DOI: 10.32347/2412-9933.2021.48.85-94
УДК 005.8

Лізунов Петро Петрович
Доктор технічних наук, професор, завідувач кафедри будівельної механіки, *orcid.org/0000-0003-2924-3025*
Київський національний університет будівництва і архітектури, Київ
Білощицький Андрій Олександрович
Доктор технічних наук, професор, проректор з науки та інновацій, *orcid.org/0000-0001-9548-1959*
Astana IT University, Нур-Султан
Кучанський Олександр Юрійович
Доктор технічних наук, доцент, доцент кафедри інформаційних систем та технологій,
orcid.org/0000-0003-1277-8031
Київський національний університет імені Тараса Шевченка, Київ
Андрашко Юрій Васильович
Кандидат технічних наук, доцент кафедри системного аналізу і теорії оптимізації,
orcid.org/0000-0003-2306-8377
ДВНЗ «Ужгородський національний університет», Ужгород

КОМБІНОВАНІ МЕТОДИ ІДЕНТИФІКАЦІЇ НЕПОВНИХ ДУБЛІКАТІВ У НАУКОВИХ ПУБЛІКАЦІЯХ

Анотація. *Розглянуто розпізнавання неповних дублікатів зображень та таблиць. З метою розпізнавання графічних даних (для класифікації та стиснення зображень) використовується вейвлет-аналіз з набором класичних характеристичних функцій: вейвлети Морле і Хаара, вейвлет мексиканський капелюх тощо. Застосовуються також особливі види фільтрів, що будуються на основі так званих риджлет-, курвлет- та бимлет-перетворень. Розглянуто основні класичні методи кластеризації колекції зображень, що можуть бути використані для пошуку неповних дублікатів у графічних даних електронних документів. Проаналізовано метод Гарріса, який дає змогу визначати опорні точки зображень за рахунок вимірювання інтенсивності яскравості зображення. Також проаналізовано технологію SIFT (масштабно-інваріантне перетворення ознак), яка є потужним засобом формування системи інваріантних структурних ознак, розглянуто ще один клас методів, які вирізняються простотою реалізації та застосування для виявлення неповних дублікатів зображень – хеш-методи. Описано, що для RGB-зображення існує три таких сигнали: яскравість у каналах Red, Green та Blue. В обробці сигналів і суміжних галузей перетворення Фур'є зазвичай розглядається декомпозиція сигналу на частоти та амплітуди. Розглянуто метод виявлення контекстно-залежних значень та індексації текстових даних, який допомагає знаходити неповні дублікати в таблицях з урахуванням текстового і числового представлення даних. Аналогічно за описаним методом можна провести індексацію даних числового і текстового типів, якщо вони розміщуються не в таблиці, а всередині контенту електронного документа. Результати дослідження використовуються в комплексі із системою виявлення неповних дублікатів у наукових документах, зокрема дисертаціях на здобуття наукового ступеня.*

Ключові слова: послідовності збігів; неповні дублікати; плагіат; наукове дослідження; наукова публікація

Вступ

Задача аналізу контенту електронних документів з метою виявлення подібностей та неповних дублікатів є актуальною для наукового співтовариства, фахових видань, спеціалізованих вчених рад, оскільки може бути цінним інструментом для уникнення зловживань та плагіату у сфері вищої освіти, сприятиме академічній доброчесності [1 – 4].

Відомо, що електронні документи можуть містити дані різних типів: текстові дані, зображення, таблиці, математичні формули, діаграми та схеми. До кожного із цих типів даних потрібно застосовувати спеціальні методи аналізу. Зокрема, аналіз зображень у електронних документах можна здійснювати з використанням стандартних програмних додатків та нестандартних додаткових модулей. Для покращення якості розпізнавання

інколи необхідно виконувати додаткові методи обробки: відновлення зашумлених зображень, аналіз форми та текстури, виділення фрагментів зображень, зіставлення двох зображень за ключовими точками тощо.

Більшість методів розпізнавання зображень базуються на результатах розпізнавання ключових точок зображень, хешування зображень, застосування стохастичної геометрії, ланцюгів Маркова, методу перцептивного хешування тощо. Ці методи доволі успішно виконують задачу пошуку неповних дублікатів зображень. Всі методи розпізнавання зображення можна поділити на дві категорії:

– методи, які базуються на виділенні ключових точок;

– методи, що базуються на аналізі окремих пікселів зображення.

Отже, в який спосіб цього можна досягти присвячено це дослідження.

Мета статті

Мета статті – ідентифікація неповних дублікатів у наукових публікаціях.

Виклад основного матеріалу

Для забезпечення застосування методів розпізнавання необхідно попередньо провести фільтрацію графічних даних, яка дає змогу спростити і підвищити чутливість методів аналізу цих даних на подібність. Особливістю методів фільтрації є те, що вони допомагають виділити особливі характеристики локальних областей зображень. Фільтрація передбачає використання деяких перетворень до всього зображення, зокрема: бінаризація із заданим пороговим значенням, фільтрація високих частот (фільтр Габора), фільтрація низьких частот (фільтр Гаусса), фільтрація Фур'є тощо. Окремим видом фільтрації зображень є ідентифікація на зображенні елементарних математичних функцій (прямої, параболи, кола тощо). До фільтрів такого типу належать фільтри Гафа, Радона тощо. У випадку складних зображень з великою кількістю фрагментів часто достатньо аналізувати не повне зображення, а тільки його контури. Основними методами фільтрації контурів є методи операторів Лапласа, Кенні, Робертса, Прюітта тощо.

Також з метою розпізнавання графічних даних (для класифікації та стиснення зображень) використовується вейвлет-аналіз з набором класичних характеристичних функцій: вейвлети Морле і Хаара, вейвлет мексиканський капелюх тощо. Існують також особливі види фільтрів, що будуються на основі так званих риджлет-, курвлет- та бимлет-перетворень.

Оскільки реалізовані системи розпізнавання, як правило, націлені на пошук подібних графічних зображень, виникає необхідність у нових дослідженнях, націлених на виявлення неповних дублікатів у зображеннях та зображень, які є дублікатами, що були оброблені та модифіковані певним чином.

Розглянемо основні класичні методи кластеризації колекції зображень, що можуть бути використані для пошуку неповних дублікатів у графічних даних електронних документів.

Метод Гаррісa. Метод дає змогу визначати опорні точки зображень. Алгоритм заснований на вимірюванні інтенсивності яскравості зображення:

$$E(u,v)=\sum_{x,\,y} w(x,y)[I(x+u,y+v)-I(x,y)]^2, \quad (1)$$

де $w(x,y)$ – функція вікна; $I(x+u,y+v)$ – майбутня інтенсивність; $I(x,y)$ – поточна інтенсивність.

Після апроксимації будується матриця часткових похідних, потім розраховується її детермінант та слід. Це допомагає знайти кутові точки, що вважатимуться ключовими:

$$R=\det M-k\left(trace\right)^2, \quad (2)$$

де R – точка; k – коефіцієнт k $\epsilon(0.04,0.06)$; M – матриця.

Значення R розраховується для кожної точки, після чого обираються тільки ті R, значення яких перевищує визначений поріг. Наприкінці обираються локальні максимуми. Множина отриманих локальних максимумів і є вихідними даними алгоритму.

Особливістю методу є інваріантність до повороту, інваріантність до зміщення інтенсивності, а також метод не має інваріантності до зміни масштабу [5 – 9].

Алгоритми, вихідними даними яких є фрагменти, у більшості полягають у боротьбі з проблемою масштабу точкових алгоритмів. Окремим випадком є алгоритм Laplacian-of-Gaussian (LoG). Цей метод було розроблено для вирішення проблем визначення неповних дублікатів зображень у разі зміни масштабу зображення. Алгоритм LoG:

1. Обирається функція, задана на фрагменті зображення, що є інваріантною до зміни масштабу (наприклад, середня інтенсивність).

2. У кожній точці зображення ця функція розглядається як функція зміни розміру фрагмента.

3. Розраховується локальний максимум цієї функції. У цьому випадку точка максимуму інваріантна до зміни масштабу.

4. Розмір фрагмента, на якому досягається локальний максимум, розраховується для кожного зображення окремо.

5. Відбувається стиснення точки разом з її оточенням. Знаходяться точки, що будуть мати максимальне значення на всіх масштабах.

Істинно особлива точка залишається такою на різних масштабах.

Отже, для кожної точки зображення її оточення «згортається» за формулою:

$$L(x,\sigma)=\sigma^2(I_{xx}(x,\sigma)+I_{yy}(x,\sigma)), \qquad (3)$$

де *x, y* – координати точки; *σ* – масштаб.

Існують алгоритми, що дають змогу не тільки знайти однакові елементи зображень на різних масштабах, але й знайти, наприклад, елемент у повороті. Алгоритми, результатом роботи яких є дескриптори, дуже широко використовуються, тому що їх вихідні дані можуть бути подані на вхід стандартним алгоритмам кластеризації, таким як: k-means, mean-shift тощо. Ці алгоритми описують фрагмент навколо точки за допомогою векторів або матриць ознак. Яскравим прикладом таких методів є SIFT-дескриптори. Технологія SIFT (масштабно-інваріантне перетворення ознак) є потужним засобом формування системи інваріантних структурних ознак [10]. Вона побудована на використанні кращих із сучасних базових принципів локального оброблення, що містять у єдиному комплексі локальну фільтрацію, формування значущих ознак, аналіз простору перетворень, апроксимацію координат ознак тощо. Головний акцент у SIFT зроблено на стійкому визначенні крайових точок (КТ) зображення у плані відділення їх від інших стійких елементів (СЕ). У процесі багатоетапного оброблення шляхом аналізу відмінностей у просторі ознак встановлюється необхідна величина відстані між СЕ, що загалом приводить до високих характеристик розпізнавання. З цих причин перспективним є застосування SIFT у випадку розпізнавання зображень об'єктів в умовах просторових перешкод. Побудова модифікацій SIFT для цих умов робить новий крок у розвитку структурних методів розпізнавання.

Згідно з технологією SIFT [10] зображення $B(x,y)$ перетворюється на множину $L(x,y,k\sigma)=B(x,y)\otimes Q(x,y,k\sigma)$, кожний елемент якої одержують згорткою зображення з маскою Гауссіана $Q(x,y,\sigma)$ (параметр σ). Величина σ змінюється дискретно в межах $[\sigma_0,\sigma_1]$.

Етап відбору стійких точок полягає у виборі найбільш стабільних КТ шляхом детального аналізу властивостей прилеглих даних у просторі $D(x,y,\sigma)$ на основі інтерполяції. Ставиться мета стабілізувати переміщення точок, викликані дискретними властивостями зображення і перетворень над ним. Для визначення більш точного положення КТ здійснюють розвинення $D(x,y,\sigma)$ в ряд Тейлора $D(c)=D+\frac{\partial D^T}{\partial c}c+\frac{1}{2}c^T\frac{\partial^2 D}{\partial c^2}c$, де $c=(x,y)$ – зсув від обраної КТ c^* у межах масштабу σ.

Положення $\hat{c}$ екстремуму функції $D(c)$ за зсувом визначається з рівності нулю похідної. Якщо одна з компонент зсуву $\hat{c}$ перевищує 0,5, то координати нової КТ обчислюються з урахуванням поправки $c^*=c^*+\hat{c}$, і для неї процедура інтерполяції повторюється. Для забезпечення точності координати обчислюються у формі з рухомою комою. Для усунення КТ з низькою контрастністю відкидаються ті, для яких член другого порядку не перевищує 0,03.

Усунення відгуків крайових точок здійснюється шляхом аналізу співвідношення між головними значеннями кривизни у напрямках для кожної виділеної КТ. Визначення кривизни зводиться до знаходження власних значень матриці $H=\begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$ другого порядку аналогічно методу Харіса, де D_{uv} – похідна за змінними u,v. Власні значення α_1,α_2 матриці H у випадку фіксованого σ пропорційні величинам просторової кривизни. Відношення $r=\alpha_1/\alpha_2$ власних значень $(\alpha_1\geq\alpha_2)$ є критерієм для виключення КТ зі списку. Відкидаються ті КТ, для яких одне з напрямків кривизни значно перевищує інше, що означає наявність «істотного» відгуку краю. Величини $\nu=tr^2(H)/det(H)$ та $\xi=(r+1)^2/r$ є близькими. Значення ν є мінімальним, якщо α_1,α_2 приблизно рівні. Критерієм відхилення КТ є умова $\nu>\xi_0$, де ξ_0 – поріг для величини r_0. У практичному використанні $r_0=10$.

Кожній КТ призначається одна чи декілька орієнтацій. Цим досягається інваріантність до обернення. Обраховується амплітуда $m(x,y)$ та орієнтація $\theta(x,y)$ градієнта:

$$m(x,y)=\sqrt{[L(x+1,y)-L(x-1,y)]^2}+\sqrt{[L(x,y+1)-L(x,y-1)]^2}, \qquad (4)$$

$$\theta(x,y)=arctg\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}.$$

На основі відгуків значень і напрямків градієнта для точок у оточенні розміром 3 σ (з центром у КТ) будується гістограма, яка містить

36 стовпчиків. Кожен стовпчик охоплює діапазон орієнтацій у 10 градусів. Значеннями гістограми є суми $v = \sum m_q$, що відповідають напряму з номером q. Для усунення викидів гістограма згладжується і нормується для забезпечення стійкості до змін яскравості, знаходиться напрямок з найбільшим відгуком. Піки гістограми відповідають панівним орієнтаціям градієнта. Стовпці гістограми з відгуком, що перевищує 80% від максимуму, формують додаткові КТ за масштабом, які відрізняються лише напрямками. Ознакова інформація про об'єкт міститься у масиві напрямків $T = \{ t_i \}, t_i = (x, y, m, \theta, \sigma)$. В оточенні 16x16 кожної точки за фіксованим масштабом аналізується множина з 16 фрагментів 4x4, що не перетинаються і утворюють розбиття. Для кожного з фрагментів будується гістограма напрямків, яка охоплює 360 градусів. Один напрямок відповідає діапазону в 40 градусів. Значеннями дескриптора є суми $v_i = \sum_{q=i} m_q$. Для забезпечення інваріантності до поворотів проводиться нормалізація шляхом зсуву щодо направлення КТ. У результаті для кожної КТ маємо дескриптор – вектор, що містить 16x8=128 значень.

Розглянемо ще один клас методів, які вирізняються простотою реалізації та застосування для виявлення неповних дублікатів зображень – хеш-методи [11]. Зображення – це двовимірний неперіодичний сигнал, що визначається залежністю яскравості від горизонтальної та вертикальної координат. Для RGB-зображення існує три таких сигнали: яскравість у каналах Red, Green та Blue. Оскільки в обробці сигналів і суміжних галузях перетворення Фур'є зазвичай розглядається декомпозиція сигналу на частоти та амплітуди, поділимо умовно зображення на 3 частоти:

– на низькій частоті будуть міститися найбільші деталі, загальний розподіл яскравості та кольору, тобто форма об'єкта;

– на середній частоті міститься середня та дрібна деталізація, яка має назву «локальний контраст» і для знятих крупним планом об'єктів є фактурою поверхні;

– на високій частоті міститься наддрібна деталізація, про яку часто говорять «мікроконтраст» і яка відповідає за різкість.

Для порівняння зображень рекомендується використовувати низькі частоти. Алгоритм в цьому разі такий:

Крок 1. Зменшується розмір зображення. Найшвидший спосіб позбутися високих частот – зменшити зображення. У цьому випадку зображення зменшується до роздільної здатності 8x8, а отже, загальне число пікселів становить 64. Можна не дбати про пропорції, зображення просто стискається в квадрат вісім на вісім. Отже, хеш буде відповідати всім варіантам зображення, незалежно від розміру та співвідношення сторін.

Крок 2. Прибирається колір зображення. Маленьке зображення переводиться в градації сірого так, що хеш зменшується втричі: з 64 пікселів (64 значення червоного, 64 зеленого і 64 синього) всього до 64 значень кольору.

Крок 3. Далі для кожного з кадрів обчислюється середнє значення пікселів.

Крок 4. Формується ланцюжок бітів. Кожен піксель порівнюється із середнім значенням, отже, якщо він більше середнього значення, то до клітинки хешу записується 1, в іншому випадку – 0.

Крок 5. Будування хешу. 64 окремих біта переводяться до єдиного 64-бітового значення. Порядок несуттєвий, якщо він зберігається незмінним.

Підсумковий хеш не зміниться, якщо картинку масштабувати, стиснути або розтягнути. Зміна яскравості або контрасту, або навіть маніпуляції з кольорами теж сильно не вплинуть на результат. І найголовніше − такий алгоритм відрізняється високою швидкістю.

Для порівняння пари зображень обчислюється відстань Гемінга (підраховується кількість різних бітів). Нульова відстань означає, що це, найвірогідніше, однакові картинки (або варіації одного зображення). Дистанція 5 означає, що картинки чимось відрізняються одна від одної, але в цілому все одно досить схожі одна на одну. Якщо дистанція 10 або більше, то це, ймовірно, зовсім різні зображення. Детальніше про аналіз зображень з метою виявлення неповних дублікатів описано в роботах [12−17].

Метод виявлення контекстно-залежних значень та індексації текстових даних

Розглянемо метод на прикладі індексації даних таблиці. Оскільки контент таблиць складається з даних текстового та числового типів, опишемо метод індексації на прикладі цих даних.

Нехай B – деяка вхідна таблиця, а $\overline{B} = \{ B_1, B_2, \ldots, B_p \}$ – таблиці, які відібрані з тексту та зберігаються в загальній базі таблиць, p – кількість таблиць у базі. Завдання полягає у визначенні такої множини таблиць $B^* = \{ B_1^*, B_2^*, \ldots, B_l^* \}$, що $B^* \subset \overline{B}, \quad l < p$, для яких виконується умова:

$$F\left(B, B_i^*\right) < \lambda, \quad i = \overline{1,l}, \tag{5}$$

де λ – порогове значення; F – відстань між таблицями.

Наявність у базі хоча б однієї такої таблиці з множини $\overline{B}$, для якої виконується умова (5), свідчить про те, що таблиця B є запозиченою.

Коміркою K_{ij} деякої таблиці B називається такий елемент таблиці, який утворюється шляхом перетину i -го рядка та j -го стовпця цієї таблиці.

Контентом або даними комірки K_{ij} деякої таблиці B називається таке значення (числове, текстове, типу дата тощо), яке відповідає цій комірці таблиці. Позначимо контент комірки K_{ij} через $Cont\left(K_{ij}\right),\ \ i=\overline{1,r_1},\ \ j=\overline{1,r_2},$ де r_1 – кількість рядків таблиці *B,* а r_2 – кількість стовпців таблиці *B.*

Як було визначено в попередньому пункті, контент комірок може вичерпуватися такими типами: числовий, текстовий, дата, рисунок, формула, комбінований. У разі, якщо в таблиці наявні рисунки та формули, вони виділяються для дослідження окремо. Контент типу дата після зведення до єдиного формату можна вважати звичайним текстовим рядком. Таким чином, спростивши можливі варіанти типів контенту комірок, можна вважати, що контент таблиці представляється у вигляді двійки [18]:

$$B=\left\langle N,S\right\rangle, \qquad (6)$$

де *N* – кортеж з числових даних комірок таблиці *B,* а *S* – кортеж з текстових даних комірок таблиці *B*.

Нехай $I=\left\{1,2,\ldots,r_1\right\}$ – множина індексів або номерів рядків таблиці, а $J=\left\{1,2,\ldots,r_2\right\}$ – множина індексів або номерів стовпців таблиці *B*. Проглянемо всі комірки таблиці і визначимо тип їх даних. Якщо дані деякої комірки належать до числового типу, то відповідний контент представляється у вигляді елемента множини $\overline{N}$, якщо текстового типу, то додається до множини $\overline{S}$ за правилом:

$$\overline{N}=\left\{k\ \middle|\ k\in Cont\left(K_{ij}\right),k\in R,i\in I,j\in J\right\}, \qquad (7)$$

$$\overline{S}=\left\{k\ \middle|\ k\in Cont\left(K_{ij}\right),k\in T,i\in I,j\in J\right\}, \qquad (8)$$

де множина T – множина символів потужності L, $card\left(T\right)=L:\ T=\left\{t_1,t_2,\ldots,t_L\right\};$ t_i – окремий символ, $i=\overline{1,N},\quad t_i\in A;\quad A$ – множина атомарних символів формальної мови або алфавіт.

Представимо множини $\overline{N}$ та $\overline{S}$ у вигляді числової послідовності та послідовності рядків відповідно довжини v та w, тобто $N=\left\{n_1,n_2,\ldots,n_v\right\}$ – послідовність із числових значень контентів комірок, $v=card\left(N\right),$ $n_i\in N,\quad i=\overline{1,v},$ а $S=\left\{s_1,s_2,\ldots,s_w\right\}$ – послідовність із рядків контентів комірок $w=card\left(S\right),\ s_j\in S,\ j=\overline{1,w}.$

Розглянемо окремо представлення цих послідовностей у вигляді, зручному для застосування моделей ідентифікації подібності та пошуку неповних дублікатів.

Розглянемо послідовність $S=\left\{s_1,s_2,\ldots,s_w\right\}.$ Кожен її елемент містить текст, що може складатися з одного або кількох слів. Проглянемо послідовно всі елементи послідовності від s_1 до s_w та виберемо з них слова. Слово довільного елемента уніграму S задається у вигляді послідовності

$$S_n^{\beta}=\left\{t_1,t_2,\ldots,t_{\beta}\right\},$$

де $n\in R$ – порядковий номер слова; β – довжина слова, $t_j\in A,\ \ t_j\notin C,\ \ j=\overline{1,\beta};$ $C=\left\{"\_","'","`","\cdot","\ ","-",":",";","\#"\right\}$ – всі небуквенні символи елементів послідовності *S*, $S_n^{\beta}\in s_j,\ \ j\in\left\{1,2,\ldots,w\right\}.$

Сформуємо новий уніграм з усіх слів – елементів уніграму рядків *S*, попередньо виключивши з розгляду так звані стоп-слова. Перелік стоп-слів задамо у вигляді множини:
$M=\left\{"i","ma","але","або","тощо","і т.п.","і т.д."\right\}.$

Тоді нова послідовність слів має такий вигляд:

$$W=\left\{S_1^{\beta_1},S_2^{\beta_2},\ldots,S_m^{\beta_m}\right\},$$

де $\beta_j,\ \ j=\overline{1,m}$ – довжини слів, а m – їх кількість.

Елементи такої послідовності є словами в канонізованій формі. Використовуючи метод плинного вікна, побудуємо сукупність послідовностей:

$$E_1=\left\{S_1^{\beta_1},S_2^{\beta_2},\ldots,S_h^{\beta_h}\right\},$$

$$E_2=\left\{S_2^{\beta_2},S_3^{\beta_3},\ldots,S_{h+1}^{\beta_{h+1}}\right\},$$

$$\ldots$$

$$E_{m-h+1}=\left\{S_{m-h+1}^{\beta_{m-h+1}},S_{m-h+2}^{\beta_{m-h+2}},\ldots,S_{m-1}^{\beta_{m-1}},S_m^{\beta_m}\right\},$$

де h – розмір вікна або кількість елементів побудованих послідовностей $E_1, E_2, \ldots, E_{m-h+1}$.

Далі за методом локально-чутливого хешування представимо сукупність послідовностей $F(W) = (E_1, E_2, \ldots, E_{m-h+1})$ у вигляді бітових рядків, тобто

$$\Delta(W) = \left(I(E_1), I(E_2), \ldots, I(E_{m-h+1})\right), \quad (9)$$

де $I(E_k)$ – елемент індексу, що задає бітовий рядок, який однозначно є послідовністю E_k, $k = \overline{1, m-h+1}$. Тобто:

$$I(E_k) = \{\delta_{k1}, \delta_{k2}, \ldots, \delta_{kc}\}, \quad (10)$$

де $\delta_{kx} \in \{0,1\}$, $k = \overline{1, m-h+1}$, $x = \overline{1, c}$, c – кількість бітів, що є бітовою послідовністю.

Розглянемо числову послідовність $N = \{n_1, n_2, \ldots, n_v\}$ вхідної таблиці *B* та побудуємо для неї набір підпослідовностей за методом плинного вікна, тобто:

$$K_1 = \{n_1, n_2, \ldots, n_g\},$$

$$K_2 = \{n_2, n_3, \ldots, n_{g+1}\},$$

…

$$K_{v-g+1} = \{n_{v-g+1}, n_{v-g+2}, \ldots, n_{v-1}, n_v\},$$

де v – кількість елементів послідовності N, а g – розмір вікна або кількість елементів підпослідовностей $K_1, K_2, \ldots, K_{v-g+1}$. Оскільки елементи побудованих підпослідовностей є дійсними числами, $n_i \in R$, $i = \overline{1, v}$, то ці підпослідовності можуть бути g -вимірними векторами. Тобто, якщо вважати, що задано простір R^g, який має евклідову структуру, можна визначити на цьому просторі метрику ρ між будь-якими двома векторами простору $a \in R^g$ та $b \in R^g : \rho(a,b)$. Причому ця метрика буде задовольняти аксіому тотожності, тобто:

$$\rho(a,b) = 0 \Leftrightarrow a = b,$$

аксіому симетрії:

$$\rho(a,b) = \rho(b,a)$$

та аксіому трикутника для деякого вектора $c \in R^g$:

$$\rho(a,c) \le \rho(a,b) + \rho(b,c).$$

Така метрика або міра близькості (подібності) між такими векторами, які є числовими значеннями контентів таблиць, – складова, яка визначатиме ступінь подібності цих таблиць.

Відповідно до постановки задачі, нехай B – вхідна таблиця, а $B_1, B_2, \ldots, B_p$ – таблиці, які відібрані та зберігаються в загальній базі таблиць, p – кількість таблиць у базі. Завдання полягає у визначенні таких таблиць з бази, для яких виконується умова (4.1).

Нехай побудовано послідовність із текстових даних $S = \{s_1, s_2, \ldots, s_w\}$ та послідовність із числових значень $N = \{n_1, n_2, \ldots, n_v\}$ для таблиці *B*. Оскільки база з таблицями вже відома, то очевидно, що кожна така таблиця є проіндексованою. Тобто для кожної з таблиць $B_1, B_2, \ldots, B_p$ відомі послідовності з текстових даних:

$$S^y = \{s_1^y, s_2^y, \ldots, s_w^y\}$$

та послідовності числових даних:

$$N^y = \{n_1^y, n_2^y, \ldots, n_v^y\}, \; y = \overline{1, p}.$$

Також, очевидно, що для послідовностей слів задано елементи індексу

$$I(E_{k_y}^y) = \{\delta_{k_y 1}^y, \delta_{k_y 2}^y, \ldots, \delta_{k_y c}^y\},$$

$\delta_{k_y x}^y \in \{0,1\}$, $k_y = \overline{1, m_y - h + 1}$, $x = \overline{1, c}$,

c – кількість бітів, що є бітовою послідовністю, а m_y – кількість слів у послідовностях:

$$W^y = \{S_1^{y,\beta_1}, S_2^{y,\beta_2}, \ldots, S_{m_y}^{y,\beta_{m_y}}\},$$

що містять слова S_j^{y,β_j} в канонізованій формі, β_j, $j = \overline{1, m_y}$ – довжини слів.

Побудуємо для рядкової послідовності $S = \{s_1, s_2, \ldots, s_w\}$ вхідної таблиці *B* уніграм зі слів в канонізованій формі:

$$W = \{S_1^{\beta_1}, S_2^{\beta_2}, \ldots, S_m^{\beta_m}\},$$

де β_j, $j = \overline{1, m}$ – довжини слів, а m – їх кількість. Далі методом плинного вікна визначимо послідовності $E_1, E_2, \ldots, E_{m-h+1}$ і за методом локально-чутливого хешування побудуємо елементи індексу:

$$I(E_k) = \{\delta_{k1}, \delta_{k2}, \ldots, \delta_{kc}\},$$

де $\delta_{kx} \in \{0,1\}$, $k = \overline{1, m-h+1}$, $x = \overline{1, c}$,

c – кількість бітів, що є послідовністю.

Розрахуємо відстані Хеммінга від елементів кожного індексу послідовностей вхідної таблиці до елементів індексу послідовностей тих таблиць, що містяться в базі, за формулою:

$$H\left(I\left(E_k\right),I\left(E_{k_y}^y\right)\right)=\frac{1}{c}\sum_{j=1}^{c}\left|\delta_{kj}-\delta_{k_y j}\right|, \quad (11)$$

$k=\overline{1,m-h+1},\ \ k_y=\overline{1,m_y-h+1},\ \ y=\overline{1,p}.$

За умови

$$H\left(I\left(E_k\right),I\left(E_{k_y}^y\right)\right)<\lambda_H \quad (12)$$

для наперед заданого значення параметра $\lambda_H\in[0,1]$, то з ймовірністю 1 можна стверджувати, що елемент індексу з номером k подібний до елемента індексу з номером k_y таблиці з номером y. Це означає, що таблиця з номером y може бути подібною до вхідної таблиці з порогом $\lambda_{\text{Н}}$, тобто в ній міститься неповний дублікат.

Побудуємо для числової скінченної послідовності $N=\{n_1,n_2,\ldots,n_v\}$ вхідної таблиці B набір підпослідовностей $K_1,K_2,\ldots,K_{v-g+1}$. Вважаємо також, що для кожної з таблиць $B_1,B_2,\ldots,B_p$ на основі їх послідовностей числових даних $N^y=\{n_1^y,n_2^y,\ldots,n_v^y\},\ \ y=\overline{1,p}$ побудовані підпослідовності $K_1^y,K_2^y,\ldots,K_{v-g+1}^y$ за методом плинного вікна:

$$K_1^y=\{n_1^y,n_2^y,\ldots,n_g^y\},$$
$$K_2^y=\{n_2^y,n_3^y,\ldots,n_{g+1}^y\},$$
$$\ldots$$
$$K_{v-g+1}^y=\{n_{v-g+1}^y,n_{v-g+2}^y,\ldots,n_{v-1}^y,n_v^y\}.$$

Якщо представити побудовані підпослідовності $K_1,K_2,\ldots,K_{v-g+1}$ та $K_1^y,K_2^y,\ldots,K_{v-g+1}^y$ у вигляді кортежів, то міри подібності між ними визначаються на основі відстані Евкліда, міської метрики або відстані Мінковського. Отримаємо такі y матрицей відстаней:

$$\rho_1\left(K_u,K_r^y\right)=\sqrt{\sum_{j=r}^{g+r-1}\left(n_{j+u-r}-n_j^y\right)^2}, \quad (13)$$

$$\rho_2\left(K_u,K_r^y\right)=\sum_{j=r}^{g+r-1}\left|n_{j+u-r}-n_j^y\right|, \quad (14)$$

$$\rho_3\left(K_u,K_r^y\right)=\left(\sum_{j=r}^{g+r-1}\left|n_{j+u-r}-n_j^y\right|^t\right)^{\frac{1}{t}}, \quad (15)$$

$y=\overline{1,p},\qquad u=\overline{1,v-g+1},\qquad r=\overline{1,v-g+1},$

t – параметр відстані Мінковського.

Далі знаходимо мінімальні значення для кожного рядка матриць $\rho_\tau\left(K_u,K_r^y\right)$ по $r=\overline{1,v-g+1}$, отримаємо для кожного $y=\overline{1,p}$ відстані:

$$\zeta_\tau\left(K_u,K_{\min}^y\right)=\min_{r=1,v-g+1}\left\{\rho_\tau\left(K_u,K_r^y\right)\right\}. \quad (16)$$

У випадку $u=\overline{1,v-g+1}$, для фіксованого $\tau=\overline{1,3}$.

Нормалізуємо значення отриманих відстаней за формулою:

$$\zeta_\tau^N\left(K_u,K_{\min}^y\right)=$$
$$=\frac{\zeta_\tau\left(K_u,K_{\min}^y\right)-\min\limits_{u=1,v-g+1}\left\{\zeta_\tau\left(K_u,K_{\min}^y\right)\right\}}{\max\limits_{u=1,v-g+1}\left\{\zeta_\tau\left(K_u,K_{\min}^y\right)\right\}-\min\limits_{u=1,v-g+1}\left\{\zeta_\tau\left(K_u,K_{\min}^y\right)\right\}}, \quad (17)$$

$$y=\overline{1,p},\ u=\overline{1,v-g+1},\ \tau=\overline{1,3}.$$

Якщо виконується умова

$$\zeta_\tau^N\left(K_u,K_{\min}^y\right)<\lambda_\rho \quad (18)$$

для наперед заданого значення параметра $\lambda_\rho\in[0,1]$, то з ймовірністю 1 можна стверджувати, що вектор з номером u подібний до вектора таблиці з номером y, тобто таблиця містить неповний дублікат. Чим більше значення λ_ρ, тим більш жорсткі вимоги до пошуку неповних дублікатів. На рисунку описано складові індексації даних в таблиці.

Описаний метод дає змогу знаходити неповні дублікати в таблицях з урахуванням текстового та числового представлення даних. Аналогічно за описаним методом можна провести індексацію даних числового і текстового типів, якщо вони розміщуються не в таблиці, а всередині контенту електронного документа.

Рисунок – Метод індексації даних в таблиці залежно від типу даних її контенту

Висновки

Здійснено огляд наявних підходів до розпізнання неповних дублікатів, з'ясовано, що більшість методів розпізнавання зображень базуються на результатах розпізнавання ключових точок зображень, хешування зображень, застосування стохастичної геометрії, ланцюгів Маркова, методу перцептивного хешування тощо. Ці методи доволі успішно виконують задачу пошуку неповних дублікатів зображень. Розглянуто метод фільтрації який допомагає виділити особливості характеристики локальних областей зображення за рахунок використання деяких перетворень бінаризації із заданим пороговим значенням, фільтрації високих частот (фільтр Габора), фільтрації низьких частот (фільтр Гаусса), фільтрації Фур'є. Також розглянуто метод виявлення контекстно-залежних значень та індексації текстових даних, який дає змогу розібрати контент вмісту даних таблиць і пошуку в них неповних дублікатів.

Список літератури

1. Hawkins J. On Intelligence [Text] / Jeff Hawkins. Times Books, 2004. 272 p.
2. Ту Дж., Гонсалес Р. Принципы распознавания образов. Москва: Мир, 1978. 411 с.
3. Яне Б. Цифровая обработка зображений. Москва: Техносфера, 2007. 587 с.
4. Гонсалес Г., Вудс Г. Цифровая обработка зображений. Москва: Техносфера, 2005. 1072 с.
5. Павлидис Т. Алгоритмы машинной графики и обработки зображений. Москва: Радио и связь, 1986. 400 с.
6. Сироджа И. Б. Квантовые модели и методы искусственного интеллекта для принятия решений и управления. Київ: Наукова думка, 2002. 420 с.
7. Фу К., Гонсалес, К. Ли. Робототехника. [пер. с англ. А. А. Сорокина, А. В. Градецкого, М. Ю. Рачкова; под. ред. В. Г. Градецкого]. Москва: Мир. 1989. 624 с.
8. Пименов В. Ю. Простые методы поиска изображений по содержанию. *Труды РОМИП, 2010.* URL: http://romip.ru/ru/2010/.
9. Mojsilović R., Kovačević J., Hu J., Safranek R. J., Ganapathy S. K. Matching and retrieval based on the vocabulary and grammar of color patterns. IEEE Trans. Image Processing, 2000, volume 9, pp. 38-54.
10. Tamura H., Mori S., Yamawaki T. Texture features corresponding to visual perception. IEEE Transactions on System, Man and Cybernatic. 1978, volume 8(6), pp. 460–473.
11. Zhang D., Lu G. Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study. In IEEE International Conference on Multimedia and Expo, 2001, pp. 289–293.
12. Quack T., Monich U., Thiele L., Manjunath B. A System for Largescale, Contentbased Web Image Retrieval. MM'04, October 1016, 2004, New York, USA. P. 120–123.
13. Волосных Д. Ф Использование визуальных особенностей восприятия компонент цветовой модели HSI при поиске изображений по содержанию. *Труды РОМИ 2010.* URL: http://romip.ru/ru/2010/.

14. Васильева Н., Гладышева Ю. Взвешенный CombMNZ для комбинирования результатов поиска изображений по цветовым признакам. *Труды РОМИП 2010*. URL: http://romip.ru/ru/2010/

15. Мельниченко А., Гончаров А. ЛММИИ на РОМИП-2009: Методы поиска изображений по визуальному подобию и детекции нечетких дубликатов изображений. *Труды РОМИП 2009.* URL: http://romip.ru/ru/2009/.

16. Стадник А. С. Анализ кадров видеоряда и вычисление продолжительности сцены используя алгоритм перцептивного хэша *Информатика и компьютерные технологии-2011*. URL: http://ea.donntu.edu.ua:8080/jspui/bitstream/123456789/3955/1/4_%D0%A1%D1%82%D0%B0%D0%B4%D0%BD%D0%B8%D0%BA.pdf

17. Чалая Л. Э., Попаденко П. Ю. Поиск неполных дубликатов в системах анализа цифровых зображений. *Вісник Кременчуцького національного університету імені Михайла Остроградського.* 2014. Вип. 5. С. 42 – 47.

18. Lizunov P., Biloshchytskyi A., Kuchansky A., Biloshchytska S., Chala L. Detection of near duplicates in tables based on the locality-sensitive hashing method and the nearest neighbor method. Eastern-European Journal of Enterprise Technologies. 2016, Vol. 6, Issue 4 (84), P. 4–10.

Стаття надійшла до редколегії 02.11.2021

Lizunov Petro
DSc (Eng.), Professor, Head of the Department of Structural Mechanics, *orcid.org/0000-0003-2924-3025*
Kyiv National University of Construction and Architecture, Kyiv
Biloshchytskyi Andrii
DSc (Eng.), Professor, Vice-Rector for Science and Innovation, *orcid.org/0000-0001-9548-1959*
Astana IT University, Nur-Sultan
Kuchansky Alexander
DSc (Eng.), Associate Professor, Department of Information Systems and Technologies, *orcid.org/0000-0003-1277-8031*
Taras Shevchenko National University of Kyiv, Kyiv
Andrashko Yurii
PhD, Associate Professor, Department of Systems Analysis and Optimization Theory, *orcid.org/0000-0003-2306-8377*
Uzzhorod National University, Uzzhorod

Combined methods for identifying incomplete duplicates in scientific publications

Abstract. *Recognition of incomplete duplicates of images and tables is considered. In order to recognize graphical data (for image classification and compression), wavelet analysis is used with a set of classic characteristic functions: Morlet and Haar wavelets, Mexican hat wavelet, etc. Special types of filters are also used, which are based on the so-called ridgelet, curvlet and beamlet transformations. The main classical methods of image collection clustering that can be used to find incomplete duplicates in the graphic data of electronic documents are considered. The Harris method is analyzed, which allows to determine the reference points of the images by measuring the intensity of the brightness of the image. SIFT (scale-invariant feature transformation) technology, which is a powerful tool for forming a system of invariant structural features, is also analyzed, another class of methods is considered, which are easy to implement and use to detect incomplete duplicate images – hash methods. It is described that there are three such signals for RGB images: brightness in Red, Green and Blue channels. In signal processing and related branches of Fourier transform, decomposition of the signal into frequencies and amplitudes is usually considered. A method for identifying context-sensitive values and indexing textual data is considered, which helps to find incomplete duplicates in tables based on textual and numerical representation of data. Similarly, the described method can be used to index data of numerical and text types, if they are not placed in a table, but inside the content of an electronic document. The results of the research are used in combination with the system of detection of incomplete duplicates in scientific documents, in particular dissertations for the degree.*

Keywords: sequence of matches; incomplete duplicates; plagiarism; scientific research; scientific publication

References

1. Hawkins, J. (2004). On Intelligence. Times Books, 272.
2. Tu, Dzh., Gonsales, R. (1978). Principles of pattern recognition. Moscow: Mir, 411.
3. Yane, B. (2007). Digital image processing. Moscow: Technosphere, 587.
4. Gonsales, G., Vuds, G. (2005). Digital image processing. Moscow: Technosphere, 1072.
5. Pavlidis, T. (1986). Algorithms for computer graphics and image processing. Moscow: Radio and Communications, 400.
6. Sirodzha, I. B. (2002). Quantum models and artificial intelligence methods for decision making and management. Kyiv: Scientific opinion, 420.
7. Fu, K., Gonsales, K. Li. (1989). Robotics. [per. from English. A. A. Sorokin, A. V. Gradetsky, M. Yu. Rachkov; under. ed. V. G. Gradetsky]. Moscow: Mir, 624.

8. Pimenov, V. Yu. (2010). Simple methods of image search by content. *Proceedings of ROMIP*. URL: http://romip.ru/ru/2010/.

9. Mojsilović, R., Kovačević, J., Hu, J., Safranek, R. J., Ganapathy, S. K. (2000). Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Trans. Image Processing*, 9, 38–54.

10. Tamura, H., Mori, S., Yamawaki, T. (1978). Texture features corresponding to visual perception. *IEEE Transactions on System, Man and Cybernatic*, 8 (6), 460–473.

11. Zhang, D., Lu, G. (2001). Content-Based Shape Retrieval Using Different Shape Descriptors: A Comparative Study. *In IEEE International Conference on Multimedia and Expo*, 289–293.

12. Quack, T., Monich, U., Thiele, L., Manjunath, B. (2004). A System for Largescale, Contentbased Web Image Retrieval. MM'04, October 1016, 2004, New York, USA, 120–123.

13. Volosnykh, D. F. (2010). Using the visual features of the perception of the components of the HSI color model when searching for images by content. Proceedings of ROMI. URL: http://romip.ru/ru/2010/.

14. Vasil'eva, N., Gladysheva, Yu. (2010). Weighted CombMNZ for combining image search results by color features. Proceedings of ROMIP-2010. URL: http://romip.ru/ru/2010/

15. Mel'nichenko, A. Goncharov, A. (2009). LMMII at ROMIP-2009: Methods for image search by visual similarity and detection of fuzzy image duplicates. Proceedings of ROMIP-*2009*. URL: http://romip.ru/ru/2009/.

16. Stadnik, A. S. Video sequence frame analysis and scene duration calculation using perceptual hash algorithm Informatics and computer Technologies-2011. URL: http://ea.donntu.edu.ua:8080/jspui/bitstream/ 123456789/3955/1/4_%D0%A1%D1%82%D0%B0%D0%B4%D0%BD%D0%B8%D0%BA.pdf

17. Chalaya, L. E., Popadenko, P. Yu. (2014). Search for incomplete duplicates in digital image analysis systems. *Bulletin of Kremenchug National University named after Mykhailo Ostrogradsky*, 5, 42–47.

18. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Biloshchytska, S., Chala, L. (2016). Detection of near duplicates in tables based on the locality-sensitive hashing method and the nearest neighbor method. *Eastern-European Journal of Enterprise Technologies*, 6, 4 (84), 4–10.

Посилання на публікацію

APA Lizunov, Petro, Biloshchytskyi, Andrii, Kuchansky, Alexander & Andrashko, Yurii. (2021). Combined methods for identifying incomplete duplicates in scientific publications. *Management of Development of Complex Systems,* 48, 85–94, dx.doi.org\10.32347/2412-9933.2021.48.85-94.

ДСТУ Лізунов П. П., Білощицький А. О., Кучанський О. Ю., Андрашко Ю. В. Комбіновані методи ідентифікації неповних дублікатів у наукових публікаціях. *Управління розвитком складних систем.* Київ, 2021. № 48. С. 85 – 94, dx.doi.org\10.32347/2412-9933.2021.48.85-94.