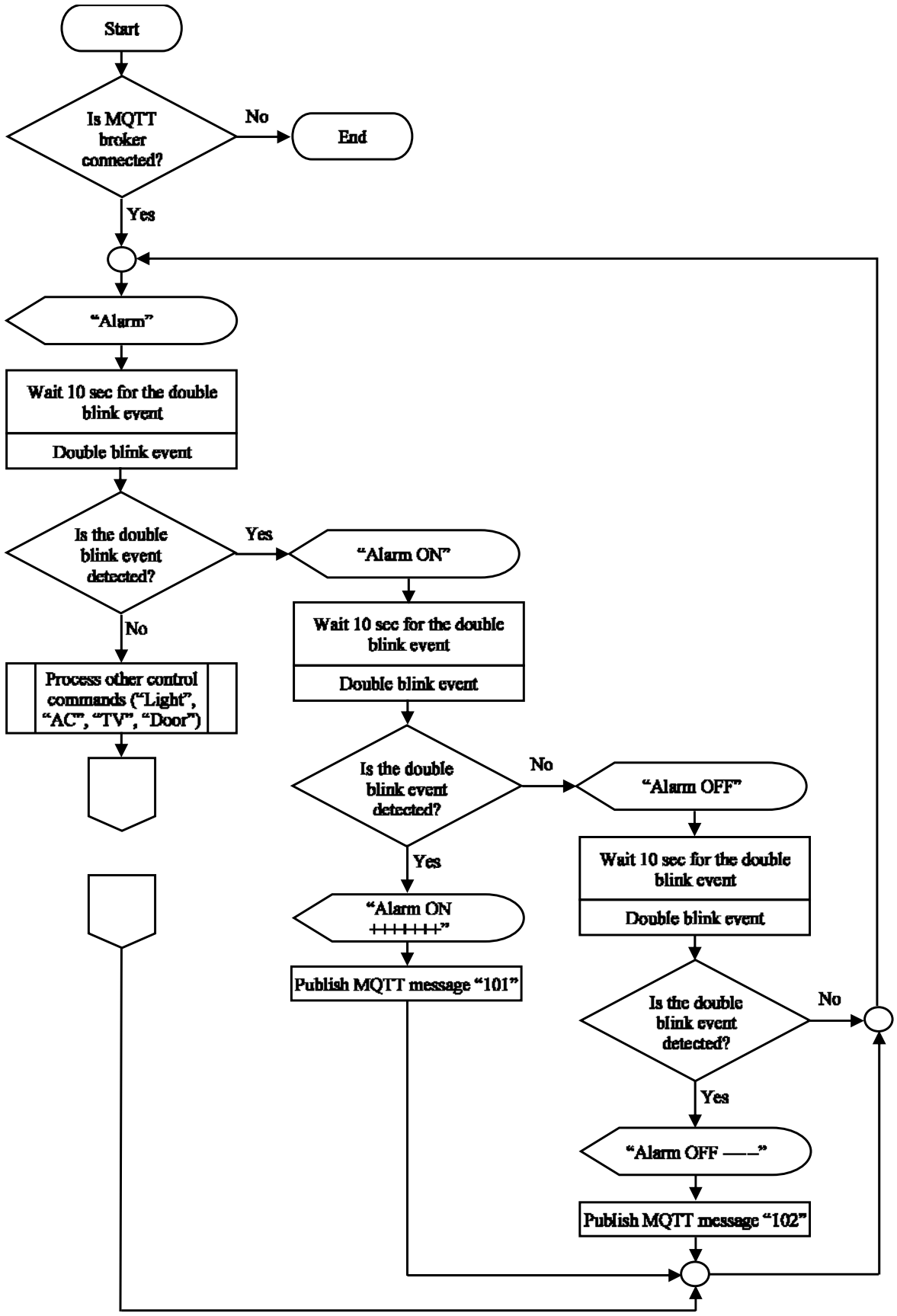

Figure 4 – A flow chart of event detection and control in the BTC smart home for disabled people without physical movement

DOI 10.15588/1607-3274-2022-2-9

The Arduino sketch is based on the code [16] but the parts related to the Ethernet connection, MQTT IoT data protocol, LCD 1602 display, and detection of the double blink event were added.

In the part on the edge wireless IoT devices, the NodeMCU ESP8266 ESP-12 boards (MQTT subscribers) with relays implement control commands such as turning the light on/off and controlling the air conditioner.

Edge wireless IoT devices, MQTT subscribers, are controlled via the relays wired to the NodeMCU ESP8266 ESP-12 boards (see Fig. 5). Arduino sketch includes two libraries, ESP8266WiFi.h and PubSubClient.h, to program NodeMCU ESP8266 ESP-12 board and receive control commands via the MQTT IoT data protocol, respectively. Arduino relays [23] are controlled via the general-purpose input/output pin 16 (D0); they can work with the 10A–250V electrical equipment that satisfies requirements to the IoT devices in a smart home.

Eclipse Mosquitto version 1.4.11 [28] was installed on the Lenovo G510 laptop with Windows 10 in this project. This open-source software includes the MQTT broker mosquitto.exe, the MQTT publisher mosquitto_pub.exe, and the MQTT subscriber mosquitto_sub.exe. The administrator of the system can receive the control commands assigned to a specific topic, "/Blink" in this project (see Fig. 6; "–q" specifies the quality of service to use for the message from 0 to 2). In this project, quality of service level 1 is applied to deliver a message at least once to the receiver. The sender stores the message until it gets an acknowledgment packet from the receiver. MQTT passwords can be applied to the system improving its security if necessary.

Table 1 illustrates how the MQTT client publishes/subscribes messages under the topic "/Blink" and controls the relay connected to the AC with control commands "301" and "302". In Arduino IDE, NodeMCU 0.9 (ESP-12 Module) was selected to program the NodeMCU ESP8266 ESP-12 board. The cost of the custom hardware is about USD 150 for the following components: Sichiray TGAM brainwave EEG sensor kit, Arduino Mega, Ethernet shield, HC-05 Bluetooth module, LCD 1602 display, five relays, and NodeMCU ESP8266 ESP-12 boards.

5 RESULTS

Ten people of different ages and gender took part in the experiment on the control of smart home appliances with the Sichiray TGAM brainwave EEG sensor kit (their preferred values $\max_{\gamma L}$ and $\max_{\gamma M}$ are shown in parentheses): five female users of 8 (30, 70), 28 (30, 70), 43 (30, 70), 46 (10, 10), and 57 (20, 20) years old; five male users of 10 (30, 70), 37 (20, 20), 42 (20, 20), 47 (20, 20), and 59 (20, 20) years old. The male software developer has preferred values (10, 10). Everyone successfully handled the Sichiray TGAM brainwave EEG sensor kit and showed satisfaction with the BTC system.

6 DISCUSSION

New users noted that it is easy to control smart home appliances if low and mid gamma waves are greater than 10. However, false detections of the double blink event appear in this case. The solution is to adjust $\max_{\gamma L}$ and $\max_{\gamma M}$ in accordance with the distinctive features of the user(s). For some new users, frequent eye blinking is a way to replace the simple double blink.

Preparation for the experiment includes the acquaintance with the mobile application "Brainwave Visualizer" [29] (see Fig. 7) that changes the on-screen shapes morph and color depending on the state of mind. Also, two parameters, attention and meditation, are shown on the smartphone screen.

Three timeslots are taken into consideration in the statistical analysis:

1. Timeslot that is needed to teach new users how to use the Sichiray TGAM brainwave EEG sensor kit and the mobile application "Brainwave Visualizer" (about five minutes).

2. Timeslot that is needed to present the BTC system for the disabled people without physical movement by the developer/instructor (about ten minutes).

3. Timeslot that is needed to study how to use the BTC system for the disabled people without physical movement by the new users (about five minutes).

New users admitted that sometimes it is much more convenient to select a specific option using the frequent eye blinking and slow eye blinking is working properly to avoid the double blinking event, as well as the explanation of the control commands helps to understand the BTC system. Also, they recommended replacing the small LCD 1602 display with the larger one.

The main objective of performing statistical analysis was to study that everyone could use this system and determine the stability and friendliness of the system. All participants admitted the stability and ease of using the proposed BTC system.

CONCLUSIONS

The urgent problem of the smart home affordable implementation for immobilized people was solved using the BTC EEG-based approach and the MQTT communication between the MQTT publisher, Sichiray TGAM brainwave EEG sensor kit connected via Bluetooth to the Arduino Mega microcontroller with Ethernet shield, and the MQTT subscribers, edge wireless IoT devices. The low-cost and reliable soft-/hardware priced at USD 150 assists people physically immobilized from neck to toes whose brain and eye control work normally.

The scientific novelty of obtained results that the method of the smart home management using low/mid gamma waves delivered from the Sichiray TGAM brainwave EEG sensor is firstly proposed. The double blink event is detected by the Sichiray TGAM brainwave EEG sensor if the quality of signal equals 0 and low/mid gamma waves less than 10 (this value is recommended for the experienced users; other values might be applied for new users) in three consecutive Bluetooth packets. This allows to implement the BTC smart home concept and reduces the workload on support staff.

DOI 10.15588/1607-3274-2022-2-9

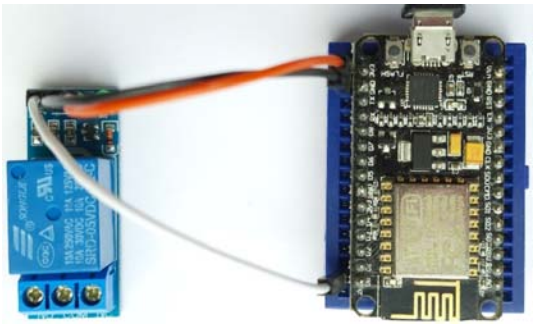

Figure 5 – NodeMCU ESP8266 ESP-12 board (on the right) with relay (on the left)

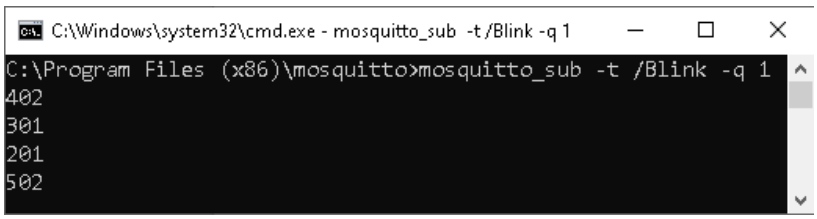

Figure 6 – Screenshot of the command-line interface window with the MQTT subscriber mosquitto_sub.exe: Control commands under the topic "/Blink" and the quality of service level 1

Table 1 – NodeMCU ESP8266 ESP-12 programming and its description: Publishing/subscribing MQTT messages under the topic "/Blink" and controlling the relay connected to the AC with control commands "301" and "302"

Microcontroller programming	Description
#include <ESP8266WiFi.h>	Include libraries to program NodeMCU ESP8266 board and work with
#include <PubSubClient.h>	MQTT IoT data protocol
String str;	Declare variable *str* of String type
byte server[] = { 192, 168, 0, 100 };	Declare array *server* with the IP address of the MQTT broker
void callback(char* topic, byte* payload, unsigned int length)	Function *callback* analyzes the Bluetooth payload and switches
{ str = String((char*)payload);	ON/OFF the relay
if (str.indexOf("301") == 0) digitalWrite (16, HIGH);	
else if (str.indexOf("302") == 0) digitalWrite (16, LOW);	
}	
WiFiClient wifiClient;	Declare WiFi and MQTT clients
PubSubClient client(server, 1883, callback, wifiClient);	
const char* ssid = "SSID";	Declare constants for SSID and password
const char* password = "Password";	
int Relay = 16;	Define the relay pin at D0
void setup() {	
WiFi.begin(ssid, password);	Connect NodeMCU ESP8266 ESP-12 board to the WiFi network
while (WiFi.status() != WL_CONNECTED)	
delay(500);	
if (client.connect("arduinoAC"))	If MQTT client connected, subscribe messages under MQTT topic
client.subscribe("/Blink");	"/Blink"
pinMode(Relay, OUTPUT);	Set pin D0 as output
}	
void loop() { client.loop(); }	Check for new MQTT messages, topic "/Blink"

Figure 7 – Screenshot of the mobile application "Brainwave Visualizer" on the Samsung Galaxy SM-M315F/DSN smartphone, Android 11 OS

DOI 10.15588/1607-3274-2022-2-9

The practical significance of obtained results is that the immobilized people can manage the smart home appliances via the double blink event to select an option shown on the LCD 1602 display with an IIC adapter using the low and mid gamma brain waves. This is achieved in two phases:

1. Headset part: Displaying information about edge IoT devices on LCD 1602 module and sending brain signals from the Sichiray TGAM brainwave EEG sensor kit to HC-05 Bluetooth board wired to the Arduino Mega with Ethernet shield (MQTT publisher).

2. Edge wireless IoT devices: NodeMCU ESP8266 ESP-12 boards with relays (MQTT subscribers).

The Arduino Mega microcontroller analyzes Bluetooth packets received from the Sichiray TGAM brainwave EEG sensor kit, and then sends control commands to the edge IoT devices via MQTT messages.

Ten people of different ages and gender took part in the experiment on the control of smart home appliances with the Sichiray TGAM brainwave EEG sensor kit: five female users of 8, 28, 43, 46, and 57 years old; five male users of 10, 37, 42, 47, and 59 years old. Everyone successfully handled the EEG headset and showed satisfaction with the BTC system.

Prospects for further research are as follows: development of the production sample of the BTC smart home for immobilized people that is ready to market; users of the BTC system recommended replacing the small LCD 1602 display with the larger one.

ACKNOWLEDGEMENTS

This paper and the research behind it could be much more complicated without the support of universities where authors have been conducting the presented project. The authors sincerely appreciate the management and colleagues of the University of Central Asia (Kyrgyzstan), the Süleyman Demirel University (Turkey), and the Kryvyi Rih National University (Ukraine) for all their patience and kind assistance in the completion of this work.

REFERENCES

1. World Health Organization, Disability and Health, 2020 [Electronic resource]. Access mode: https://www.who.int/news-room/fact-sheets/detail/disability-and-health
2. COVID-19 and disabled people snapshot – March 2021 [Electronic resource]. Access mode: http://www.activityalliance.org.uk/how-we-help/research/5854-covid19-and-disabled-people-snapshot-september-2020
3. Ukraine: Authorities to tighten COVID-19 restrictions in Kyiv and Lviv March 20-April 9 /update 20, 2021 [Electronic resource]. Access mode: https://www.garda.com/crisis24/news-alerts/457186/ukraine-authorities-to-tighten-covid-19-restrictions-in-kyiv-and-lviv-march-20-april-9-update-20
4. Teles A., Cagy M., Silva F. et al.Using brain-computer interface and Internet of Things to improve healthcare for wheelchair users, *Mobile Ubiquitous Computing, Systems, Services, and Technologies: 11th international conference, Barcelona, 12–16 November 2017, proceedings*. Wilmington: IARIA XPS Press, 2017, pp. 92–94.
5. Zhang X. , Yao L. , Zhang S. et al.] Internet of Things meets brain-computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity, *IEEE Internet of Things*, 2019, Vol. 6, pp. 2084–2092.
6. Jafri S.R.A., Hamid T., Mahmood R. et al. Wireless brain computer interface for smart home and medical system, *Wireless Personal Communications*, 2019, Vol. 106, pp. 2163–2177. DOI: 10.1007/s11277-018-5932-x
7. Lee W. T., Nisar H., Malik A. S. et al. A brain computer interface for smart home control, *Consumer Electronics: 17th IEEE international symposium, Hsinchu, 3–6 June 2013, proceedings.* Los Alamitos, IEEE, 2013, pp. 35–36. DOI: 10.1109/ISCE.2013.6570240
8. Qin L. Y., Nasir N. M. , Huq M. S. et al. Smart home control for disabled using brain computer interface, *International Journal of Integrated Engineering*, 2020, Vol. 12, No. 4, pp. 74–82.
9. Madoš B., Ádám N., Hurtuk J. et al. Brain-computer interface and Arduino microcontroller family software interconnection solution, *Applied Machine Intelligence and Informatics, 14th IEEE International Symposium, Herlany, 21–23 January 2016, proceedings.* NY, Curran Associates, 2016, pp. 217–221. DOI: 10.1109/SAMI.2016.7423010
10. Sharma V., Sharma A. Review on: Smart home for disabled using brain-computer interfaces, *Information Sciences and Computing Technologies,* 2015, Vol. 2, No. 2, pp. 142–146.
11. Kim M., Kim M.-K., Hwang M. et al. Online home appliance control using EEG-based brain-computer interfaces, *Electronics,* 2019, Vol. 8, No. 10: 1101, 13 p. DOI: 10.3390/electronics8101101
12. Nanda P., Kamath A. Survey on home automation system using brain computer interface paradigm based on auditory selection attention, *International Research Journal of Engineering and Technology*, 2019, Vol. 6, pp. 3100–3111.
13. Belkacem A. N., Jamil N., Palmer J. A. et al. Brain computer interfaces for improving the quality of life of older adults and elderly patients, *Frontiers in Neuroscience,* 2020, Vol. 14, No. 692, pp. 1–11. DOI: 10.3389/fnins.2020.00692
14. Gu X., Cao Z., Jolfaei A. et al. EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, Vol. 18, No. 5, p. 1645–1666. DOI: 10.1109/TCBB.2021.3052811
15. Wu Z. P., Zhang W., Zhao J. et al. Optimized complex network method (OCNM) for improving accuracy of measuring human attention in single-electrode neurofeedback system, *Computational Intelligence and Neuroscience*, 2019, Vol. 2019, No. 2167871. 10 p. DOI: 10.1155/2019/2167871
16. Irfan G., Zubov D. EEG-based brain-computer interface: connecting Sichiray brainwave sensor kit and Arduino Uno with HC-05 Bluetooth module, *Computer Intelligent Systems and Networks, 14th international conference, Kryvyi Rih, 23–25 March 2021, proceedings*. Kryvyi Rih, Kryvyi Rih National University, 2021, pp. 21–28.
17. Dizdarevic J. , Carpio F. , Jukan A. et al. Survey of communication protocols for Internet-of-Things and related challenges of fog and cloud computing integration, *ACM Computing Surveys*, 2019,Vol. 51, No. 6, Article 116, pp. 1–29. DOI: 10.1145/3292674
18. Trent M., Abdelgawad A., Yelamarthi K. et al. Gaggi O., Manzoni P., Palazzi C. et al. (Eds.) A smart wearable navi-

DOI 10.15588/1607-3274-2022-2-9

gation system for visually impaired, *Smart objects and technologies for social good*, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Cham, Springer, 2017, Vol. 195, pp. 333–341. DOI: 10.1007/978-3-319-61949-1_35
19. ScalAgent Distributed Technologies: Benchmark of MQTT servers, ActiveMQ 5.10.0, Apollo 1.7, JoramMQ 1.1.3 (based on Joram 5.9.1), Mosquitto 1.3.5, RabbitMQ 3.4.2. ScalAgent Distributed Technologies, Échirolles, 2015 [Electronic resource]. Access mode: http://www.scalagent.com/IMG/pdf/Benchmark_MQTT_servers-v1-1.pdf
20. Tijms H., Boucherie R., Van Dijk N. M. (Eds.). One-step improvement ideas and computational aspects, *Markov decision processes in practice*. Cham, Springer, 2017, pp. 3–32. DOI: 10.1007/978-3-319-47766-4.
21. Watkins C. J., Dayan P. Technical Note: Q-Learning, *Machine Learning*, 1992, Vol. 8, pp. 279–292. DOI: 10.1023/A:1022676722315
22. Getting Started w/ NodeMCU ESP8266 on Arduino IDE © GPL3+ [Electronic resource]. Access mode: https://create.arduino.cc/projecthub/electropeak/getting-started-w-nodemcu-esp8266-on-arduino-ide-28184f
23. Guide for Relay Module with Arduino [Electronic resource]. Access mode: https://randomnerdtutorials.com/guide-for-relay-module-with-arduino/
24. Dawaga M.E. Automatic detection of eye blinking using the generalized Ising model: thesis ... master of science. London: University of Western Ontario, 2016, 54 p.
25. Jia X., Kohn A. Gamma rhythms in the brain, *PLOS Biology*, 2011, Vol. 9, No. 4, 4 p. DOI: 10.1371/journal.pbio.1001045
26. Brain Wave Frequencies [Electronic resource]. Access mode: https://nhahealth.com/brainwaves-the-language/
27. Arduino library for reading Neurosky EEG brainwave data [Electronic resource]. Access mode: https://www.pantechsolutions.net/blog/arduino-library-for-reading-neurosky-eeg-brainwave-data/
28. Eclipse Mosquitto™: An open-source MQTT broker [Electronic resource]. Access mode: http://mosquitto.org
29. Brainwave Visualizer (Android&iOS) [Electronic resource]. – Access mode: https://store.neurosky.com/products/brainwave-visualizer-1

Received 20.02.2022.
Accepted 23.03.2022.

УДК 004.35

ПРОТОТИП РОЗУМНОГО БУДИНКУ ДЛЯ ПАРАЛІЗОВАНИХ ЛЮДЕЙ: ВЗАЄМОДІЯ МОЗОК-РІЧ НА ОСНОВІ EEG/MQTT

Зубов Д. А. – д-р техн. наук, доцент кафедри комп'ютерних наук Університету Центральної Азії, Бішкек, Киргизька Республіка.

Кureші М. Ш. – д-р техн. наук, доцент кафедри комп'ютерних наук Університету Центральної Азії, Бішкек, Киргизька Республіка.

Козе У. – д-р техн. наук, доцент кафедри комп'ютерної інженерії Університету імені Сулеймана Демірeля, Іспарта, Турецька Республіка.

Купін А. І. – д-р техн. наук, професор, завідувач кафедри комп'ютерних систем і мереж Криворізького національного університету, Кривий Ріг, Україна.

АНОТАЦІЯ

Актуальність. Паралізовані люди мають додаткові перешкоди в багатьох сферах життя, включаючи такі прості дії як вмикання/вимикання освітлення та керування повітряним кондиціонером. Мета роботи – розробка взаємодії мозок-річ в низькій ціновій категорії для контролю підсистем розумного будинку людьми, які паралізовані нижче шиї.

Метод. Неінвазивний прилад Sichiray TGAM вимірює активність мозку людей за допомогою датчика електроенцефалограми і потім передає інформацію через Bluetooth на модуль HC-05, який підключений до мікроконтролера Arduino Mega. Інформація щодо пристроїв розумного будинку показується паралізованій людині на екрані модуля LCD 1602, який підключений до того ж Arduino Mega. Паралізована людина вибирає опцію за допомогою подвійного моргання, що відображується в нульовому значенні якості сигналу та величинах нижніх і середніх гамма хвиль менше ніж десять в трьох послідовних Bluetooth пакетах. Команди керування надсилаються від Arduino Mega (MQTT-видавець) до пристроїв (MQTT-підписники) розумного будинку, які аналізують їх і виконують певну операцію, наприклад відкриття дверей та вмикання/вимикання сигналізації.

Результат. П'ять чоловіків і п'ять жінок віком від 8 до 59 років тестували комплекс керування підсистемами розумного будинку на базі приладу Sichiray TGAM. Результати показали успішне освоєння і зацікавленість у використанні системи мозок-річ.

Висновки. У даній роботі представлена концепція розумного будинку для паралізованих людей на базі принципу мозок-річ і MQTT взаємодії між MQTT-видавцем (прилад Sichiray TGAM з датчиком електроенцефалограми, який через Bluetooth підключений до мікроконтролера Arduino Mega) і пристроями розумного будинку загальною ціною близько USD 150. Перспективою подальшого розвитку є виробництво пристроїв готових до масового використання.

КЛЮЧОВІ СЛОВА: мозок-річ, паралізовані люди, датчик електроенцефалограми, MQTT.

DOI 10.15588/1607-3274-2022-2-9

УДК 004.35

ПРОТОТИП УМНОГО ДОМА ДЛЯ ПАРАЛИЗОВАННЫХ ЛЮДЕЙ: ВЗАИМОДЕЙСТВИЕ МОЗГ-ВЕЩЬ НА ОСНОВЕ EEG/MQTT

Зубов Д. А. – д-р техн. наук, доцент кафедры компьютерных наук Университета Центральной Азии, Бишкек, Киргизская Республика.

Куреши М. Ш. – д-р техн. наук, доцент кафедры компьютерных наук Университета Центральной Азии, Бишкек, Киргизская Республика.

Козе У. – д-р техн. наук, доцент кафедры компьютерной инженерии Университета имени Сулеймана Демиреля, Испарта, Турецкая Республика.

Купин А. И. – д-р техн. наук, профессор, заведующий кафедрой компьютерных систем и сетей Криворожского национального университета, Кривой Рог, Украина.

АННОТАЦИЯ

Актуальность. Парализованные люди встречают дополнительные препятствия во многих сферах жизни включая такие простые действия как включение/выключение освещения и управление воздушным кондиционером. Цель работы – разработка взаимодействия мозг-вещь в низкой ценовой категории для контроля подсистем умного дома людьми, которые парализованы ниже шеи.

Метод. Неинвазивный прибор Sichiray TGAM измеряет активность мозга людей при помощи датчика электроэнцефалограммы и затем передает информацию по Bluetooth на модуль HC-05, который подключен к микроконтроллеру Arduino Mega. Информация об устройствах умного дома показывается парализованному человеку на экране модуля LCD 1602, подсоединенному к тому же Arduino Mega. Парализованный человек выбирает показываемую опцию при помощи двойного моргания, что отражается в нулевом значении качества сигнала и величинах нижних и средних гамма волн меньше десяти в трех последовательных Bluetooth пакетах. Управляющие команды посылаются от Arduino Mega (MQTT-издатель) к устройствам (MQTT-подписчики) умного дома, которые анализируют их и выполняют определенную операцию, например открытие двери и включение/выключение сигнализации.

Результаты. Пять человек мужского и пять женского полов возрастом от 8 до 59 лет тестировали комплекс управления подсистемами умного дома на базе прибора Sichiray TGAM. Результаты показали успешное освоение и заинтересованность в использовании системы мозг-вещь.

Выводы. В данной работе представлена концепция умного дома для парализованных людей на базе принципа мозг-вещь и MQTT взаимодействия между MQTT-издателем (прибор Sichiray TGAM с датчиком электроэнцефалограммы, который через Bluetooth подсоединяется к микроконтроллеру Arduino Mega) и устройствами умного дома общей ценой около USD 150. Перспективой дальнейшего развития является производство изделия готового к массовому использованию.

КЛЮЧЕВЫЕ СЛОВА: мозг-вещь, парализованные люди, датчик электроэнцефалограммы, MQTT.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. World Health Organization, Disability and Health, 2020 [Electronic resource]. – Access mode: https://www.who.int/news-room/fact-sheets/detail/disability-and-health
2. COVID-19 and disabled people snapshot – March 2021 [Electronic resource]. – Access mode: http://www.activityalliance.org.uk/how-we-help/research/5854-covid19-and-disabled-people-snapshot-september-2020
3. Ukraine: Authorities to tighten COVID-19 restrictions in Kyiv and Lviv March 20–April 9 /update 20, 2021 [Electronic resource]. – Access mode: https://www.garda.com/crisis24/news-alerts/457186/ukraine-authorities-to-tighten-covid-19-restrictions-in-kyiv-and-lviv-march-20-april-9-update-20
4. Using brain-computer interface and Internet of Things to improve healthcare for wheelchair users / [A. Teles, M. Cagy, F. Silva et al.] // Mobile Ubiquitous Computing, Systems, Services, and Technologies: 11th international conference, Barcelona, 12–16 November 2017: proceedings. – Wilmington: IARIA XPS Press, 2017. – P. 92–94.
5. Internet of Things meets brain-computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity / [X. Zhang, L. Yao, S. Zhang et al.] // IEEE Internet of Things. – 2019. – Vol. 6. – P. 2084–2092.
6. Wireless brain computer interface for smart home and medical system / [S.R.A. Jafri, T. Hamid, R. Mahmood et al.] // Wireless Personal Communications. – 2019. – Vol. 106. – P. 2163–2177. doi:10.1007/s11277-018-5932-x
7. A brain computer interface for smart home control / [W.T. Lee, H. Nisar, A.S. Malik et al.] // Consumer Electronics: 17th IEEE international symposium, Hsinchu, 3–6 June 2013: proceedings. – Los Alamitos : IEEE, 2013. – P. 35–36. DOI: 10.1109/ISCE.2013.6570240
8. Smart home control for disabled using brain computer interface / [L.Y. Qin, N.M. Nasir, M.S. Huq et al.] // International Journal of Integrated Engineering. – 2020. – Vol. 12, № 4 – P. 74–82.
9. Brain-computer interface and Arduino microcontroller family software interconnection solution / [B. Madoš, N. Ádám, J. Hurtuk et al.] // Applied Machine Intelligence and Informatics: 14th IEEE International Symposium, Herlany, 21–23 January 2016: proceedings. – NY: Curran Associates, 2016. – P. 217–221. DOI: 10.1109/SAMI.2016.7423010
10. Sharma V. Review on: Smart home for disabled using brain-computer interfaces / V. Sharma, A. Sharma // Information Sciences and Computing Technologies. – 2015. – Vol. 2, №. 2. – P. 142–146.
11. Online home appliance control using EEG-based brain-computer interfaces / [M. Kim, M.-K. Kim, M. Hwang et al.] // Electronics. – 2019. – Vol. 8, № 10: 1101. – 13 p. DOI: 10.3390/electronics8101101
12. Nanda P. Survey on home automation system using brain computer interface paradigm based on auditory selection attention / P. Nanda, A. Kamath // International Research

DOI 10.15588/1607-3274-2022-2-9

Journal of Engineering and Technology. – 2019. – Vol. 6. – P. 3100–3111.
13. Brain computer interfaces for improving the quality of life of older adults and elderly patients / [A.N. Belkacem, N. Jamil, J.A. Palmer et al.] // Frontiers in Neuroscience. – 2020. – Vol. 14, № 692. – P. 1–11. DOI: 10.3389/fnins.2020.00692
14. EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications / [X. Gu, Z. Cao, A. Jolfaei et al.] // IEEE/ACM Transactions on Computational Biology and Bioinformatics. – 2021. – Vol. 18, № 5. – P. 1645–1666. DOI: 10.1109/TCBB.2021.3052811
15. Optimized complex network method (OCNM) for improving accuracy of measuring human attention in single-electrode neurofeedback system / [Z.P. Wu, W. Zhang, J. Zhao et al.] // Computational Intelligence and Neuroscience. – 2019. – Vol. 2019, № 2167871. – 10 p. DOI: 10.1155/2019/2167871
16. EEG-based brain-computer interface: connecting Sichiray brainwave sensor kit and Arduino Uno with HC-05 Bluetooth module / G. Irfan, D. Zubov // Computer Intelligent Systems and Networks: 14th international conference, Kryvyi Rih, 23–25 March 2021: proceedings. – Kryvyi Rih: Kryvyi Rih National University, 2021. – P. 21–28.
17. Survey of communication protocols for Internet-of-Things and related challenges of fog and cloud computing integration / [J. Dizdarevic, F. Carpio, A. Jukan et al.] // ACM Computing Surveys. – 2019. – Vol. 51, № 6. – Article 116. – P. 1–29. DOI: 10.1145/3292674
18. Trent M. A smart wearable navigation system for visually impaired / [M. Trent, A. Abdelgawad, K. Yelamarthi et al.] // Smart objects and technologies for social good / [O. Gaggi, P. Manzoni, C. Palazzi et al.] (Eds.) // Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. – Cham: Springer, 2017. – Vol. 195. – P. 333–341. DOI: 10.1007/978-3-319-61949-1_35
19. ScalAgent Distributed Technologies: Benchmark of MQTT servers: ActiveMQ 5.10.0, Apollo 1.7, JoramMQ 1.1.3 (based on Joram 5.9.1), Mosquitto 1.3.5, RabbitMQ 3.4.2. ScalAgent Distributed Technologies, Échirolles, 2015 [Electronic resource]. – Access mode: http://www.scalagent.com/IMG/pdf/Benchmark_MQTT_servers-v1-1.pdf
20. H. Tijms. One-step improvement ideas and computational aspects / H. Tijms // Markov decision processes in practice / R. Boucherie, N.M. Van Dijk (Eds.). – Cham: Springer, 2017. – P. 3–32. doi: 10.1007/978-3-319-47766-4.
21. Watkins C.J. Technical Note: Q-Learning / C.J. Watkins, P. Dayan // Machine Learning. – 1992. – Vol. 8. – P. 279–292. DOI: 10.1023/A:1022676722315
22. Getting Started w/ NodeMCU ESP8266 on Arduino IDE © GPL3+ [Electronic resource]. – Access mode: https://create.arduino.cc/projecthub/electropeak/getting-started-w-nodemcu-esp8266-on-arduino-ide-28184f
23. Guide for Relay Module with Arduino [Electronic resource]. – Access mode: https://randomnerdtutorials.com/guide-for-relay-module-with-arduino/
24. Dawaga M.E. Automatic detection of eye blinking using the generalized Ising model: thesis ... master of science / Dawaga Marwa Elsayh. London: University of Western Ontario, 2016. – 54 p.
25. Jia X. Gamma rhythms in the brain / X. Jia, A. Kohn // PLOS Biology. – 2011. – Vol. 9, № 4. – 4 p. DOI: 10.1371/journal.pbio.1001045
26. Brain Wave Frequencies [Electronic resource]. – Access mode: https://nhahealth.com/brainwaves-the-language/
27. Arduino library for reading Neurosky EEG brainwave data [Electronic resource]. – Access mode: https://www.pantechsolutions.net/blog/arduino-library-for-reading-neurosky-eeg-brainwave-data/
28. Eclipse Mosquitto™: An open-source MQTT broker [Electronic resource]. – Access mode: http://mosquitto.org
29. Brainwave Visualizer (Android&iOS) [Electronic resource]. – Access mode: https://store.neurosky.com/products/brainwave-visualizer-1

DOI 10.15588/1607-3274-2022-2-9

e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2022. № 2
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2022. № 2

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

UDC 004.658.3

USING THE ANALYTIC HIERARCHY PROCESS WITH FUZZY LOGIC ELEMENTS TO OPTIMIZE THE DATABASE STRUCTURE

Dvoretskyi M. L. – PhD, Senior Lecturer of the Department of Software Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine.

Savchuk T. O. – PhD, Professor, Professor of the Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, Ukraine.

Fisun M. T. – Dr. Sc., Professor, Professor of the Department of Software Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine.

Dvoretska S. V. – Senior Lecturer of the Department of Software Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine.

ABSTRACT

Context. Informational systems are very common and use databases to store information that users need. Many different data models can be used but the relational model is still relevant. The last decade show tendency of using distributed databases while working with relational data model and this approach requires a specially designed module to synchronize data of all separate databases. Considering optimizing the database structure, researchers didn't pay much attention to the potential of users' SQL-queries history. The optimal structure of all the distributed nodes could reduce the necessity of synchronization while the data access speed and its actuality would remain stable. The object of the research is the process of optimizing the structure of the distributed database of corporate information systems, which are based on the relational database's model.

Objective. The research aims at improving the accuracy of the data representation marker's value on the distributed corporate information system's (DCIS) node, obtained using the analytic hierarchy process by applying the fuzzy logic elements while processing the alternatives' global priority vector.

Method. The research's authors in the set of their previous works emphasize the potential of using the collected history of users' SQL queries. Firstly presented technology of users' queries parsing. Then, the idea of using the multidimensional database for analyzing users' queries by slices of workstation type, application, user, and his/her position was considered. Finally, the authors gave the full-scaled mathematical model for formalizing database and query models, and criteria of database structure's optimality.

The current research continues the given sequence and tries to increase the efficiency of the decision support system, by introducing elements of fuzzy logic to the analytic hierarchy process algorithm. The approach's main idea is in presenting the global priorities vector in the form of a series of fuzzy sets of one variable with subsequent transformation to the exact value. This approach made it possible to maintain the accuracy of the obtained result while decreasing the number of solution alternatives.

For new tuples added to the database's tables after all calculations had been performed, the classification problem was formalized. After obtaining the probability of a tuple belonging to the class "needed" and performing the normalization of the value, it is taken as the level of the representation marker. Accordingly, the data is loaded onto the node if this value is greater than the optimal level of the representation marker for the DCIS node.

Results. After calculating and obtaining the alternatives global priorities' vector in order to improve the accuracy of the obtained result, the apparatus of fuzzy sets was used. The obtained vector of global priorities was presented as a vector of fuzzy digits for the data representation marker with subsequent transformation to the exact value. This approach made it possible to maintain the accuracy of the obtained result while decreasing the number of solution alternatives.

Conclusions. While working on the research, the concept of a data representation marker on the DCIS node for the elements of the SQL query model was introduced. An aggregation function has been developed that allows determining the level of need for attributes and tuples in the database's relation for the DCIS node based on the statistics of SQL queries. A model of the dependence of the database structure's optimality criteria on the value of the data representation marker is built. Received further development method of analytic hierarchy process. The initialization of the alternatives' pairwise comparisons matrix can be performed automati-

DOI 10.15588/1607-3274-2022-2-10

cally according to the obtained mathematical models. Representation of the obtained result in the form of the vector of fuzzy numbers with the reduction to the exact value allows increasing the accuracy of the obtained results.

KEYWORDS: corporate information system, database management system, distributed database, SQL-query, data replication, multicriteria problem, analytic hierarchy process, fuzzy logic, classification problem, naive Bayes algorithm.

ABBREVIATIONS

CIS is a corporate information system;

DB is a database;

DBMS is a database management system;

DCIS is a distributed corporate information system;

ORM is an object-relational mapping;

SQL is a structured query language.

NOMENCLATURE

R is a database relation or database table;

$R[P]$ is a database relation's or database table's projection;

$R[S]$ is a table selection;

tup is a tuple of the relation R;

R'' is a result of sequential execution of selection and projection operations to the base relation (table);

$F(tup, S)$ is a conditional boolean function on the tuple of the relation;

R''_{union} is a subset of the relation R that is defined as the union of R'' subsets from the DCIS remote node;

P_{union} is a union of the relation's R attributes from the set of projections $R[P]$;

S_{union} is a union of the relation's R selection conditions from the set of selections $R[S]$;

R_{shema} is a set of relations' attributes;

$R_{primary}$ is a subset of relations' attributes that uniquely identify cortege;

$F_a(Node, A)$ is an evaluation function that determines whether the A attribute is needed on the remote node *Node*;

R_{schema}^{remote} is a subset of the R_{shema}, part of relations' attributes which are presented on the remote node;

$F_{tup}(Node, tup)$ is an evaluation function that determines whether the *tup* tuple is needed on the remote node *Node*;

$R_{data}^{remote-dep}$ is a set of the relation tuples which are linked from other relations' foreign keys;

Q is a set of dimensions of a user's query;

Q_{set}^{inner} is a set of inner SQL-queries of outer SQL-query Q;

R''_{shema} is a resulting relations' set;

Mrk is a data representation marker that reflects the level of data representation necessity at the DCIS node;

$koef_{repr}^{node}$ is a data representation's threshold coefficient, defined on the range [–1, 1];

$Repr(Node,R,A,tup)$ is a conditional boolean function that determines the necessity of the data slice $<R, A, tup>$ on the remote node *Node*;

$F_{availab}$ is an estimation function that determines the value of the data availability criterion (independence from the central database node);

F_{size} is a value relative value of remote node's database size that is the database size criterion;

$F_{synchro}$ is a need for data synchronization criterion;

W^{global} is an alternatives' global priority vector;

W_5^{global} is a global priority vector in case of five alternatives;

μ is a belonging function;

a_r is a center of mass.

INTRODUCTION

Nowadays informational systems are everywhere around us, and their users to even notice to use them. This list of examples can be very long, from reading news on the internet, e-commerce, remote education to e-banking, accounting systems, decision support, and so on. All these systems use databases to store information that users need, process it, and present it in appropriate form. Historically several data models are used to present informational system data among which can be mentioned hierarchical, network, relational, object, and document models. Two of them (relational and document) for the set of reasons are particularly popular [1].

Some works positioned relational model as an old and irrational way of storing data while document model is shown to be easily scaled and more productive and sufficient [2]. However, deeper considering which of these two is the most appropriate to use makes it clear that there is no precise answer. It depends on many conditions. For example, among them can be given the necessity of transactional data processing and extracting objects from the database not only by their unique identifier values [3].

Meanwhile, most of the accounting systems were historically built on relational database management systems this fact makes that model particularly useful. The possibility of using object-relational mapping (ORM) technologies [4] also makes the convenience of the relational model almost equal to documental while working with object-oriented methodology.

All this highlighted that the relational model is still relevant but according to the facts of rising the storage capacities, increasing data access speed, using of the ORM technologies, and so on, normalization now is not the primary trend and data duplication can be justified. In this context, the question of database structure optimizing can be viewed from a different angle.

The object of study is the process of optimizing the database structure of the distributed corporate information systems based on the relational data model

The key factor influencing the reliability and availability of the database is the localization of links. The high degree of localization of links can be made by the presentation on the node of the data that is needed exclusively by the current node's users. Database relations are pre-

DOI 10.15588/1607-3274-2022-2-10

sented at the DCIS node after applying the projection and selection operations. That is, for optimal presentation of data it is necessary to use elements of vertical and horizontal data fragmentation.

The subject of study is the analytic hierarchy process for choosing the best alternative of the DCIS node's structure based on the created model of SQL-queries to the relational database

The method allows picking the most optimal decision from the set of alternatives. Increasing the number of alternatives can improve the accuracy of the obtained numeric solution but leads to rising the size of the matrices of pairwise comparison. It was suggested using elements of fuzzy logic while working with the obtained vector of alternatives' global priority.

The purpose of the work is to improve the accuracy of the data representation marker's value on the distributed corporate information system's (DCIS) node, obtained using the analytic hierarchy process by applying the fuzzy logic elements while processing the alternatives' global priority vector.

1 PROBLEM STATEMENT

The last decade shows the tendency of using distributed databases while working with the relational data model. And there are several reasons for this. There are several arias of accounting within one company [5]. For example, it can be a warehouse, human resource, access control, and other types of accounting. The attempt of combining them in one "universal" information system (corporate information systems, CIS) provides a single accounting environment and gives access to all company data for future analysis and decision-making. Nevertheless, presenting all data in one database is connected with the set of potential problems [6], among which productivity, reliability, and safety should be mentioned first.

To solve part of them, the special accounting systems can be separated and each of them will use its own database. This technic is also known in modern application development as the use of the microservices approach in high-load projects [7]. It is clearly understandable that this approach requires a specially designed module to synchronize data of all separate databases.

Company structure also can be geographically distributed and some parts of data have to be presented locally so that data consumers won't rely on the availability of the remote database server. According to this, the set of company databases $D = \{D_1, D_2, \dots D_n\}$, or its subset $D` \subset D$, or maybe some subset of tables $R`_{set} \subset R_{set} = \{R_1, R_2, \dots R_m\}$ should be placed on the local database server and periodically synchronized with the central database version.

Considering optimizing the database structure [8–11], researchers didn't pay much attention to the potential of users' SQL-queries history. In works [9–11] took into account increasing productivity by using the materialized views, database restructuring, and denormalization. But the problem wasn't considered in the context of the single distributed CIS node.

According to the given above, the tendency of using the distributed databases is justified and the need for their parts to be synchronized is clear. And the optimal structure of all the distributed nodes could reduce the necessity of synchronization while the data access speed and its actuality would remain stable. Therefore, the task of optimizing the remote node's database structure is quite relevant.

While defining the objective function, three optimality criteria can be defined. These are independence from the central database node $F_{availab}$,local database size F_{size}, and the need for data synchronization $F_{synchro}$. So the objective function has three input variables $F_{objective}(F_{availab}, F_{size}, F_{synchro})$. Taking into account the goal of minimizing the F_{size} and $F_{synchro}$ criteria, and maximizing the $F_{availab}$, objective function hypothetically can be defined as $F_{objective} = \frac{F_{availab} \times W_{availab}}{F_{size} \times W_{size} + F_{synchro} \times W_{synchro}} \to \max$. But the definition of the weight coefficients could be a very difficult task for the decision-maker, so the analytic hierarchy process is considered rational.

2 REVIEW OF THE LITERATURE

The research's authors in the set of their previous works emphasize on potential of using the collected history of users' SQL-queries [12–15]. The queries should previously be parsed to extract the sets of entities (table), attributes (columns), and tuples (rows). Then, this data can be used to determine the level of necessity for this data to be presented on the node of the distributed database. And finally after the optimal level of data representation on the distributed database node will be found the optimal structure of its database can be built.

In work [12] authors firstly presented technology of users' queries parsing. Then, in [13] the idea of using multidimensional database for analysis users' queries by slices of workstation type, application, user and his/her position was considered. Finally, in [14] authors gave the full scaled mathematical model for formalizing database and query models, and criteria of database structure's optimality. It was presented the "data representation marker" term, which determined the level of their need at the node of the distributed corporate information system (DCIS). The multicreterial decision support system that was also presented in [15], searches the optimal value of data representation marker based on the following criteria: independence from the central node of the database, the size of the local database and the level of needed data synchronization.

The current research continues the given sequence and tries to increase the efficiency of decision support system, given in [14], by introduction to the analytic hierarchy process algorithm (that was used previously) elements of fuzzy logic. The approach's main idea is in presenting the global priorities vector in the form of series of fuzzy sets of one variable. The following representation this set in

DOI 10.15588/1607-3274-2022-2-10

the form of exact value (defasification) will increase the accuracy of the obtained solution without the necessity of dividing the range of solution's valid values to more number of intervals. Also, in order to reduce the necessity of re-conducting the analysis after some time passes, the solution of classification task is considered that will help to determine the "data representation marker" for the new data in the database.

3 MATERIALS AND METHODS

While working on [14] authors proposed the model of users' SQL-queries according to which subsets of data can be extracted from the main database to be presented on the remote node of the DCIS. It was suggested following legends: R – database relation (table); $R[P]$ – table projection; $R[S]$ – table selection; *tup* –a tuple of the relation R. Within the SQL-query for data selecting, a number of relations can be involved, all of which are the result of sequential execution of selection and projection operations to the base relation (table). $R'' = R'[P]$, where $R' = R[S]$, i.e. $R'' = \{tup[P] \mid tup[P] \in R[P]_{data} \wedge F(tup, S) = true\}$ [14]. When working with the sequence of the database's queries, the subset R''_{union} of the relation R was defined as the union of subsets R' of each SQL-query that came from the DCIS remote node. $R''_{union} = \bigcup_{i=1}^{n} R''_i$, or $R''_{union} = \{tup[P_{union}] \mid tup[P_{union}] \in R[P_{union}]_{data} \wedge F(tup, S_{union}) = true\}$, where $tup[P_{union}] = \bigcup_{i=1}^{n} tup[P_i]$, and $S_{union} = \bigcup_{i=1}^{n} S_i$.

Some part of data can be presented locally to avoid the necessity of remote queries and another part is placed at the remote node to reduce the amount of data that should be replicated between the nodes. The subset of the base relation R that describes the relation of the remote node was represented as follows: $R^{remote}_{schema} = \{A \mid A \in R_{shema}, R_{primary} \subset R^{remote}_{schema}, A \in R_{primary} \vee F_a(Node, A) = true\}$. And the set of table's tuples was determined by the formula $R^{remote}_{data} = \{ tup \mid tup \in R_{data}, tup_{primary} \in R^{remote-dep}_{data} \vee F_{tup}(Node, tup) = true\}$. The need for the data on the remote node is determined mostly by the evaluation function $F_{tup}(Node, tup)$.

The model of SQL-queries supports further classification according to belonging to the workplace, location, user role and other potential dimensions $Q = <Workplace, User, Application, R''_{shema}, Q^{inner}_{set}>$. In this formula $R''_{set} = \{ R'' \mid \{tup[P] \mid tup[P] \in R[P]_{data} \wedge F(tup, S) = true \}$ – is the resulting relations' set; Q^{inner}_{set} – the nested queries' set. Based on this query model the multidimensional database [13] was created. It has the following set of dimensions: $<DateTime, WorkplaceType, Location, UserRole, Application, R, A, tup>$. Then, the term of data representation marker was proposed that reflects the level of data representation necessity at the DCIS node. For each dimension's element value of the marker is determined from the following set: {"*obligatorily*", "*necessary*", "*neutral*", "*not required*", "*forbidden*"}. After been converted to a numeric values ("*obligatorily*" – "2", "*necessary*" – "1", "*neutral*" – "0", "*not required*" – "–1", "*forbidden*" – "–2"), it was defined the aggregation function for the marker:

$$Aggr^{n}_{i=1} Mrk_i = \begin{cases} 2, if\ \exists Mrk_i = 2 \\ -2, if\ \exists Mrk_i = -2 \wedge \nexists Mrk_i = 2. \\ \sum_{i=1}^{n} (Mrk_i \times \frac{vol_i}{\sum_{i=1}^{n} vol_i}) \end{cases} \quad (1)$$

The decision about the necessity of the data slice $<R, A, tup>$ on the remote node is made according to the following condition:

$$Repr(Node, R, A, tup) = \\ = Aggr^{n}_{i=1}(R, A, tup) Mrk_i > koef^{node}_{repr}, \quad (2)$$

where $koef^{node}_{repr}$ – the data representation's threshold coefficient, defined on the range [–1, 1].

The optimality of the $koef^{node}_{repr}$ is defined by three criteria. These are independence from the central database node, local database size, and the need for data synchronization. The value of each criterion is defined by (3), (4), and (5) accordingly

$$F_{availab} = \frac{\sum_{i=1}^{n} F_{availab}(Q_n)}{n}, \quad (3)$$

$$F_{size} = \sum_{i=1}^{n} \frac{size(R''_i)}{size(R_i^{DBMS})}, \quad (4)$$

$$F_{synchro} = \frac{p^{\mathrm{mod}\, if}_{node}}{p_{node}}. \quad (5)$$

A deeper explanation of given formulas can be found in [15].

The obtained multicriteria problem was solved using the analytic hierarchy process. When compiling the hierarchy, the following relationship between the levels' elements was used: goal – stakeholders – criteria – alternatives. The value of the data representation marker (alternative) is a real number in the interval [–1, 1]. It leads to the potentially large number of alternatives at the 4th level of the hierarchy and therefore the matrices of pairwise comparisons by criteria can become very big. This complicates the estimation process for the decision-makers. It is proposed to simplify the task by reducing the number of alternatives to 5: "*low*" (*L*) – "–1", "l*ower them medium*" (*LM*) – "–0.5", "*medium*" (*M*) – "0", "*higher then medium*" (*HM*) – "0.5", and "*high*" (*H*) – "1". The level of "*decision-makers*" is represented by the elements "*Owner*", "*Database Administrator*", "*Database*

DOI 10.15588/1607-3274-2022-2-10

Developer", and "*CIS Operator*". The obtained hierarchical model is shown in Fig. 1.

When splitting the set of alternatives of the data representation marker's value (determined on the set of real numbers in the interval [–1; 1]) into a larger number of intervals, it is possible to increase the accuracy of the obtained solution.

Based on the difficulties of the larger number of intervals implementation and the need to improve the accuracy of the method, it was suggested to use the elements of the fuzzy logic apparatus [16–19]. This makes it possible to obtain the same result's accuracy as in the case of the larger number of alternatives, using a fewer number of intervals of the data representation marker.

In the theory of sets, an element either belongs to the set or not. A fuzzy set is defined using a membership function that corresponds to the concept of a characteristic function in classical logic. The membership function can take any form, but the piecewise linear form is used most often to represent it. Piecewise linear membership functions are traditionally used for several reasons: they are characterized by simplicity; they contain points that allow defining areas where the concept is true and where it is false, which simplifies the system's description [17].

In Fig. 2 the selected intervals of the data representation marker level values ("*low*", "*lower than medium*", "*medium*", "*higher than medium*", "*high*") are presented in the form of a series of fuzzy sets of one variable with piecewise linear membership functions.

The classical process of fuzzy inference consists of the following stages: fuzzification (presenting the exact numeric value in a fuzzy form), fuzzy inference itself, usually based on a set of rules, and defuzzification (numerical expression of a fuzzy result) [20].

In our case, the solution of the multicriteria analysis problem was performed previously using the analytic hierarchy process method, and the following vector of alternatives' global priorities was obtained (10).

Further, the found vector of global priorities is represented as a vector of fuzzy numbers for the data representation marker. That is, to obtain the numerical value of the data presentation marker's optimal level, should be accomplished mapping of the results with the next defuzzification phase. Defuzzification is the process of trans-

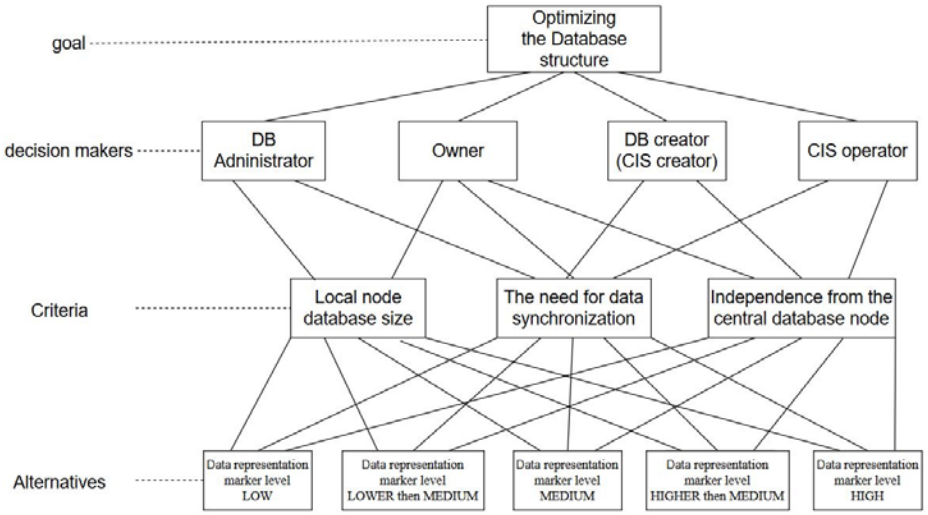

Figure 1 – Hierarchical model of the distributed CIS node structure optimization problem

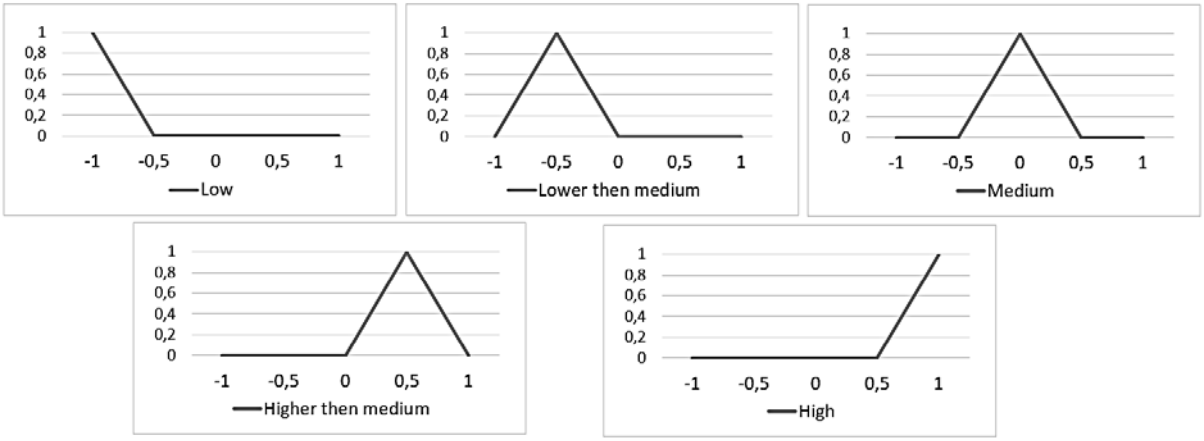

Figure 2 – The series of fuzzy sets for describing the data representation marker's level

DOI 10.15588/1607-3274-2022-2-10

forming a fuzzy number to its exact numeric interpretation. In the theory of fuzzy sets, defuzzification is similar to finding characteristics of the random variables' position (mean, mode, median) in the theory of probability. Among the defuzzification methods, there is the choice of an exact number with the maximum value of the membership function. Alternatively, it can be methods based on the idea of finding the point of concentration (center of mass) of the area located between the membership function's graph and the abscissa axis. This is done in order to "average" the possible values taken by a fuzzy number, taking into account the belonging function value.

Among the most common methods is the center of mass method $a_r = \int_{\min}^{\max} u\mu_{\varepsilon}(u)du \Big/ \int_{\min}^{\max} \mu_{\varepsilon}(u)du$ and the median method $a_r : \int_{\min}^{\max} u\mu_{\varepsilon}(u)du = \int_{\min}^{\max} \mu_{\varepsilon}(u)du$.

In the process of using the described technology, there is a problem related to adding new tuples to the database tables. The level of data presentation's need is determined in accordance with the accumulated statistics of SQL queries and the range of primary keys that meet the selection condition. The range of the primary key of the recently added to the database tuples was not available during the analysis. Therefore, they were not marked according to the statistics of queries in the database. For these tuples, there is no value for the level of the need to represent data in the DCIS node. The simplest solution to the problem is to replicate all new data to the remote DCIS node (since there is no data about the degree of their usefulness for the node), that is, setting for them the value of the need for presentation at the level of "obligatory". This approach is not fully justified, because, firstly, it does not take into account the influence of data (added and specially modified) rows on other rows due to their intersection within the same query (software or host type). In addition, when the amount of data reaches the critical point, based, for example, on load during selecting or data synchronization operations, there comes a time when it is necessary to re-initialize the multidimensional database of SQL-queries and update the data of using the attributes and tuples of the database tables by the DCIS node [13, 21].

Accomplishing the complete analysis of using the attributes and tuples of database tables is a resource-expensive operation and cannot be performed frequently. Therefore, the given above approach is unacceptable for large and frequently changing databases. At the same time, secondary and subsequent monitoring of the database users' activity is complicated by the fact that the database of the remote node is already in use, and requests to it must also be taken into account.

In most cases, the main influence on the level of representation in the analysis of user queries to the database makes the combination of the attributes' values of the relation's tuple. Taking into account this fact, it is proposed to present the problem with determining the representation marker's level of the new data in the form of a classification problem. The input parameters are the name of the table (relationship) and the list of tuple attributes' values and the result will be a decision of presenting the tuple for the DCIS node.

Among the many algorithms for solving the classification problem by means of machine learning, the most popular approaches are identified, including linear and logical regression, discriminant analysis, decision trees, the Naive Bayes algorithm, k-nearest neighbors, and the use of various types of neural networks [22]. Linear and logical regression are some of the most well-known methods, but according to the non-numerical characteristics of most of the input variables (table attributes), they are not optimal in this case. For the same reasons, and also because of the complexity of determining the distance, the k-nearest neighbors aren't considered either.

The use of neural networks is now the most popular direction in solving the classification problem [23–24]. However, not all tables contain a sufficient number of rows to carry out a high-quality training stage (insufficient amount of data for training and testing the results). In addition, each table (relation) has a different number of attributes, and each of them is defined on its own domain. Based on this, the problem of classifying new data for each relation must be solved using a separately trained neural network. According to the given above, the use of neural networks isn't also the appropriate approach.

To solve the problem, the Bayes naive algorithm [25] was used, which is simple but effective and allows to quickly determine the probability of an object belonging to a particular class. The algorithm is known to be based on the assumption that each input variable is independent, which is often not true. But production versions of the relational databases in most cases are in the 3rd normal form, which indicates the absence of transitive dependencies inside the relation [26]. This fact allows asserting that the main part of the input variables corresponds to the basic algorithm's assumption.

The algorithm is based on Bayes' theorem, which allows calculating the probability of an object belonging to a particular class [25]. For the current task, the probability that the tuple X is needed to be presented at a remote node according to the value of the i-th attribute x_i, was found using the following formula:

$$P_x(needed \mid x_i) = \frac{P(x_i \mid needed) \times P(needed)}{P(x_i)}, \qquad (6)$$

where $P(needed)$ – the total probability of the relation that the tuple has to be represented on a remote node; $P(x_i)$ – the probability of the x_i value of the *i-th* attribute; $P(x_i \mid needed)$ – the probability of the x_i value of the *i-th* attribute on the subset of the relation's tuples of the remote node.

Also, the same way calculate the probability that the tuple X is not needed to be presented at a remote node:

DOI 10.15588/1607-3274-2022-2-10

$$P_x(not\ needed \mid x_i) = \frac{P(x_i \mid not\ needed) \times P(not\ need}{P(x_i)} \quad (7)$$

After obtaining probabilities $P_x(not\ needed \mid x_i)$ for all tuple's attributes based on (6) and (7), the calculation of the probability that the whole tuple will be needed is performed:

$$P_x(needed \mid X) = \\ = \frac{\prod_{i=1}^{n} P_x(needed \mid x_i)}{\prod_{i=1}^{n} P_x(needed \mid x_i) + \prod_{i=1}^{n} P_x(not\ needed \mid x_i)}. \quad (8)$$

If the value, obtained according to (6) is greater than the optimal value of the data representation marker, then the tuple is accepted as the one that has to be presented on the remote node of the DCIS:

$$F_X^{needed} = \begin{cases} true, \dfrac{koef_{distrib}^{node} + 1}{2} > P(needed \mid X) \\ false, \dfrac{koef_{distrib}^{node} + 1}{2} \le P(needed \mid X) \end{cases}. \quad (9)$$

So, using (8) and (9) it is possible to make a decision about presenting a new table's tuples on the remote node of DCIS without the necessity of re-processing the SQL-queries statistic. It simply can be done based on the data for which the decision about its presenting on the remote node was already made.

4 EXPERIMENTS

Using SQL-queries statistics and models (1–5) the global alternatives' vector was obtained:

$$W^{global} = \begin{bmatrix} 0.000 \\ 0.273 \\ 0.334 \\ 0.393 \\ 0.000 \end{bmatrix}. \quad (10)$$

While working with pairwise comparison matrices the ranges of available values of optimality criteria were also taken into account. The full description of obtaining the given result can be found in [15].

The analytic hierarchy process method that was given in [14], was used to obtain the optimal solution from 5 and 21 alternatives (intervals of the data representation marker's values). For the case of 21 alternatives the task has to be simplified, so no restrictions on the optimality criteria were introduced and a three-level hierarchical model without the "*stakeholder*" level was used. As a matrix of pairwise comparisons of optimality criteria, the matrix that was filled by the "*Owner*" person was taken [15]. Also, the initial state of the alternatives' advantages matrices according to the optimality criteria, which were obtained in accordance with the mathematical models (3), (4), and (5) is accepted as final.

Taking into account the introduced simplifications of the model, the calculation of the global priorities' vector for five alternatives ("*low*" (*L*) – "–1", "*lower than medium*" (*LM*) – "–0.5", "*medium*" (*M*) – "0", "*higher than medium*" (*HM*) – "0.5", and "*high*" (*H*) – "1") gives the following result with the optimal alternative "*high*", which corresponds to the level of the data representation marker equal to "1":

$$W_5^{global} = \begin{bmatrix} 0.09 \\ 0.09 \\ 0.15 \\ 0.30 \\ 0.37 \end{bmatrix}. \quad (11)$$

After dividing the range of marker's values into *21* intervals with a step of 0.1, the 21 alternatives were obtained. The set of the data representation marker's values is following: A = {–1.0; –0.9; –0.8; –0.7; –0.6; –0.5; –0.4; –0.3; –0.2; –0.1; 0; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9; 1.0}. The obtained alternatives are designated as *A1*, *A2*, *A3*,…, *A21.* After calculating the values of the optimality criteria, according to (3), (4), and (5), normalizing the obtained values, filling in the matrices of pairwise comparisons, and calculating the vector of global priorities in accordance with $W_i = k_{size} \times W_{size,i} + k_{availab} \times W_{availab,i} + k_{synchro} \times W_{synchro,i}$ and Barkly's formula [16], the corresponding value of the alternatives' global priorities vector, given in Table 1, was obtained.

In this case, the decrease of the interval's step and, accordingly, an increase in the accuracy reveals the alternative A18 as the best. It corresponds to the data representation marker's value = 0.7. However, the use of this approach leads to the need for further filling of the pairwise comparisons' matrix of size 21x21 by the decision-maker, i.e. requires to perform 210 operations of pairwise comparisons of alternatives for each criterion of optimality by each decision-maker. In addition, it is very difficult for a person to prioritize alternatives with insignificant changes in their qualitative characteristics.

In Fig. 3 is shown the graphical representation of the fuzzy numbers' vector. This vector was obtained for the simplified version of the problem and given in the form of the global priorities matrix (11) for the data representation marker.

DOI 10.15588/1607-3274-2022-2-10

Table 1 – Alternatives' global priorities vector for 21 intervals of the data representation marker's values

Alternative	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
Data representation marker	–1.0	–0.9	–0.8	–0.7	–0.6	–0.5	–0.4	–0.3	–0.2	–0.1	0.0
Priorities vector's element	0.008	0.012	0.009	0.013	0.01	0.012	0.009	0.01	0.013	0.021	0.034
Alternative	A12	A13	A14	A15	A16	A17	**A18**	A19	A20	A21	–
Data representation marker	0.1	0.2	0.3	0.4	0.5	0.6	**0.7**	0.8	0.9	1.0	–
Priorities vector's element	0.05	0.054	0.063	0.076	0.082	0.101	**0.117**	0.111	0.103	0.093	–

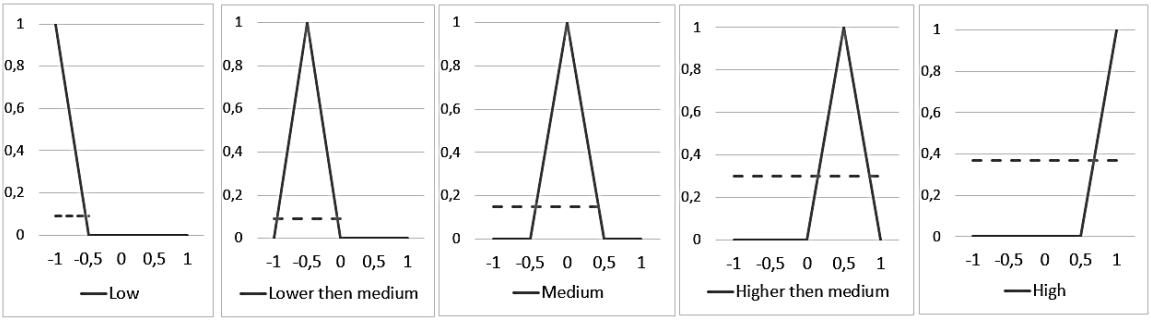

Figure 3 – Fuzzy numbers' vector of alternatives' global priorities

After performing the aggregation and defuzzification operations of the obtained results, the data representation marker's value at the DCIS node at the level of 0.7 was received (Fig. 4).

The obtained optimal solution is the same as for the case of 21 intervals of the data representation marker's values (table 1). According to the received results, it can be concluded that it is possible to use elements of fuzzy inference to increase the solution accuracy with a small number of intervals of the data representation marker's value. The "elements of fuzzy inference" mean the representation of the alternatives global priorities' vector in the form of the fuzzy numbers' vector with subsequent aggregation and defuzzification of the result.

5 RESULTS

After making sure on test solution that method works the research applies the modified method to the alternatives global priorities' vector that was got after solving the task with all available value's area restrictions and filling pairwise comparison matrix by the decision-makers. In Fig. 5 shows the result obtained previously according to the data of the alternatives global priorities' vector (10) and additionally processed using fuzzy logic elements. According to it, the optimal value of the data representation marker of the DCIS node for the work's implementation subject area of the results is obtained at the level of 0.2.

Consequently, the use of the analytic hierarchy process method based on a limited set of alternatives makes it possible to construct advantages matrices and perform the necessary calculations to obtain the values of the alternatives' global advantages. At that, in some cases with the involvement of a decision-maker, and in some cases using mathematical models of optimality criteria presented previously. Using elements of fuzzy logic, specifically the defuzzification of the fuzzy numbers' vector for the data representation marker (vector of global priorities), makes it possible to increase the accuracy of the result while determining the optimal value of the data representation marker's level.

6 DISCUSSION

To select the best alternative of the data representation marker's level the analytic hierarchy process was used. To construct matrices of the alternatives' advantages, their set was reduced to five alternatives. The matrix of the optimality criteria's advantages was obtained classically with the involvement of a decision-maker and subsequent concordance index estimation. The matrices of the alternatives' advantages were filled in automatically, without the participation of a decision-maker, based on the models of optimality criteria. Also, restrictions for the range of valid values of the optimality criteria were introduced. Were presented maximum and minimum values for each criterion, which leads to reducing the number of alternatives at the last level of the hierarchy model.

After calculating and obtaining the alternatives global priorities' vector in order to improve the accuracy of the obtained result, the apparatus of fuzzy sets was used. The obtained vector of global priorities was presented as a vector of fuzzy digits for the data representation marker with subsequent transformation to the exact value. This approach made it possible to maintain the accuracy of the obtained result while decreasing the number of solution alternatives.

For new tuples added to the database's tables after all calculations had been performed, the classification problem was formalized. This task assumes determining the belonging of the new tuple to, one of two classes – "*needed*" at a remote node and "*not needed*". After comparing the most popular approaches to solving the classification problem, the Naïve Bayes algorithm was chosen,

DOI 10.15588/1607-3274-2022-2-10

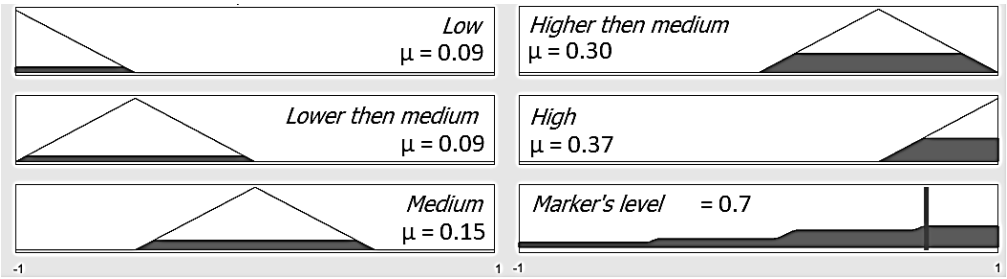

Figure 4 – Stage of aggregation and dephasification

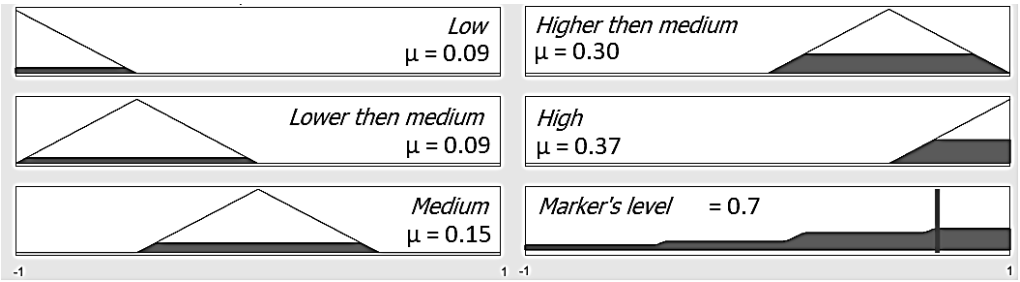

Figure 5 – Aggregation and dephasification of the obtained results

since, in the absence of transitive dependencies (the requirement for the third normal form of a relational database), all the attributes of the relation are independent. After obtaining the probability of a tuple belonging to the class "*needed*" and performing the normalization of the value, it is taken as the level of the representation marker. Accordingly, the data is loaded onto the node if this value is greater than the optimal level of the representation marker for the DCIS node.

While working on the research, the concept of a data representation marker on the DCIS node for the elements of the SQL query model was introduced. An aggregation function has been developed that allows determining the level of need for attributes and tuples in the database's relation for the DCIS node based on the statistics of SQL queries. A model of the dependence of the database structure's optimality criteria of the DCIS node on the value of the data representation marker is built. It, in contrast to existing approaches, allows determining the optimal value of the marker based on the statistics of SQL queries. Received further development method of analytic hierarchy process. The initialization of the alternatives' pairwise comparisons matrix can be performed automatically according to the obtained mathematical models. Representation of the obtained result in the form of the vector of fuzzy numbers with the reduction to the exact value allows increasing the accuracy of the obtained results.

CONCLUSIONS

While resolving the problem of optimizing the structure of the database node in the corporate information systems solved the urgent problem of determining the optimal value of the data representation marker on the DCIS node. Due to the use of fuzzy logic elements in solving the problem by the analytic hierarchy process, the accuracy of the obtained result is increased without the need to increase the number of solution alternatives.

The scientific novelty of obtained results is that the concept of data representation marker of DCIS node for dimension's elements of SQL-query model was firstly introduced with following developing of the aggregation function, presenting the model of dependence of DCIS node's database structure optimality criteria on the data representation marker's value.

The analytic hierarchy process has received further development. It happened due to the automatic initialization of the alternatives' pairwise comparison matrix according to the received mathematical models. The vector of alternatives' global priorities was given in the form of the vector of fuzzy numbers with the future reduction to the numeric value, which increased the accuracy of the obtained value.

The practical significance of obtained results is that the software supporting the decision-making in determining the representation of data on the DCIS node was developed. It allows optimizing the structure of the database node. Developed models and information technology were implemented at "Elite Building TOV" to identify the dependence of the optimal database structure's criteria and the level of data representation marker. The effect of the implementation is to increase the speed of requests' execution by 14%

Prospects for further research are to study the results of experiments of implementing the suggested modification in the analytic hierarchy process on different subject areas. It is also will be interesting to try the model on the DCIS with no central data node.

ACKNOWLEDGEMENTS

This study was funded and supported by Petro Mohyla Black Sea National University (PM BSNU) in Mykolaiv (Ukraine), and also financed in part of the PM BSNU Science-Research Work by the Ministry of Education and Science of Ukraine "Development of the modules for automatization of wireless devices for recovery of postinfarction, post-stroke patients in individual conditions of

DOI 10.15588/1607-3274-2022-2-10

remote individual rehabilitation" (State Reg. No. 0121U109898).

REFERENCES

1. Hamouda S., Zainol Z. Document-Oriented Data Schema for Relational Database Migration to NoSQL, *2017 International Conference on Big Data Innovations and Applications (Innovate-Data), Czech Republic*, 2017, pp. 43–50. DOI: 10.1109/Innovate-Data.2017.13
2. Hows D., Membrey P., Plugge E., Hawkins T. The Definitive Guide to MongoDB. Berkeley, CA, Apress, 2015, 343 p. DOI: 10.1007/978-1-4842-1182-3
3. Thakur N., Gupta N. Relational and Non Relational Databases: A Review, *Journal of University of Shanghai for Science and Technology*, 2021, Vol. 23, No. 8, pp. 117–121. DOI: 10.51201/jusst/21/08341
4. Kundu P., Arora T. Research of Persistence Solution Based on ORM and Hibernate Technology, *International Journal of Advanced Research in Computer Science and Software Engineering*, 2017, Vol. 7, No. 4, pp. 359–362. DOI: 10.23956/ijarcsse/v7i3/0154
5. Becker J., Uhr W., Vering O. Systems for the Support of the Company Management, Retail Information Systems Based on SAP Products. Berlin, Springer Berlin Heidelberg, 2013, Chapter 5, pp. 121–150. DOI: 10.1007/978-3-662-09760-1_5
6. Petrova E. Overview of modern automation information systems activities of trade enterprises, *Journal of management studies*, 2018, Vol. 4, No. 9, pp. 76–85. DOI: 10.12737/article_5d68d5afb331c1.42407139
7. Christudas B. Practical Microservices Architectural Patterns. Berkeley, CA, Apress, 2019, 812 p. DOI: 10.1007/978-1-4842-4501-9
8. Peterson C., Wilson A., Pirkelbauer P. et al. Optimized Transactional Data Structure Approach to Concurrency Control for In-Memory Databases, *2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD),* 2020, pp. 107–115. DOI: 10.1109/SBAC-PAD49847.2020.00025
9. Perez L. L., Jermaine C. M. History-aware query optimization with materialized intermediate views, *2014 IEEE 30th International Conference on Data Engineering,* 2014, pp. 520–531, DOI: 10.1109/ICDE.2014.6816678
10. Tsegelyk G. G., Krasniuk R. P. The optimization of databases replication in distributed information systems, *Information Extraction and Processing*, 2017, Vol. 45, No. 121, pp. 104–112. DOI:https://doi.org/10.15407/vidbir2017.45
11. Korniyenko B. Y., Galata L. P. Optimization of the Information System of the Corporate Network, *MCM-TECH, Kamianets-Podilskyi National Ivan Ohiienko University*, 2019, pp. 56–62. DOI: 10.32626/2308-5916.2019-19.56-62
12. Fisun M., Dvoretskyi M., Shved A. et al. Query parsing in order to optimize distributed DB structure, *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems, Technology and Applications (IDAACS), Bucharest, 2017, proceeding.* Bucharest, IEEE, 2017, pp. 172–178. DOI: 10.1109/IDAACS.2017.8095071
13. Dvoretskyi M., Dvoretska S., Nezdoliy Y. et al. Data Utility Assessment while Optimizing the Structure and Minimizing the Volume of a Distributed Database Node, *1st International Workshop on Information-Communication Technologies & Embedded Systems (ICTES), 2516, 2019, proceeding*, CEUR Workshop, 2019, pp. 128–137. Available online: http://ceur-ws.org/Vol-2516/paper10.pdf
14. Dvoretskyi M., Dvoretska S., Horban H. et al. Optimization of the database structure of a distributed corporate information system node using the analytic hierarchy process, *T&I Workshops, 2845, 2020, proceeding, CEUR Workshop*, 2020, pp. 193–203. Available online: http://ceur-ws.org/Vol-2845/Paper_19.pdf
15. Fisun M., Dvoretskiy M., Dvoretska S. Building a model to optimize the database structure of the node in corporate information systems, *Information technology and computer engineering: International Scientific and Technical Journal of Vinnytsia National Technical University*, 2020, Vol. 48, No. 2, pp. 52–60. DOI: 10.31649/1999-9941-2020-48-2-52-60
16. Zadeh L. A., Klir G. J., Yuan B. Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems, World scientific, 1996, 840 p. DOI: 10.1142/2895
17. Alang-Rashid N. K., Heger A. S. A general purpose fuzzy logic code, *IEEE International Conference on Fuzzy Systems, 1992, proceeding*, IEEE, 1992, pp. 733–742. DOI: 10.1109/FUZZY.1992.2587
18. Gozhyj A., Kalinina I., Gozhyj V. Fuzzy cognitive analy-sis and modeling of water quality, *9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, proceeding*. IEEE, 2017, pp. 289–293. DOI: 10.1109/IDAACS.2017.8095092
19. Yager R. R. On inference structures for fuzzy systems modeling, *IEEE 3rd International Fuzzy Systems Conference*. – 1994, Vol. 2, pp. 1252–1256. DOI: 10.1109/FUZZY.1994.343642
20. Nakamura K., Sakashita N., Nitta Y. et al. Fuzzy inference and fuzzy inference processor, *IEEE Micro,* 1993, Vol. 13, No. 5, pp. 37–48. DOI: 10.1109/40.238000
21. Dvoretskiy M., Dvoretska S., Davidenko E. Information technology for determining useful data while optimizing the structure and minimizing the volume of the distributed database node, *Bulletin of Cherkasy State Technological University*, 2019, No. 4, pp. 26–35. DOI: 10.24025/2306-4412.4.2019.184808
22. [Hegde R., Anusha G. V., Madival S. et al. Review on Data Mining and Machine Learning Methods for Student Scholarship Prediction, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, proceeding,* IEEE, 2021, pp. 923–927. DOI: 10.1109/ICCMC51019.2021.9418376
23. Zaki M. J., Meira W. J. Neural Networks, *Data Mining and Machine Learning. Cambridge University Press,* 2020, pp. 637–671. DOI: 10.1017/9781108564175.031
24. Graupe D. Deep Learning Neural Networks. World scientific, 2016, 280 p. DOI: 10.1142/10190
25. Janssen J., Laatz W. Naive Bayes, *Statistische Datenanalyse mit SPSS*. Springer Berlin Heidelberg, 2017, pp. 557–569. DOI: 10.1007/978-3-662-53477-9_25
26. Krishna S. Introduction to Database and Knowledge-Base Systems, World scientific, 1992, 344 p. DOI: 10.1142/1374

Received 20.02.2022.
Accepted 23.04.2022.

УДК 004.658.3

ВИКОРИСТАННЯ МЕТОДУ АНАЛІЗУ ІЄРАРХІЙ ТА ЕЛЕМЕНТІВ НЕЧІТКОЇ ЛОГІКИ ДЛЯ ОПТИМІЗАЦІЇ СТРУКТУРИ БАЗИ ДАНИХ

Дворецький М. Л. – канд. техн. наук, доцент б.в.з. кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна.

DOI 10.15588/1607-3274-2022-2-10

Савчук Т. О. – канд. техн. наук, професор, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, Вінниця, Україна.

Фісун М. Т. – д-р техн. наук, професор, професор кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна.

Дворецька С. В. – старший викладач кафедри інженерії програмного забезпечення, Чорноморський національний університет імені Петра Могили, Миколаїв, Україна.

АНОТАЦІЯ

Актуальність. Інформаційні системи дуже поширені і використовують бази даних для зберігання інформації. Для використання доступні різні моделі даних, але реляційна модель залишається популярною. Останнє десятиліття демонструє тенденцію використання розподілених баз даних під час роботи з реляційною моделлю, і цей підхід вимагає спеціально розробленого модуля для синхронізації даних усіх окремих частин БД. Оптимальна структура всіх розподілених вузлів могла б зменшити необхідність синхронізації, а швидкість доступу до даних та її актуальність залишалися б стабільними.

Метод. Автори дослідження в серії своїх попередніх робіт акцентують увагу на можливості використання зібраної історії SQL-запитів користувачів. Спочатку представлена технологія розбору запитів користувачів. Потім була розглянута ідея використання багатовимірної бази даних для аналізу запитів користувачів за зрізами типу робочої станції, програми, користувача та його посади. Також автори надали математичну модель формалізації моделі бази даних і запитів, а також критерії оптимальності структури бази даних. Дослідження продовжує наведену послідовність і намагається підвищити ефективність системи підтримки прийняття рішень шляхом введення в метод аналізу ієрархій елементів нечіткої логіки. Основна ідея підходу полягає в представленні вектору глобального пріоритету у вигляді серії нечітких множин однієї змінної з подальшим перетворенням до точного значення. Для нових кортежів, доданих до таблиць бази даних після виконання всіх обчислень, була сформульована задача класифікації.

Результати. Після розрахунку та отримання вектору глобального пріоритету альтернатив з метою підвищення точності отриманого результату було використано апарат нечітких множин. Отриманий вектор глобальних пріоритетів був представлений у вигляді вектору нечітких множин для маркера представлення даних з подальшим перетворенням до точного значення. Такий підхід дозволив зберегти точність отриманого результату при зменшенні кількості альтернатив рішення.

Висновки. Під час роботи над дослідженням було введено поняття маркера представлення даних на вузлі РКІС для елементів моделі запиту SQL. Розроблено функцію агрегації, яка на основі статистики SQL-запитів дозволяє визначити рівень необхідності атрибутів і кортежів відношення бази даних на вузлі РКІС. Побудовано модель залежності критеріїв оптимальності структури бази даних вузла РКІС від значення маркера представленості даних. Отримав подальший розвиток метод аналізу ієрархій. Ініціалізація матриці попарних порівнянь альтернатив може виконуватися автоматично відповідно до отриманих математичних моделей. Представлення отриманого результату у вигляді вектору нечітких чисел із приведенням до точного значення дозволяє підвищити точність отриманих результатів.

КЛЮЧОВІ СЛОВА: корпоративна інформаційна система, система управління базами даних, розподілена база даних, SQL-запит, реплікація даних, багатокритеріальна задача, метод аналізу ієрархій, нечітка логіка, задача класифікації, наївний алгоритм Байєса.

УДК 004.658.3

ИСПОЛЬЗОВАНИЕ МЕТОДА АНАЛИЗА ИЕРАРХИЙ И ЭЛЕМЕНТОВ НЕЧЕТКОЙ ЛОГИКИ ДЛЯ ОПТИМИЗАЦИИ СТРУКТУРЫ БАЗЫ ДАННЫХ

Дворецкий М. Л. – канд. техн. наук, и.о. доцента кафедры инженерии программного обеспечения, Черноморский национальный университет имени Петра Могилы, Николаев, Украина.

Савчук Т. А. – канд. техн. наук, профессор, профессор кафедры компьютерных наук, Винницкий национальный технический университет, Винница, Украина.

Фисун Н. Т. – д-р техн. наук, профессор, профессор кафедры инженерии программного обеспечения, Черноморский национальный университет имени Петра Могилы, Николаев, Украина.

Дворецкая С. В. – старший преподаватель кафедры инженерии программного обеспечения, Черноморский национальный университет имени Петра Могилы, Николаев, Украина.

АННОТАЦИЯ

Актуальность. Информационные системы широко распространены и используют базы данных для хранения информации. Для использования доступны разные модели данных, но реляционная модель остается популярной. Последнее десятилетие демонстрирует тенденцию использования распределенных баз данных при работе с реляционной моделью, и этот подход требует специально разработанного модуля для синхронизации данных всех отдельных частей БД. Оптимальная структура всех распределенных узлов могла бы снизить необходимость синхронизации, а скорость доступа к данным и ее актуальность оставались бы стабильными.

Метод. Авторы исследования в серии своих предыдущих работ акцентируют внимание на возможности использования собранной истории SQL запросов пользователей. Первоначально представлена технология разбора запросов пользователей. Затем была рассмотрена идея использования многомерной базы данных для анализа запросов пользователей по срезам типа рабочей станции, программы, пользователя и его должности. Также авторы предоставили математическую модель формализации модели базы данных и запросов, а также критерии оптимальности структуры базы данных. Исследование продолжает приведенную последовательность и пытается повысить эффективность системы поддержки принятия решений путем введения в метод анализа иерархий элементов нечеткой логики. Основная идея подхода заключается в представлении вектора глобального приоритета в виде серии нечетких множеств одной переменной с последующим превращением в точное значение. Для новых кортежей, добавленных в таблицы базы данных после выполнения всех вычислений, была сформулирована задача классификации.

Результаты. После расчета и получения вектора глобального приоритета альтернатив с целью повышения точности полученного результата был использован аппарат нечетких множеств. Полученный вектор глобальных приоритетов был представлен в виде вектора нечетких множеств для представления данных маркера с последующим превращением в точное значение. Такой подход позволил сохранить точность получаемого результата при уменьшении количества альтернатив решения.

Выводы. При работе над исследованием было введено понятие маркера представления данных на узле РКИС для элементов модели запроса SQL. Разработана функция агрегации, которая на основе статистики SQL-запросов позволяет определить уровень необходимости атрибутов и кортежей отношения базы данных на узле РКИС. Построена модель зависимости критериев оптимальности структуры базы данных узла РКИС от значения маркера представленности данных. Получил дальнейшее развитие метод

DOI 10.15588/1607-3274-2022-2-10

анализа иерархий. Инициализация матрицы попарных сравнений альтернатив может выполняться автоматически в соответствии с полученными математическими моделями. Представление полученного результата в виде вектора нечетких чисел с приведением к точному значению позволяет повысить точность полученных результатов.

КЛЮЧЕВЫЕ СЛОВА: корпоративная информационная система, система управления базами данных, распределенная база данных, SQL-запрос, репликация данных, многокритериальная задача, метод анализа иерархий, нечеткая логика, задача классификации, наивный алгоритм Байеса.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Hamouda S. Document-Oriented Data Schema for Relational Database Migration to NoSQL / S. Hamouda, Z. Zainol // 2017 International Conference on Big Data Innovations and Applications (Innovate-Data), Czech Republic. – 2017. – P. 43–50. DOI: 10.1109/Innovate-Data.2017.13
2. The Definitive Guide to MongoDB / [D. Hows, P. Membrey, E. Plugge, T. Hawkins]. – Berkeley, CA : Apress, 2015. – 343 p. DOI: 10.1007/978-1-4842-1182-3
3. Thakur N. Relational and Non Relational Databases: A Review / N. Thakur, N. Gupta // Journal of University of Shanghai for Science and Technology. – 2021. – Vol. 23, № 8. – P. 117–121. DOI: 10.51201/jusst/21/08341
4. Kundu P. Research of Persistence Solution Based on ORM and Hibernate Technology / P. Kundu, T. Arora // International Journal of Advanced Research in Computer Science and Software Engineering. – 2017. – Vol. 7, № 4. – P. 359–362. DOI: 10.23956/ijarcsse/v7i3/0154
5. Becker J. Systems for the Support of the Company Management, Retail Information Systems Based on SAP Products / Becker J., Uhr W., Vering O. – Berlin : Springer Berlin Heidelberg, 2013. Chapter 5. – P. 121–150. DOI: 10.1007/978-3-662-09760-1_5
6. Petrova E. Overview of modern automation information systems activities of trade enterprises / E. Petrova // Journal of management studies. – 2018. – Vol. 4, № 9. – P. 76–85. DOI: 10.12737/article_5d68d5afb331c1.42407139
7. Christudas B. Practical Microservices Architectural Patterns / B. Christudas. – Berkeley, CA : Apress, 2019. – 812 p. DOI: 10.1007/978-1-4842-4501-9
8. Optimized Transactional Data Structure Approach to Concurrency Control for In-Memory Databases / [C. Peterson, A. Wilson, P. Pirkelbauer et al.] // 2020 IEEE 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD). – 2020. – P. 107–115. DOI: 10.1109/SBAC-PAD49847.2020.00025
9. Perez L. History-aware query optimization with materialized intermediate views / L. L. Perez, C. M. Jermaine // 2014 IEEE 30th International Conference on Data Engineering. – 2014. – P. 520–531, DOI: 10.1109/ICDE.2014.6816678
10. Tsegelyk G. G. The optimization of data-bases replication in distributed information systems / G. G. Tsegelyk, R. P. Krasniuk // Information Extraction and Processing. – 2017. – Vol. 45, № 121. – P. 104–112. DOI:https://doi.org/10.15407/vidbir2017.45
11. Korniyenko B. Y. Optimization of the Information System of the Corporate Network / B. Y. Korniyenko, L. P. Galata // MCM-TECH, Kamianets-Podilskyi National Ivan Ohiienko University. – 2019. – P. 56–62. DOI: 10.32626/2308-5916.2019-19.56-62
12. Query parsing in order to optimize distributed DB structure / [M. Fisun, M. Dvoretskyi, Shved A. et al.] // 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, 2017 : proceeding. – Bucharest: IEEE, 2017. – P. 172–178. DOI: 10.1109/IDAACS.2017.8095071
13. Data Utility Assessment while Optimizing the Structure and Minimizing the Volume of a Distributed Database Node / [M. Dvoretskyi, S. Dvoretska, Y. Nezdoliy et al.] // 1st International Workshop on Information-Communication Technologies & Embedded Systems (ICTES), 2516, 2019 : proceeding. – CEUR Workshop, 2019. – P. 128–137. Available online: http://ceur-ws.org/Vol-2516/paper10.pdf
14. Optimization of the database structure of a distributed corporate information system node using the analytic hierarchy process / [M. Dvoretskyi, S. Dvoretska, H. Horban et al.] // T&I Workshops, 2845, 2020 : proceeding. – CEUR Workshop, 2020. – P. 193–203. Available online: http://ceur-ws.org/Vol-2845/Paper_19.pdf
15. Fisun M. Building a model to optimize the database structure of the node in corporate information systems / M. Fisun, M. Dvoretskiy, S. Dvoretska // Information technology and computer engineering: International Scientific and Technical Journal of Vinnytsia National Technical University. – 2020. – Vol 48, № 2. – P. 52–60. DOI: 10.31649/1999-9941-2020-48-2-52-60
16. Zadeh L. A. Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems / L. A. Zadeh, G. J. Klir, B. Yuan. – World scientific, 1996. – 840 p. DOI: 10.1142/2895
17. Alang-Rashid N. K. A general purpose fuzzy logic code / N. K. Alang-Rashid, A. S. Heger // IEEE International Conference on Fuzzy Systems, 1992 : proceeding. – IEEE, 1992. – P. 733–742. DOI: 10.1109/FUZZY.1992.2587
18. Gozhyj A. Fuzzy cognitive analysis and modeling of water quality / A. Gozhyj, I. Kalinina, V. Gozhyj // 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017 : proceeding. – IEEE, 2017. – P. 289–293. DOI: 10.1109/IDAACS.2017.8095092
19. Yager R. R. On inference structures for fuzzy systems modeling / R. R. Yager // IEEE 3rd International Fuzzy Systems Conference. – 1994. – Vol. 2. – P. 1252–1256. doi: 10.1109/FUZZY.1994.343642
20. Fuzzy inference and fuzzy inference processor / [Nakamura K., Sakashita N., Nitta Y. et al.] // IEEE Micro. – 1993. – Vol. 13, № 5. – P. 37–48. DOI: 10.1109/40.238000
21. Dvoretskiy M. Information technology for determining useful data while optimizing the structure and minimizing the volume of the distributed database node / M. Dvoretskiy, S. Dvoretska, E. Davidenko // Bulletin of Cherkasy State Technological University. – 2019. – № 4. – P. 26–35. DOI: 10.24025/2306-4412.4.2019.184808
22. Review on Data Mining and Machine Learning Methods for Student Scholarship Prediction / [R. Hegde, G. V. Anusha, S. Madival et al.] // 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021: proceeding. – IEEE, 2021. – P. 923–927. DOI: 10.1109/ICCMC51019.2021.9418376
23. Zaki M. J. Neural Networks / M. J. Zaki, W. J. Meira // Data Mining and Machine Learning. Cambridge University Press. – 2020. – P. 637–671. DOI: 10.1017/9781108564175.031
24. Graupe D. Deep Learning Neural Networks / D. Graupe. – World scientific, 2016. – 280 p. DOI: 10.1142/10190
25. Janssen J. Naive Bayes / J. Janssen, W. Laatz // Statistische Datenanalyse mit SPSS. – Springer Berlin Heidelberg, 2017. – pp. 557–569. DOI: 10.1007/978-3-662-53477-9_25
26. Krishna S. Introduction to Database and Knowledge-Base Systems / S. Krishna. – World scientific, 1992. – 344 p. DOI: 10.1142/1374

DOI 10.15588/1607-3274-2022-2-10

e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2022. № 2
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2022. № 2

UDC 004.8

DEVELOPMENT OF METHOD FOR IDENTIFICATION THE COMPUTER SYSTEM STATE BASED ON THE DECISION TREE WITH MULTI-DIMENSIONAL NODES

Gavrylenko S. Y. – Dr. Sc., Professor, Professor at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine.

Chelak V. V. – Post-graduate Student at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine.

Semenov S. G. – Dr. Sc., Professor, Head at Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine.

ABSTRACT

Context. The problem of identifying the state of a computer system is considered. The object of the research is the process of computer system state identification. The subject of the research is the methods of constructing solutions for computer system state identification.

Objective. The purpose of the work is to develop a method for decision trees learning for computer system state identification.

Method. A new method for constructing a decision tree is proposed, combining the classical model for constructing a decision tree and the density-based spatial clustering method (DBSCAN). The simulation results showed that the proposed method makes it possible to reduce the number of branches in the decision tree, which will increase the efficiency of identifying the state of the computer system. Belonging to hyperspheres is used as a criterion for decision-making, which enables to increase the identification accuracy due to the nonlinearity of the partition plane and to perform a more optimal adjustment of the classifier. The method is especially effective in the presence of initial data with high correlation coefficients, since it combines them into one or more multivariate criteria. An assessment of the accuracy and efficiency of the developed method for identifying the state of a computer system is carried out.

Results. The developed method is implemented in software and researched in solving the problem of identifying the state of the functioning of a computer system.

Conclusions. The carried out experiments have confirmed the efficiency of the proposed method, which makes it possible to recommend it for practical use in order to improve the accuracy of identifying the state of a computer system. Prospects for further research may consist in the development of an ensemble of decision trees.

KEYWORDS: computer system, abnormal state, identification, decision tree, clustering, DBSCAN algorithm, hypersphere.

ABBREVIATIONS

CS is a computer system;

OS is an operating system;

DT is a decision tree;

DBSCAN is a density-based spatial clustering of applications with noise (a data clustering algorithm).

NOMENCLATURE

X is the source data (OS events);

m is a number of the object features;

n is a number of classes in the source subset;

N_i is a number of samples of the i-th class;

N is a total number of samples in the subset;

p_k is a probability of belonging to the k-th class;

w is classifier settings;

I is information gain;

MinPts is a minimum number of neighbors for creating a cluster;

$|C|$ is a number of objects in the largest cluster;

ε is a radius of the neighborhood hypersphere;

d is a distance between objects that are clustered;

xc is a set of coordinates of the cluster center;

η is a radius of the hypersphere that bounds the cluster.

INTRODUCTION

Today, computer technology is an integral part of any state and determines its economic and political role internationally. Despite a number of promising developments in information security, the number of man-made disasters and accidents and attempts to destabilize the functioning of computer systems is increasing [1]. This is due to the imperfection of methods and means of data protection, as well as increased interest in the CS on the part of attackers. That is why the issue of CS state identification in order to spot and localize the destabilizing actions of its functioning in an increasing number of external influences is an urgent task.

The computer system is characterized by a large amount of performance criteria, which leads to difficulties in choosing the most informative criteria and the development of methods for identifying its condition under external influences [2, 3].

Researches of existing methods have revealed a number of limitations in their use [2, 4]. Thus, when the CS operates on the border between normal and abnormal states, modern methods do not always remain effective and require a long time, along with software and hardware resources, which leads to a decrease in efficiency and accuracy of its state identification [5, 6].

In addition, such tasks become especially relevant when the initial data are heterogeneous, absent or insufficient, but there are some observations in the functionality

DOI 10.15588/1607-3274-2022-2-11

of the CS, which is under the condition identification [7]. For this class of tasks, a highly effective tool is DTs and their ensembles [8–10] – a way to represent the rules in a hierarchical structure, where each object corresponds to a single node, which gives the resulting solution. A rule means a logical construction, presented in the form of if-then construct.

The object of the research is the process of computer system state identification.

DT (classification tree, regression tree) is one of the methods of automatic data analysis. DTs allow you to find repetitive patterns in the data, and to perform training to recognize patterns. The fundamental work that gave impetus to the development of this area was the book by E.B. Hunt, J. Marin and P.J. Stone in "Experiments in Induction", which was published in 1966. DT has a number of advantages [4, 5], namely: easy to understand and interpret, able to work with both numerical and categorical data, requires little data preparation, uses a white box model and is easily explained by Boolean logic. The correctness of the model can be verified by statistical tests, which makes it possible to verify its reliability. In addition, during the construction of DT, less informative features will be used to a lesser extent, which makes it possible to either remove them from subsequent runs or use special algorithms for taking into account less informative features [11].

However, DTs have a number of disadvantages [4, 5]. The problem of constructing an optimal DT is NP-complete in terms of some aspects of optimality even for simple tasks. The DT construction algorithm is based on a greedy algorithm, where the only optimal solution is selected locally in each node, which cannot ensure the optimality of the whole tree. DT also has a high sensitivity to noise and changes in the source data, which can lead to the construction of a completely different DT, even with small changes in the source data.

The subject of the research is the methods of constructing solutions for computer system state identification.

There are various methods of constructing DTs, the process of which is a consistent, recursive division of the learning set into subsets using the decision rules in the nodes [12]. The process of partitioning continues until all the nodes at the end of all the branches are declared as leaves. At the same time, when constructing a DT, partitioning in nodes forms rectangular clusters in the feature space, the shape of which may not coincide with the shape of real clusters, which leads to a decrease in the accuracy of decision-making. This is especially important when the functioning of the CS lies on the verge of distinguishing between normal and abnormal states, is characterized by highly correlated data, or is presented by fuzzy data that requires the development of new models of DT construction [6, 13–16].

The purpose of the work is to develop a method for decision trees learning for computer system state identification.

1 PROBLEM STATEMENT

We will assume that the functioning of a CS is characterized by the set of its performance criteria $X = \{x_{i1}, x_{i2}, ..., x_{im}\}$. Input data is the set of marked pairs $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the CS state criteria set and y_i is a classifying label. There exists an unknown fitness function – a mapping $\boldsymbol{f : X \rightarrow Y}$ the values of which are only known for a finite set of training samples $(X,Y)=\{(x_1, y_1),\ldots,(x_m,y_m)\}$. The structure of a DT $\boldsymbol{f}$, must be formed, which should be able to classify an arbitrary object $x \in X$ and adjust its parameter w: $F(f(w, x), y) \rightarrow opt$.

2 REVIEW OF THE LITERATURE

Most popular decision tree learning algorithms are based on the divide-and-conquer principle [17].

During the construction of the DT, it is necessary to solve several key problems, each of which is associated with the corresponding step of the learning process:

1) Choice of the partition attribute for a given node.
2) Choice of termination criteria for learning.
3) Choice of decision tree pruning method.
4) Assessment of the accuracy of the constructed tree.

Currently, a significant number of algorithms have been developed for choosing the next partition attribute (DT learning algorithms): ID3, CART, C4.5, C5.0, Newld, ITrule, CHAID, CN2, etc. The most widespread and popular are the following algorithms:

1) ID3 (Iterative Dichotomized) – the algorithm can only use a discrete target variable, so DTs that are built using this algorithm are classifiers [18, 19]. Attribute choice is based on information gain:

$$I = -\sum_{i=1}^{p} \frac{N_i}{N} \log\left(\frac{N_i}{N}\right),$$

or based on Gini impurity:

$$I = \sum_{k=1}^{n} p_k (1 - p_k).$$

For this algorithm the number of children of a tree node is not limited. The algorithm does not support training samples with incomplete data.

2) C4.5 – an improved version of ID3, which adds the ability to work with missing data values. Attribute choice is based on information gain [19,20].

3) CART (Classification and Regression Tree) – a decision tree learning algorithm, which allows the use of both discrete and continuous target variables, i.e. it can solve both classification and regression problems. The algorithm builds trees that have only two children in each node, i.e. it builds a binary DT. Works slowly on large input data with lots of noise [21, 22].

4) Chi-square automatic interaction detection (CHAID). Performs multiway splits during the DT classification calculation DT.

DOI 10.15588/1607-3274-2022-2-11

5) MARS*:* extends DT to improve digital data processing.

No algorithm for constructing a DT can a priori be considered the best or perfect. Feasibility of a particular algorithm should be verified and confirmed experimentally.

Since CS is characterized by a large number of performance criteria, the significance of which is uncertain and which correlate with each other and can characterize the CS state as being in between normal and abnormal states, it is necessary to improve existing or develop new methods of identifying the CS state.

3 MATERIALS AND METHODS

In this study, in accordance with the problem statement, a DT construction method was developed, which differs from the known methods by combining the classical model of DT construction based on the *C4.5* algorithm with the DBSCAN method.

The DBSCAN algorithm was proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996, as a solution to the problem of partitioning data into clusters of arbitrary shape.

The algorithm is based on the idea that inside each cluster the density of objects is significantly higher than outside, and that the density in areas with noise is lower than the density of any of the clusters.

The algorithm requires two parameters: *MinPts* – the minimum number of neighbors for creating a cluster; ε – the radius of the neighborhood hypersphere.

The first step of the algorithm is to compute a matrix of distances *d* between objects that are being clustered, either using the squared Euclidean distances:

$$d(x_i, x_j) = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2,$$

or the Manhattan distances:

$$d(x_i, x_j) = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|.$$

In the next step, a neighbor matrix is computed using the distance matrix, each element in which determines whether the object x_i is a neighbor of x_j:

$$Neighbour_{ij} = \begin{cases} 0, & d(x_i, x_j) > \varepsilon \\ 1, & d(x_i, x_j) \le \varepsilon \end{cases},$$
$$i = \overline{1..N},\ j = \overline{1...N}.$$

Subsequently, clusters are formed based on the neighbor matrix. Initially, all objects are considered undetermined. The clustering procedure is iterative, and starts with an arbitrary object x_i which has not been determined yet. For a given object x_i, a list of neighbors is created, which contains all objects x_j, that have the corresponding element of the *i*-th row of the neighbor matrix set to one. The number of neighbors *K* is counted and compared with *MinPts*. When the count of neighbors is less than *MinPts*, the object is labeled as unclustered, and the next arbitrary undetermined object x_i. is processed. When the count of neighbors is greater or equals to *MinPts*, the current object x_i is considered to be a core object. Objects x_j, which were included in the list are considered reachable in terms of density (also core objects). The current object x_i and its neighbors x_j form a new cluster and are labeled with its number. Next, the iterative process of finding new neighbors is started. Objects that are either undetermined or unclustered are analyzed, and those that are reachable for the x_i objects of the cluster (have the corresponding element of the *j*-th row of the neighbor matrix set to one), are added to the cluster. The iterative process of joining new neighbors is repeated until no more objects can be added to the cluster.

The process of forming new clusters is repeated until all objects are determined. Objects that were labeled as unclustered and were not subsequently placed into a cluster are considered noise and remain unused.

The following procedure of finding the decision parameter η for a multidimensional DT node is developed:

– the cluster with the maximum number of elements *C* is found:

$$C = \max_{A_i \in A}(|A_i|),$$

where $|A_i|$ – is the number of elements in the *i*-th cluster;

– The center of the cluster *xc* is found using each feature of the x_i object:

$$xc_k = \frac{\sum_{i=1}^{|C|} x_{ik}}{|C|}.$$

After obtaining the center of the cluster, η is defined to be the maximum distance from the object to the centers:

$$\eta = \max_{x_i \in C}(d(xc, x_i)).$$

The value of η is the radius of the hypersphere that bounds the cluster and is further used as a decision parameter for the multidimensional DT node.

The process of constructing a DT is as follows. The source data of the DT are the indicators of the functioning of the CS (CPU load, memory load, network traffic, number of read/write operations to disk, intrusion signatures; statistical data based on system events: number of operations with the registry or file system, number of processes, etc.). The source data is divided into clusters using the DBSCAN algorithm. For example, when identifying the CS state, a singular multidimensional criterion can combine features representing the load of individual CPU

DOI 10.15588/1607-3274-2022-2-11

cores. Each of clusters can be further considered as a multidimensional criterion in the construction of the next DT node.

Further process of constructing a DT consists of sequential, iterative division of the learning set into subsets using the decision rules in the nodes. When a given DT node is formed, the feature that gives the best partitioning out of the whole set of features is selected as the partition feature, providing the maximum entropy reduction of the resulting subset relative to the parent. The feature can be either one-dimensional or multidimensional. If the partition feature is one-dimensional, the partition criterion is a comparison with a given threshold value. If the partition feature is multidimensional, the partition criterion belongs to a hypersphere of a given radius η.

Fig. 1 shows a construction of a tree with multidimensional decision nodes. Fig. 2 shows a construction of a decision tree which uses a one-dimensional feature and two features, combined into a single two-dimensional criterion.

Thus, the method of constructing DT can be formulated as follows:

1. To form a training sample of labeled data <x, y>.

2. To divide the source data into clusters using the DBSCAN algorithm

3. Determine the termination criteria of DT construction to avoid overlearning.

To do this, we considered the following approaches:

– Early termination – the algorithm will be aborted as soon as the specified value of a criterion is reached, such as the percentage of correctly recognized samples. The advantage of the approach is the reduction of training time and reduction of variance error, and the disadvantage is the reduction of the accuracy of DT classification;

– Limiting the tree depth – the establishment of the maximum number of partitions in the branches, after which the training stops. This method also leads to a decrease in the accuracy of DT classification;

– Establishment of the minimum admissible number of leaves in a node, which will allow to avoid creation of trivial splits and, consequently, insignificant rules.

4. Determine the information gain Ii of all one-dimensional and multidimensional features in relation to the result value y_i. and select the attribute that will be partitioned in this node.

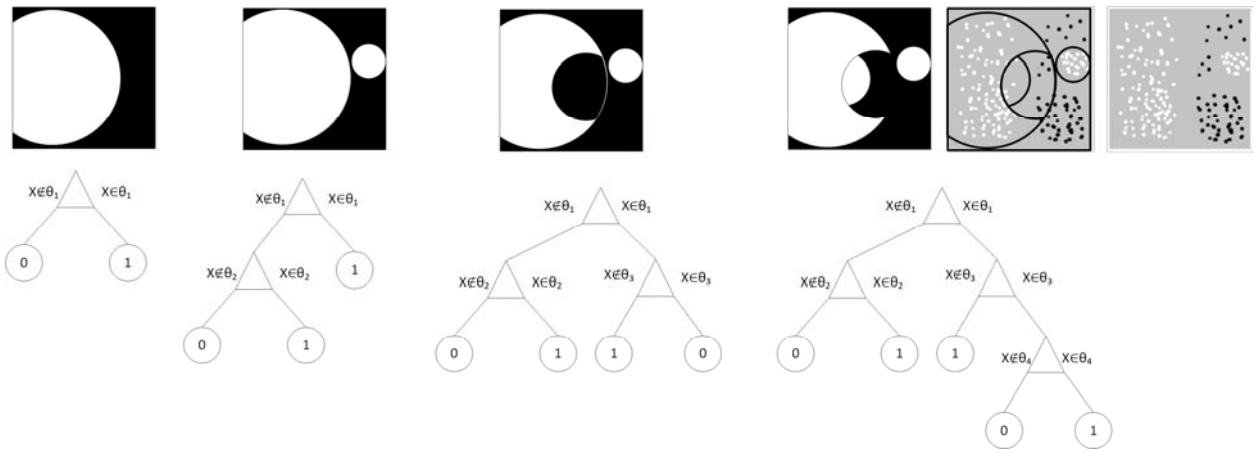

Figure 1 – Example of constructing a tree with multidimensional decision nodes

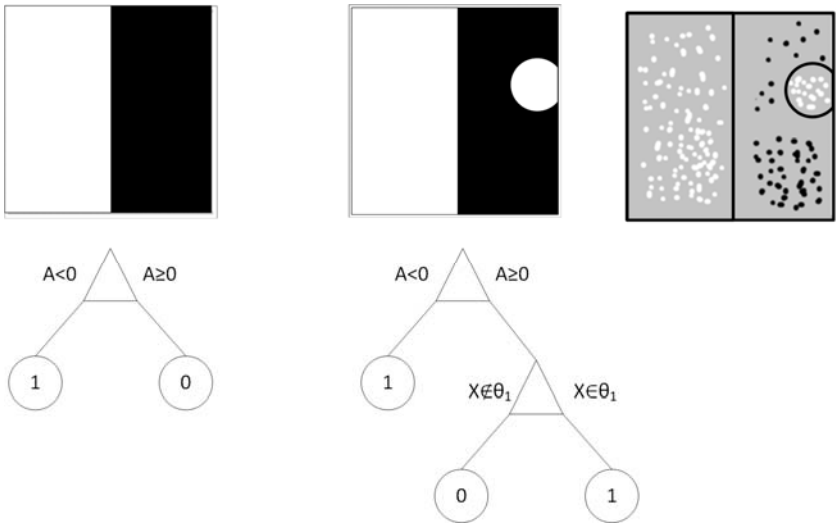

Figure 2 – Example of constructing a tree with multidimensional and one-dimensional nodes

DOI 10.15588/1607-3274-2022-2-11

5. Construct the current DT node based on the selected feature and form leaves with the appropriate set of samples.

6. Check the defined tree termination criteria. Complete the DT construction procedure if at least one of the termination criteria meets the requirements, or return to step 4.

7. If necessary, prune the DT, using the following algorithm:

– Identify two indicators: the relative accuracy of the model (the ratio of the number of correctly recognized samples to the total number of samples) and the absolute error value (the number of incorrectly classified samples);

Remove leaves and nodes from the tree, the cutting of which will not significantly reduce the accuracy of the model or increase the error. Cutting branches, carried out from bottom to top, by successively transforming the nodes into leaves.

4 EXPERIMENTS

According to the proposed algorithm, the DT is constructed (Fig. 3). A comparative analysis of classification accuracy was performed. The following DT construction algorithms have been examined as DT-based classification methods: Fine Tree, Medium Tree, Coarse Tree. Classifiers based on support-vector machines (SVM) and *k*-nearest neighbors (KNN) were also considered. Table 1 shows a comparative estimate of the classification error in the training set (Bias) and the test set (Variance).

As can be seen from Table 1, the method of DT construction, which combines the classical model of DT construction and density-based method of spatial clustering, makes it possible to achieve the accuracy of up to 100% for the training set, while the classification error on the test data set does not exceed 9.1%.

The evaluation of the performance of the classifier based on the proposed method in comparison with previously known methods is shown in Fig. 4. As can be seen from the figure, the proposed method leads to an increase in the efficiency of identification of the state of the CS by 50% compared to the Medium Tree method, which was shown to be the most efficient in previous studies [23].

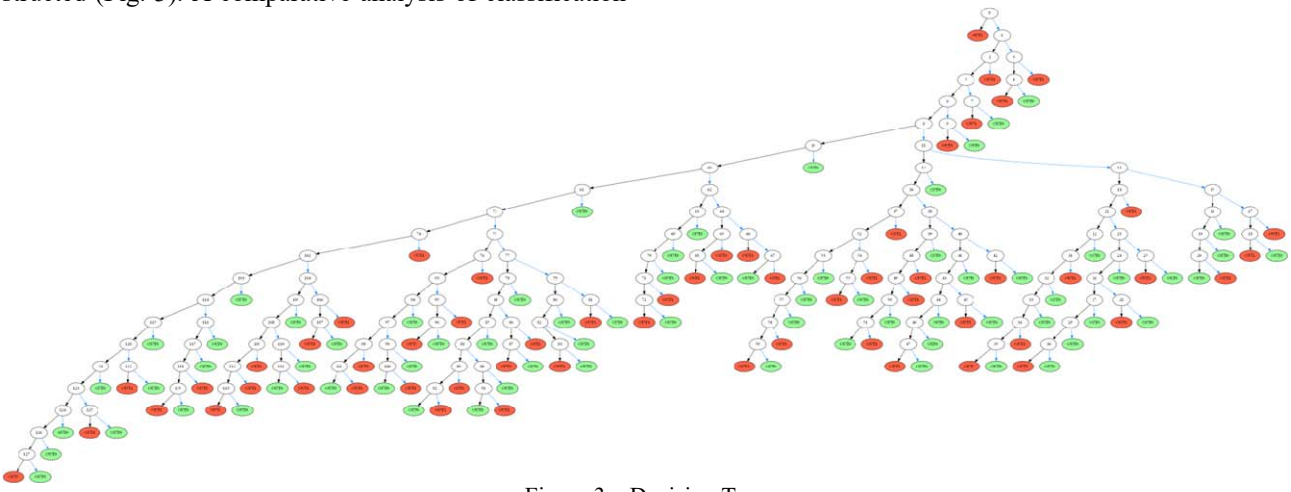

Figure 3 – Decision Tree

Table 1 – Assessment of classification accuracy

Method	Bias, %	Variance %	Method	Bias, %	Variance %
Fine Tree	0.13	31.97	Fine KNN	0	8.63
Medium Tree	0.13	35.03	Medium KNN	0.03	19.57
Coarse Tree	0.23	46.87	Coarse KNN	0.37	45.7
Decision tree with multi-dimensional nodes	0	9.1	Cosine KNN	0.1	28.77
Linear SVM	0.87	33.73	Cubic KNN	0.03	43.73
Quadratic SVM	0.07	48	Weighted KNN	0.03	10.97
Fine Gaussian SVM	0.17	28.87	Subspace Discriminant	3.47	56.43
Medium Gaussian SVM	0.03	42.1	Subspace KNN	0	10.37
Coarse Gaussian SVM	1.87	43.13			

DOI 10.15588/1607-3274-2022-2-11

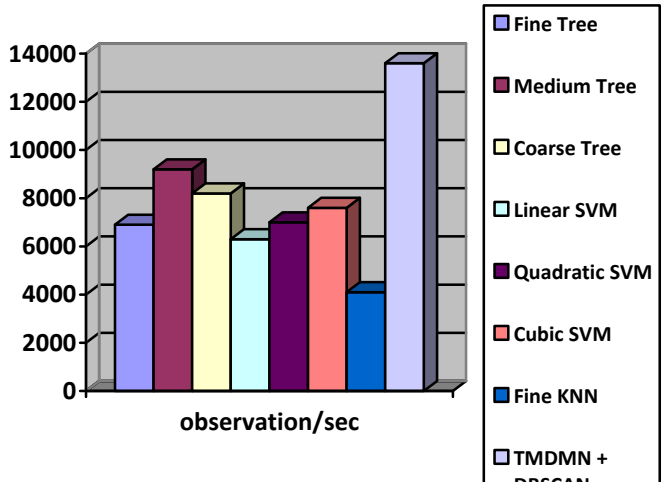

Figure 4 – Performance of CS state identification

5 RESULTS

The simulation results have shown that the proposed method makes it possible to reduce the number of branches in the DT, which leads to an increase in the efficiency of CS state identification by 50%. The use of belonging to the hyperspheres as a decision criterion makes it possible to increase the accuracy of identification (and achieve classification error rates as low as 0% on the training set, and 9.1% on the test data set) due to the nonlinearity of the partitioning plane. Furthermore, a larger set of hyperparameters allows for a more optimal and flexible fine-tuning of the classifier. This method is especially effective when used with source data samples that have high correlation coefficients, as it combines them into one or more multidimensional criteria. The disadvantage of this method is the increase in the training time of the classifier and a slight increase in the amount of resources needed for storage of the obtained models.

6 DISCUSSION

A number of limitations in the use of existing methods have been identified while solving problems related to the identification and protection of computer systems. Thus, anomalies caused by intrusions into the CS with unidentified or fuzzy properties, given the large number of parameters of the functioning of the CS, can not always be identified, which leads to increased damage as a result of cyberattacks.

That is why the conducted research has led us to a method of CS state identification based on DTs. The conducted experiments have allowed us to assess the accuracy and efficiency of the CS state identification, the practical significance and prospects of further research.

CONCLUSIONS

Hence, the problem of increasing the efficiency and accuracy of CS state identification is solved in this work.

The scientific novelty of the obtained results is that for the first time a method of identifying the CS state based on DTs is proposed, which differs from the known methods of DT construction by combining the classical model of tree construction with the DBSCAN method.

The initial data of the DT are CS performance criteria, processed by a special algorithm, namely: criteria that are highly correlated are combined into one or more multidimensional criteria.

A comparative analysis of the accuracy of the proposed algorithm for constructing DTs and the following algorithms: Fine Tree, Medium Tree, Coarse Tree. In addition, classifiers based on SVM and KNN methods were studied.

The simulation results showed that the proposed method makes it possible to reduce the number of branches in the DT, which leads to an increase in the efficiency of CS state identification by 50%. The use of belonging to the hyperspheres as a decision criterion makes it possible to increase the accuracy of identification (and achieve classification error rates as low as 0% on the training set, and 9.1% on the test data set) due to the nonlinearity of the partition plane. In addition, the presence of more hyperparameters allows for a more optimal fine-tuning of the classifier. This method is especially effective when used with source data samples that have high correlation coefficients, as it combines them into one or more multidimensional criteria. The disadvantage of this method is the increase in training time of the classifier. The method also requires more memory.

The practical significance lies in the fact that the developed method is implemented in software and has been researched while solving the problem of CS state identification.

The experiments confirmed the efficiency of the proposed method, which makes it possible to recommend it for practical use as a state identification method.

Prospects for further research may consist of development of an ensemble of trees with multidimensional decision nodes.

DOI 10.15588/1607-3274-2022-2-11

ACKNOWLEDGEMENTS

This work is supported by National Technical University "Kharkiv Polytechnic Institute" in the field of research "Models and methods of processing and protection of information in computer systems" (№60028).

REFERENCES

1. Daniel Schatz, Bashroush Rabih, Wall Julie. Towards a More Representative Definition of Cyber Security, *The Association of Digital Forensics, Security and Law* (ADFSL), 2017, Vol. 12, No. 2, pp. 53–74. DOI: 10.15394/jdfsl. 2017.1476
2. Farooq Anjum and Petros Mouchtaris. Intrusion Detection Systems, *Security for Wireless Ad Hoc Networks*. Wiley, 2007, pp. 120–159. DOI: 10.1002/ 9780470118474.ch5.
3. Kelleher J., Namee B., Archi A. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case, Dublin: The MIT Press, 2015, 642 p.
4. Iqbal H. Sarker, Shahriar Badsha, Hamed Alqahtani, Paul Watters, Alex Ng. Cybersecurity data science: an overview from machine learning perspective, *Journal of Big Data*, 2020, Vol. 7 (41) pp. 1–29. DOI: 10.1186/s40537-020-00318-5
5. Xavier Larriva-Novo, Mario Vega-Barbas, Victor A. Villagra, Diego Rivera, Manuel Alvarez-Campana, Julio Berrocal. Efficient Distributed Preprocessing Model for Machine Learning-Based Anomaly Detection over Large-Scale Cybersecurity Datasets, *Applied Sciences*, 2020, Vol. 10, pp. 30–34. DOI: 10.3390/app10103430
6. Gavrylenko S., Semenov S., Sira O., Kuchuk N. Identification of the state of an object under conditions of fuzzy input data, *Eastern-European Journal of Enterprise Technologies,* 2019, Vol. 1, No. 4 (97), pp. 22–29. DOI: 10.15587/1729-4061.2019.157085
7. Alpaydin E. Introduction to Machine learning, London: The MIT Press, 2010, 400 p.
8. Kaminski B., Jakubczyk M., Szufel P. A framework for sensitivity analysis of decision trees, *Central European Journal of Operations Research,* 2018, Vol. 26, pp. 135–159. DOI: 10.1007/s10100-017-0479-6
9. Gavrylenko S., Sheverdin I, Kazarinov M. The ensemble method development of classification of the computer system state based on decision trees, *Advanced Information Systems*, Vol. 4, No. 3, pp. 5–10. DOI:10.20998/2522-9052.2020.3.01
10. SubbotIn S. Podannya y obrobka znan u sistemah shtuchnogo Intelektu ta pidtrimki priynyattya, Zaporizhzhya, ZNTU, 2008, 341 p.
11. Subbotin S. O. Postroenie derevev resheniy dlya sluchaya maloinformativnyih, *Radio Electronics, Computer Science, Contro*l, 2019, No. 1, pp. 122–130. DOI: 10.15588/1607-3274-2019-1-12
12. Mitrofanov S., E. Semenkin. An Approach to Training Decision Trees with the Relearning of Nodes, *International Conference on Information Technologies* (*InfoTech*), 2021, pp. 1–5, DOI: 10.1109/InfoTech52438.2021.9548520
13. Wang S., Wang, L., Jiang, C. Adapting naive Bayes tree classification, *Knowledge and Information system*, 2015, Vol. 44, No. 1, pp. 77–89. DOI: 10.1007/s10115-014-0746-y
14. Kornienko Y., Borisov A. A hybrid algorithm for decision tree generation, *International Scientific Journal of Computing*, 2004, Vol. 3, Issue 3, pp. 51–57. DOI: 10.47839/ijc.3.3.305
15. Irad Ben-Gal, Alexandra Dana, Niv Shkolnik, Gonen Singer. Efficient Construction of Decision Trees by the Dual Information Distance Method, *Quality Technology & Quantitative Management,* 2014, Vol. 11, No. 1, pp. 133–147. DOI: 10.1080/16843703.2014. 11673330
16. Geurts P., Ernst D., Wehenkel L. Extremely randomized trees, *Machine Learning*, 2006, Vol. 63, No. 1, pp. 3–42. DOI: 10.1007/s10994-006-6226-1
17. Kesinee Boonchuay. Krung Sinapiromsaran, Chidchanok Lursinsap Boundary expansion algorithm of a decision tree induction for an imbalanced dataset, *Songklanakarin Journal of Science and Technology* (SJST), 2017, Vol. 39, No. 5, pp. 665–673. DOI: 10.14456/sjst-psu.2017.82
18. Quinlan J. R. Induction of Decision Trees, *Machine Learning*, 1986, No. 1, pp. 81–106.
19. Hssina B.,Merbouha A., Ezzikouri H., Erritali M. comparative study of decision tree ID3 and C4.5, *International Journal of Advanced Computer Science and Applications,* 2014, Vol. 4(2), pp. 13–19. DOI: 10.14569/SpecialIssue.2014.040203
20. Idris Mochamad, Mustafid, Suseno Jatmiko Endro. Implementation of C4.5 Algorithm and Forward Chaining Method for Higher Education Performance Analysis, *The 4th International Conference on Energy, Environment, Epidemiology and Information System,* 2019, Vol. 125. DOI: 10.1051/e3sconf/ 201912521002
21. Painsky A., Rosset S. Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2017, Vol. 39, No. 11, pp. 2142–2153. DOI: 10.1109/TPAMI.2016.2636831.
22. Maddeh M., Ayouni S., Alyahya S., Hajjej F. Decision tree-based Design Defects Detection, *IEEE Access,* 2021, Vol. 9, pp. 71606–71614. DOI: 10.1109/ACCESS.2021.3078724.
23. Gavrylenko S., Chelak V., Hornostal O. Ensemble approach based on bagging and boosting for Identification the Computer System State, Proceedings of the 31th International Scientific Symposium Metrology and Metrology Assurance.–Sozopol, Bulgaria IEEE Access, 2021. DOI:10.1109/ MMA52675.2021.9610949

Received 03.12.2021.
Accepted 10.12.2021.

УДК 004.8

РОЗРОБКА МЕТОДУ ІДЕНТИФІКАЦІЇ СТАНУ КОМП'ЮТЕРНОЇ СИСТЕМИ НА ОСНОВІ ДЕРЕВА РІШЕНЬ З БАГАТОВИМІРНИМИ ВУЗЛАМИ

Гавриленко С. Ю. – д-р техн. наук, професор, професор кафедри «Обчислювальна техніка та програмування», Національний технічний університет «Харківський політехнічний інститут», Харків, Україна.

Челак В. В. – аспірант кафедри «Обчислювальна техніка та програмування», Національний технічний університет «Харківський політехнічний інститут», Харків, Україна.

DOI 10.15588/1607-3274-2022-2-11

Семенов С. Г. – д-р техн. наук, професор, завідувач кафедри «Обчислювальна техніка та програмування», Національний технічний університет «Харківський політехнічний інститут», Харків, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу ідентифікації стану комп'ютерної системи. Об'єктом дослідження є процес ідентифікації стану комп'ютерної системи. Предметом дослідження є методи побудови дерев рішень для ідентифікації стану КС

Мета. Розробка методу побудови дерев рішень для ідентифікації стану комп'ютерної системи.

Метод. Запропоновано новий метод побудови дерева рішень, який поєднує класичну модель побудови дерева рішень та оснований на щільності метод просторової кластеризації (DBSCAN). Результати моделювання показали, що запропонований метод надає можливість зменшити кількість розгалужень в дереві рішень, що дозволяє підвищити оперативність ідентифікації стану комп'ютерної системи. Використання приналежності до гіперсфер у якості критерію прийняття рішень, надає можливість підвищити точність ідентифікації за рахунок нелінійності площині розбиття та виконати більш оптимальне налаштування класифікатору. Метод є особливо ефективним за наявності вихідних даних, які мають високі кореляційні коефіцієнти, так як поєднує їх в один або декілька багатомірних критеріїв. Проведено оцінку точності та оперативності розробленого методу ідентифікації стану комп'ютерної системи.

Результати. Розроблений метод реалізований програмно і досліджений під час розв'язання задачі ідентифікації стану функціонування комп'ютерної системи.

Висновки. Проведені експерименти підтвердили працездатність запропонованого методу, що надає можливість рекомендувати його для практичного використання з метою підвищення точності ідентифікації стану комп'ютерної системи. Перспективи подальших досліджень можуть полягати в розробці ансамблю дерев рішень.

КЛЮЧОВІ СЛОВА: комп'ютерна система, аномальний стан, ідентифікація, дерево рішень, кластеризація, алгоритм DBSCAN, гіперсфера.

УДК 004.8

РАЗРАБОТКА МЕТОДА ИДЕНТИФИКАЦИИ СОСТОЯНИЯ КОМПЬЮТЕРНОЙ СИСТЕМЫ НА ОСНОВЕ ДЕРЕВА РЕШЕНИЙ С МНОГОМЕРНЫМИ УЗЛАМИ

Гавриленко С. Ю. – д-р техн. наук, профессор, профессор кафедры «Вычислительная техника и программирование», Национальный технический университет «Харьковский политехнический институт», Харьков, Украина.

Челак В. В. – аспирант кафедры «Вычислительная техника и программирование», Национальный технический университет «Харьковский политехнический институт», Харьков, Украина.

Семенов С. Г. – д-р техн. наук, профессор, заведующий кафедры «Вычислительная техника и программирование», Национальный технический университет «Харьковский политехнический институт», Харьков, Украина.

АННОТАЦИЯ

Актуальность. Рассмотрена задача идентификации состояния компьютерной системы. Объектом исследования является процесс идентификации состояния компьютерной системы. Предметом исследования являются методы построения решений для идентификации состояния КС

Цель. Разработка метода построения деревьев решений для идентификации состояния компьютерной системы.

Метод. Предложен новый метод построения дерева решений, сочетающий классическую модель построения дерева решений и основанный на плотности метод пространственной кластеризации (DBSCAN). Результаты моделирования показали, что предложенный метод позволяет уменьшить количество ветвлений в дереве решений, что позволит повысить оперативность идентификации состояния компьютерной системы. Использование принадлежности к гиперсферам в качестве критерия принятия решений позволяет повысить точность идентификации за счет нелинейности плоскости разбиения и выполнить более оптимальную настройку классификатора. Метод особенно эффективен при наличии исходных данных, имеющих высокие корреляционные коэффициенты, так как объединяет их в один или несколько многомерных критериев. Проведена оценка точности и оперативности разработанного метода идентификации состояния компьютерной системы.

Результати. Разработанный метод реализован в виде программного обеспечения и исследован при решении задачи идентификации состояния функционирования компьютерной системы.

Выводы. Проведенные эксперименты подтвердили работоспособность предлагаемого метода, что позволяет рекомендовать его для практического использования с целью повышения точности идентификации состояния компьютерной системы. Перспективы дальнейших исследований могут состоять в разработке ансамбля деревьев решений.

КЛЮЧЕВЫЕ СЛОВА: компьютерная система, аномальное состояние, идентификация, дерево решений, кластеризация, алгоритм DBSCAN, гиперсфера.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Daniel Schatz. Towards a More Representative Definition of Cyber Security / Schatz Daniel, Bashroush Rabih, Wall Julie // The Association of Digital Forensics, Security and Law (ADFSL). – 2017. – Vol. 12, No. 2. – P. 53–74. DOI: 10.15394/jdfsl.2017.1476
2. Farooq Anjum. Intrusion Detection Systems / Farooq Anjum and Petros Mouchtaris // Security for Wireless Ad Hoc Networks. – Wiley, 2007. – P. 120–159. DOI: 10.1002/9780470118474.ch5
3. Kelleher J. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies / J. Kelleher, B. Namee, A. Archi. – Dublin : The MIT Press, 2015. – 642 p.
4. Cybersecurity data science: an overview from machine learning perspective / [Iqbal H. Sarker, A. S. M. Kayes,

DOI 10.15588/1607-3274-2022-2-11

Shahriar Badsha et al.] // Journal of Big Data. – 2020. – Vol. 7 (41). – 29 p. DOI: 10.1186/s40537-020-00318-5
5. Xavier Larriva-Novo. Efficient Distributed Preprocessing Model for Machine Learning-Based Anomaly Detection over Large-Scale Cybersecurity Datasets / [Xavier Larriva-Novo, Mario Vega-Barbas, Victor A. Villagra et al.] // Applied Sciences. – 2020. – Vol. 10. – 19 p. DOI: 10.3390/app10103430
6. Identification of the state of an object under conditions of fuzzy input data / [S. Semenov, O. Sira, S. Gavrylenko, N. Kuchuk] // Eastern-European Journal of Enterprise Technologies. – 2019. – Vol. 1, No. 4 (97). – P. 22–29. DOI: 10.15587/1729-4061.2019.157085
7. Alpaydin E. Introduction to Machine learning / E. Alpaydin. – London : The MIT Press, 2010. – 400 p.
8. Bogumil Kaminski. A framework for sensitivity analysis of decision trees / B. Kaminski, M. Jakubczyk, P. Szufel // Central European Journal of Operations Research. – 2018, Vol. 26. – P. 135–159 DOI: 10.1007/s10100-017-0479-6
9. Gavrylenko S. The ensemble method development of classification of the computer system state based on decision trees / S. Gavrylenko, I. Sheverdin, M. Kazarinov // Advanced Information Systems. – 2020. – P. 5–10. DOI:10.20998/2522-9052.2020.3.01
10. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень / С. О. Субботін. – Запоріжжя : ЗНТУ, 2008. – 341 с.
11. Субботин С. А. Построение деревьев решений для случая малоинформативных признаков / С. О. Субботин // Радиоэлектроника, информатика, управление. – 2019. – № 1. – С. 122–130. DOI: 10.15588/1607-3274-2019-1-12
12. Sergei Mitrofanov. An Approach to Training Decision Trees with the Relearning of Nodes / S. Mitrofanov and E. Semenkin // 2021 International Conference on Information Technologies (InfoTech). – 2021 .– P. 1–5. DOI: 10.1109/Info Tech52438.2021.9548520
13. Wang S. Adapting naive Bayes tree classification / S. Wang, L. Jiang, C. Li // Knowledge and Information system. – 2015. – Vol. 44, № 1. – P. 77–89. DOI: 10.1007/s10115-014-0746-y
14. Kornienko Y. A hybrid algorithm for decision tree generation / Y. Kornienko, A. Borisov // International Scientific Journal of Computing. – 2004. – Vol. 3, Issue 3. – P. 51–57. DOI: 10.47839 /ijc.3.3.305
15. Efficient Construction of Decision Trees by the Dual Information Distance Method / [Irad Ben-Gal, Alexandra Dana, Niv Shkolnik, Gonen Singer] // Quality Technology & Quantitative Management. – 2014. – Vol. 11, № 1. – P. 133–147. DOI: 10.1080/16843703.2014.11673330
16. Geurts P. Extremely randomized trees / P. Geurts, D. Ernst, L. Wehenkel // Machine Learning. – 2006. – Vol. 63, No. 1. – P. 3–42. DOI:10.1007/s10994-006-6226-1
17. Kesinee Boonchuay. Boundary expansion algorithm of a decision tree induction for an imbalanced dataset / Kesinee Boonchuay, Krung Sinapiromsaran, Chidchanok Lursinsap // Songklanakarin Journal of Science and Technology (SJST). – 2017. – Vol. 39, No. 5. – P. 665–673. DOI: 10.14456/sjst-psu.2017.82
18. Quinlan J. R. Induction of Decision Trees Machine Learning. / J. R. Quinlan // Kluwer Academic Publishers. – 1986. – № 1. – P. 81–106.
19. A comparative study of decision tree ID3 and C4.5 / [B. Hssina, A. Merbouha, H. Ezzikouri, M. Erritali] // International Journal of Advanced Computer Science and Applications. – 2014. – Vol. 4 (2). – P. 13–19. DOI: 10.14569/SpecialIssue. 2014.040203
20. Idris Mochamad. Implementation of C4.5 Algorithm and Forward Chaining Method for Higher Education Performance Analysis / Idris Mochamad, Mustafid, Suseno Jatmiko Endro // The 4th International Conference on Energy, Environment, Epidemiology and Information System (ICENIS 2019). – 2019. – Vol. 125. DOI: 10.1051/e3sconf/201912521002
21. Painsky A. Cross-Validated Variable Selection in Tree-Based Methods Improves Predictive Performance / A. Painsky, S. Rosset // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39, No. 11. – P. 2142–2153. DOI: 10.1109/TPAMI.2016.2636831.
22. Decision tree-based Design Defects Detection / [M. Maddeh, S. Ayouni, S. Alyahya and F. Hajjej] // IEEE Access. – 2021. – Vol. 9. – P. 71606–71614. DOI: 10.1109/ACCESS.2021. 3078724.
23. Gavrylenko S. Ensemble approach based on bagging and boosting for Identification the Computer System State / S. Gavrylenko, V. Chelak, O. Hornostal // Proceedings of the 31th International Scientific Symposium Metrology and Metrology Assurance. – Sozopol, IEEE Access, 2021. DOI:10.1109/MMA52675.2021.961094

DOI 10.15588/1607-3274-2022-2-11