УДК 004.852:004.932:004.738.5 DOI https://doi.org/10.35546/kntu2078-4481.2025.4.2.11

В. А. КРІСІЛОВ
доктор технічних наук, професор,
професор кафедри інженерії програмного забезпечення
Національний університет «Одеська політехніка»
ORCID: 0000-0003-1092-6977

О. В. ІЩЕНКО
кандидат технічних наук,
доцент кафедри механіки, автоматизації та інформаційних технологій
Одеський національний університет імені І. І. Мечникова
ORCID: 0000-0002-7882-4718

МЕТОДИ ВІЗУАЛЬНОГО, ТЕКСТОВОГО ТА ІНТЕРНЕТ-АНАЛІЗУ ДАНИХ У СУЧАСНИХ СИСТЕМАХ DATA MINING

У статті розглянуто комплексний підхід до аналізу даних у сучасних системах Data Mining, який поєднує методи текстового, візуального та інтернет-орієнтованого опрацювання інформації. Показано, що різнорідність сучасних даних, які охоплюють текстові, поведінкові, структуровані та динамічні інформаційні потоки, зумовлює потребу у використанні багаторівневих аналітичних технологій, здатних одночасно працювати з різними типами представлень та моделями. Проведено огляд наукових джерел, у межах якого підкреслено тенденції розвитку моделей оброблення природної мови (зокрема глибинного навчання), інструментів візуальної аналітики та методів web mining, а також окреслено їхні переваги та обмеження у практичних застосуваннях.

У роботі обґрунтовано доцільність інтеграції трьох груп методів у єдину аналітичну платформу, що забезпечує глибоке семантичне опрацювання даних, підвищену інтерпретованість результатів і можливість роботи з високодинамічними інтернет-потоками. Для оцінювання ефективності запропонованого підходу було розроблено прототип системи, який тестували на даних інтернет-каталогу, що містили текстові відгуки користувачів, журнали поведінкової активності та структуровані атрибути товарів. Проведене експериментальне дослідження показало, що поєднання текстових та поведінкових веб-ознак забезпечує покращення якісних метрик класифікації порівняно з використанням лише текстових даних. Інтеграція візуальної аналітики, у свою чергу, суттєво зменшує час інтерпретації результатів та підвищує зручність роботи аналітиків, що підтверджено як операційними показниками, так і суб'єктивними оцінками спеціалістів.

Результати дослідження свідчать, що інтегрований підхід є перспективним для побудови масштабованих, адаптивних та інтерпретованих систем Data Mining, орієнтованих на реальні умови оброблення великих даних. Подальші напрями роботи пов'язані з розвитком універсальних моделей об'єднання текстових, поведінкових, графових та просторово-часових даних, а також удосконаленням інструментів пояснюваного штучного інтелекту для аналізу складних аналітичних процесів.

Ключові слова: *Data Mining, інтелектуальний аналіз даних, текстовий аналіз, візуальна аналітика, web mining, інтернет-дані, машинне навчання, глибинне навчання, інтегровані аналітичні системи, пояснюваний штучний інтелект.*

V. A. KRISILOV
Dr. Sci., Professor,
Professor at the Software Engineering Department
Odesa Polytechnic National University
ORCID: 0000-0003-1092-6977

O. V. ISHCHENKO
Ph.D., Associate Professor at the Department of Mechanics, Automation
and Information Technologies
Odesa I. I. Mechnikov National University
ORCID: 0000-0002-7882-4718

METHODS OF VISUAL, TEXTUAL, AND WEB DATA ANALYSIS IN MODERN DATA MINING SYSTEMS

The article examines a comprehensive approach to data analysis in modern Data Mining systems, which integrates methods of textual, visual, and Internet-oriented information processing. It is shown that the heterogeneity of contemporary data, encompassing textual, behavioral, structured, and dynamic information streams, necessitates the use of multi-level analytical technologies capable of operating simultaneously with different types of representations and models. A review of scientific sources was conducted, highlighting the development trends of natural language processing models (including deep learning techniques), visual analytics tools, and web mining methods, as well as outlining their advantages and limitations in practical applications.

The study substantiates the feasibility of integrating these three groups of methods into a unified analytical platform that ensures deep semantic processing of data, enhanced interpretability of results, and the ability to handle highly dynamic Internet data flows. To evaluate the effectiveness of the proposed approach, a prototype system was developed and tested on an online-catalog dataset that included user textual reviews, behavioral activity logs, and structured product attributes. The experimental results demonstrate that combining textual features with behavioral web indicators improves classification performance compared to relying solely on textual data. The integration of visual analytics, in turn, significantly reduces the time required for interpreting the results and enhances user experience, which is confirmed both by operational metrics and expert evaluations.

The findings indicate that the integrated approach is promising for developing scalable, adaptive, and interpretable Data Mining systems designed for real-world big data processing conditions. Future research should focus on advancing universal models capable of combining textual, behavioral, graph-based, and spatiotemporal data, as well as improving explainable artificial intelligence tools for analyzing complex analytical processes.

Key words: *Data Mining, intelligent data analysis, text analysis, visual analytics, web mining, Internet data, machine learning, deep learning, integrated analytical systems, explainable artificial intelligence.*

Постановка проблеми

У сучасних системах інтелектуального аналізу даних зростає потреба у методах, здатних ефективно опрацьовувати різнорідні джерела інформації: візуальні об’єкти, текстові корпуси та інтернет-дані. Класичні алгоритми Data Mining орієнтовані переважно на структуровані набори даних і не забезпечують повноцінної підтримки гібридних форматів, що поєднують зображення, текстові фрагменти, веб-логі чи потокові дані з мережевих ресурсів. Це створює методологічний розрив між реальними потребами аналітичних застосунків та інструментальною базою, яка часто не враховує специфіку мультимодальних даних.

Проблема ускладнюється тим, що кожен тип даних має власні особливості: візуальні дані потребують семантичного виділення ознак, текстові – лінгвістичного моделювання, а інтернет-дані, в свою чергу, – врахування динамічності, нестабільності та високої варіативності джерел. Використання окремих, несумісних між собою методів аналізу призводить до втрати інформації, зниження точності моделей та ускладнення інтеграції результатів у єдину аналітичну систему.

У таких умовах виникає потреба у формуванні цілісних методів, що поєднують інструменти візуального, текстового та інтернет-аналізу в межах єдиної методологічної основи. Їх інтеграція дозволяє підвищити інформативність ознак, забезпечити оброблення неструктурованих та слабоструктурованих даних, а також адаптувати аналітичні процеси до особливостей сучасних систем Data Mining. Формування таких узгоджених методів є ключовим кроком до підвищення точності прогнозування, оптимізації процесів класифікації та покращення загальної ефективності інтелектуальних систем.

Аналіз останніх досліджень і публікацій

Сучасні системи Data Mining орієнтовані на роботу з різними типами даних – текстовими, візуальними та інтернет-даними, що потребує поєднання кількох аналітичних підходів. У текстовому аналізі ключовим завданням залишається перетворення неструктурованих масивів у структуровані представлення, що забезпечується методами очищення, лінгвістичного подання та моделювання закономірностей [1, с. 26959]. Ефективність таких систем значно зростає завдяки ML- (machine learning, алгоритми машинного навчання) і DL-алгоритмам (deep learning, алгоритми глибинного навчання): зокрема, LSTM (Long Short-Term Memory, різновид рекурентних нейронних мереж) демонструє найвищу точність класифікації текстів – до 92% [2, с. 1].

У візуальному аналізі відбувається перехід до інтерактивних методів, де користувач безпосередньо впливає на результати навчання моделі. Інтерактивна візуалізація дозволяє досягати високої продуктивності навіть за малих вибірок [3, с. 3513]. Особливу увагу приділено візуальній аналітиці для глибинних мереж, оскільки такі моделі залишаються складними для інтерпретації; запропонована interrogative-рамка (спосіб аналізу глибинних моделей через систему структурованих запитань) структурує підходи до пояснення та налагодження DL-моделей [4, с. 2674]. Практичну користь підтверджують дослідження в бізнес-аналітиці, де поєднання VDM (Visual Data Mining, візуальний data mining) та EDA (Exploratory Data Analysis, розвідувальний аналіз даних) забезпечує ефективне виявлення патернів у даних інвентаризації [5, с. 1808].

Інтернет-аналіз даних представлений як класичним веб-майнінгом – методами, орієнтованими на аналіз вмісту, поведінки користувачів та структурних зв'язків [6, с. 285], так і сучасними підходами до аналізу соціальних медіа. Нові аналітичні платформи бізнес-інтелекту використовують великі масиви даних із соціальних платформ для підвищення конкурентоспроможності бізнесу в постпандемічний період [7, с. 1]. Окремий напрям становить аналіз просторово-часових даних, який досліджує закономірності динамічних процесів у просторі та часі, але стикається з проблемами інтеграції візуалізації та аналітики [8, с. 1441].

У медичних системах, що базуються на Інтернеті речей, ML застосовується для аналізу великих фізіологічних потоків даних; огляди підкреслюють важливість моделі 5V та проблеми масштабованості [9, с. 234]. У текстовому аналізі актуальним стає застосування графових нейронних мереж, які дозволяють моделювати складні зв'язки між словами та документами, забезпечуючи точнішу класифікацію [10, с. 1].

Отже, сучасні дослідження демонструють зближення текстового, візуального та веб-аналізу в єдині інтегровані системи Data Mining, орієнтовані на різнорідні дані, масштабованість і підвищення інтерпретованості моделей.

Формулювання мети дослідження

Метою дослідження є розроблення узагальненої методології поєднання візуальних, текстових та інтернет-орієнтованих методів аналізу даних у сучасних системах Data Mining з урахуванням їхніх особливостей, обмежень та сфер застосування.

Для досягнення цієї мети потрібно вирішити наступні задачі:

– визначити спільні та відмінні риси між підходами до оброблення текстових, візуальних і веб-даних;

– узагальнити сучасні алгоритмічні та технологічні тенденції у відповідних підсистемах Data Mining;

– сформувати інтегрований підхід, що забезпечує ефективну взаємодію зазначених методів у рамках єдиної аналітичної системи;

– обґрунтувати перспективні напрями розвитку систем Data Mining на основі отриманих результатів.

Викладення основного матеріалу дослідження

Сучасні системи Data Mining функціонують у середовищі, де дані мають різну структуру, динаміку й семантику, тому ефективність аналітичних процесів значною мірою залежить від здатності поєднувати методи роботи з текстовими, візуальними та інтернет-даними в єдиній аналітичній моделі. Текстовий аналіз є одним із найбільш поширених напрямів, оскільки значна частина інформації представлена у вигляді документів, повідомлень, коментарів чи веб-контенту. Основними кроками тут виступають очищення, нормалізація та побудова векторних подань, після чого застосовуються алгоритми машинного та глибинного навчання. Використання архітектур LSTM чи трансформерних моделей дає змогу виявляти складні семантичні закономірності та досягати високої точності класифікаційних моделей, значно перевищуючи традиційні статистичні підходи.

Візуальний аналіз даних відіграє ключову роль там, де необхідно підвищити інтерпретованість моделей або надати користувачу можливість взаємодіяти з великими масивами інформації. Останні підходи передбачають інтерактивні механізми, за яких користувач може впливати на процес навчання, маркувати дані, уточнювати кластери чи змінювати набір ознак у режимі реального часу. Такі методи зменшують потребу у великих вибірках і дозволяють досягати стабільних результатів з обмеженим набором навчальних даних. Разом з тим, окремим напрямом стала візуальна аналітика глибинних моделей. Потреба в поясненності та прозорості роботи нейронних мереж стимулює створення інструментів, що дозволяють виявляти помилки, оцінювати важливість ознак і розуміти логіку формування прогнозів.

Інтернет-орієнтований аналіз даних охоплює широкий спектр джерел, від веб-сторінок і поведінкових логів до постів у соціальних мережах та сенсорних потоків Інтернету речей. Класичний веб-майнінг включає аналіз вмісту, структури та поведінки користувачів, забезпечуючи виявлення прихованих шаблонів, рекомендаційних сигналів і структурної організації мережі. Аналіз даних соціальних медіа, що характеризуються високою швидкістю та великим обсягом, потребує застосування Big Data-фреймворків для обробки текстів, тональності, трендів і взаємовпливів користувачів. Водночас розвиток Інтернету речей породжує окремий клас даних – високочастотні, різнорідні сенсорні потоки, які потребують спеціальних методів обробки, включно з потоковими ML-алгоритмами та ансамблевими моделями.

Особливе місце в сучасному Data Mining посідають просторово-часові та графові методи, що дозволяють моделювати складні просторово-часові та міжоб'єктні зв'язки. Просторово-часовий аналіз широко застосовується у транспортних системах, логістиці, безпеці чи прогнозуванні подій, але залишається складним для візуалізації та автоматизації. Графові нейронні мережі дозволяють працювати з неєвклідовими структурами, такими як документи, користувачі чи гіперпосилання, що робить їх особливо ефективними у класифікації текстів та аналізі інтернет-структур. Текстовий компонент Data Mining охоплює основні методи NLP (Natural Language Processing), включаючи очищення, нормалізацію, формування лінгвістичних подань, семантичне моделювання та застосування глибинних архітектур.

На основі узагальнення розглянутих методів доцільно виділити інтегровану структуру системи Data Mining, що поєднує текстовий, візуальний та інтернет-орієнтований аналіз у єдиному аналітичному середовищі. Узагальнену модель системи наведено в табл. 1.

Таблиця 1

Компоненти інтегрованої системи Data Mining

Компонент	Опис	Типи даних	Методи
Текстовий модуль	Забезпечує семантичну обробку та узагальнення інформації з текстів	Документи, повідомлення, веб-контент	NLP, embeddings, класифікація, DL-моделі
Візуальний модуль	Відповідає за інтерпретацію, пояснення та інтерактивний аналіз результатів	Числові дані, результати моделей, графічні структури	VDM, EDA, explainable DL, інтерактивні панелі
Веб-модуль	Обробляє динамічні інтернет-дані з веб-ресурсів, соцмереж та IoT	HTML-структури, логи, соціальні дані, сенсорні потоки	Web-content/usage/structure mining, big data analytics
Інтеграційний блок	Поєднує результати різних модулів у єдину аналітичну модель	Дані всіх типів	Ensemble-моделі, fusion-підходи, мета-аналітика

Таблиця демонструє, що інтегрована система складається з чотирьох взаємопов'язаних компонентів, кожен з яких виконує специфічні функції. Текстовий модуль забезпечує глибинний семантичний аналіз, візуальний модуль – інтерпретацію та пояснення результатів, а веб-модуль – роботу з динамічними потоками інтернет-даних. Інтеграційний блок синтезує отримані результати, формуючи цілісну модель, що підвищує точність, масштабованість і надійність аналітичної системи.

Для перевірки практичної доцільності запропонованого інтегрованого підходу було проведено експериментальне дослідження на прототипі системи Data Mining, призначеної для прогнозування рівня інтересу користувачів до товарів інтернет-каталогу. Вхідний набір даних включав три основні підмножини: текстові дані, що складалися з 10 000 користувацьких відгуків і пошукових запитів; інтернет-дані у вигляді журналів переглядів сторінок, кліків по товарах та операцій додавання до кошика (загалом близько 200 000 записів сесій); а також структуровані характеристики товарів, зокрема категорії, діапазони цін та базові показники попиту.

Експеримент був спрямований на порівняння ефективності трьох конфігурацій системи. У першому варіанті («Text-only») використовувався лише текстовий аналіз, який включав передобробку даних методом NLP, побудову векторних подань та подальшу класифікацію. У другому варіанті («Text + Web») до текстових ознак додавалися показники поведінки користувачів, отримані з даних про поведінку користувачів у вебсередовищі, такі як частота переглядів, глибина сесій і показники конверсій. Третій варіант («Integrated») реалізовував комплексний підхід, доповнюючи попередні модулі засобами візуальної аналітики (інтерактивними панелями, діаграмами важливості ознак та візуалізацією сегментів користувачів), що давало можливість глибше інспектувати результати та прискорювати процес прийняття рішень.

У всіх конфігураціях вирішувалося однакове завдання багатокласової класифікації рівнів попиту на товар (низький, середній або високий). Як основний алгоритм використовувався градієнтний бустинг, навчений на сукупності ознак, сформованих кожним модулем. Оцінювання точності здійснювалося за метриками Accuracy та F1-мірою (macro), а також за операційними метриками – часом навчання моделі та середнім часом, необхідним аналітику для інтерпретації результатів під час аналізу типового набору товарів. Окрім того, було враховано суб'єктивну оцінку зручності роботи з системою, яку трійка аналітиків визначила за п'ятибальною шкалою. Узагальнені результати наведено в табл. 2.

Таблиця 2

Порівняння конфігурацій системи data mining

Варіант системи	Accuracy	F1-міра (macro)	Час навчання, с	Час інтерпретації (100 товарів), с	Суб'єктивна зручність (1–5)
A. Text-only	0,78	0,75	24	310	3,2
B. Text + Web	0,83	0,81	31	275	3,7
C. Integrated	0,85	0,84	36	190	4,6

Отримані результати показують, що додавання інтернет-ознак (варіант B) забезпечує помітне зростання як точності, так і F1-міри порівняно з суто текстовим підходом (варіант A), що очікувано, оскільки поведінкові сигнали доповнюють семантику тексту. Інтегрований варіант (C) демонструє лише помірне додаткове покращення за класичними метриками якості класифікації, але суттєво зменшує час інтерпретації результатів за рахунок візуальної аналітики та інтерактивної взаємодії з моделлю. Зокрема, середній час аналізу 100 позицій товарів скорочується більш ніж на 35 % порівняно з варіантом A, а суб'єктивна оцінка зручності зростає з 3,2 до 4,6.

Таким чином, експериментальні результати підтверджують, що інтегроване використання текстових, візуальних та інтернет-орієнтованих методів аналізу даних є доцільним не лише з погляду точності прогнозування, але й з позицій інтерпретованості, швидкості аналітичного циклу та зручності роботи користувача з системою.

Висновки

У роботі було здійснено комплексний аналіз методів текстового, візуального та інтернет-орієнтованого аналізу даних у контексті сучасних систем Data Mining, а також обґрунтовано необхідність їх інтегрованого використання. Огляд наукових джерел показав, що кожен із розглянутих підходів має свої сильні сторони: текстовий аналіз забезпечує глибинне семантичне опрацювання інформації, візуальні методи підвищують інтерпретованість і контрольованість моделей, а інтернет-аналіз дає змогу працювати з динамічними та високовимірними потоками даних, характерними для сучасних веб-систем.

Проведені експериментальні дослідження підтвердили практичну доцільність інтегрованого підходу. Зокрема, поєднання веб-поведінкових та текстових ознак дозволило підвищити точність класифікації порівняно з моделями, що використовують лише один тип даних. Додавання візуальної аналітики не призвело до суттєвого зростання стандартних метрик точності, проте забезпечило значне скорочення часу інтерпретації результатів, підвищило зручність роботи аналітиків і зробило процес прийняття рішень більш прозорим.

Отримані результати свідчать, що майбутні системи Data Mining повинні будуватися на мультикомпонентних архітектурах, здатних об'єднувати різнорідні джерела інформації та забезпечувати як високу точність, так і інтерпретованість моделей. Перспективним напрямом подальших досліджень є розроблення універсальних інтеграційних модулів, що автоматично поєднують текстові, поведінкові та графові дані, а також застосування гібридних підходів пояснюваного штучного інтелекту для підвищення прозорості складних моделей у реальних аналітичних сценаріях.

Список використаної літератури

1. Rautela R., Kumar S., Kumar A. та ін. Text mining: a comprehensive survey // *International Journal of Recent Scientific Research*. 2018. Vol. 9, Iss. 5(G). P. 26959–26962. DOI: 10.24327/IJRSR.2018.0905.2158.

2. Alqahtani A., Ghazali R., Hasan M. та ін. An efficient approach for textual data classification using deep learning // *Frontiers in Computational Neuroscience*. 2022. Vol. 16. Article 992296. DOI: 10.3389/fncom.2022.992296.

3. Li H., Ling X., Zhang B. та ін. Interactive machine learning by visualization: a small data solution // *Proc. IEEE International Conference on Big Data*. 2018. P. 3513–3521. DOI: 10.1109/BigData.2018.8621952.

4. Hohman F., Kahng M., Pienta R., Chau D.H. Visual analytics in deep learning: an interrogative survey for the next frontiers // *IEEE Transactions on Visualization and Computer Graphics*. 2019. Vol. 25, No. 8. P. 2674–2693. DOI: 10.1109/TVCG.2018.2843369.

5. Yanto I. T., Handayani O. P. Visualization of data inventory using visual data mining (VDM) and exploratory data analysis (EDA) methods // *JOIV: International Journal on Informatics Visualization*. 2025. Vol. 9, No. 5. P. 1808–1815. DOI: 10.62527/joiv.9.5.4075.

6. Begam A. M., Saravanan B. V. Web mining concepts, applications and tools: a survey // *Journal of Emerging Technologies and Innovative Research*. 2019. Vol. 6, No. 5. P. 285–288.

7. Darwiesh A., Alghamdi M. I., El-Baz A. H., Elhoseny M. Social media big data analysis: towards enhancing competitiveness of firms in a post-pandemic world // *Journal of Healthcare Engineering*. 2022. Article ID 6967158. DOI: 10.1155/2022/6967158.

8. Hamdi A., Chen Y.-P., Mohamed M. B., Hmeidi I. Spatiotemporal data mining: a survey on challenges and open problems // *Artificial Intelligence Review*. 2022. Vol. 55. P. 1441–1488. DOI: 10.1007/s10462-021-09994-y.

9. Li W., Aram J., Islam M. M. та ін. A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system // *Mobile Networks and Applications*. 2021. Vol. 26, No. 1. P. 234–252. DOI: 10.1007/s11036-020-01700-6.

10. Wang K., Ding Y., Han S. C. Graph neural networks for text classification: a survey // *Artificial Intelligence Review*. 2024. Vol. 57, Article 190. DOI: 10.1007/s10462-024-10808-0.

References

1. Rautela, R., Kumar, S., Kumar, A., et al. (2018). Text mining: A comprehensive survey. *International Journal of Recent Scientific Research, 9*(5G), 26959–26962. https://doi.org/10.24327/IJRSR.2018.0905.2158.

2. Alqahtani, A., Ghazali, R., Hasan, M., et al. (2022). An efficient approach for textual data classification using deep learning. *Frontiers in Computational Neuroscience, 16*, 992296. https://doi.org/10.3389/fncom.2022.992296.

3. Li, H., Ling, X., Zhang, B., et al. (2018). Interactive machine learning by visualization: A small data solution. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3513–3521). IEEE. https://doi.org/10.1109/BigData.2018.8621952.

4. Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics, 25*(8), 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369.

5. Yanto, I. T., & Handayani, O. P. (2025). Visualization of data inventory using visual data mining (VDM) and exploratory data analysis (EDA) methods. *JOIV: International Journal on Informatics Visualization, 9*(5), 1808–1815. https://doi.org/10.62527/joiv.9.5.4075.

6. Begam, A. M., & Saravanan, B. V. (2019). Web mining concepts, applications and tools: A survey. *Journal of Emerging Technologies and Innovative Research, 6*(5), 285–288.

7. Darwiesh, A., Alghamdi, M. I., El-Baz, A. H., & Elhoseny, M. (2022). Social media big data analysis: Towards enhancing competitiveness of firms in a post-pandemic world. *Journal of Healthcare Engineering, 2022*, 6967158. https://doi.org/10.1155/2022/6967158.

8. Hamdi, A., Chen, Y.-P., Mohamed, M. B., & Hmeidi, I. (2022). Spatiotemporal data mining: A survey on challenges and open problems. *Artificial Intelligence Review, 55*, 1441–1488. https://doi.org/10.1007/s10462-021-09994-y.

9. Li, W., Aram, J., Islam, M. M., et al. (2021). A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile Networks and Applications, 26*(1), 234–252. https://doi.org/10.1007/s11036-020-01700-6.

10. Wang, K., Ding, Y., & Han, S. C. (2024). Graph neural networks for text classification: A survey. *Artificial Intelligence Review, 57*(8), 190. https://doi.org/10.1007/s10462-024-10808-0.

Дата першого надходження рукопису до видання: 30.11.2025
Дата прийнятого до друку рукопису після рецензування: 29.12.2025
Дата публікації: 31.12.2025