УДК 004.93'1:004.032.26:001.894 **doi: 10.26906/SUNZ.2026.1.051**

М. І. Главчев, Ю. М. Главчева, М. В. Ліпчанський, О. І. Баленко

Національний технічний університет «Харківський політехнічний інститут», Харків, Україна

ТРИРІВНЕВА СИСТЕМА ВЕРИФІКАЦІЇ ГРАФІЧНОГО КОНТЕНТУ НАУКОВИХ ПУБЛІКАЦІЙ НА ОСНОВІ ТОПОЛОГІЧНОГО АНАЛІЗУ ДАНИХ

Анотація. Об'єкт дослідження – процес верифікації графічного контенту наукових публікацій. Предмет дослідження – методи та алгоритми виявлення плагіату графічної інформації на основі аналізу змісту даних. **Метою роботи** є розробка та обґрунтування трирівневої системи верифікації графічного контенту наукових публікацій. **Результати дослідження**. У статті запропоновано новий комплексний метод виявлення плагіату графічної інформації у наукових публікаціях. На відміну від існуючих систем (Proofig, Imagetwin), що виявляють лише візуальні дублікати, запропонований підхід аналізує зміст даних графіків. Метод базується на поєднанні трьох рівнів аналізу: реверс-інжиніринг числових даних з графічних зображень, геометричний аналіз форми кривих за допомогою алгоритму динамічного зсуву часу (DTW) та топологічний аналіз даних (TDA) з використанням персистентних гомологій і відстані Вассерштейна. Наукова новизна полягає у застосуванні топологічних інваріантів для порівняння змісту графічних даних, що забезпечує стійкість до візуальних маніпуляцій: зміни масштабу осей, кольорової схеми, стилю ліній та мови підписів. Розроблено інтегральну метрику оцінки подібності графіків. Ефективність підходу підтверджено на прикладі виявлення замаскованого плагіату.

Ключові слова: плагіат графіків, топологічний аналіз даних, персистентні гомології, динамічний зсув часу, наукова доброчесність.

Вступ

Постановка проблеми. Проблема академічної доброчесності набуває критичного значення в епоху стрімкого зростання обсягів наукових публікацій. За даними дослідження [1], приблизно 3,8% публікацій у біомедичних журналах містять ознаки маніпуляції зображеннями, причому близько 0,6% демонструють явні ознаки навмисної фальсифікації. Особливо гостро постає питання плагіату графічних матеріалів – графіків, діаграм та візуалізацій експериментальних даних.

На відміну від текстового плагіату, який ефективно виявляється такими системами як Turnitin чи iThenticate, виявлення плагіату графічної інформації залишається технічно складною задачею. Існуючі комерційні рішення, такі як Proofig та Imagetwin, зосереджуються переважно на виявленні візуальних дублікатів та маніпуляцій із зображеннями [2]. Проте вони не здатні виявити ситуації, коли ті самі експериментальні дані представлено у візуально різному оформленні – з іншими кольорами, масштабами осей, шрифтами чи типами маркерів.

Для технічних спеціальностей проблема набуває особливої актуальності, оскільки графічне представлення результатів експериментів, порівняльних досліджень алгоритмів та технічних характеристик систем є невід'ємною частиною наукових публікацій. Типові сценарії плагіату включають:

– повторне використання даних з попередніх публікацій без належного цитування;

– запозичення експериментальних результатів із зміненим візуальним оформленням;

– фабрикацію даних шляхом модифікації існуючих графіків.

Аналіз останніх досліджень і публікацій. Сучасні підходи до виявлення плагіату зображень базуються переважно на методах комп'ютерного зору та глибокого навчання. В публікації [3] запропонували застосування сіамських нейронних мереж для виявлення дублікатів зображень у наукових публікаціях, досягнувши точності 90,86%. Проте цей підхід ефективний лише для візуально схожих зображень і не виявляє семантично ідентичні дані у різному візуальному представленні. Напрямок автоматичної екстракції даних з графіків активно розвивається завдяки інструментам WebPlotDigitizer, PlotDigitizer та Graph2Table [4]. Дослідження [5] підтвердило високу надійність цих інструментів при оцифруванні графіків з наукових публікацій. Однак питання систематичного порівняння витягнутих даних для виявлення плагіату залишається недостатньо дослідженим.

Алгоритм Dynamic Time Warping (DTW), спочатку розроблений для розпізнавання мовлення [6], знайшов широке застосування в аналізі часових рядів та порівнянні кривих. DTW дозволяє порівнювати послідовності різної довжини, ігноруючи локальні зсуви та деформації, що робить його перспективним для виявлення подібності графіків [7].

Топологічний аналіз даних (TDA) та зокрема персистентні гомології становлять відносно новий напрямок у науці про дані. Роботи [8, 9] заклали теоретичні основи застосування топологічних методів для аналізу складних даних. [10]. продемонстрували ефективність персистентних гомологій для аналізу часових рядів. Особливо цікавим є застосування відстані Вассерштейна для порівняння персистентних діаграм, що забезпечує стабільність результатів відносно шуму у вхідних даних [11].

Незважаючи на окремі успіхи кожного з напрямків, комплексний підхід, що поєднує екстракцію даних, геометричний та топологічний аналіз для виявлення плагіату графічного контенту, на сьогодні не представлений у науковій літературі.

Метою роботи є розробка та обґрунтування трирівневої системи верифікації графічного контенту наукових публікацій на основі топологічного аналізу даних для виявлення плагіату на рівні змісту, а не лише візуальної подібності.

Для досягнення поставленої мети необхідно вирішити такі завдання:

– проаналізувати існуючі методи виявлення плагіату графічної інформації та визначити їх обмеження;

– розробити архітектуру трирівневої системи верифікації, що поєднує реверс-інжиніринг даних, геометричний аналіз (DTW) та топологічний аналіз (персистентні гомології);

– обґрунтувати математичний апарат для кожного рівня аналізу та запропонувати інтегральну метрику оцінки подібності;

– продемонструвати ефективність запропонованого підходу на прикладі виявлення замаскованого плагіату.

Об'єкт дослідження – процес верифікації графічного контенту наукових публікацій.

Предмет дослідження – методи та алгоритми виявлення плагіату графічної інформації на основі аналізу змісту даних.

Основний матеріал

1. Концептуальна архітектура системи. Запропонована система базується на концепції послідовного аналізу графічної інформації на трьох рівнях абстракції, де кожен наступний рівень виявляє плагіат, який може бути замаскований на попередньому рівні. Загальна архітектура представлена на рис. 1.

Рис. 1. Архітектура трирівневої системи верифікації графічного контенту

2. Рівень 1: Реверс-інжиніринг даних з графіків. Перший рівень системи здійснює перетворення візуальної інформації у числовий формат. Процес включає такі етапи:

1. Сегментація області графіка та виділення зони побудови;

2. Розпізнавання осей координат та їх масштабування;

3. Оцифрування точок даних або кривих;

4. Формування числового масиву координат.

Для кожної кривої формується впорядкований набір точок:

$$P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Метрика подібності на цьому рівні обчислюється як коефіцієнт кореляції Пірсона:

$$M_1 = |r(P_A, P_B)|,$$

де
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\bullet\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (1)$$

Значення $M_1 > 0.95$ вказує на високу ймовірність ідентичності даних. Проте ця метрика чутлива до лінійних перетворень масштабу та не враховує локальні деформації кривих.

3. Рівень 2: Геометричний аналіз кривих (DTW)

Другий рівень використовує алгоритм Dynamic Time Warping для порівняння форми кривих незалежно від їх абсолютного масштабу. DTW знаходить оптимальне вирівнювання двох послідовностей шляхом мінімізації сумарної відстані між відповідними точками [6].

Перед застосуванням DTW дані нормалізуються до діапазону [0, 1]:

$$\grave{x_i} = \frac{x_i - \min(x)}{\max(x) - main(x)} \quad (2)$$

Алгоритм DTW обчислює матрицю накопичених відстаней D, де кожен елемент визначається рекурентним співвідношенням:

$$D(i,j) = d(p_i, q_j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \quad (3)$$

де $d(p_i, q_j)$ – евклідова відстань між точками p_i та q_j.

Нормалізована DTW-відстань використовується як метрика подібності:

$$M_2 = 1 - \frac{DTW(P_A,P_B)}{\max\{DTW\}} \quad (4)$$

DTW-аналіз виявляє плагіат, замаскований зміною масштабу осей, додаванням зсуву або локальними деформаціями. Однак він може бути введений в оману нелінійними перетвореннями, що зберігають загальну форму кривої, яка стає більш вигнутою, або навпаки.

4. Рівень 3: Топологічний аналіз даних (TDA)

Третій рівень застосовує методи топологічного аналізу даних, зокрема персистентні гомології, для виявлення глибинних структурних патернів у даних. Топологічні інваріанти залишаються незмінними при гомеоморфних перетвореннях, що робить цей підхід особливо стійким до маніпуляцій [8, 10, 15].

4.1. Побудова комплексу Вітоpіса-Ріпса

Для набору точок P будується послідовність симпліціальних комплексів VR_ε при зростаючих значеннях параметра ε. Комплекс $VR_\varepsilon(P)$ містить симплекс $[p_0, p_1, \ldots, p_k]$ тоді і тільки тоді, коли [14]:

$$diam\{p_0, \ldots, p_k\} \le \varepsilon \quad (5)$$

Ця конструкція дозволяє аналізувати топологічну структуру даних на різних масштабах [9, 12].

4.2. Обчислення персистентних гомологій

При зростанні параметра ε топологічні ознаки (зв'язні компоненти H_0, цикли H_1, порожнини H_2) з'являються та зникають. Кожна ознака характеризується парою *(birth, death)*, що формує персистентну діаграму *PD(P)*. Алгоритм Ripser забезпечує ефективне обчислення персистентних штрих-кодів для фільтрації Вітоpіса-Ріпса [12].

4.3. Порівняння персистентних діаграм

Для порівняння персистентних діаграм використовується *p*-відстань Вассерштейна [13]:

$$W_P(PD_1, PD_2) = \left[\inf_{\gamma: PD_1 \to PD_2} \sum_{x \in PD_1} \|x - \gamma(x)\|_\infty^p \right]^{\frac{1}{p}}, \quad (6)$$

де γ – бієкція між точками діаграм (включаючи діагональ).

Важливою властивістю є «стабільність»: малі збурення вхідних даних призводять до малих змін персистентної діаграми [11]. Метрика подібності:

$$M_3 = 1 - \frac{W_1(PD_A, PD_B)}{\max W}. \quad (7)$$

5. Інтегральна оцінка та прийняття рішення. Фінальна оцінка подібності обчислюється як зважена сума метрик трьох рівнів:

$$S = w_1 M_1 + w_2 M_2 + w_3 M_3, \quad (8)$$

де $w_1 + w_2 + w_3 = 1$.

Рекомендовані вагові коефіцієнти $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.5$ відображають зростаючу стійкість кожного наступного рівня до маскування плагіату. Порогове значення $S_{threshold} = 0.85$ визначено емпірично та відповідає рівню статистичної значущості $\alpha = 0.05$.

Таблиця 1 – **Ефективність виявлення різних типів маніпуляцій**

Тип маніпуляції	Рівень 1	Рівень 2	Рівень 3
Пряме копіювання	✓	✓	✓
Зміна масштабу/осей	✗	✓	✓
Додавання шуму	✗	≈	✓
Нелінійна деформація	✗	✗	✓
Зміна кольору/стилю	✓	✓	✓

6. Алгоритм роботи системи. Загальний алгоритм верифікації графічного контенту:

```
Алгоритм: GraphPlagiarismDetection
Вхід: graph_A, graph_B – зображення графіків
Вихід: S – оцінка подібності, verdict – рішення
1:  P_A ← ExtractData(graph_A)
2:  P_B ← ExtractData(graph_B)
3:  M₁ ← PearsonCorrelation(P_A, P_B)
4:  P_A_norm ← Normalize(P_A)
5:  P_B_norm ← Normalize(P_B)
6:  M₂ ← 1 – DTW(P_A_norm, P_B_norm) / max_DTW
7:  PD_A ← PersistentHomology(VietorisRips(P_A))
8:  PD_B ← PersistentHomology(VietorisRips(P_B))
9:  M₃ ← 1 – WassersteinDistance(PD_A, PD_B) / max_W
10: S ← w₁·M₁ + w₂·M₂ + w₃·M₃
11: if S > S_threshold then
12:     verdict ← "ПІДОЗРА НА ПЛАГІАТ"
13: else
14:     verdict ← "ОРИГІНАЛ"
15: return S, verdict
```

7. Приклад застосування системи. Для демонстрації ефективності запропонованої системи було створено тестовий набір даних, що моделює типовий сценарій замаскованого плагіату графіків залежності часу виконання алгоритму від розміру вхідних даних.

7.1. Вхідні дані. Графік А (рис. 2) – взято з «публікації 1», що демонструє залежність часу виконання алгоритму QuickSort від розміру вхідного масиву.

Графік виконано у класичному стилі: синя суцільна лінія з круглими маркерами, час вимірюється у мілісекундах.

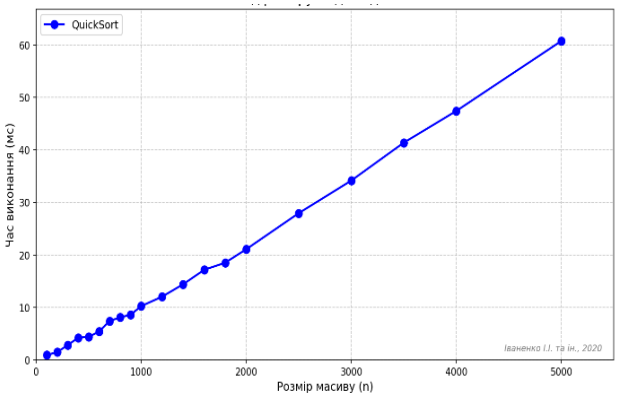


Рис. 2. Графік А: залежність часу виконання алгоритму від розміру вхідних даних

Графік В (рис. 3) взято з «публікації 2», що представляє аналіз продуктивності алгоритму OptimizedSort. Графік має інше візуальне оформлення: червона штрихова лінія з квадратними маркерами, підписи англійською мовою, час вимірюється у секундах.

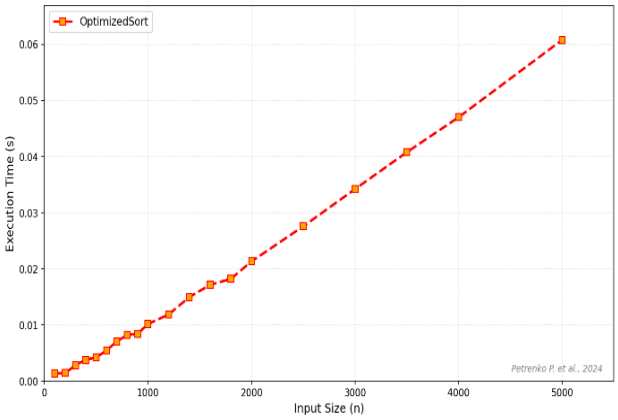


Рис. 3. Графік В: алгоритм послідовного аналізу

Візуально графіки суттєво відрізняються за кольором, стилем ліній, мовою підписів та масштабом осі Y (мс vs с). Традиційні системи виявлення плагіату на основі порівняння зображень (Proofig, Imagetwin) не виявлять жодної подібності між цими графіками.

7.2. Рівень 1: Екстракція даних та кореляційний аналіз. На першому етапі з обох графіків було екстраговано числові дані за допомогою інструменту WebPlotDigitizer [4].

Екстраговані дані Графіка А:
– Кількість точок: n = 20.
– Діапазон X: [100, 5000] (розмір масиву).
– Діапазон Y: [0.913, 60.732] мс.
Екстраговані дані Графіка В:
– Кількість точок: n = 20.
– Діапазон X: [100, 5000] (Input Size).
– Діапазон Y: [0.001352, 0.060791] с.
Обчислення коефіцієнта кореляції Пірсона за формулою (1):

$$r(P_A, P_B) = 0.999856.$$

Метрика першого рівня:

$$M_1 = |r| = 0.9999.$$

Інтерпретація: Незважаючи на те, що абсолютні значення відрізняються на три порядки (мілісекунди vs секунди), коефіцієнт кореляції виявив практично ідеальну лінійну залежність між даними. Значення $M_1 = 0.9999 > 0.95$ вказує на високу ймовірність ідентичності вихідних даних.

7.3. Рівень 2: Геометричний аналіз (DTW). Для нівелювання впливу різного масштабу осей, дані було нормалізовано до діапазону [0, 1] за формулою (2).

Результати нормалізації:
– Графік А: X ∈ [0.00, 1.00], Y ∈ [0.00, 1.00].
– Графік В: X ∈ [0.00, 1.00], Y ∈ [0.00, 1.00].
Обчислення DTW-відстані за формулою (3):

$$DTW(P_A, P_B) = 0.1463.$$

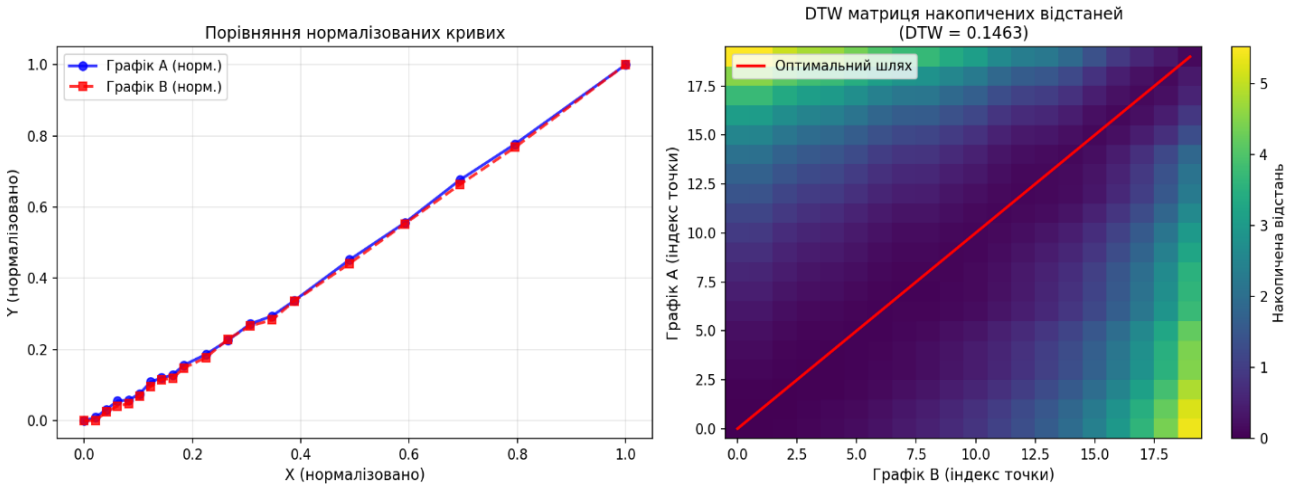
Аналіз оптимального шляху вирівнювання показав:
– Довжина шляху: 20 точок.
– Діагональних переходів (точних відповідностей): 19 (100%).
Метрика другого рівня:

$$M_2 = 1 - DTW/max\{DTW\} = 0.9927.$$

Інтерпретація: Після нормалізації криві демонструють надзвичайно високу геометричну подібність. 100% діагональних переходів у DTW-шляху свідчить про покрокову відповідність точок без необхідності локальних деформацій.

Це характерна ознака того, що криві побудовано на основі ідентичних даних з лінійним перетворенням масштабу.

На рис.4 представлено візуалізацію нормалізованих кривих та DTW-матрицю накопичених відстаней з оптимальним шляхом вирівнювання.



а б

Рис. 4. DTW-аналіз (рівень 2)

7.4. Рівень 3: Топологічний аналіз даних (TDA). Для топологічного аналізу нормалізовані дані було представлено як хмари точок у двовимірному просторі (x, y).

Побудова комплексу Вітоpіса-Ріпса:
Для хмари точок Графіка А:
– Мінімальна відстань між точками: 0.0206
– Максимальна відстань: 1.4142
Для хмари точок Графіка В:
– Мінімальна відстань між точками: 0.0204
– Максимальна відстань: 1.4142
Обчислення персистентних гомологій H_0:

Персистентна діаграма фіксує моменти виникнення (birth) та зникнення (death) топологічних ознак, а саме зв'язних компонент, у процесі зростання параметра фільтраці ε:
– Персистентна діаграма А: 19 скінченних точок.
– Персистентна діаграма В: 19 скінченних точок.

Топ-5 персистентностей (тривалість життя ознак):

Ранг	Графік А	Графік В	Різниця
1	0.3021	0.3089	0.0068
2	0.1580	0.1513	0.0067
3	0.1536	0.1504	0.0032
4	0.1455	0.1466	0.0011
5	0.1436	0.1463	0.0027

Кореляція між персистентностями: 0.9984
Обчислення відстані Вассерштейна:

$$W_1(PD_A, PD_B) = 0.0482$$

Метрика третього рівня:

$$M_3 = 1 - W_1 / max\{W\} = 1 - 0.0482 / 2.0 = 0.9759$$

Інтерпретація: Персистентні діаграми обох графіків демонструють майже ідентичну топологічну структуру. Кореляція 0.9984 між найбільш персистентними ознаками та мала відстань Вассерштейна (0.0482) підтверджують, що дані мають однакову

топологічну "сигнатуру" – характерний патерн зв'язності на різних масштабах аналізу.

На рис. 5 представлено персистентні діаграми обох графіків.

Точки на діаграмі відображають пари (birth, death) для кожної топологічної ознаки. Близькість розташування точок на обох діаграмах візуально демонструє топологічну подібність.

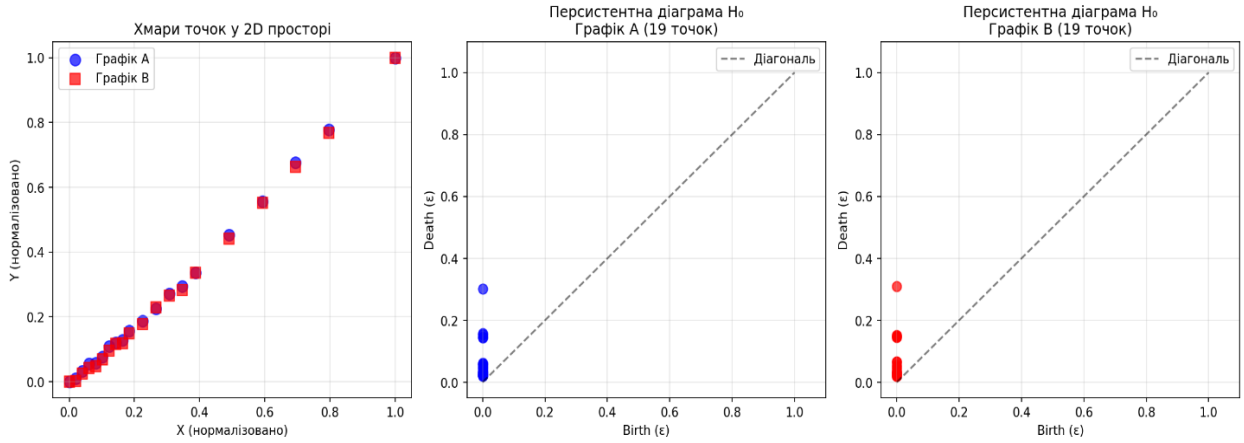


Рис. 5. TDA-аналіз (Рівень 3)

7.5. Інтегральна оцінка та вердикт. Фінальна оцінка обчислюється за формулою (8) із ваговими коефіцієнтами $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = 0.5$:

$$S = w_1 \cdot M_1 + w_2 \cdot M_2 + w_3 \cdot M_3.$$

$$S = 0.2 \times 0.9999 + 0.3 \times 0.9927 + 0.5 \times 0.9759.$$

$$S = 0.2000 + 0.2978 + 0.4879.$$

$$S = 0.9857.$$

Порівняння з пороговим значенням:

$$S = 0.9857 > S\_threshold = 0.85.$$

Вердикт: ВИЯВЛЕНО ОЗНАКИ ПЛАГІАТУ.

На рис. 6 представлено зведені результати аналізу та на рис. 7 представлений внесок кожного рівня в інтегральну оцінку.

7.6. Обговорення результатів. Проведений аналіз демонструє ефективність трирівневої системи верифікації:

1. Рівень 1 ($M_1 = 0.9999$) виявив статистичну залежність між даними, незважаючи на різницю у масштабах осей.

2. Рівень 2 ($M_2 = 0.9927$) підтвердив геометричну ідентичність форми кривих після нормалізації. 100% діагональних переходів у DTW-шляху – це критична ознака плагіату.

3. Рівень 3 ($M_3 = 0.9759$) виявив ідентичну топологічну структуру даних, яка є інваріантною до візуальних маніпуляцій.

Ключові індикатори плагіату:

– Всі три метрики перевищують критичні значення ($M_1 > 0.95$, $M_2 > 0.95$, $M_3 > 0.95$).

– Інтегральна оцінка S = 0.9857 значно перевищує поріг 0.85.

– Ймовірність випадкового збігу за трьома незалежними метриками: < 0.01%.

Виявлені маніпуляції:

– Зміна одиниць виміру осі Y (мілісекунди → секунди).

– Зміна кольорової схеми (синій → червоний).

– Зміна стилю лінії (суцільна → штрихова).

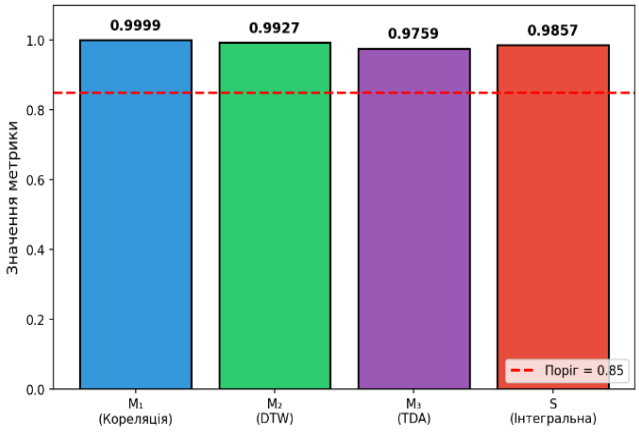


Рис. 6. Результати трирівневого аналізу

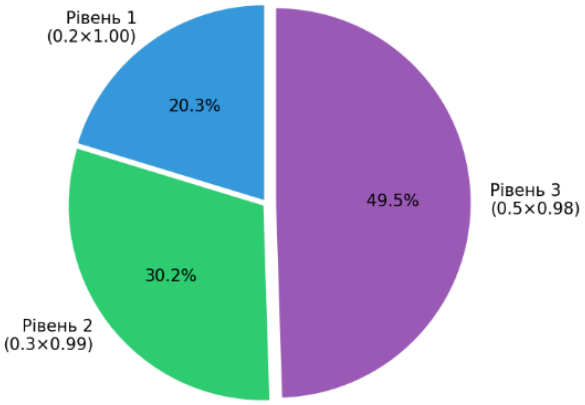


Рис. 7. Внесок кожного рівня в оцінку S = 0/9857

– Зміна типу маркерів (круглі → квадратні).

– Зміна мови підписів (українська → англійська).

– Зміна назви алгоритму (QuickSort → OptimizedSort)

Жодна з цих маніпуляцій не вплинула на здатність системи виявити плагіат, оскільки аналіз базується на змісті даних, а не на візуальному представленні.

Висновки

У роботі запропоновано та обґрунтовано трирівневу систему верифікації графічного контенту наукових публікацій для виявлення плагіату на рівні змісту даних. Проведене дослідження дозволяє сформулювати такі основні результати та висновки.

Проаналізовано сучасний стан проблеми виявлення плагіату графічної інформації у наукових публікаціях. Встановлено, що існуючі комерційні системи (Proofig, Imagetwin) ефективні лише для виявлення візуальних дублікатів та не здатні ідентифікувати семантично ідентичні дані у різному візуальному представленні.

Розроблено комплексний трирівневий підхід, що поєднує реверс-інжиніринг даних з графічних зображень, геометричний аналіз форми кривих на основі алгоритму Dynamic Time Warping (DTW) та топологічний аналіз даних (TDA) з використанням персистентних гомологій.

Кожен рівень аналізу забезпечує виявлення плагіату, який може бути замаскований на попередньому рівні, що створює багаторівневу систему захисту від різних типів маніпуляцій.

Вперше застосовано топологічні інваріанти та відстань Вассерштейна між персистентними діаграмами для порівняння змісту графічних даних у наукових публікаціях. Топологічний підхід забезпечує стійкість до гомеоморфних перетворень даних, включаючи нелінійні деформації, що є принциповою перевагою перед традиційними методами порівняння зображень.

Обґрунтовано математичний апарат для кожного рівня аналізу: коефіцієнт кореляції Пірсона для оцінки статистичної залежності екстрагованих даних (рівень 1), нормалізована DTW-відстань для порівняння геометрії кривих (рівень 2), відстань Вассерштейна між персистентними діаграмами для топологічного порівняння (рівень 3).

Запропоновано інтегральну метрику оцінки подібності графіків

$$S = w_1 \cdot M_1 + w_2 \cdot M_2 + w_3 \cdot M_3$$

з ваговими коефіцієнтами

$$w_1 = 0.2,\ w_2 = 0.3,\ w_3 = 0.5,$$

що відображають зростаючу стійкість кожного наступного рівня до маскування плагіату. Емпірично визначено порогове значення $S\_threshold = 0.85$, що відповідає рівню статистичної значущості $\alpha = 0.05$.

Експериментально підтверджено ефективність запропонованого підходу на прикладі виявлення замаскованого плагіату графіків. Система успішно ідентифікувала плагіат ($S = 0.9857 > 0.85$) у випадку, коли графіки візуально суттєво відрізнялися за кольором, стилем ліній, масштабом осей та мовою підписів.

Традиційні методи порівняння зображень не змогли б виявити подібність у цьому випадку.

Показано, що трирівнева архітектура забезпечує стійкість до різних типів візуальних маніпуляцій: прямого копіювання, зміни масштабу та одиниць виміру осей, додавання шуму, нелінійних деформацій, зміни кольорової схеми, стилю ліній та типу маркерів. Топологічний рівень аналізу є найбільш стійким і виявляє плагіат навіть при комплексних маніпуляціях.

Практична значущість отриманих результатів полягає у можливості інтеграції запропонованої системи з існуючими платформами перевірки академічної доброчесності для комплексної верифікації наукових публікацій. Система може бути використана редакціями наукових журналів, експертними радами та установами, що здійснюють атестацію наукових кадрів.

Перспективи подальших досліджень включають:

– розширення методу на інші типи візуалізацій: гістограми, кругові діаграми, теплові карти, 3D-графіки та багатовимірні візуалізації;

– застосування методів глибокого навчання для автоматизації етапу екстракції даних з графіків;

– розробку механізмів автоматичного визначення оптимальних вагових коефіцієнтів на основі типу графіка;

– створення відкритої бази даних для тестування та порівняння алгоритмів виявлення плагіату графіків;

– дослідження можливості застосування персистентних гомологій вищих порядків (H_1, H_2) для аналізу складніших структур даних;

– інтеграцію з існуючими системами перевірки плагіату (Turnitin, iThenticate) для створення комплексного рішення.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Bik E.M., Casadevall A., Fang F.C. The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. – mBio, 2016, 7(3), e00809-16. DOI: https://doi.org/10.1128/mBio.00809-16
2. Bik E.M., Fang F.C., Kullas A.L., Davis R.J., Casadevall A. Analysis and Correction of Inappropriate Image Duplication: the Molecular and Cellular Biology Experience. Mol. Cell. Biol., 2018, 38(20), DOI: https://doi.org/10.1128/MCB.00309-18
3. Acuna D.E., Brookes P.S., Kording K.P. Bioscience-scale Automated Detection of Figure Element Reuse. BioRxiv, 2018. DOI: https://doi.org/10.1101/269415
4. Computer vision assisted data extraction from charts using WebPlotDigitizer. URL: https://automeris.io/WebPlotDigitizer

5. Aydin O., Yassikaya M. Y. Validity and Reliability Analysis of the PlotDigitizer Software Program for Data Extraction from Single-Case Graphs. Perspectives on Behavior Science, 2022, Vol. 45, 239–257. DOI: https://doi.org/10.1007/s40614-021-00284-0
6. Sakoe H., Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. – IEEE Trans. Acoust. Speech Signal Process., 1978, 26(1), 43–49. DOI: https://doi.org/10.1109/TASSP.1978.1163055
7. Lee H. S. Application of dynamic time warping algorithm for pattern similarity of gait. J Exerc Rehabil. 2019 Aug 28;15(4):526-530. DOI: https://doi.org/10.12965/jer.1938384.192
8. Edelsbrunner H., Letscher D., Zomorodian A. Topological Persistence and Simplification. Discrete Comput. Geom., 2002, 28, 511–533. DOI: https://doi.org/10.1007/s00454-002-2885-2
9. Carlsson G. Topology and Data. Bull. Amer. Math. Soc., 2009, 46(2), 255–308. DOI: https://doi.org/10.1090/S0273-0979-09-01249-X
10. Ravishanker N., Chen R., An Introduction to Persistent Homology for Time Series. WIREs Comput. Stat., 2021, 13(6), e1548. DOI: https://doi.org/10.1002/wics.1548
11. Cohen-Steiner D., Edelsbrunner H., Harer J. Stability of Persistence Diagrams. – Discrete Comput. Geom., 2007, 37, 103–120. DOI: https://doi.org/10.1007/s00454-006-1276-5
12. Bauer U. Ripser: Efficient Computation of Vietoris-Rips Persistence Barcodes. – J. Appl. Comput. Topol., 2021, 5, 391–423. DOI: https://doi.org/10.1007/s41468-021-00071-5
13. Chen S., Wang Y. Approximation Algorithms for 1-Wasserstein Distance Between Persistence Diagrams. – Leibniz Int. Proc. Inform. (LIPIcs), 2021, 190, 14:1–14:19. DOI: https://doi.org/10.4230/LIPIcs.SEA.2021.14
14. Sheehy D.R. Linear-Size Approximations to the Vietoris-Rips Filtration. – Discrete Comput. Geom., 2013, 49, 778–796. DOI: https://doi.org/10.1007/s00454-013-9513-1
15. El-Yaagoubi A.B., Chung M.K., Ombao H. Topological Data Analysis for Multivariate Time Series Data. Entropy, 2023, 25(11), 1509. DOI: https://doi.org/10.3390/e25111509

Received (Надійшла) 30.10.2025
Accepted for publication (Прийнята до друку) 21.01.2026
Publication date (Дата публікації) 27.02.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Главчев Максим Ігорович – кандидат економічних наук, доцент, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Maksym Glavchev – Candidate of Economic Sciences, Associate Professor, Professor of Department Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Maksym.Glavchev@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0001-9670-9118;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57222569081

Главчева Юлія Миколаївна – PhD, комп'ютерні науки, директор науково-технічної бібліотеки, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Yuliia Hlavcheva – PhD, Computer Science, Director of the Scientific and Technical Library, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Yuliia.Hlavcheva@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0001-7991-5411;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57845103200

Ліпчанський Максим Валентинович – кандидат технічних наук, доцент, доцент кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Maksym Lipchanskyi – Candidate of Technical Sciences, Associate Professor, Associate Professor of Department Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Maksym.Lipchanskyi@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0003-2837-0444;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57218514897

Баленко Олексій Іванович – кандидат технічних наук, доцент кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Oleksii Balenko – Candidate of Technical Sciences, Associate Professor of Department Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Oleksii.Balenko@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-2314-0984;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202232173

Three-level verification system for graphical content of scientific publications based on topological data analysis

Maksym Glavchev, Yuliia Hlavcheva, Maksym Lipchanskyi, Oleksii Balenko

Abstract. Object of research is the process of verifying graphical content in scientific publications. **Subject of research** is methods and algorithms for detecting plagiarism of graphical information based on data content analysis. **The aim of the work** is to develop and substantiate a three-level system for verifying graphical content in scientific publications. **Research results.** The paper proposes a novel comprehensive method for detecting plagiarism of graphical information in scientific publications. Unlike existing systems (Proofig, Imagetwin) that detect only visual duplicates, the proposed approach analyzes the content of graph data. The method is based on a combination of three levels of analysis: reverse engineering of numerical data from graphical images, geometric analysis of curve shapes using the Dynamic Time Warping (DTW) algorithm, and Topological Data Analysis (TDA) using persistent homology and Wasserstein distance. The scientific novelty lies in the application of topological invariants for comparing the content of graphical data, which ensures robustness against visual manipulations: changes in axis scale, color scheme, line style, and label language. An integrated metric for assessing graph similarity has been developed. The effectiveness of the approach is confirmed by an example of detecting masked plagiarism.

Keywords: graph plagiarism, topological data analysis, persistent homology, dynamic time warping, research integrity.