УДК 004.75

DOi 10.30837/ bi.2024.2(101).01

С.Г. Удовенко[1], Є.М. Грабовський[1], Д.О. Донський [2], Л.Е. Чала [2]

[1]ХНЕУ ім. С. Кузнеця, м. Харків, Україна, serhiy.udovenko@hneu.net,
ORCID iD: 0000-0001-5945-8647

[1]ХНЕУ ім. С. Кузнеця, м. Харків, Україна, yevhen.hrabovskyi @hneu.net,
ORCID iD: 0000-0001-7799-7249

[2]ХНУРЕ, м. Харків, Україна, dmytro.donskyi@nure.ua, ORCID iD: 0009-0003-0558-499X

[2]ХНУРЕ, м. Харків, Україна, larysa.chala@nure.ua, ORCID iD: 0000-0002-9890-4790

МУЛЬТИМОДАЛЬНА ТЕХНОЛОГІЯ ПОШУКУ ТА КЛАСТЕРИЗАЦІЇ СЛАБОСТРУКТУРОВАНИХ ТЕКСТОВО-ГРАФІЧНИХ ДОКУМЕНТІВ

Досліджено проблему пошуку та кластеризації слабоструктурованих текстово-графічних документів (ТГД) з використанням нейромережевих технологій. Запропоновано підхід до побудови мультимодальної системи пошуку та аналізу ТГД, що передбачає використання гібридного критерія порівняння текстових та графічних фрагментів в аналізованих документах. Розглянуто процедуру поєднання процесів пошуку текстово-графічних фрагментів електронних документів за характеристиками зображення та текстовими підписами (ключовими словами). Запропоновано метод кластеризації та індексації ТГД за результатами аналізу їх текстової частини, заснований на застосуванні алгоритму SOINN, лінгвістичних дескрипторів та модульної системи обробки файлових масивів. Метод дозволяє формувати динамічну структуру кластерів ТГД зі створенням єдиних індексів. Наведено результати тестування і рекомендації щодо використання нейромережевих моделей для практичної реалізації запропонованого підходу до пошуку та кластеризації ТГД.

ТЕКСТОВО-ГРАФІЧНІ ДОКУМЕНТИ, МУЛЬТИМОДАЛЬНА СИСТЕМА, КЛАСТЕРИЗАЦІЯ, АЛГОРИТМ SOINN, ПОРІВНЯННЯ ТЕКСТОВО-ГРАФІЧНИХ ФРАГМЕНТІВ, АНОТАЦІЯ ЗОБРАЖЕНЬ

S.G. Udovenko, Ye.M. Hrabovskyi, D.O. Donskyi, L.E. Chala. Multimodal technology for searching and clustering weakly structured text-graphic documents.The problem of searching and clustering poorly structured text-graphic documents (TGD) using neural network technologies is addressed. A multimodal approach to building a system for searching and analyzing TGD is proposed, utilizing a hybrid criterion to compare text and graphic fragments within the analyzed documents. Integrating processes for identifying text-graphic fragments in electronic documents based on image characteristics and text signatures (keywords) is described.

A method for clustering and indexing TGD is introduced, leveraging the SOINN algorithm, linguistic descriptors, and a modular file array processing system. This method enables the formation of a dynamic cluster structure for TGD and the creation of unified indexes. The testing results and practical recommendations for implementing neural network models in TGD search and clustering are presented.

TEXT-GRAPHIC DOCUMENTS, MULTIMODAL SYSTEM, CLUSTERIZATION, SOINN ALGORITHM, COMPARISON OF TEXT-GRAPHIC FRAGMENTS, IMAGE ANNOTATION

Вступ

В сучасних системах автоматичного пошуку та аналізу електронних документів часто виникає необхідність оброблення графічних та текстово-графічних об'єктів, які або вже проіндексовані та проанотовані, або лише знаходяться в черзі на індексацію та анотацію (нещодавні надходження). Оскільки така база не анотована рівномірно, то необхідним є здійснення доступу до графічної інформації через текстові запити. На цей час існують різні методи взаємодії, навігації та пошуку графічних та текстово-графічних об'єктів в ресурсах мережі Інтернет або існуючих базах зображень що є вагомою частиною текстово-графічних документів (ТГД). Насамперед, це модель короткотермінової взаємодії, що використовується для підвищення точності системи, та модель довготривалої взаємодії, що допомагає пов'язати текстові слова та візуальні характеристики для пошуку зображень за текстом, за візуальним наповненням чи за змішаними характеристиками (текст/зображення). Модель пошуку зображень, що присутні в ТГД, дозволяє ітеративно формувати та уточнювати анотації до зображень.

При цьому перспективним є застосування мультимодальної системи пошуку зображень (або інших графічних фрагментів (ГФ), таких як схеми, діаграми тощо) за різними показниками. Така система може об'єднувати різні джерела даних та аналізувати вміст зображення і текстові пояснення, що присутні, зокрема, в підрисункових підписах. Це дозволяє формувати гібридні запити за методикою зворотного зв'язку, що поєднує класичні методи, які використовуються при пошуку інформації (використання руху точки запиту і формування розширених запитів).

З іншого боку, завдання пошуку ТГД за запитом з метою їх подальшої кластеризації передбачає необхідність використання методів аналізу текстових фрагментів (ТФ) в аналізованих документах.

Мультимодальна системи пошуку ТГД може доповнюватися інтерфейсом, що дозволяє візуалізувати

змішані текстово-графічні запити. Навіть якщо на даний момент два використовуються лише окремі типи інформації, тобто текстовий або візуальний, загальна пропонована модель дозволяє її розширити з залученням зовнішніх даних, таких як місцеположення (GPS) та час.

Відмінною особливістю методів кластеризації, які у сучасних системах аналізу корпусів електронних документів, є здатність автоматично виділяти групи у потоці вхідних даних. У контексті обробки текстів природною мовою ця властивість є особливо привабливою, коли виникає необхідність оперативно виділити тематичні кластери у великому масиві текстових документів. Така кластеризація, наприклад, актуальна для інформаційно-пошукових систем, коли користувачеві необхідно отримати загальне уявлення про весь список знайдених документів, не переглядаючи їх більшу частину. При цьому виникає завдання синтезу анотацій кластерів, що коротко відображають тематику їх документів, важливою особливістю якої є необхідність модифікації структурованих кластерів, що формуються в реальному часі, що дозволяє враховувати зміну характеру аналізованих даних.

До найефективніших підходів до анотування кластерів, здійснюваному їхнього подальшого використання у завданнях аналізу текстів, є присвоєння унікального індексу кожному документу з кластера. Даний спосіб дозволяє однозначно та оперативно визначати ступінь тематичної близькості документів з можливістю швидкого оцінювання наслідків входження нових документів у структуру кластера та модифікувати цю структуру за істотної зміни характеру даних. При кластеризації великих обсягів текстової інформації ефективним є виділення та аналіз лінгвістичних дескрипторів. Лінгвістичний дескриптор – лексична одиниця (слово, словосполучення, абревіатура), що служить для опису основного змістового документа або формулювання запиту при пошуку документа в інформаційно-пошуковій або класифікуючій системі [1]. Актуальною є завдання розробки методу кластеризації даних з аналізом лінгвістичних дескрипторів у масивах з розподіленим зберіганням інформації, що дозволяє враховувати можливість зміни місцезнаходження файлу у файловій структурі із ієрархією, що зберігається. Для вирішення такого завдання можна використовувати розміщення додаткової інформації в системному розділі метаданих файлу, що дозволяє реалізувати зручне та компактне зберігання передопрацьованих даних цього файлу з маркуванням за датою витяжки. У комбінації з такими стандартними властивостями, як дата створення та дата зміни, цей підхід дозволяє отримати мінімально необхідну розподілену систему котролю версії попередньо обробленого вектора дескрипторів для кожного текстового документа в бібліотеці, що формується.

Метою цієї статті є розроблення та дослідження мультимодальної технології пошуку та кластеризації слабоструктурованих текстово-графічних документів (ТГД) з використанням нейромережевих моделей.

Реалізація такої технології дозволить із застосуванням лінгвістичних дескрипторів вирішити проблему кластеризації та індексації великих обсягів електронної текстової інформації за користувальницькими категоріями, враховуючи можливість розподіленого зберігання масивів інформації.

Відповідно до поставленої мети, вирішуються наступні завдання:

– аналіз існуючих підходів до оброблення політематичних слабоструктурованих ТГД;

– розроблення модуля кластеризації ТГД (в текстовій частині) з використанням самоорганізованої інкрементної нейронної мережі;

– розроблення технології пошуку близьких за мультимодальним критерієм ТГД (в графічній частині) з комбінованим використанням трансформерів та методів хешування;

– експериментальне дослідження запропонованого підходу.

1. Технології пошуку та аналізу політематичних слабоструктурованих ТГД

Завдання пошуку та оброблення великого масиву ТГД, що мають політематичний характер, зазвичай, передбачає необхідність подальшого формування відповідних кластерів близьких за мультимодальними критеріями ТГД. Насамперед такий масив формується за результатами пошуку в ресурсах мережі Інтернет з використанням пошукових систем. Крім того, таким масивом може бути, наприклад, електронна збірка матеріалів конференції, що містять анотації та основний текст; масив документів великої електронної бібліотеки, присвячених деякому загальному науковому напрямку, тощо.

Сучасні технології, що застосовуються для вирішення такого завдання, базуються на формуванні динамічних запитів за обраними критеріми, які беруть до уваги необхідність поточного порівняння як текстових так і графічних фрагментів ТГД. Остаточна процедура визначення близькості аналізованих ТГД передбачає необхідність зваженого поєднання результатів такого багатоознакового порівняння.

Розглянемо спочатку завдання аналізу текстової частини, що є присутня в загальній структурі ТГД.

Проблемами, що виникають під час реалізації такого завдання, є: наявність службових елементів та сторонніх блоків тексту, які не належать до основної тематики ТГД; наявність у навчальному масиві аномальних документів (порожніх, у невідомих кодуваннях тощо); складність автоматичного формування вирішальних правил для рубрик через негативний вплив

сторонньої інформації; зниження якості кластеризації через накладання кількох рубрик одна на олну; складність інтерпретації результатів аналізу ТГД через невизначеність розташування у тексті інформації [2].

Перспективним напрямом підвищення ефективності реалізації підходу, пов'язаного з автоматичною кластеризацією політематичних ТГД, є застосування нейромережевих методів, що ґрунтуються на можливості застосування у схемі кластеризації операцій виділення з текстів дескрипторів та ключових слів. Відповідно до поставленої задачі, для великого обсягу різнорідних ТФ ТГД необхідно здійснити формування масивів, що містять документи певної тематики. Після цього слід застосувати процедуру побудови відповідного модуля кластеризації з попередньою фільтрацією вихідного масиву текстових документів та виділенням лінгвістичних дескрипторів. Зазвичай, дескриптор може бути однозначно поставлений у відповідність до групи ключових слів природної мови, відібраних з тексту, що відноситься до певної галузі знань. Це дозволяє створити унікальні словники дескрипторів для врахування їх у схемі обробки вхідних текстів з метою підвищення точності кластеризації. При цьому для характеристики текстів, що аналізуються, доцільно використовувати векторну модель подання документів вихідного масиву:

$$d = (t_1, t_2, ..., t_n), \quad (1)$$

де $t_1, t_2, ..., t_n$ – дескрипторні терміни (ознаки) текстів; n – загальна кількість термінів, що беруться до уваги.

Компонентам вектора (1) можуть бути поставлені у відповідність бінарна або частотна функції зважування на основі кількості входження певного терміна до класів документів.

Текстову частину документа D, що має порівнюватися з текстовою частиною запиту Q, представимо як зважений вектор термінів цього документу [3]:

$$d_i = (w_{i,1}, w_{i,2}, ...w_{i,t}). \quad (2)$$

Для обчислення подібності між запитом Q і документом D визначається скалярний добуток між відповідними векторами:

$$S(d_i, q) = \sum_j w_{ij} \times w_{qj} S(d_i, q) = \sum_j w_{ij} \times w_{qj}. \quad (3)$$

Різні методи оцінки вагових коефіцієнтів дозволяють створити чимало функцій ранжирування для моделі векторного простору. Розглянемо підхід з використанням класичної моделі TF-IDF та техніки зважування і нормалізації. Класична векторна модель обчислює вагу терміна в документі як добуток терміну частоти (TF) і зворотної частоти документів (IDF) [4]:

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}}; \quad idf_i = \log \frac{N}{\left|\{d : t_i \in d\}\right|}, \quad (4)$$

де $n_{i,j}$ – число входжень розглянутих термінів t_i в документ d_j, а знаменник – сума числа входжень усіх членів документа d_j; N – загальна кількість документів в корпусі, $|\{d_j : t_i \in d\}|$ є числом документів, де з'являється термін t_i (який має $n_{i,j} = 0$).

Така модель дозволяє отримати функцію подібності між запитом Q і документом D у наступному вигляді:

$$S(d_i, q) = \frac{\sum_j D_i \times D_q}{\left|D_i\right|\left|D_q\right|}, \quad (5)$$

де $D_k = tf_k \cdot idf_k$ та $\left|D_i\right| \left|D_q\right|$ – знаменник для стандартизації.

При реалізації процедури кластеризації для кожного присутнього в текстах слова розраховується та зберігається його вага – оцінка ймовірності того, що текст із цим словом належить до одного з можливих класів (одна з можливих апроксимацій такої оцінки полягає у заміні бінарної функції зважування термінів t_i у векторній моделі (1) частотною функцією, що ставить у відповідність цим термінам частоту їх появи в аналізованому документі). На основі даних про класифікацію лінгвістичних дескрипторів при складанні локальних словників проводиться розрахунок доцільності вибору тих чи інших його варіантів з урахуванням вимог до системи, по кожному обраному класу дескрипторів у різних класифікаціях.

При цьому враховується загальне співвідношення слів, розподілених за категоріями обраних підкласів дескрипторів. Для формування локальних словників дескрипторів кожного з формованих надалі класів розглядаються такі види класифікації, як «Частини промови» і «Частотність». На основі аналізу співвідношення числа дескрипторів і можливостей їх вилучення з неструктурованої текстової інформації можна стверджувати, що для підвищення точності класифікації найраціональніше використовувати іменники. В той же час за частотністю присутності в тексті доцільно використовувати високочастотні та низькочастотні дескриптори. Найбільш відповідною цим вимогам структурою є абревіатури. Абревіатури – це іменники, що складаються з усічених слів, що входять у вихідне словосполучення, або з усічених частин вихідного складного слова, а також назв початкових літер цих слів (або їх частин).

Абревіатури є низькочастотними дескрипторами, які легко отримати з текстів (особливо їх анотацій). У випадку, якщо в текстах відсутні абревіатури, доцільно використовувати високочастотні дескриптори для максимального охоплення масиву текстової інформації.

Для формування вхідних даних системи кластеризації можна, наприклад, використовувати алгоритм виділення дескрипторів з тексту, запропонований в [5]. Векторні дескриптори різних типів, що надходять на вхід блоку кластеризації, накопичуються у верхньому суперкласі програмної системи (для

забезпечення доступу до списка всіх унікальних дескрипторів при кластеризації документа) і зберігаються в метаданих файлах для виключення можливості повторного аналізу одного і того ж файлу.

Сформовані вектори є основою для подальшого аналізу та складання кластерної структури даних, подальшої індексації та організації роботи пошукової машини. Розглянута в [5] схема попередньої обробки даних характеризується універсальною застосовністю незалежно від використовуваного формату файлів і дозволяє ефективно використовувати інформацію, що міститься в кожному файлі. У результаті обробки у текстового файла з'являються такі додаткові властивості: дата індексації; повне індексне ім'я; вхідний вектор файлів дескрипторів у скороченому вигляді.

У випадку, якщо у файлі відсутня можливість запису в блок метаінформації, всі метадані записуються в перший рядок файлу. Зазначимо, що існують проблемні формати, модифікація яких може призвести до пошкодження. Серед них:

– формати сканованих або автогенерованих книг-зображень (наприклад, .pdf та .djvu);

– формати зображень без можливості створення особливих метатегів (наприклад, .jpg, .png, .gif тощо);

– медіа формати, що відповідають за відео та звук (наприклад, .mp3, .mpg, .mkv, .avi тощо);

– робочі файли візуальних редакторів та вихідні коди додатків (наприклад, .crd, .cpp, .h, .rb тощо);

– системні та виконувані файли додатків: .exe, .dll, .config, .db.

Враховуючи ці обмеження, основну увагу приділимо розгляду текстових форматів, що підтримуються в процесі аналізу структури файлів з метою коректної організації пошуку. Для побудови ефективної масштабованої системи кластеризації та індексації файлів необхідно передбачити перспективи розширення програми, збільшення обсягу даних, можливість модифікації без значних змін у коді системи та зручний порядок взаємодії користувача з системою [6, 7].

Відзначимо, що для кластеризації масивів текстових документів останнім часом наболо поширення використання нейромережевих моделей, зокрема, самоорганізованих інкрементних нейромереж SOINN (Self-Organizing Incremental Neural Network) [8, 9].

Розглянемо далі векторну модель визначення подібності ТГД за візуальними характеристиками ГФ ТГД (зображень), автоматично витягнутих з візуального вмісту. Така модель залежить від функції подібності та використовуваних візуальних сигнатур, які можуть бути векторами ознак або узагальненими локальними характеристиками. При цьому функція оцінювання подібності ГФ залежить від обраних характеристик. Наприклад, подібність текстурних ознак часто вимірюється за допомогою відстані Мінковського або відстані Махаланобіса. Розглянемо два зображення, що індексовані відповідними векторами $I=(I_1,I_2,...,I_n)$ та $J=(J_1,J_2,...,J_n)$. Оцінювання подібності між двома зображеннями полягає в обчисленні подібності між I та J. Зокрема, метрики Мінковського L_p, які є найбільш поширеними геометричними відстанями, мають такий загальний вигляд:

$$L_p=\sqrt[p]{\sum_{i=1}^{n}(I_i-J_i)^p}\ . \qquad (6)$$

Відзначимо, що база аналізованих ТГД може корегуватися в реальному часі, тобто обсяг ГФ документів цієї бази не є фіксованим.

Для використання та розвитку технології пошуку та порівняння ГФ ТГД на етапі їх індексації будемо використовувати такі припущення:

– екземпляри ГФ ТГД бази не завжди містять повну текстово-графічну інформацію;

– база знань технології пошуку та порівняння ТФ ТГД ґрунтується на анотаціях зображень, доповнених текстом, а також на їх візуальних характеристиках (колір, текстура, форма тощо);

– навчання нейромережевої схеми пошуку та порівняння ГФ ТГД може здійснюватися з використанням традиційних методів навчання (зокрема, навчання з підкріпленням). Бажаною є взаємодія між користувачами та експертами схеми пошуку та порівняння ГФ ТГД для фільтрації не релевантних запитам ГФ під час пошуку;

– схема може формувати анотації зображень в режимі реального часу.

Зважаючи на ці припущенн, в кожний поточний момент можуть розглядатися три типи ГФ ТГД в пропонованій технології: зображення без анотації; зображення з додатковою інформацією (наприклад, додатковий опис або довідкові дані); зображення з автоматично доданою інформацією (з «розширеними» анотаціями).

Якщо частина ГФ ТГД з аналізованої бази не анотована, то необхідно використовувати зв'язок між текстовими словами та візуальними характеристиками для мультимодального пошуку та анотування ГФ, які містять текстову частину.

В значній мірі швидкодія методів порівняння ГФ ТГД залежить від тривалості попередньої обробки, що дозволяє представити анотовані зображення в придатному для швидкого порівняння форматі. При цьому доцільно використовувати хеш-методи в комбінації з деякими більш точними методами [10].

Метод мультимодального визначення подібності ГФ ТГД має поєднувати візуальний пошук зразків зображень та текстовий пошук за ключовими словами. Візуальна подібність між запитом, пов'язаним з зображенням в запиті, та текстово-графічним зображенням в базі даних оцінюється за величиною скалярного добутку відповідних векторів з використанням комбінованого критерія.

На рис. 1 наведено варіант пропонованої схеми пошуку та порівняння ГФ ТГД.

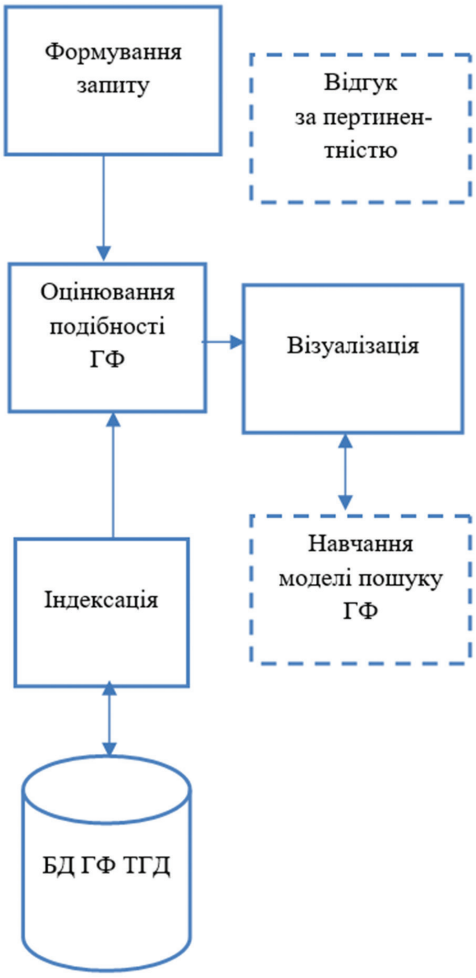

Рис. 1. Схема пошуку та порівніння ГФ ТГД

Ця схема може бути доповнена додатковим блоком, що реалізує можливість створення та інтерактивного розширення анотацій до ГФ ТГД з використанням зворотного зв'язку за пертинентністю (рис.2).

Концепція специфікації запитів, що використовується в пропонованій технології, передбачає здійснення таким типів пошуку:

– пошук за запитами зображень: користувач надає приклади зображень до системи;

– пошук за ключовими словами: користувач надає ключові слова, після чого здійснюється пошук відповідних зображень на основі текстових підписів (або анотацій), де система перетворює ключові слова у візуальні зображення та шукає відповідні зображення на основі візуальних характеристик. Цей тип запиту вимагає реалізацію перетворень між зображеннями та текстом;

– пошук за комбінацією візуальних характеристик;

– пошук за локалізованим запитом: користувач локалізує конкретні області складних зображень;

– пошук за комбінованим запитом (використання комбінації різних типів пошуку.

Специфікація запитів дозволяє користувачеві уточнювати вектор пошуку за допомогою різних типів даних.

Рис. 2. Схема розширення анотацій для пошуку ГФ ТГД

Ця технологія дозволяє покращити результати інтерактивного пошуку зображень в ГТД за рахунок застосування моделі KVR (Keyword Visual Representation). Модель KVR використовується для створення зв'язків між семантичними поняттями та візуальними уявленнями [10]. Завдяки послідовній стратегії навчання ця модель забезпечує зв'язок міжсемантичних понять та візуальних характеристик, що допомагає поліпшити точність анотації зображень та результати пошуку.

Розглянемо можливість поєднання процесів порівняння ТГД за характеристиками (ознаками) зображення та текстовими підписами (ключовими словами) [3]. Візуальна подібність V_d між запитом, пов'язаним з зображенням Q, та ТФ ГТД I_i оцінюється за величиною скалярного добутку відповідних векторів:

$$V_d(\overline{I}_i, \overline{Q}) = \sum_j w_{ij} \times w_{qj} \,. \qquad (9)$$

Текстова подібність S_d між запитом, пов'язаним з зображенням Q, та ТФ ГТД I_i визначається так:

$$S_d(I_i, q) = \frac{\left|K_{i,q}\right|}{\left|K_q\right|}, \qquad (10)$$

де $\left|K_{i,q}\right|$ – кількість однакових ключових слів для зображення Q та зображень I_i; $\left|K_q\right|$ – загальна кількість ключових слів у зображенні Q.

Інтегроване значення подібності ГТД визначається наступним чином:

$$S_{final} = r * S_q + (1-r) * V_d \quad S_{final} = r * S_q + (1-r) * V_d \,, \quad (11)$$

де r – ваговий коефіцієнт.

До найбільш важливих завдань реалізації запропонованої технології слід віднести розроблення нейромережевого модуля кластеризації ТГД (в текстовій частині) та розроблення нейромережевої технології пошуку близьких за мультимодальним критерієм ТГД (в графічній частині).

2. Модуль кластеризації ТГД

Структура пропонованого модуля кластеризації «CLASTFIND» (в текстовій частині) містить чотири основні блоки: «Інтерфейс», «Парсер», «Обробник» та «Пошукова машина». Такий поділ дозволяє вносити зміни в модуль в кожному з блоків при збереженні інтерфейсу взаємодії з користувачем. Схему передачі даних у модулі наведено на рис. 1.

Рис. 3. Схема передачі даних в модулі «CLASTFIND»

Блок «Інтерфейс» забезпечує отримання необхідної інформації від користувача та логіку розподілу завдань, залежно від поточного стану модуля. У цьому блоці здійснюється взаємодія з користувачем від отримання посилання на бібліотеку даних щодо ТГД, що кластеризуються, до видачі користувачеві даних про належність до певного файлу його пошукового запиту. Крім того, він формує повідомлення про помилки послідовності взаємодії із модулем та координує видачу релевантних результатів перегляду текстів.

Блок «Парсер» здійснює попереднє опрацювання ТГД бібліотеки та подальший аналіз ТФ ТГД з метою складання повної карти розподілу вхідних дескрипторів за кластерами. Даний модуль, отримуючи адресу бібліотеки текстів, аналізує вміст масивів ТФ та формує універсальний базовий блок-вектор, який є основою для подальших аналітичних та алгоритмічних перетворень та висновків. Універсальний блок-вектор включає найменування слова і його частотну характеристику щодо аналізованого документа. Цей блок-вектор абсорбується загальним частотним словником, який складається з усіх документів, що входять до базового кластера. Модуль «Обробник» визначає структуру кластерів, а також здійснює їх індексацію, зберігання та подальше використання для класифікації вхідних запитів.

Отримана з парсера інформація надходить у модуль «Обробник» для формування кластерів та їх послідовного індексування. Під індексацією кластерів розуміється присвоєння кластеру унікального номера та створення для нього відповідного блоку векторних дескрипторів шляхом злиття векторів вхідних документів.

Завдання формування кластерів пов'язана, насамперед, із знаходженням топологічної структури розподілу вхідних даних. Існують різні алгоритми розв'язання цього завдання. Наприклад, алгоритм самоорганізованих карт Кохонена є методом проєкції багатовимірного простору в простір з нижчою розмірністю з визначеною структурою. При цьому виникають дефекти проєктування, аналіз яких є достатньо складним завданням. Однією з альтернатив цьому підходу є гібридне застосування методів конкурентного навчання Хебба та нейронного газу. Однак для практичної реалізації такого підходу необхідні апріорні знання про розмір мережі та адаптацію швидкості її навчання. Алгоритм кластеризації документів за векторними дескрипторами повинен задовольняти таким вимогам: навчання в режимі онлайн; поділ на класи без попередніх знань про вхідні дані; поділ даних з нечітким кордоном та виявлення структури кластерів.

В основу процедури онлайн кластеризації, яка використовується в модулі «Обробник», покладено алгоритм SOINN (Self-Organizing Incremental Neural Network), який частково вирішує зазначені вище проблеми [8, 9]. SOINN є нейронною мережею з двома шарами. Перший шар здійснює визначення топологічної структури кластерів, а другий визначення числа кластерів і виявлення вузлів-прототипів. Спочатку навчається перший шар мережі, а потім, за наслідками цього навчання, навчається другий шар мережі. Для завдання класифікації без вчителя необхідно визначити, чи належить вхідний образ одному з раніше отриманих кластерів. Нехай два образи належать одному кластеру, якщо відстань (за заздалегідь заданою метрикою) між ними менше порога відстані T. Якщо T буде занадто великим, всі вхідні образи будуть віднесені до одного кластеру. Якщо T занадто маленький, кожен образ буде ізольований як окремий кластер. Для отримання прийнятного розбиття на кластери T має бути більше внутрішньокласової відстані та менше міжкласової. Поріг T обчислюється для кожного з шарів мережі SOINN. Перший шар не має апріорних знань про структуру вихідних даних,

тому T для нього підбирається адаптивно з урахуванням знань структури вже побудованої мережі та поточного вхідного образу. При навчанні другого шару можна розрахувати внутрішньокласову та міжкласову відстані та підібрати постійне значення порога T. Для представлення топологічної структури в онлайновому режимі навчання зростання мережі є важливим фактором зниження помилки та адаптації до змінних умов при збереженні старих даних. У той же час неконтрольоване збільшення числа вузлів призводить до перевантаження мережі та її «перенавчання» загалом. Додавання вузлів SOINN здійснює в регіоні з максимальною помилкою. При цьому для оцінки корисності такої вставки використовують так званий «радіус помилки». Застосування цієї оцінки контролює приріст вузлів і зрештою стабілізує їх кількість. Для формування зв'язків між нейронами використовується конкурентне правило Хебба, запропоноване для визначення топологій в мережах. Відповідно до цього правила для кожного вхідного сигналу об'єднуються два найближчих вузли. Доведено, що граф, який формується при цьому, оптимально представляє топологію вхідних даних. Вузли, які є сусідніми на ранній стадії, можливо, не будуть близькими у пізнішій стадії. При цьому виникає необхідність видалення з'єднань, які останнім часом не оновлювалися, що може привести до перекриття існуючих кластерів. Щоб визначити кількість кластерів точно, передбачається, що вхідні дані розділяються: щільності ймовірності в центральній частині кожного кластера вище, ніж щільність в частині між кластерами, а перекриття кластерів має низьку щільність ймовірності. Поділ кластерів відбувається шляхом видалення вузлів із регіонів із низькою щільністю ймовірності. Алгоритм SOINN пропонує наступну стратегію видалення вузлів: якщо число сигналів, що генеруються до поточного кроку, є цілим кратним заданим параметром, то слід видалити ті вузли, які мають тільки одного топологічного сусіда або не мають сусідів зовсім. Якщо локальний накопичений рівень сигналу є маленьким, то вважаємо, що такі вузли лежать у області низької щільності ймовірності і також видаляємо їх. Для вирішення завдання кластеризації текстових документів в запропонованій технології було використано модифікацію базової версії алгоритму (SOINN-index). У базовій версії як міра відстані вибирається евклідова метрика, а в модифікованої версії – косинусна метрика і міра Жаккара для оцінювання відстаней між документами, представленими векторною дескрипторною моделлю (1).

Наведемо формальний опис алгоритму навчання як першого і другого шарів мережі SOINN-index. Зазначимо, що вхідні дані навчання другого шару породжуються першим шаром і під час навчання другого шару використовуються знання топологічної структурі першого шару для обчислення постійного порога подібності T.

В алгоритмі використовуються такі позначення: A – множина, що використовується для зберігання вузлів; N_A – кількість вузлів у множині A; C – множина зв'язків між вузлами; N_C – кількість ребер в C; W_i – n-вимірний вектор ваг для вузла i; E_i – локальний акумулятор помилки для вузла i; M_i – локальний акумулятор сигнала для вузла i; R_i – радіус помилки для вузла i ($R_i = E_i / M_i$); C_i – мітка кластера для вузла i; Q – число кластерів; T_i – поріг подібності для вузла i; N_i – набор топологичних сусідів для вузла i; $age(i, j)$ – вік зв'язку между вузлами i та j.

Алгоритм SOINN-index:

1. Ініціалізувати множину вузлів A вузлами c_1 та c_2: $A = \{c_1, c_2\}$. Ініціалізувати множину ребер C порожньою множиною: $C = \{\}$.

2. Подати новий сигнал x з R^n.

3. Знайти в множині A вузли переможця s_1 та другого переможця s_2, як вузли з найближчим та наступним за ним векторами ваг (за деякою заданою метрикою). Якщо відстань між x, s_1 та s_2 більше порогів поддібності Ts_1 або Ts_2, то створити новий вузол і перейти до кроку 2.

4. Якщо ребро між s_1 та s_2 відсутнє, то створити його та прийняти вік ребра між ними рівним 0.

5. Збільшити на одиницю вік всіх дуг, що виходять із s_1.

6. Додати відстань між вхідним сигналом та переможцем до локальної сумарної помилки Es_1.

7. Збільшити локальну кількість сигналів вузла s_1: $Ms_1 = Ms_1 + 1$.

8. Адаптувати вектори ваг переможця та його прямих топологічних сусідів:

$$Ws1 = Ws_1 + e_1(t)(x \quad Ws_1) Ws_i = Ws_i + e_2(t)(x \quad Ws_i),$$

де $e_1(t)$ и $e_2(t)$ – коефіцієнти навчання переможця та його сусідів.

9. Видалити ребра з віком, який перевищує задане граничне значення.

10. Якщо число вхідних сигналів, що генеруються на поточний момент, кратно параметру lambda, вставити новий вузол і видалити вузли в областях низької щільності ймовірності за такими правилами:

10.1. Знайти вузол q з максимальною помилкою.

10.2. Знайти серед сусідів q вузол f з максимальною помилкою.

10.3. Додати вузол r таким чином, щоб його ваговий вектор був середнім арифметичним вагових коефіцієнтів q та f.

10.4. Розрахувати накопичену помилку *Er*, сумарну кількість сигналів *Mr* та успадкований радіус помилки:

$$R = alpha_1 * \left(E_q + E_f\right) M_r =$$

$$= alpha_2 * \left(M_q + M_f\right) R_r = alpha_3 * \left(R_q + R_f\right)$$

10.5. Зменшити сумарну помилку вузлів q та r: $q = beta * E_q E_f = beta * E_f$.

10.6. Зменшити накопичену кількість сигналів: $M_q = gamma * M_q M_f = gamma * M_f$.

10.7. Визначити, чи успішно відбулася вставка нового вузла. Якщо вставка не може зменшити середню похибку даної області, додана вершина видаляється, а всі параметри повертаються у вихідні стани. Інакше оновлюється радіус помилки всіх задіяних у вставці вузлів:

10.8. $R_q = E_q / M_q R_f = E_f / M_f R_r = E_r / M_r$.

10.9. Якщо вставка пройшла успішно, створити зв'язки $q \leftrightarrow r$ и $r \leftrightarrow f$ та видалити зв'язок $q \leftrightarrow f$.

10.10. Серед усіх вузлів в A знайти такі, які мають тільки одного сусіда, потім порівняти накопичену кількість вхідних сигналів із середнім значенням всіх вузлів. Якщо вузол містить лише одного сусіда і лічильник сигналів не перевищує адаптивний поріг, видалити його з набору вузлів. Наприклад, якщо $L_i = 1$ и $M_i < c * sum\{j = 1...N_A\}\ M_j / N_A$ (где $0 < c < 1$), то видалити вершину i.

10.11. Видалити всі ізольовані вузли.

10.12. Перевірити умову зупинення алгоритму: якщо кожному класу із вхідних даних відповідає компонента зв'язаності у побудованому графі, то оновити мітки класів та завершити навчання; інакше перейти до пункта 2 та продовжити навчання.

Розглянемо окремо процедуру обчислення порогового значення T алгоритму SOINN-index.

Для першого шару поріг подоби T обчислюється адаптивно за таким алгоритмом:

1. Ініціалізувати поріг подібності для нових вузлів рівним $+\infty$.

2. Якщо вузол є першим або другим переможцем, оновити значення порогу подібності за таким правилом: якщо вузол має прямих топологічних сусідів ($L_i > 0$), оновити значення T_i як максимальну відстань між вузлом та його сусідами: $T_i = max\left\|W_i - W_j\right\|$; якщо вузол не має сусідів, T_i встановлюється як мінімальна відстань між вузлом та іншими вузлами у множині A: $T_i = \min\left\|W_i - W_c\right\|$.

Для другого шару розраховується постійний поріг подібності:

1. Обчислити внутрішньокласову відстань таким чином:

$$dw = 1 / N_C * sum\{(i,j)\ in\ C\}\left\|W_i \quad W_j\right\|.$$

2. Обчислити міжкласові відстані. Відстань між класами C_i та C_j обчислюється таким чином:

$$db\left(C_i, C_j\right) = max\left\{i\ in\ C_i, j\ in\ C_j\right\}\left\|W_i - W_j\right\|.$$

3. Як постійний поріг подоби T_c прийняти мінімальну міжкласову відстань, що перевищує внутрішньокласову відстань.

Блок «Пошукова машина» виконує попередню обробку запитів користувача, направлення цих запитів на одержання індексу в блок «Обробник» та виведення релевантних результатів відповідно до індексу. Цей блок частково використовує результати роботи блоків «Парсер» та «Обробник» із внесенням своєї специфіки до процесів обробки та аналізу. Пошуковий запит спочатку проходить попередню обробку у блоці «Парсер», потім кластеризується блоком «Обробник», після чого з поточної ієрархії кластерів формується шляхом складання індексів кластерів шуканий індекс файлу. За індексами файлу проводиться пошук в індексній базі та формується кінцевий шлях до необхідних даних.

Пошукові індекси кластерів зберігаються у базі як зворотні пошукові індекси з урахуванням контексту, що дозволяє економити простір на диску з допомогою використання числового позначення груп. Оскільки одному кластеру відразу асоційований повний набір файлів (з урахуванням вкладеності), це виключає необхідність проведення додаткових операцій над множинами в межах бази. При необхідності актуалізації кластерного індексу для мінімізації обчислювальних навантажень та підтримки високої якості пошуку у системі передбачено алгоритм актуалізації. У разі зміни проміжних кластерів модифікація індексів відбувається лише в межах зміненої групи, що підвищує загальну стабільність системи. Дані про структуру та вміст файлу зберігаються у властивостях метаінформації файлу у форматі «.doc».

Якщо у файлі відсутня можливість запису в блок метаданих, метадані записуються в перший рядок файлу у форматі «.txt» (рис. 4).

Рис. 4. Структура метаданих файлу в форматі «.txt»

3. Технологія пошуку близьких ТГД за мультимодальним критерієм (в графічній частині)

Проблема пошуку подібних за змістом ГФ ТГД (ND-зображень (near duplicates – англ.)) в деяких колекціях ТГД або виявлення високого рівня подібності аналізованих зображень (наприклад, для виявлення плагіату графічних частин в порівнюваних ТГД) пов'язана, насамперед, з необхідністю реалізації

процедур порівняння зображень за сукупністю визначених ознак.

Хоча задача порівняння ГФ ТГД є подібною до багатьох інших задач класифікації чи розпізнавання графічних об'єктів, вона відрізняється від них рівнем необхідної точності та суб'єктивності рішення. Наприклад, задача розпізнавання осіб за фотографіями передбачає необхідність отримання цілком об'єктивних результатів. Задача ж пошуку подібних за змістом зображень в колекціях є більш загальною і передбачає можливість отримання дещо быльш суб'єктивних результатів такого пошуку. Ця суб'єктивність полягає в тому, що можуть існувати різні точки зору щодо того, чи є відмінності між двома зображеннями суттєвими, щоб вважати зображення різними.

До відмінностей, які можуть бути присутні на порівнюваних ND-зображеннях, відносять наявність шуму, артефактів стиснення, зміни колірного балансу, яскравості, контрасту, геометричних перетворень (масштабування, повороту, обрізання) та більш складні зміни сцени (пересування камери, зміна умов освітлення) [11].

Методи виявлення ND-зображень мають визначити, чи є два порівнювані зоброження неповними дублікатами один одного. В даному випадку дуже важливо не тільки виявити всі неповні дублікати, але і помилково не ідентифікувати їх як дублікати зображень, що недостатньо близькі за змістом. Таким чином, вдосконалення методів виявлення ND-зображень безпосередньо пов'язане зі зниженням ймовірності виникнення помилок першого і другого роду.

Існує декілька напрямків пошуку ND-зображень: пошук за змістом (знайти неповністю подібні зображення в колекціях графічних документів), пошук за візуальним зразком (знайти зображення, подібні заданому), пошук за описом (наприклад, знайти зображення, помічене як «Архітектурні об'єкти Харкова») тощо. Кожен з напрямків пошуку має свої особливості та сфери використання.

В рамках пропонованої технології основну увагу будемо приділяти пошуку зображень за візуальним зразком, що полягає у вилученні істотних властивостей зображень та побудові на їх основі дескрипторів, які використовуватимуться для порівняння пар зображень ТГД. При цьому до кожної пари мають належати зображення з колекції та зображення-зразок (еталон для пошуку).

Результатом порівняння є величина, яка називається візуальною подібністю (тематичною близькістю) зображень. З точки зору інформаційного пошуку таке оцінювання, здійснене людиною-експертом, називають, зазвичай, змістовною релевантністю, а розраховане в системі – формальною релевантністю. Властивості ГФ ТГД оцінюються алгоритмами обчислення значень глобальних та локальних ознак. До глобальних ознак (дескрипторів) належать основні кольори, текстури, форми, значущі елементи всього зображення тощо. Локальні ознаки вираховуються для окремих фрагментів зображення. Вектор дескрипторів будемо називати сигнатурою.

Під час пошуку зображень ТГД за зразком аналізуються не окремі екземпляри, а пари зображень, які зіставляються. Тому від ознак зображень доцільно перейти до ознак пар, значення яких вираховують як абсолютні різниці значень відповідних ознак кожного з зображень пари.

Такий підхід дозволяє швидко розрахувати їх для проіндексованих зображень і реалізувати на практиці пошук в реальному часі. В ефективних методах пошуку в колекціях зображень з обмеженими ресурсами системи мають використовуватися компактні описи кожного зображення і швидкодіючі процедури порівняння пар зображень.

Отже, подібні за змістом зображення (або ND-зображення (near duplicates)) мають визначатися з урахуванням відмінностей, які можуть бути присутні на одному з зображень. Приклади ND-зображень наведено на рис. 5.

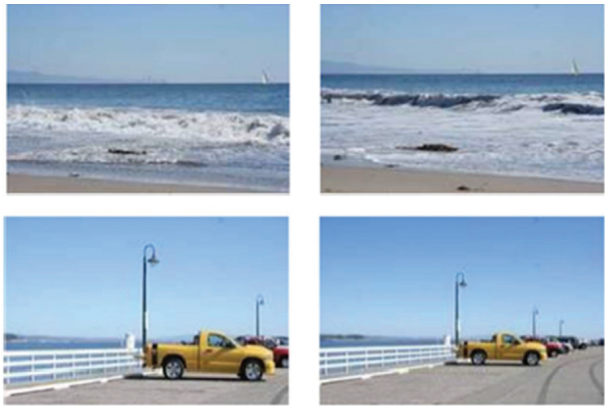

Рис. 5. Приклади ND-зображень

Відповідно, зображення, які не є достатньо подібними, будемо називати NND-зображеннями (not near duplicates). Приклади NND-зображень (візуально подібних, але не ND-зображень) наведено на рис. 6.

Рис. 6. Приклади NND-зображень

На сьогодні існує низка традиційних методів порівняння зображень за дескрипторами, що можуть бути використані для пошуку ND-зображень в колекціях ТГД, зокрема: методи піксельного порівняння зображень; методи порівняння зображень за заданими ознаками; хешування порівнюваних зображень; нейромережеві методи порівняння зображень.

У кожного з цих методів є як переваги, так і недоліки, тому той чи інший метод краще вибирати залежно від вхідних даних та умов, для яких вирішується задача (необхідність врахування цифрового шуму, розмиття, масштабу, наявності поворотів зображення в процесі побудови вектора опису тощо).

Хешування зображень та порівняння хешів також є одним зі способів швидкого зіставлення зображень. Такий тип хешування в загальному випадку відрізняється від криптографічного, оскільки тут хеші для подібних фрагментів зображення мають також бути схожими. Методи криптографічного хешування (SHA1, MD5 та ін.) дозволяють генерувати різні значення хешу навіть для дуже близьких, але не однакових послідовностей. Ця властивість є припустимою для пошуку точних дублікатів, але не підходить для виявлення ND-зображень.

Для порівняння ND-зображень більш придатним є метод хешування LSH (Locality-sensitive hashing), призначений для зниження розмірності багатовимірних даних за ймовірнісним підходом. Цей метод дозволяє будувати структуру для швидкого імовірнісного пошуку багатовимірних векторів, близьких за заданою метрикою до вхідного шаблону. Він є особливо ефективним, коли є у наявності великі обсяги даних (наприклад, колекції зображень ТГД), які необхідно порівняти між собою.

Ефективним підходом до пошуку близьких за змістом зображень є використання згорткових (конволюційних) нейронних мереж (CNN). Згорткові нейронні мережі показують високу точність у задачах класифікації зображень завдяки здатності автоматично витягувати ознаки з зображень на основі згортки з ядрами різних розмірів і форм. Однак, згорткові нейронні мережі мають такі недоліки, як висока кількість параметрів та складність навчання.

Незважаючи на значний успіх згорткових нейронних мереж у вирішенні задачі класифікації зображень, вони все ще стикаються з певними проблемами, наприклад, їх здатністю до узагальнення. У зв'язку з цим виникає потреба у нових моделях, які можуть справлятися з складнішими завданнями та мають більш високу здатність до узагальнення.

Одним із таких нових підходів є використання візуального трансформера (Visual Transformer). Visual Transformer (ViT) – це нейронна мережа, яка використовує механізм уваги обробки вхідних зображень [12,13]. Візуальний трансформер складається з блоків, кожен з яких включає безліч багаторівневих самоуважних (self-attention) механізмів, які можуть шукати взаємодії між різними частинами зображення.

В даній роботі здійснено дослідження можливості комбінованого використання трансформерів та методів хешування для підвищення ефективності виявлення ND-зображень в колекціях та базах даних ТГД.

Пошук ND-зображень за пропонованим алгоритмом передбачає послідовне виконання таких операцій: завантаження колекції зображень ТГД; створення бази даних хешів зображень; пошук дублікатів за зображенням-запитом (еталоном), редукція колекції для пошуку, порівняння пар зображень з використанням ViT.

Робота алгоритму, заснованого лише на використанні ViT, на великих базах даних зображень потребує значних часових ресурсів. Тому пропонований гібридний алгоритм передбачає небхідність попередньої редукції колекції можливих дублікатів заданого зображення. Для цього за допомогою перцептивних хешів будується редукована колекція зображень, в якій представлені всі можливі кандидати на внесення до бази ND-зображень, а також шумові зображення, які помилково були визначені перцептивним хешем як подібні до заданого. Далі на зменшеній колекції реалізується подальша фаза пошуку (з використанням модифікованої моделі Bag of Words, яка фільтрує шуми, отримані на етапі перцептивного хешування). Для побудови словника використовується модифікована модель Bag of Words, на основі якого відбувається побудова гістограми зображень для порівняння.

Остаточне порівняння пар зображень та формування множини виявлених неповних дублікатів здійснюється з використанням ViT (базової або модифікованої архітектури).

Процес пошуку ND-зображень в аналізованому просторі зображень передбачає послідовну реалізацію таких етапів:

– хешування кожного зображення аналізованої колекції за допомогою перцептивного хешу DCTBH (етап 1);

– занесення отриманих значень хешів до буферної бази даних зображень ББДЗ (етап 2);

– визначення відстані Хеммінга для кожної пари зображень в ББДЗ (ця операція є найбільш трудомісткою з точки зору процесорного часу, тому доцільно розпаралелити її виконання для зменшення загального часу роботи алгоритму) (етап 3);

– за результатами аналізу значень відстаней Хеммінга приймається рішення про віднесення пар близьких зображень до редукованої буферної бази даних зображень (РББДЗ), що можуть бути з високою ймовірністю неповними дублікатами (етап 4);

– зображення з РББДЗ аналізуються з використанням модифікованої моделі Bag of Words, що

фільтрує шуми, отримані на етапі перцептивного хешування зображення, та дозволяє спростити остаточне виділення ND-зображень. Ця модель передбачає: компактне представлення зображень РББДЗ у вигляді вектора ознак (локальних дескрипторів); побудуву словника візуальних слів заданого розміру; побудува гістограми (глобального дескриптора) для кожного зображення на основі словника (етап 5);

– здійснюється порівняння зображень за результатами аналізу кореляції їхніх гістограм (значення кореляції є прямо пропорційним подібності зображень за дескрипторами); якщо кореляція перевищує задане порогове значення, то відповідні зображення відносять до подвійно редукованої бази даних зображень РББДЗ 2 (етап 6);

– здійснюється остаточне порівняння пар зображень РББДЗ 2 та формується множина виявлених неповних дублікатів з використанням візуального трансформера (базової або модифікованої архітектури) (етап 7).

Вибір архітектури системи пошуку ND-зображень ТГД було здійснено згідно з завданнями реалізації етапів алгоритму (рис. 7) .

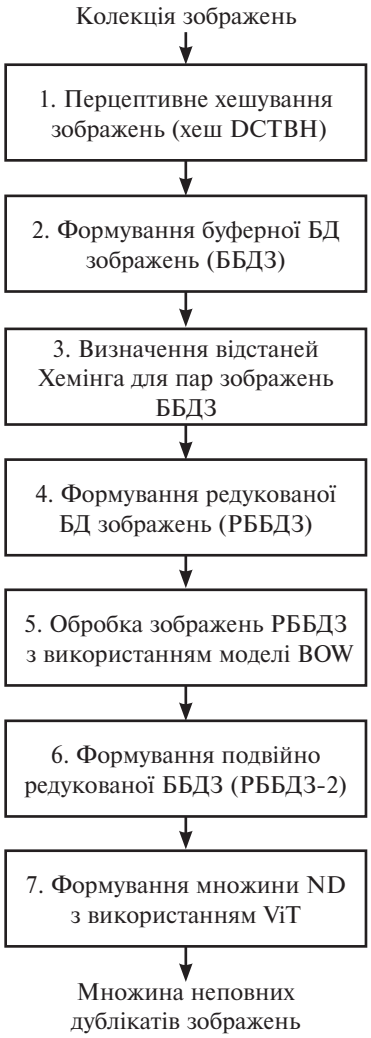

Рис. 7. Схема пошуку ND-зображень з використанням ViT та перцептивного хешування

Операції попередньої редукції колекції зображень здійснюються з метою зменшення обчислювальної складності виявлення можливих неповних дублікатів дублікатів заданого зображення.

Зупинимося докладніше на моделях ViT, які були застосовані в подальшому для остаточного порівняння пар зображень ND-зображень РББДЗ 2 та формування множини виявлених неповних дублікатів (етап 7).

Базова модель ViT у загальному вигляді наведена на рис. 8. До цієї моделі входять, зокрема, механізми розбиття вхідного зображення на патчі, енкодер та шари ViT.

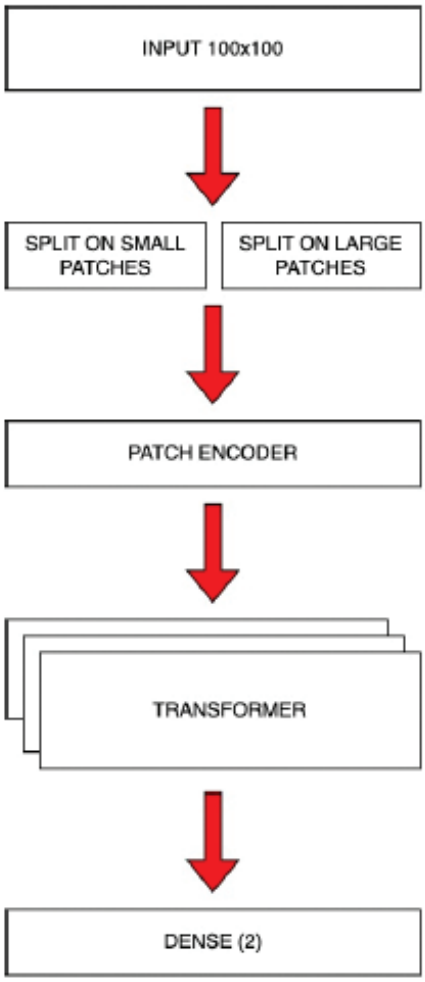

Рис. 8. Базова модель ViT з патчами різних розмірів

На вхід моделі подається набір трьохканальних зображень 100x100, що розбиваються на патчі (кількість та розмір патчів належать до настроюваних параметрів моделі), кожен з яких розгортається в послідовність пікселів. Під час кодування патчів до їх атрибутів додається позиція на зображенні.

На рис. 9 наведено схему використання в структурі ViT механізму уваги (Multi Head Attention).

Механізм уваги частково вирішує проблему обмеженого доступу декодувальника до повної інформації, представленої на вході ViT.

При обчисленні функції уваги використовуються три ключові компоненти: запит (Query), ключ (Key) та значення (Value). Значення уваги розраховується з використанням матричного множення «запиту» на «ключ», ділення на корень з розмірності «ключа» та матричного множення на «значення». В даній роботі був застосований механізм мультиуваги, який використовує кілька функцій уваги для виділення різних особливостей зображення.

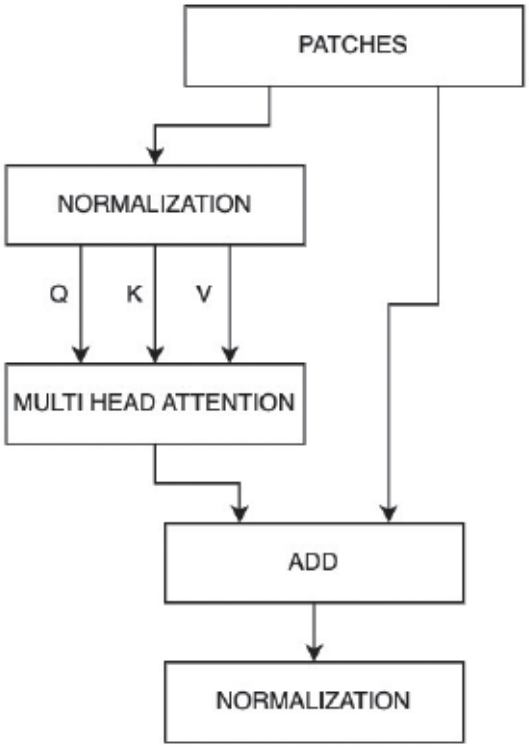

Рис. 9. Схема застосування в ViT механізму уваги

В подальшому будемо вживати такі позначення для досліджуваних архітектур системи пошуку ND-зображень в базах ТГД:

– VIT_BASE_HB – архітектура системи пошуку з ViT, перцептивним хешуванням та використанням BOW-моделі;

– VIT_BASE_H – архітектура системи пошуку з ViT та перцептивним хешуванням;

– VIT_BASE – архітектура системи пошуку з ViT;

– VIT_BASE_L – архітектура системи пошуку з ViT з патчами різних розмірів.

Для усіх варіантів системи пошуку ND-зображень були обрані такі настроювані параметри ViT:

– img_size – розмір, до якого зменшується зображення під час навчання;

– patch_size – розмір однієї сторони квадратних патчів, на які розбивається зображення;

– tfrm_layers – кількість застосувань функції мультиуваги до зображення;

– proj_dim – розмірність, до якої може бути зменшено патч;

– att_heads – кількість застосувань одиничної функції уваги до зображення.

Для ViT з двома розмірами патчів (VIT_BASE_L) параметр patch_size задається у форматі (x, y), де x та y – розміри звичайного та розширеного патчів відповідно.

4. Тестування запропонованої технології

Результати тестування модуля кластеризації та індексування «CLASTFIND» (для ТФ ТГД) із застосуванням алгоритму SOINN-index, наведені на рис. 10, дозволяють зробити висновок, що пропонований підхід є більш ефективним, ніж застосування стандартного повнотекстового пошуку, вбудованого в операційну систему, та повнотекстового пошуку MySQL. При проведенні тестування були наступні параметри алгоритму SOINN-index: Для проведення експериментів були вибрані наступні параметри алгоритм: lambda = 100, c = 1, alpha1 = 0.16, alpha2 = 0.25, alpha3 = 0.25, beta = 0.67, gamma = 0.75.

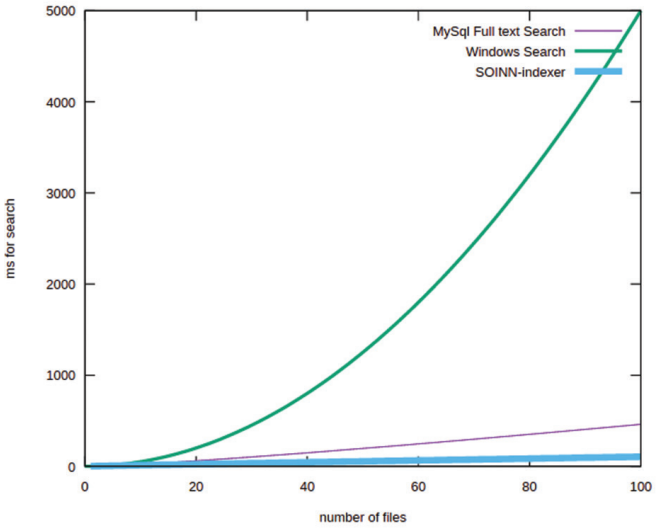

Рис. 10. Результати тестування модуля «CLASTFIND»

Для моделювання пропонованої системи пошуку ND-зображень (ГФ ТГД), були використані зображення з набору даних INRIA Holidays, який гарантовано містив зображення, що можна віднести до класів ND та NND [14].

Здійснювався пошук ND-зображень серед екземплярів заданої колекції, що передбачав застосування лише моделі хешування зображень аналізованої колекції за допомогою перцептивного хешу DCTBH та модифікованої моделі Bag of Words, яка фільтрує шуми, отримані на етапі перцептивного хешування.

Було проаналізовано 1000 зображень датасету, що передбачало необхідність здійснення близько 50000 порівнянь пар зображень (при цьому в тестовій виборці були присутні близько 1000 пар, кваліфікованих в датасеті, як ND-зображення).

Результати моделювання, що містять стандартні показники якості виявлення неповних дублікатів (точність (accuracy), повнота (recall) та F-міра) для

різних архітектур системи пошуку та різних методів попереднього пошуку наведено в таблиці 1.

Таблиця 1

Результати пошуку ND-зображень

Архітектура системи пошуку ND-зображень	Повнота (recall)	Точність (accuracy)	F-міра
VIT_BASE_H	0,96	0,95	0,954
VIT_BASE_HB	0,927	0,978	0,952
VIT_BASE	0,956	0,963	0,959
VIT_BASE_L	0,962	0,97	0,96

Результати моделювання свідчать про поліпшення показників якості пошуку ND-зображень внаслідок гібридного використання методів хешування та ViT. Додаткове використання ViT для порівняння пар зображень з редукованої буферної бази даних дозволяє поліпшити значення повноти, точності та F-міри (разом з підвищенням середнього часу обробки пари зображень). При цьому до остаточно сформованої множини ND-зображень було віднесено 1020 пар, 125 пар було віднесено до множини NND-зображень, а 5 пар – до множини різних зображень.

Для визначення інтегрованих показників подібності ГТД за формулою (7) значення r приймались від 0,1 до 0,8 для різних типів ГФ ТГД: зображення з розширеними підписами: $r = 0{,}8$; зображення зі короткими підписами: $r = 0{,}1$. Як базу ТГД було використано колекцію авторефератів дисертаційних робіт.

Результати тестування запропонованого підходу підтверджують доцільність його використання в системах пошуку та кластеризації текстово-графічних документів.

Висновки

В статті наведено результати розробки та дослідження мультимодальної технології пошуку та кластеризації слабоструктурованих текстово-графічних документів (ТГД) з використанням нейромережевих моделей.

До найбільш важливих завдань реалізації запропонованої технології слід віднести: розроблення нейромережевого модуля кластеризації ТГД (в текстовій частині) з використанням самоорганізованої інкрементної нейронної мережі; розроблення нейромережевої технології пошуку близьких за мультимодальним критерієм ТГД (в графічній частині) з комбінованим використанням трансформерів та методів хешування.

Структура запропонованого модуля кластеризації «CLASTFIND» містить чотири основні блоки: «Інтерфейс», «Парсер», «Обробник» та «Пошукова машина». Для вирішення завдання кластеризації текстових документів було використано модифікацію базової версії алгоритму SOINN (Self-Organizing Incremental Neural Network), який заснований на побудові нейронної мережі з двома шарами. Перший шар використовується визначення топологічної структури кластерів, а другий визначення числа кластерів і виявлення вузлів-прототипів.

Результати тестування системи кластеризації та індексування із застосуванням алгоритму SOINN-index підтверджують доцільність його використання при вирішенні широкого класу завдань політематичної класифікації текстово-графічних документів. Перспективним розвитком методу є проведення експериментів щодо удосконалення процедури пошуку електронних документів шляхом модифікації алгоритму індексації текстових даних та зберігання індексів.

В роботі досліджено також завдання пошуку неповних дублікатів зображень за допомогою нейромережевих візуальних трансформерів, які аналізують характер взаємодії між різними частинами зображення. Розглянуто можливість комбінованого застосування архітектури ViT разом з перцептивним хешуванням та використанням BOW-моделі для формування кластерів подібних зображень.

Наведено загальну характеристику та методи вирішення проблеми пошуку подібних за змістом зображень (ND-зображень) в колекціях текстово-графічних документів або виявлення високого рівня подібності аналізованих зображень (наприклад, для виявлення плагіату графічних частин в порівнюваних документах).

Список літератури

[1] A Guide on Word Embeddings in NLP. URL: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp (last accessed: 15.05.2024).

[2] Kumar J., Ye P., Doermann D. Structural Similarity for Document Image Classification and Retrieval. Pattern Recognition Letters, 2014, vol 43, pp. 119-126. DOI: 10.1016/j.patrec.2013.10.030

[3] Удовенко С.Г. Метод порівняння текстово-графічних фрагментів в електронних документах за гібриднимкритерієм / С.Г. Удовенко, Л.Е. Чала, Є.С. Кушвід // Біоніка интелекту. – 2019. – Вып. 1 (92). – С. 71-76.

[4] Text Normalization for Natural Language Processing (NLP) [Електронний ресурс]. – 2021. – Режим доступу: https://towardsdatascience.com/text-normalization-for-natural-language-processing-nlp-70a314bfa646

[5] Удовенко С. Мультимодальна система порівняння текстово-графічних фрагментів в електронних документах / С. Удовенко, Л. Чала, Є. Кушвід // Матеріали Міжнародної наукової конференції «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту (ISDMCI'2019)». – Залізний Порт, 2019 – С. 184-186.

[6] Veena R, Ramesh D., Automatic text summarization – A systematic literature review, World Journal of Advanced Engineering Technology and Sciences, 2023, 08(02), pp. 126–129. DOI: https://doi.org/10.30574/wjaets.2023.8.2.0080

[7] B. Kumar, B. Sathya, P. Anima. Automatic Keyword Extraction for Text Summarization in Multi-document Articles (2017). European Journal of Advances in Engineering and Technology, Vol. 4 (6), P. 410 – 427.

[8] Wiwatcharakoses C., Berrar D. (2021) A self-organizing incremental neural network for continual supervised learning. Expert Systems with Applications. Volume 185, 115662. https://doi.org/10.1016/j.eswa.2021.115662

[9] Wiwatcharakoses C. , Berrar D. (2020) SOINN+, a Self-Organizing Incremental Neural Network for Unsupervised Learning from Noisy Data Streams. Expert Systems with Applications. Volume 143, 113069. https://doi.org/10.1016/j.eswa.2019.113069.

[10] Nguyen N,·Coustaty M., Ogier J. (2020). An adaptive document recognition system for latrines. International Journal on Document Analysis and Recognition (IJDAR) (2020) 23:115–128. https://doi.org/10.1007/s10032-019-00346-9

[11] Лізунов, П. П. Автоматичний аналіз подібностей схем та діаграм в електронних текстових документах / П. П. Лізунов, А. О. Білощицький, Л. Е. Чала, С.В. Білощицька, О. Ю. Кучанський, С. Г. Удовенко // Управління розвитком складних систем. – 2017. – № 28. – С. 147 – 1568. Кузнецов О.В.

[12] Liu Z., Lin Y., Cao Y., Hu H., Wei Y., Zhang Z., Lin S., Guo B. Swin. Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021. P. 9992-10002. DOI: doi.org/10.1109/ICCV48922.2021.00986

[13] Chen C., Fan Q., Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. ICCV. 2021. P.1-10. DOI: doi.org/10.1109/ICCV48922.2021.00041

[14] Misra R. News Category Dataset (2018). URL: https://www.kaggle.com/rmisra/news-category-dataset (last accessed: 15.05.2024).

Надійшла до редколегії 18.09.2024