

УДК 004

О.Ширяєв, магістр гр. КН-23М,
Центральноукраїнський національний технічний університет

ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ МОНІТОРИНГУ РОБОТИ СОЦІАЛЬНОГО СЕРВІСУ З ВИКОРИСТАННЯМ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ

У статті розроблено програмне забезпечення, яке призначено для системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. Метою розробки є дослідження та програмна реалізація системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. Об'єктом дослідження є процес моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. Предметом дослідження є методи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. Методи дослідження базуються на методах кластерного аналізу, методах математичної статистики, методах розробки програмного забезпечення. Результат роботи – програмна реалізація системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. В процесі роботи над програмною моделлю виконано аналіз існуючих апаратних та програмних засобів. В повній мірі описані всі компоненти розробленого програмного забезпечення.

Постановка проблеми. В сучасному світі люди вже не можуть уявити свого життя без інтернету та соціальних сервісів. Одним із видів соціальних сервісів є веб-сервіси, які надають можливість користувачам задати питання або відповісти на вже задані питання. Таким чином формується соціальна мережа, учасником якої може стати кожен. У таких системах кожен користувач стикався з необхідністю швидко знайти необхідну інформацію. Тому використання «інтелектуального» групування результатів при пошуку необхідної інформації надає можливість вирішення проблеми релевантного та швидкого пошуку в надзвичайно великих обсягах інформації.

Аналіз останніх досліджень і публікацій. При аналізі останніх досліджень і публікацій [1-10] було виявлено певні прогалини у забезпеченні системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Мета й завдання дослідження. Метою роботи є дослідження та програмна реалізація системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Для досягнення поставленої мети визначена програма дослідження, що складається з наступних завдань:

- Огляд існуючих систем моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.
- Дослідження системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.
- Програмна реалізація системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Об'єктом дослідження є процес моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Предметом дослідження є методи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Методи дослідження базуються на методах кластерного аналізу, методах математичної статистики, методах розробки програмного забезпечення.

Виклад основного матеріалу. Технічним завданням магістерського проекту передбачена розробка програмного забезпечення моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Глобальний ринок програмного забезпечення для кластерного аналізу переживає стрімке зростання, зумовлене зростаючим впровадженням аналітики даних у різних секторах. Хоча точні дані щодо розміру ринку та CAGR не надаються, обґрунтована оцінка, заснована на галузевих тенденціях та звітному періоді дослідження (2019-2033), свідчить про значний ринок. Враховуючи широке застосування кластерного аналізу в різних галузях, таких як BFSI (банківська справа, фінансові послуги та страхування), державний сектор та комерційні підприємства, а також зростання впровадження хмарних та веб-рішень, вартість ринку, ймовірно, становитиме мільярди доларів до 2033 року. Складний річний темп зростання (CAGR) за цей період оцінюється від 12% до 15%, що відображає високий попит на передові аналітичні інструменти та постійну цифровізацію бізнес-процесів. Це зростання підживлюється кількома факторами, включаючи необхідність кращої сегментації клієнтів, покращеного управління ризиками, виявлення шахрайства та оптимізованого розподілу ресурсів. Наявність потужних та зручних програмних рішень, а також зростаюча доступність великих наборів даних, ще більше стимулює розширення ринку. Однак, такі фактори, як висока вартість впровадження, потреба в спеціалізованій експертизі та занепокоєння щодо конфіденційності та безпеки даних, створюють певні труднощі для ширшого впровадження. Ринок сегментовано за типом розгортання (хмарні та веб-інструменти) та застосування (BFSI, урядові, комерційні та інші), причому очікується, що хмарний сегмент домінуватиме завдяки своїй масштабованості, економічній ефективності та доступності. Очікується, що сектор BFSI займатиме значну частку ринку через критичну потребу в прийнятті рішень на основі даних у фінансовій галузі.

Конкурентне середовище характеризується поєднанням таких визнаних гравців, як TIBCO Software, Microsoft, IBM та Oracle, а також менших спеціалізованих постачальників. Присутність численних компаній, як міжнародних, так і регіональних, підкреслює зростаючу зрілість та фрагментацію цього сегмента ринку. Майбутнє зростання, ймовірно, залежатиме від здатності постачальників впроваджувати інновації, пропонувати зручні інтерфейси та надавати комплексні можливості підтримки та інтеграції. Крім того, очікується, що стратегічні партнерства та придбання відіграватимуть значну роль у формуванні майбутньої траєкторії ринку. Географічно різноманітний ринок охоплює Північну Америку, Європу, Азіатсько-Тихоокеанський регіон та інші регіони, причому Північна Америка та Європа, як очікується, збережуть лідируючі позиції завдяки високому впровадженню передових аналітичних технологій та присутності основних гравців. Однак, за прогнозами, країни, що розвиваються, в Азіатсько-Тихоокеанському регіоні зазнають швидкого зростання, зумовленого зростанням цифровізації та розширенням галузей, що інтенсивно використовують дані.

Концентрація та характеристики програмного забезпечення для кластерного аналізу

Глобальний ринок програмного забезпечення для кластерного аналізу оцінюється в 2,5 мільярда доларів у 2024 році, з прогнозованим сукупним річним темпом зростання (CAGR) 12% до 2029 року. Концентрація ринку помірна, при цьому кілька основних гравців займають значну частку ринку, але також робить свій внесок значна кількість менших спеціалізованих постачальників. Інновації в першу чергу зосереджені на підвищенні ефективності алгоритмів, інтеграції з платформами великих даних, включенні розширених можливостей візуалізації та розширенні хмарних рішень.

Зони концентрації:

- Хмарні рішення швидко набирають обертів.
- Інтеграція з машинним навчанням та штучним інтелектом є ключовим акцентом.
- Акцент на зручних інтерфейсах для ширшої доступності.

Характеристики інновацій:

- Розробка передових алгоритмів кластеризації (наприклад, DBSCAN, HDBSCAN).
- Покращена масштабованість для обробки великих наборів даних.
- Автоматизовані методи вибору ознак та попередньої обробки.

Вплив нормативних актів: Нормативні акти щодо конфіденційності даних (GDPR, CCPA) суттєво впливають на розвиток методів кластеризації, що зберігають конфіденційність.

Замінники продуктів: Хоча прямих замінників не існує, альтернативні методи аналізу даних, такі як дерева рішень або нейронні мережі, можуть частково замінити кластерний аналіз у деяких застосуваннях.

Концентрація кінцевих користувачів: Сектори BFSI та державного сектору представляють значну концентрацію кінцевих користувачів, що зумовлено необхідністю сегментації клієнтів та виявлення шахрайства (BFSI) та розподілу ресурсів/оцінки ризиків (уряд).

Рівень злиттів та поглинань: Рівень активності злиттів та поглинань (M&A) є помірним, при цьому більші постачальники купують менші, спеціалізовані фірми для розширення свого портфеля продуктів та технологічних можливостей. Ми прогнозуємо приблизно 5-7 значних угод M&A у цьому секторі щорічно.

Аналітика продукту програмного забезпечення для кластерного аналізу

Програмні продукти для кластерного аналізу варіюються від базових автономних програм до складних інтегрованих платформ, що пропонують повний набір інструментів для аналізу та аналізу даних. Багато продуктів зараз містять розширені функції візуалізації, що дозволяє користувачам інтерактивно досліджувати структури та шаблони кластерів. Хмарні рішення стають все більш поширеними, пропонуючи масштабованість та доступність, тоді як локальні рішення залишаються актуальними для організацій із суворими вимогами безпеки. Зростаючою тенденцією є інтеграція можливостей автоматизованого машинного навчання (AutoML), що спрощує процес створення та розгортання моделей кластерного аналізу навіть для користувачів з обмеженим досвідом. Моделі ціноутворення широко варіюються, від доступу на основі підписки до безстрокових ліцензій, при цьому витрати масштабуються залежно від функціональності та можливостей користувача. Ринок спостерігає перехід до гнучких моделей оплати за використання.

Сегментація ринку:

- Тип розгортання: Хмарні рішення переживають швидке зростання завдяки підвищеній доступності, масштабованості та економічній ефективності. Веб-рішення пропонують аналогічні переваги, але можуть мати обмеження щодо налаштування та безпеки даних порівняно з хмарними розгортаннями.

- Застосування: BFSI (банківська справа, фінансові послуги та страхування) – це значний сегмент ринку, що використовує кластерний аналіз для сегментації клієнтів, виявлення шахрайства та управління ризиками. Урядові програми включають розподіл ресурсів, громадську безпеку та демографічний аналіз. Комерційні програми охоплюють різні галузі, від маркетингу та продажів до оптимізації ланцюга поставок. Сегмент «Інші» охоплює нішеві програми в різних секторах.

Північна Америка займає найбільшу частку ринку, що зумовлено раннім впровадженням передових аналітичних технологій та сильною присутністю основних постачальників програмного забезпечення. Європа є важливим ринком, де зростає впровадження в різних секторах. Азіатсько-Тихоокеанський регіон – це регіон з високим потенціалом зростання, що підживлюється розширенням цифровізації та збільшенням інвестицій в аналітику даних. Однак проникнення на ринок залишається відносно нижчим порівняно з Північною Америкою та Європою.

Тенденції програмного забезпечення для кластерного аналізу

Кілька ключових тенденцій формують ринок програмного забезпечення для кластерного аналізу. Зростаючий обсяг і швидкість передачі даних стимулюють попит на рішення, здатні обробляти великі та складні набори даних. Розвиток технологій великих

даних, таких як Hadoop і Spark, впливає на розробку масштабованих та розподілених алгоритмів кластеризації. Досягнення в галузі штучного інтелекту (ШІ) та машинного навчання (МН) призводять до інтеграції автоматизованих можливостей розробки ознак та вибору моделей. Зростаюча увага до конфіденційності та безпеки даних впливає на розробку методів кластеризації, що зберігають конфіденційність. Впровадження хмарних рішень продовжує прискорюватися, пропонуючи економічну ефективність та масштабованість. Нарешті, попит на зручні інтерфейси стимулює розробку інтуїтивно зрозумілих інструментів, що робить кластерний аналіз доступним для ширшого кола користувачів.

Рушійні сили: що рухає ринок програмного забезпечення для кластерного аналізу:

- Вибух великих даних: експоненціальне зростання даних вимагає ефективних інструментів аналізу.

- Досягнення у сфері штучного інтелекту/машинного навчання: інтеграція можливостей штучного інтелекту/машинного навчання підвищує точність моделі та автоматизацію.

- Впровадження хмарних обчислень: Хмарні рішення забезпечують масштабованість та доступність.

- Зростаючий попит на прийняття рішень на основі даних: Бізнес дедалі більше покладається на аналітичні дані.

- Посилена увага до сегментації клієнтів: розуміння поведінки клієнтів визначає маркетингові стратегії.

Проблеми та обмеження програмного забезпечення для кластерного аналізу:

- Складність даних: Обробка великих, високимірних наборів даних залишається проблемою.

- Вибір алгоритму: Вибір відповідного алгоритму кластеризації вимагає досвіду.

- Інтерпретація результатів: Розуміння наслідків результатів кластерного аналізу може бути складним.

- Вартість впровадження: Високі початкові інвестиції можуть бути перешкодою для деяких організацій.

- Брак кваліфікованих фахівців: Існує нестача фахівців з обробки даних, які добре знають методи кластерного аналізу.

Нові тенденції в програмному забезпеченні для кластерного аналізу:

- Пояснювальний штучний інтелект (ХАІ): зростаючий попит на розуміння рішень моделей.

- Федеративне навчання: Забезпечує спільне навчання моделей без обміну даними.

- Периферійні обчислення: обробка даних ближче до джерела для підвищення швидкості та ефективності.

- Інтеграція з даними Інтернету речей: аналіз даних з підключених пристроїв.

- Покращені засоби візуалізації: покращені методи візуалізації складних кластерних структур.

Каталізатори зростання в індустрії програмного забезпечення для кластерного аналізу

Ринок програмного забезпечення для кластерного аналізу переживає значне зростання завдяки збігу факторів. Експоненціальне зростання обсягу та швидкості даних вимагає ефективних інструментів для аналізу та інтерпретації. Інтеграція передових можливостей штучного інтелекту та машинного навчання підвищує точність та автоматизацію моделей кластерного аналізу, а хмарні рішення забезпечують масштабованість та доступність. Зростаюча увага до прийняття рішень на основі даних у різних галузях стимулює попит. Крім того, зростаюча потреба в ефективній сегментації клієнтів та цілеспрямованих маркетингових стратегіях стимулює розширення ринку. Урядові ініціативи, що сприяють аналітиці даних, ще більше сприяють загальному зростанню.

Значні розробки в секторі програмного забезпечення для кластерного аналізу:

– 2021: Кілька великих постачальників випускають нові версії свого програмного забезпечення для кластерного аналізу, включаючи розширені можливості штучного інтелекту/машинного навчання.

– 2022: Посилена увага до хмарних рішень та моделей ціноутворення «оплата за використання».

– 2023: Значна активність у сфері злиттів та поглинань, коли більші постачальники придбали менші спеціалізовані фірми.

– 1 квартал 2024 року: Запуск кількох нових продуктів, що включають функції штучного інтелекту (XAI).

– 3 квартал 2024 року: Впровадження інноваційних методів кластеризації, що зберігають конфіденційність.

У сучасному світі, керованому даними, вибір правильних інструментів аналізу даних є критично важливим для компаній, які прагнуть ефективно використовувати їхню силу. Цей посібник пропонує глибокий огляд популярних інструментів аналізу даних, пропонуючи уявлення про їхні характеристики, можливості, використання та ціни, допомагаючи вам зробити обґрунтований вибір для вашої організації.

Microsoft Power BI

Microsoft Power BI – це провідна платформа бізнес-аналітики, яка пропонує інтерактивні візуалізації та можливості звітності. Вона підтримує широкий спектр джерел даних і дозволяє користувачам безперешкодно створювати та ділитися інформаційними панелями.

Основні характеристики:

– Інтеграція даних: підключається до численних джерел даних, включаючи користувацькі API.

– Автоматизоване машинне навчання: Користувачі можуть створювати моделі машинного навчання без глибоких технічних знань.

– Інструменти для співпраці: легко діліться інформаційними панелями та звітами в організації.

Ціноутворення:

– Power BI Pro: \$9.99 за користувача/місяць.

– Power BI Premium: ціна починається від 20 доларів США за користувача/місяць або 4995 доларів США за ресурс/місяць.

Підходить для: Ідеально підходить для підприємств, яким потрібні надійні можливості візуалізації даних та обміну звітами. Він ідеально підходить для таких галузей, як фінанси, охорона здоров'я та роздрібна торгівля, яким потрібна детальна та інтерактивна звітність.

Не підходить для: Організації, яким потрібні розширені функції маніпулювання даними та глибокі функції аналізу даних, окрім візуалізації та звітності, можуть вважати Power BI обмеженим.

Приклад використання: Роздрібна компанія використовує Power BI для інтеграції даних про продажі з різних регіонів. За допомогою інтерактивних інформаційних панелей менеджери можуть відстежувати ключові показники ефективності (KPI) у режимі реального часу, покращуючи процеси прийняття рішень.

Tableau

Tableau – це потужний інструмент візуалізації даних, який робить складні дані доступними та зрозумілими.

Основні характеристики:

– Зручний інтерфейс: функція перетягування допомагає швидко створювати візуалізації.

– Дані в режимі реального часу: підключення до джерел даних у реальному часі для негайного отримання аналітичної інформації.

– Міжплатформне підключення: працює з хмарними сервісами, такими як Google BigQuery.

Ціноутворення:

- Tableau Creator: \$70 за користувача/місяць.
- Tableau Explorer: \$42 за користувача/місяць.
- Переглядач Tableau: 15 доларів США за користувача/місяць.

Підходить для: компаній, що зосереджені на інтуїтивному розповіді історій на основі даних, ідеально підходить для таких секторів, як маркетинг і продажі, де візуальне розуміння тенденцій є критично важливим.

Не підходить для: Фірми, яким потрібні розширені можливості маніпулювання даними та моделювання, можуть вважати функції Tableau обмеженими.

Приклад використання: Маркетингова команда використовує Tableau для візуалізації тенденцій даних клієнтів, покращуючи аналіз аудиторії та таргетування кампанії.

R

R – це мова статистичного програмування, відома своїми всебічними аналітичними можливостями.

Основні характеристики:

- Широкий вибір пакетів: тисячі пакетів для різноманітних завдань аналізу даних.
- Розширені статистичні функції: адаптовані для поглибленого статистичного моделювання.

– Багата візуалізація: пропонує потужні графічні інструменти для візуалізації даних.

Ціноутворення: Безкоштовно та з відкритим вихідним кодом.

Підходить для: фахівців з обробки даних та статистиків, що займаються академічними дослідженнями та моделюванням складних даних.

Не підходить для: користувачів без технічних знань, яким потрібен інтуїтивно зрозумілий інтерфейс без глибоких знань статистики.

Приклад використання: Науково-дослідний інститут впроваджує R для епідеміологічних досліджень, створюючи передові статистичні моделі для прогнозування поширення хвороби.

Python

Python – це універсальна мова програмування, яка широко використовується для аналізу даних, пропонуючи потужні бібліотеки та підтримку спільноти.

Основні характеристики:

– Комплексні бібліотеки: Бібліотеки, такі як NumPy, Pandas та Matplotlib, сприяють різноманітним завданням роботи з даними.

– Гнучкість: Підтримує написання сценаріїв для швидких завдань та розробки складних програм.

– Фреймворки машинного навчання: такі інструменти, як TensorFlow та scikit-learn, забезпечують розширені можливості машинного навчання.

Ціноутворення: Безкоштовно та з відкритим вихідним кодом.

Підходить для: фахівців з обробки даних та розробників, які шукають комплексний та гнучкий інструмент для різних завдань з даними, включаючи машинне навчання та візуалізацію.

Не підходить для: користувачів без знань програмування, які шукають просте, готове до використання рішення для аналізу даних.

Приклад використання: Технологічний стартап використовує Python для розробки моделей машинного навчання з метою покращення рекомендацій щодо своїх продуктів, використовуючи алгоритми для аналізу моделей поведінки користувачів.

Looker

Looker – це хмарний аналітичний інструмент, який пропонує комплексне дослідження та візуалізацію даних.

Основні характеристики:

– LookML: Дозволяє користувачам визначати моделі даних та ефективно виконувати аналіз.

– Перегляд даних у режимі реального часу: забезпечує негайне отримання аналітичних даних з потоків даних у реальному часі.

– Налаштовувані інформаційні панелі: спрощують обмін аналітичними даними.

Ціноутворення: Індивідуальне ціноутворення залежно від потреб організації. Зверніться до Looker для отримання цінової пропозиції.

Підходить для: компаній, які надають пріоритет середовищам спільної роботи з даними та потребують моделювання структурованих даних.

Не підходить для: компаній, яким потрібні інструменти офлайн-аналізу, або тих, хто не має значних інвестицій у хмарну інфраструктуру.

Приклад використання: Технологічний стартап використовує Looker для моніторингу показників залученості користувачів, покращуючи співпрацю команди над стратегіями розробки продукту.

SAP BusinessObjects

SAP BusinessObjects – це набір інструментів бізнес-аналітики, призначених для виявлення даних та створення звітів.

Основні характеристики:

– Інтеграція з Microsoft Office: безперешкодна інтеграція робочих процесів з такими інструментами, як Excel, для розширених можливостей.

– Самообслуговування прогнозу аналітики: дозволяє користувачам створювати прогнозні моделі без серйозного технічного втручання.

– Панелі інструментів на основі ролей: пропонують персоналізовані звіти на основі ролей користувачів в організації.

Ціноутворення: Ціноутворення налаштовується відповідно до потреб підприємства та вимагає безпосереднього контакту з SAP для отримання цінкових пропозицій.

Підходить для: Найкраще підходить для підприємств, які значною мірою покладаються на інструменти Microsoft Office та прагнуть інтегрованої автономії в дослідженні даних.

Не підходить для: Невеликі підприємства з обмеженою IT-підтримкою можуть вважати складність платформи надто складною без виділених ресурсів.

Приклад використання: Фінансова компанія використовує SAP BusinessObjects для створення та розповсюдження складних фінансових звітів. Аналітики можуть порівнювати дані Excel з аналітикою BusinessObjects для оптимізації звітності.

Sisense

Sisense – це аналітична платформа, спрямована на спрощення аналізу складних даних за допомогою інтуїтивно зрозумілих панелей інструментів.

Основні характеристики:

– Технологія In-Chip: використовує кешування процесора для швидшої обробки даних.

– Інтерфейс перетягування: дозволяє користувачам без технічних знань легко створювати візуалізації.

– Вбудована аналітика: інтегруйте аналітику в клієнтські додатки.

Ціноутворення: індивідуальне ціноутворення залежно від бізнес-вимог. Зверніться до Sisense для отримання детальної цінової пропозиції.

Підходить для: підприємств, яким потрібна швидка інтеграція та налаштування аналітики даних у додатки, особливо в таких секторах, як SaaS та фінтех.

Не підходить для: фірм, яким потрібні розширені можливості ETL або глибші ручні маніпуляції даними.

Приклад використання: Технологічна компанія використовує Sisense для покращення своєї SaaS-платформи шляхом інтеграції аналітики в режимі реального часу, яка покращує взаємодію з користувачем та аналітику.

Apache Spark

Apache Spark – це уніфікований аналітичний механізм, відомий для обробки великомасштабних даних.

Основні характеристики:

- Багатомовна підтримка: Працює з Java, Scala, Python та R.
- Аналітика в режимі реального часу: здатна обробляти великі потоки даних за допомогою Spark Streaming.

- Розподілені обчислення: Ефективно обробляє величезні набори даних у кластерах.

Ціноутворення: Як інструмент з відкритим кодом, Apache Spark сам по собі є безкоштовним. Однак можуть виникнути витрати, пов'язані з інфраструктурою та хмарними сервісами.

Підходить для: організацій, що займаються проектами з великими даними, що потребують високошвидкісних обчислень та широкої масштабованості, таких як технологічні гіганти у сфері штучного інтелекту та науки про дані.

Не підходить для: малого та середнього бізнесу з простими потребами в обробці даних через складність та вимоги до ресурсів.

Приклад використання: Телекомунікаційна компанія використовує Spark для аналізу потоків мережових даних у режимі реального часу, виявлення перебоїв та оптимізації надання послуг.

Аналітична платформа KNIME

KNIME – це платформа з відкритим кодом, відома створенням програм для обробки даних.

Основні характеристики:

- Робочі процеси на основі вузлів: полегшують візуальне створення робочих процесів без кодування.

- Різноманітні інтеграції: підключається до багатьох форматів та джерел даних.

- Розширені вузли спільноти: Отримайте доступ до вузлів, створених спільнотою, для гнучких робочих процесів.

Ціноутворення: Безкоштовно для версії з відкритим кодом; ціни на серверні рішення та інші професійні послуги визначаються залежно від потреб підприємства.

Підходить для: Дослідників та спеціалістів з обробки даних, які шукають настроювану аналітичну платформу з мінімальними вимогами до кодування.

Не підходить для: користувачів, яким потрібні прості аналітичні рішення без необхідності масштабного налаштування або співпраці зі спільнотою.

Приклад використання: Екологічне агентство використовує KNIME для синтезу даних з датчиків та проведення прогнозової аналітики впливу зміни клімату.

Вибір інструменту аналізу даних значною мірою залежить від конкретних потреб та технічних можливостей вашої організації. Від потужних можливостей візуалізації даних Tableau та Power BI до розширених аналітичних можливостей R та Python, кожен інструмент має унікальні сильні сторони та варіанти використання. Для підприємств, яким потрібні масштабовані хмарні рішення, Looker пропонує розуміння середовищ для спільної роботи, тоді як платформа з відкритим кодом KNIME обслуговує тих, хто шукає гнучкості та вдосконалень, орієнтованих на спільноту. Apache Spark відмінно підходить для обробки великих даних для підприємств з надійною технічною інфраструктурою, тоді як Sisense забезпечує швидку аналітику для розгортання SaaS.

Ретельно оцінюючи перелічені інструменти, компанії можуть ефективно узгодити свої стратегії роботи з даними, забезпечуючи використання своїх даних для прийняття обґрунтованих рішень у 2025 році та надалі.

Кластерним підходом до виявлення структури в базах даних без міток є кластерний аналіз. У кластерному аналізі передбачається, що випадки або одиниці можуть бути згруповані у значущі групи, які називаються «кластерами», на основі даних спостереження.

Кластерний аналіз використовувався в природничих науках у ранніх спробах класифікації видів і отримав назву чисельної таксономії.

Неієрархічний

Основною формою кластерного аналізу є неієрархічний. У неієрархічному кластерному аналізі ми просто намагаємося сформувати групи, але немає порядку між групами. Наприклад, при порівнянні загальних показників смертності та народжуваності для країн світу часто виділяються три групи: країни з високим рівнем народжуваності та високим рівнем смертності (країни, що розвиваються, позначені «1»), країни з високим рівнем народжуваності і низькі загальні показники смертності (країни, що проходять через «демографічний перехід», позначені «3»), і країни з низьким загальним рівнем народжуваності та зростаючими загальними показниками смертності (розвинені країни, позначені «2»; зростання загального рівня смертності спричинено старіння населення).

Ієрархічний

На відміну від неієрархічного кластерного аналізу, ієрархічний кластерний аналіз формує кластери ітеративно, шляхом послідовного приєднання або розбиття груп. Є два типи: розбіжний, який починається з усього набору даних в одній великій групі, а потім послідовно розбиває його на менші групи, поки кожне спостереження не стане окремою групою; і агломераційний, у якому кожне спостереження починається з власної групи, а групи послідовно об'єднуються в пари, поки в кінці кожне спостереження не буде в одній великій групі. Розділові методи потребують обчислень і мають обмежене застосування в соціальних науках. Агломераційні методи реалізовані в багатьох стандартних програмних пакетах.

В ієрархічному кластерному аналізі виникає проблема, коли два (або більше) спостереження поміщено в групу: якщо я порівнюю нове спостереження з групою, чи вибираю спостереження (у групі), яке найближче до мого нового спостереження, чи я вибираю спостереження (у групі), яке є найдалшим від мого спостереження, чи я вибираю якусь середню точку (скажімо, середнє значення) для порівняння з моїм новим спостереженням? Перший варіант відомий як одиночне з'єднання, другий – як повне з'єднання, а третій – як середнє з'єднання.

Під час проходження передмагістерської практики був проведений аналіз повсякденного використання користувачами подібних веб-сервісів, який показав, що такі програми найчастіше використовуються в тих випадках, коли потрібно вирішити якусь проблему і отримати компетентні відгуки та інформацію від експертів в даній сфері.

Веб-сервіс, що підлягає розробці, повинен виконувати всі вищеназвані функції. Розроблене ПЗ повинне бути компактним і зручним у користуванні. Це планується досягти за рахунок використання Yii 2 і всіх його прикладних засобів.

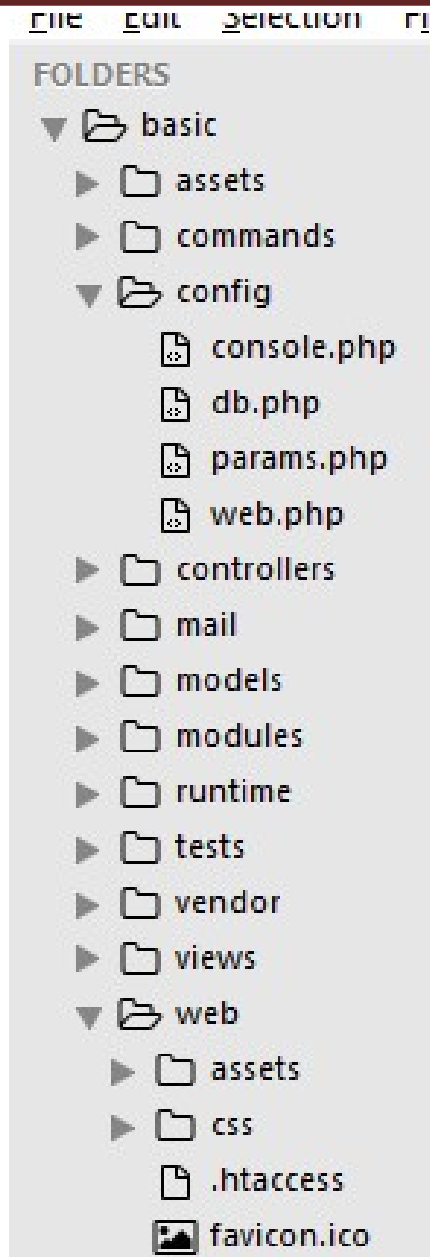


Рисунок 1 – Структура проекту в Yii2

Ієрархічна багаторівнева побудова системи дозволить подолати складність розв’язання проблеми масштабування системи при подальшій розробці. Структура проекту в Yii 2 зображена на рисунку 1

Система, що розробляється, буде побудована використовуючи шаблон проектування Модель-Представлення-Контролер (MVC, Model-View-Controller), який широко використовується в веб-програмуванні.

Розробка структурної схеми

Кластерний (або класифікаційний) аналіз упорядковує місця, види або змінні в групи (кластери), у яких подібність усередині групи є максимальною, а між групами – мінімізованою. Кластерний аналіз є явним способом пошуку структури в даних (наприклад, визначення регіональної структури). Кластерний аналіз накладає групову структуру, навіть якщо існує безперервна структура даних (наприклад, відповідь уздовж градієнта). Кластерний аналіз може бути особливо корисним при розробці схем регіональної класифікації (блокування) об’єктів і встановленні еталонних умов (Gerritsen et al., 2000).

Кластерний аналіз (класифікація) зазвичай використовується для:

1. Надати інформацію про збіг видів (внутрішня структура даних).

2. Встановити типи спільнот для описових досліджень (синтаксономія та відображення).

3. Виявлення зв'язків між спільнотами та середовищем шляхом аналізу груп, сформованих кластерним аналізом щодо змінних середовища (зовнішній аналіз).

Кластерний аналіз може бути розділовим або агломеративним методом. Методи розділення починаються з усіх об'єктів в одній групі, а потім послідовно ділять кожну групу на дві менші групи, доки не буде виконано якесь «правило зупинки». Агломераційні методи починаються з окремих об'єктів, які об'єднуються в групи шляхом колекції об'єктів або групуються в більші групи, поки всі об'єкти не потраплять в одну групу. Кластеризація також може бути ієрархічною або неієрархічною. Методи ієрархічної кластеризації впорядковують групи в ієрархічну систему, яка визначає зв'язки (подібності) між групами (рис. 13). Неієрархічні методи поділяють дані на групи без визначення зв'язків між групами.

Статистичну значущість кластерів можна оцінити за допомогою таких методів, як дискримінантний аналіз, SIMPROF, ANOSIM (Clarke and Gorley, 2006) і MRPP (McCune and Mefford, 1999). Ці методи перевіряють гіпотезу про відсутність різниці між двома або більше групами об'єктів. SIMPROF, ANOSIM і MRPP особливо підходять для аналізу даних спільноти, оскільки вони є непараметричними методами, які не припускають, що дані є багатовимірними нормальними або дисперсіями однорідними. Усі ці методи вимагають, щоб групи (кластери) були визначені апріорі. Отже, ці методи не слід використовувати для перевірки значущості класифікації, отриманої з кластерного аналізу. Статистичний аналіз апріорних кластерів та аналіз видів-індикаторів може бути корисним у розробці регіональних еталонних ділянок та/або для блокування дизайну (тобто визначення груп ділянок, які мають подібні спільноти) програми моніторингу. Аналіз видів-індикаторів (Dufrene and Legendre, 1997; McCune and Mefford, 1999) можна використовувати для ідентифікації видів, характерних для конкретних кластерів. Для перевірки стабільності кластерного аналізу були розроблені такі методології, як «відбійний ніж», тобто ці методології оцінюють, чи призводить підхід до кластеризації до статистично значущих кластерів, які є економними і на які відносно не впливає видалення сайту з аналізу (Omernik, 2004).

Прикладами алгоритмів кластерного аналізу є:

- Ієрархічні алгоритми.
- Алгоритм К-середніх.
- Мінімальне покриваюче дерево.
- Метод найближчого сусіда.
- Алгоритми нечіткої кластеризації.
- Застосування нейронних мереж.
- Генетичні алгоритми.

В ході розробки магістерського проекту було прийняте рішення застосувати для кластеризації результатів пошуку алгоритм k-Means. Даний алгоритм простий у розробці та є доволі швидким.

Структурна схема зображена на рисунку 2 відображає основні вузли програмного забезпечення та їх взаємодію між собою. Веб-сервіс, що розробляється, буде побудований використовуючи шаблон проектування MVC.

Структурна схема відображає основні блоки системи, що розробляється та їх взаємодію між собою. Головний модуль являє собою головну сторінку веб-сервісу, з якого при натисненні відповідних кнопок можна виконувати різні операції. Основними структурними складовими системи, що розробляється є:

- Головна сторінка
- Реєстрація і авторизація;
- Профіль користувача (історія активності користувача та налаштування)
- Запитання (створення, редагування, коментарі, помічення тегами)
- Інформація про автора

Сформуємо та розглянемо структурну схему побудови системи, яка показана на рисунку 2.

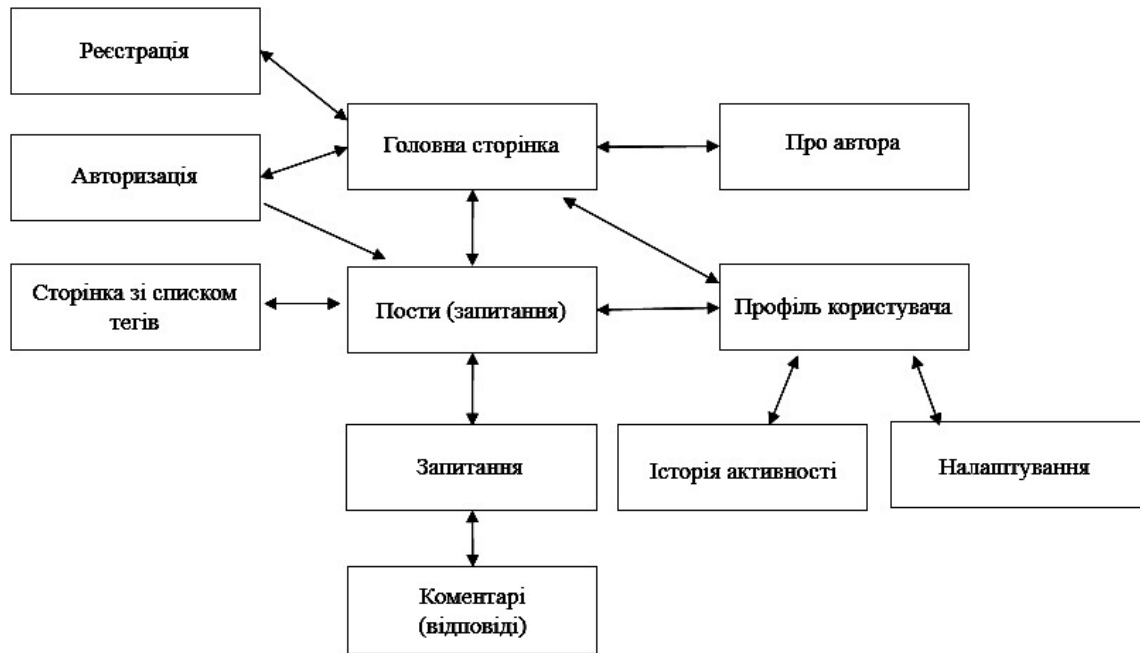


Рисунок 2 – Структурна схема веб-сервісу

Висновки. У статті наведені теоретичне узагальнення й рішення наукового завдання дослідження методів моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Рішення даного завдання полягало у вирішенні наступних задач:

- Був проведений огляд існуючих систем моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.
- Досліджена система моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.
- На основі отриманих результатів досліджень створена програмна реалізація системи моніторингу роботи соціального сервісу з використанням методів кластерного аналізу.

Розроблені під час виконання випускної кваліфікаційної роботи за другим (магістерським) рівнем вищої освіти алгоритми дозволяють успішно вирішувати завдання моніторингу роботи соціального сервісу з використанням методів кластерного аналізу. Проведено аналіз предметної галузі в ході якого були виявлені об'єкти, взаємодія яких носить істотний характер для функціональної діяльності предметної галузі, і їхні основні характеристики; побудована алгоритм і вибраний середовище розробки.

Список літератури

1. Smirnov, O., Lakhno, V., Akhmetov, B., Chubaievskyi, V., Khorolska, K., Bebesko, B. «Selection of a Rational Composition of Information Protection Means Using a Genetic Algorithm». In: Rajakumar, G., Du, KL., Vuppapapati, C., Beligiannis, G.N. (eds) Intelligent Communication Technologies and Virtual Mobile Networks. Lecture Notes on Data Engineering and Communications Technologies, vol 131. 2023. Springer, Singapore. pp. 21-34.
2. Kuznetsov, A., Oleshko, I., Chernov, K., Bagmut, M., Smirnova, T. «Biometric authentication using convolutional neural networks». Lecture Notes in Networks and Systems. Volume 152, 2021, Pages 85-98.
3. Smirnov O., Kuznetsov A., Kryvinska N., Kiian A., Kuznetsova K. «Full Non-Binary Constant-Weight Codes». SN Computer Science, Vol 2, 337, 2021. <https://doi.org/10.1007/s42979-021-00739-w>.
4. Smirnov O., Neskorodieva T., Fedorov E., Rymar P. «Neural Network Modeling Method of Transformations Data of Audit Production with Returnable Waste». CEUR Workshop Proceedings Volume 3101, 2021, Pages 192-207.

5. Smirnov, O., Kuznetsov, A., Potii, O., Poluyanenko, N., Stelnyk, I., Mialkovsky, D. «Combining and filtering functions in the framework of nonlinear-feedback shift register». International Journal of Computing; 2020, Volume 19, Issue 2 – Research Institute for Intelligent Computer Systems – 2020. – P. 247-256.
6. Smirnov O., Kuznetsov A., Kiian A., Kuznetsova T. «Non-binary constant weight coding technique». CEUR Workshop Proceedings. Volume 2740, 2020, Pages 102-114.
7. Smirnov O., Kuznetsov A., Kiian A., Cherep A., Kanabekova M., Chepurko I. «Testing of code-based pseudorandom number generators for post-quantum application». 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT), Ukraine, Kyiv, May 14-18. 2020. P. 172-177.
8. Smirnov O., Kuznetsov A., Pushkar'ov A., Serhiienko R., Babenko V., Kuznetsova T., «Representation of Cascade Codes in the Frequency Domain». In: Radivilova T., Ageyev D., Kryvinska N. (eds) Data-Centric Business and Applications. Lecture Notes on Data Engineering and Communications Technologies, vol 48. Springer, Cham. 2021. pp 557-587.
9. Smirnov, O., Drieieva, H., Drieiev, O., Polishchuk, Y., Brzhanov, R., Aleksander, M. «Method of fractal traffic generation by a model of generator on the graph». CEUR Workshop Proceedings Volume 2616, 2020, Pages 366-379.
10. Smirnov, O., Drieieva, H., Drieiev, O., Simakhin, V., Bondar, S., Odarchenko, R. «Managing multifractal properties of the binary sequence generated with the Markov chains», CEUR Workshop Proceedings Volume 2608, 2020, Pages 633-645.
11. Smirnov O. Kuznetsov A., Zaichenko Yu., Pastukhov M., Oleshko O., Kuznetsova K., «Formation of Discrete Signals with Special Correlation Properties». International Conference on Information and Telecommunication Technologies and Radio Electronics, UkrMiCo 2019; Odessa; Ukraine; 9-13 September 2019. P.22-28.
12. Smirnov, O., Kuznetsov, A., Kolovanova, I., Kuznetsova, T., «Noise immunity of the algebraic geometric codes». International Journal of Computing; 2019, Volume 18, Issue 4 – Research Institute for Intelligent Computer Systems – 2019. – P. 393-407.
13. Smirnov, O., Kuznetsov, A., Reshetniak, O., Ivko, N., Katkova, T., Kuznetsova, T., «Generators of Pseudorandom Sequence with Multilevel Function of Correlation». 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine, 8 – 11 October 2019 . P.517-522.
14. Smirnov, O., Ulichev, O., Meleshko, Y., Khokh, V., Goncharenko, I. «Method of Choosing Objects for Informational Influence in Social Networks during Information Campaign Based on the Analytic Hierarchy Process». CEUR Workshop Proceedings, Vol 2588, P. 215-227, 2019.
15. Smirnov, O., Krasnobayev, V., Yanko, A., Kuznetsova, T. «Methods of nulling numbers in the system of residual classes». CEUR Workshop Proceedings, Vol 2588, P. 90-106, 2019.
16. Smirnov, O., Kuznetsov, A., Kovalchuk, D., Averchev, A., Pastukhov, M., Kuznetsova, K., «Formation of Pseudorandom Sequences with Special Correlation Properties», 2019 3rd International Conference on Advanced Information and Communications Technologies, AICT -2019/ Lviv, Ukraine, 2-6 July, 2019, P. 395-399.
17. Smirnov, O., Kuznetsov, A., Kiian, A., Zamula, A., Rudenko, S., Hryhorenko, V., «Variance Analysis of Networks Traffic for Intrusion Detection in Smart Grids», 2019 IEEE 6th International Conference On Energy Smart Systems (2019 IEEE ESS), Kyiv, Ukraine April 17-19, 2019 P. 353-358.
18. Smirnov, O., Kuznetsov, A., Kavun, S., Babenko, B., Nakisko, O., Kuznetsova, K., «Malware Correlation Monitoring in Computer Networks of Promising Smart Grids», 2019 IEEE 6th International Conference On Energy Smart Systems (2019 IEEE ESS), Kyiv, Ukraine April 17-19, 2019 P. 347-352.
19. Smirnov, O., Kuznetsov, A., Kovalchuk, D., Pastukhov, M., Kuznetsova, K., Prokopovych-Tkachenko, D., «Discrete Signals with Special Correlation Properties», CEUR Workshop Proceedings Volume 2353, CEUR Workshop Proceedings 2019, Pages 618-629.
20. Smirnov, S., Bulekbaeva, G., Kikvidze, O.G., Lakhno, V., Brzhanov, R., Tabylov, A. «Computer simulation in the MathCAD package of plastic deformation of the deposited layer on the flat surface of the part». Journal of Theoretical and Applied Information Technology Volume 97, Issue 20, 2019, Pages 2467-2484. (Scopus).
21. Smirnov A.A., Kuznetsov A.A., Danilenko D.A., Berezovsky A., «The statistical analysis of a network traffic for the intrusion detection and prevention systems», Telecommunications and Radio Engineering. – Volume 74, Issue 1. – Begel House Inc. – 2015. – P. 61-78.
22. Аль-Мудхафар Акіл Абдулхуссейн М., Смірнова Т.В., Буравченко К.О., Смірнов О.А. «Метод оцінки та підвищення користувальницького досвіду абонентів в програмно-конфігурованих мережах на основі використання машинного навчання». Сучасні інформаційні системи, 2023, том 7, № 2, С. 49-56.
23. Смірнов О.А., Смірнова Т.В., Буравченко К.О., Кравченко С.С., Горбов В.О., «Хмарна система підтримки прийняття рішень технологічного процесу відновлення поверхонь конструкцій і деталей машин». Сучасні інформаційні системи. 2021. Т. 5, № 4. С. 79-95