МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет «Одеська юридична академія»

Кафедра кібербезпеки

МЕТОДИЧНІ ВКАЗІВКИ
виконання лабораторних та самостійних робіт
з дисципліни

«Аналіз великих даних»

для підготовки здобувачів вищої освіти
галузі знань 12 «Інформаційні технології»

Одеса 2020

УДК 004.6

Укладач:
Кухаренко С.В. – кандидат технічних наук, доцент кафедри кібербезпеки Національного університету «Одеська юридична академія»

**Рекомендовано навчально-методичною радою
Національного університету «Одеська юридична академія»,
протокол № 2 від 4 грудня 2020 р.**

Рецензенти:
Яценко В.О. – керівник проекту Освітній Фонд KeepSolid;

Логінова Н.І. – доцент, кандидат педагогічних наук, завідувач кафедри інформаційних технологій Національного університету «Одеська юридична академія»

Розглянуто поняття великі данні (Bing Data) на основі аналізу матеріалів з різних джерел. Даються визначальні характеристики даних, що збираються, зберігаються, обробляються та аналізуються за допомогою методу Big Data, наводиться їх класифікація, коротко описується історія виникнення та розвитку, представлені основоположні принципи роботи, коротко викладаються методи і технології аналізу та візуалізації, описується життєвий цикл управління даними з використанням технології Big Data. Описані основні елементи статистичного та математичного інструментарію, які використовуються в рамках технології Big Data.

Призначено для студентів з метою закріплення лекційного матеріалу і підготовки до лабораторних занять та самостійних робіт з дисципліни «Аналіз великих даних».

Методичні вказівки до виконання лабораторних та самостійних робіт з дисципліни «Аналіз великих даних» для здобувачів вищої освіти галузі знань 12 «Інформаційні технології» / Уклад.: С.В.Кухаренко. – О.: НУ «ОЮА», 2020. – 30 с.

ПЕРЕДМОВА

Дисципліна «Аналіз великих даних» вивчається студентами академії за напрямом бакалаврської підготовки з галузі знань 12 «Інформаційні технології», спеціальності 125 «Кібербезпека» на четвертому курсі навчання.

Метою викладання навчальної дисципліни «Аналіз великих даних» є оволодіння студентами базовими поняттями щодо великих даних, питаннями аналізу великих даних та пов’язаних з ними технічних, концептуальних та етичних проблем, а також отримання теоретичних знань та практичних умінь і навиків, достатніх для успішного виконання професійних обов’язків в області комп’ютерних наук, видобутку й інтелектуального аналізу великих даних різної природи в розподілених інформаційних системах.

Дисципліна знайомить з предметною областю великих даних (big data), а також показує взаємозв’язки з наукою про дані (data science) та аналізом даних (data analytics). Розвиває уміння працювати з великими даними з урахуванням їх ключових характеристик: обсягу, різноманітності та мінливості. Надає практичні навички щодо використання математичного, алгоритмічного та програмного забезпечення для розв’язання основних задач предметної області великих даних та вміння аналізувати й ефективно застосовувати хмарні системи опрацювання великих даних.

Оволодіння курсом «Аналіз великих даних» передбачає читання лекцій, проведення лабораторних занять, виконання самостійної роботи та складання екзамену. Особлива увага приділяється прищепленню студентам практичних навичок із застосування сучасних інформаційно-комунікаційних технологій у майбутній професійній діяльності.

Згідно з вимогами освітньо-професійної програми після вчення дисципліни «Аналіз великих даних » здобувачі вищох освіти повинні:

– **володіти** методологією визначення і методами отримання соціально-економічних даних;

– **володіти** методами системного аналізу, методами математичного та інформаційного моделювання для побудови та дослідження моделей об’єктів і процесів;

– **вміти** застосовувати аналітичний та методичний інструментарій для обґрунтування пропозицій та прийняття управлінських рішень;

– **вміти** видобувати знання шляхом інтеграції та аналізу великих даних, отриманих з різноманітних та різнорідних джерела інформації;

– **вміти** застосовувати на практиці мережеві технології для експериментальної та аналітичної роботи.

Лабораторна робота № 1.
ВІКОРИСТАННЯ ПРОГРАМНОГО ПАКЕТА WEKA ДЛЯ АНАЛІТИКИ ВЕЛИКИХ ДАНИХ

Навчальні питання:

1. Основи роботи з програмним пакетом WEKA.
2. Створення набору даних для завантаження в WEKA.
3. Створення регресійної моделі в WEKA.

Завдання

1. Ознайомитись з програмним пакетом з відкритим вихідним кодом для аналізу даних Waikato Environment for Knowledge Analysis (WEKA).
2. Провести регресивний аналіз для інтелектуального аналізу даних.
3. Розглянути основні поняття: метод найменших квадратів, нормальний розподіл, коефіцієнт детермінації R-квадрат.
4. Перетворити набір даних у формат, зрозумілий для програмного пакету WEKA.
5. Завантажити дані до програмного пакету WEKA.
6. Створити регресівну модель в програмному пакеті WEKA.
7. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Waikato Environment for Knowledge Analysis (WEKA), є вільно поширюваним програмним пакетом з відкритим вихідним кодом для аналізу даних. WEKA забезпечує графічний користувальницький інтерфейс для роботи з файлами даних і генерації візуальних результатів (у вигляді таблиць і графіків). Крім того, можливо інтегрувати WEKA, як і будь-яку іншу бібліотеку, у свої

власні додатки, наприклад, для автоматизації аналізу даних на стороні сервера, використовуючи стандартний API.

WEKA поширюється по ліцензії GNU General Public License (GPL).

WEKA має чотирі графічних інтерфейсу для роботи з даними.

Рис. 1. Вікно вибору інтерфейсу в WEKA

Метод регресійного аналізу є найпростішим, але одним з найменш ефективнх методів інтелектуального аналізу даних.

Метод найменших квадратів (МНК, OLS, Ordinary Least Squares) – математичний метод, який застосовується для вирішення різних задач, заснований на мінімізації суми квадратів деяких функцій від шуканих змінних. Він може використовуватися зокрема для апроксимації точкових значень деякою функцією. МНК є одним з базових методів регресійного аналізу для оцінки невідомих параметрів регресійних моделей за вибірковими даними.

Нормальний закон розподілу (normal law of distribution) (закон Гаусса) відіграє виключно важливу роль в теорії ймовірностей і займає серед інших законів розподілу особливий стан. Це закон, який найчастіше зустрічається на практиці.

Коефіцієнт детермінації (позначається як R2 – R-квадрат) – статистичний показник, що використовується в статистичних моделях як міра залежності варіації залежної змінної від варіації незалежних змінних. Вказує наскільки отримані спостереження підтверджують модель.

Для завантаження даних в WEKA, їх слід перетворити у формат, зрозумілий для цього програмного пакету. Найбільш підходящим форматом для завантаження даних в WEKA є формат Attribute-Relation File Format (ARFF), який спочатку визначає тип завантажуваних даних, а потім вказує власне дані.

Контрольні запитання

1. Що таке WEKA?
2. Поясніть метод регресійного аналізу.
3. Поясніть метод найменших квадратів.
4. Що таке закон нормального розподілу?

Лабораторна робота № 2.
ВИКОРИСТАННЯ ПРОГРАМНОГО ПАКЕТА WEKA: КЛАСИФІКАЦІЯ І КЛАСТЕРИЗАЦІЯ

Навчальні питання:

1. Класифікація в WEKA.
2. Кластеризація в WEKA.
3. Метод найближчих сусідів.

Завдання

1. Ознайомитись з методом класифікації – алгоритмом аналізу даних для визначення покрокового способу прийняття рішень.
2. Створити набір даних для аналізу за різними атрибутами.
3. Створити конкретну моділь аналізу засобами пакету WEKA.
4. Ознайомитися з методом кластерного аналізу та розділити дані на окреми групи за певними ознаками.
5. Підготувати дані для кластерного аналізу засобами пакету WEKA.
6. Створити модель кластерізації.

7. Проаналізувати отримані результати.

8. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Метод класифікації (також відомий як метод класифікаційних дерев або дерев прийняття рішень) – це алгоритм аналізу даних, який визначає покроковий спосіб прийняття рішення в залежності від значень конкретних параметрів. Дерево цього методу має наступний вигляд: кожен вузол являє собою точку прийняття рішення на підставі вхідних параметрів. Залежно від конкретного значення параметра переход здійснюється до наступного вузла, від нього – до наступного вузла, і так далі, поки не дійде до листка, який і дає остаточне рішення.

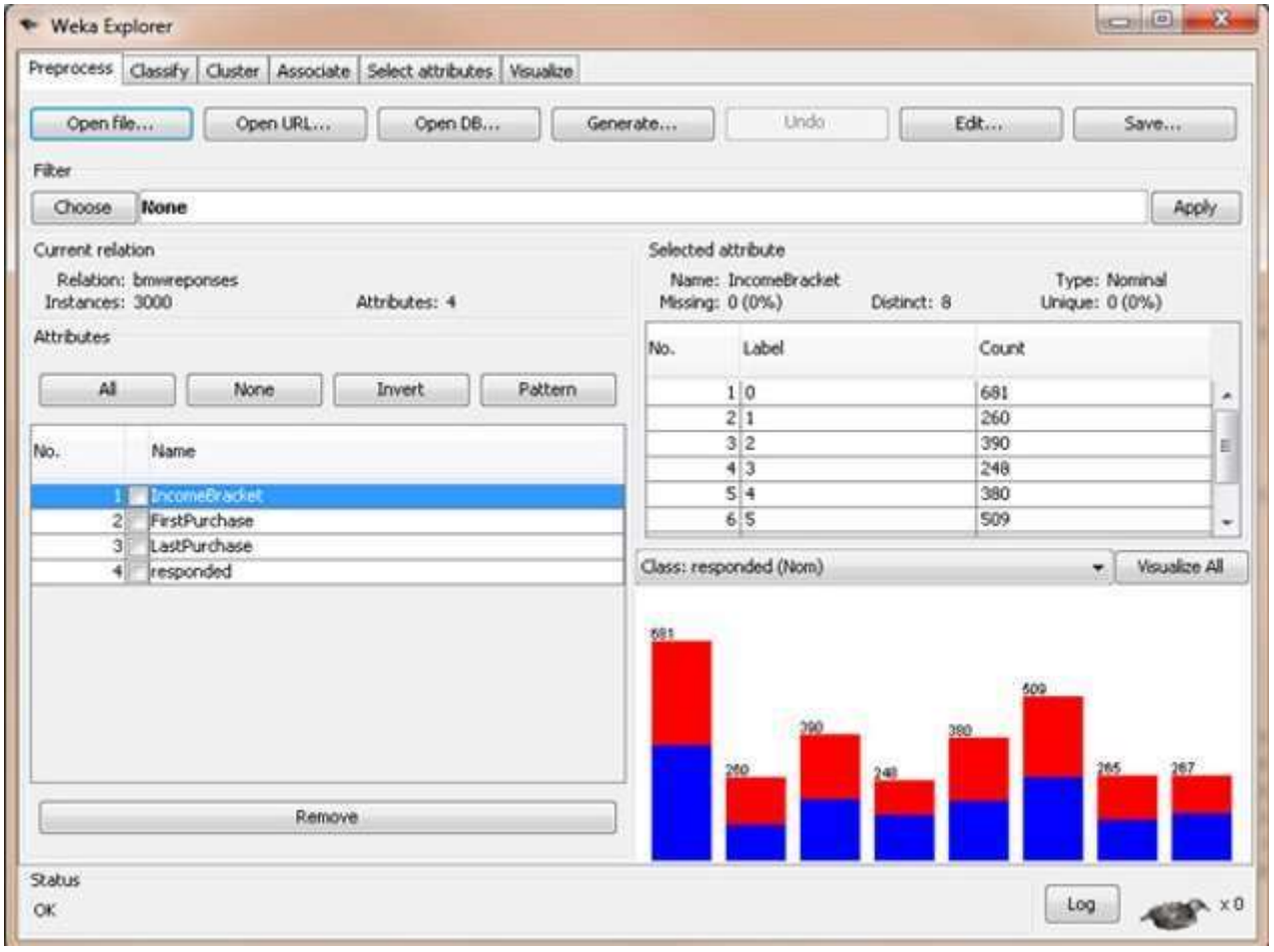

Рис. 2. Класифікація даних в WEKA

Кластеризація дозволяє розбити дані на групи, кожна з яких має певні ознаки. Метод кластерного аналізу використовується в тих випадках, коли необхідно виділити деякі правила, взаємозв'язки або тенденції у великих наборах даних. Залежно від потреб, ви можете виділити кілька різних груп даних. Одна з явних переваг кластеризації в порівнянні з класифікацією полягає в тому, що для розбиття множини на групи може використовуватися будь-який атрибут (метод класифікації використовує тільки певну підмножину атрибутів).

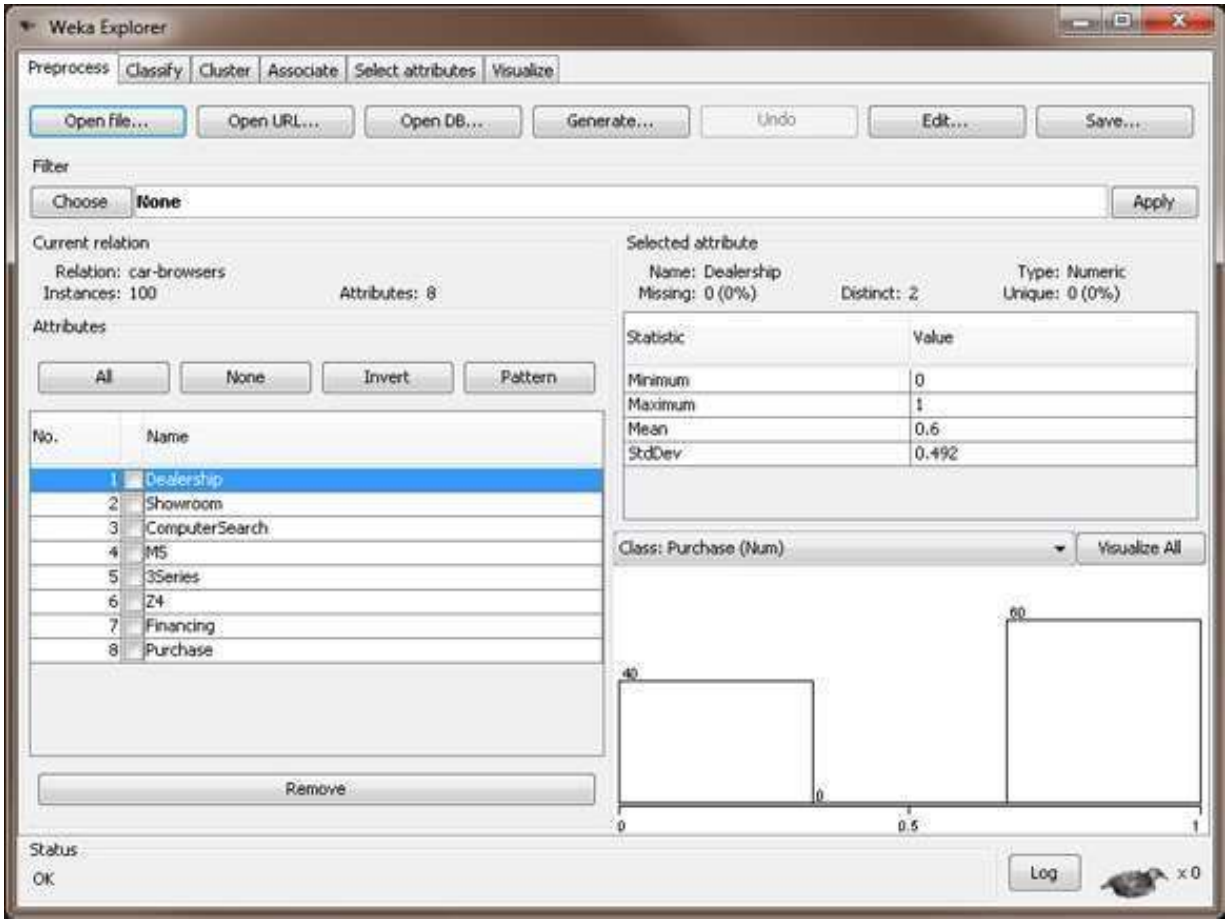

Рис. 3. Кластерізація даних в WEKA

Контрольні запитання

1. Які існують методи інтелектуального аналізу даних?
2. Під якими іншими назвами відомий метод класифікації?
3. У чому полягає основна перевага класифікаційних дерев?
4. В чому полугає сутність підходу створення моделі на основі навчальної послідовності?

5. З якою метою в методі класифікаційного аналізу набір даних ділиться на дві частини?
6. Описати принцип відсікання гілок.
7. Охарактеризувати сенс проблеми помилкового розпізнавання.
8. В результаті аналізу для тестового і навчального набору даних одержана приблизно однакова точність. Як це вплине на нові дані які будуть використовуватися в цій моделі в майбутньому?
9. Коли використовується метод кластерного аналізу?
10. Описати основний недолік методу кластеризації.

Лабораторна робота № 3.
ВИКОРИСТАННЯ ПРОГРАМНОГО ПАКЕТА WEKA: МЕТОД НАЙБЛИЖЧИХ СУСІДІВ

Навчальні питання:

1. Математичний апарат методу найбліжчих сусідів.
2. Реалізація методу найближчих сусідів в WEKA.

Завдання

1. Ознайомитись з методом найближчих сусідів для аналізу великих даних.
2. Створити математичну модель методу найближчих сусідів.
3. Підготувати файл з даними для аналізу методом найближчих сусідів за допомгою пакету WEKA.
4. Створити модель методу найближчих сусідів для набору даних.
5. Проаналізувати результат обчислення моделі.
6. Змінити кількість найбліжчих сусідів в моделі та виконати обчислення.
7. Визначити недоліки методу найближчих сусідів.

8. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Метод найближчих сусідів, так само відомий як метод колаборативної фільтрації або метод навчання на прикладах, є досить корисним способом інтелектуального аналізу даних, що дозволяє використовувати накопичені приклади з відомим результатом для прогнозування очікуваного результату для нових зразків даних.

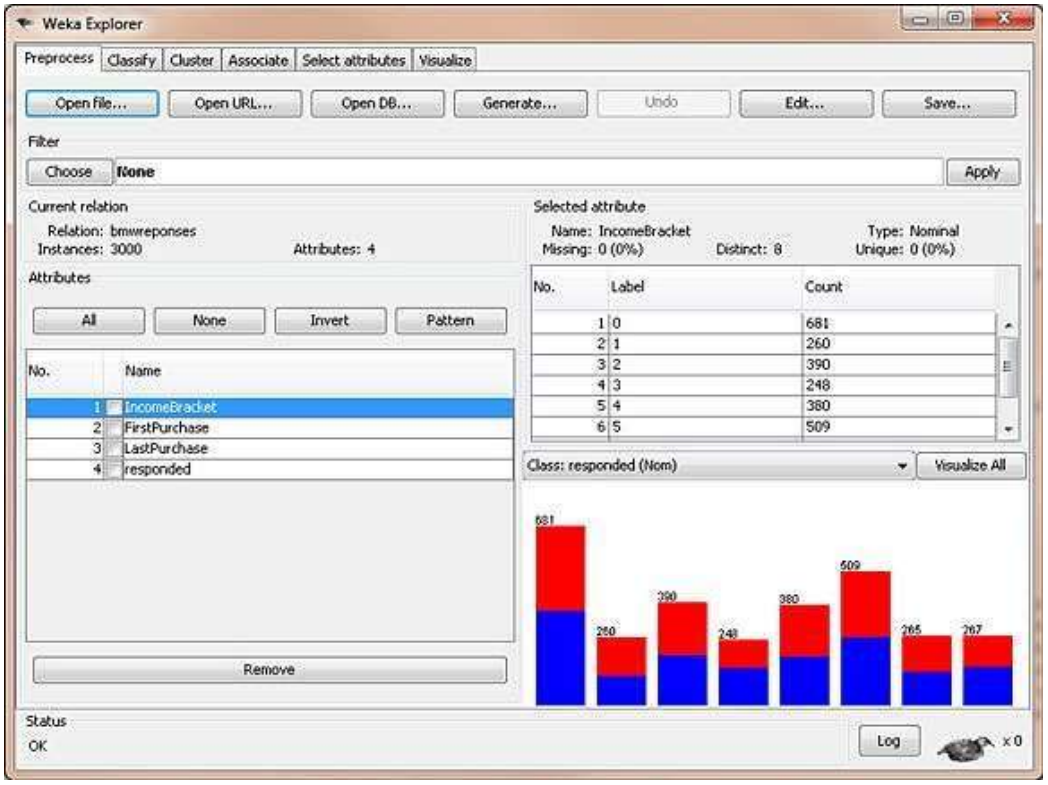

Рис. 4. Метод найближчих сусідів в WEKA

Алгоритм методу найближчих сусідів в чомусь схожий з алгоритмом, використовуваним в методі кластеризації. Метод визначає відстань між невідомою точкою і всіма відомими точками даних. Визначення відстані - цілком тривіальна процедура, що легко виконується в рамках електронних таблиць, так що досить потужний комп'ютер впорається з цим завданням практично миттєво. Найпростіший і найбільш поширений спосіб визначення відстані – це нормалізована евклідова відстань.

Контрольні запитання

1. Які інші назви використовуються для методу найближчих сусідів?
2. В чом полягає основна відмінність регресійного аналізу від методу найближчих сусідів?
3. Описати математичний алгоритм методу найближчих сусідів.
4. Яким чином можна розширити алгоритм методу найближчих сусідів?

Лабораторна робота № 4.
ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ СИСТЕМИ STATISTICA

Навчальні питання:

1. Основні поняття та принципи роботи з таблицями в системі Statistica.
2. Типи графічних даних в системі Statistica.
3. Використання функції для аналізу великиї даних в системі Statistica.

Завдання

1. Ознайомитися з принципом роботи системи Statistica.
2. Створити новий аркуш у системі Statistica.
3. Ознайомитися з принципами роботи з рядками та стовпцями на аркуші.
4. Заповнюти таблицю довільними даними та виконати обчислення стандартних функцій.
5. Додати новий аркуш у системі Statistica.
6. Заповнюти таблицю з трьома змінними. Перейменувати їх на x, y і z.
7. Побудувати графік по змінними з допомогою пункту меню Graphics-Histograms.
8. Змінити властивості графіка – колір, товщину ліній тощо.

9. Використовуючи вікно з налаштування властивостей графіка змінити тип графіка.
10. Аналогічно побудувати інший тип графіка – Graphs-2D Graphs-Scatterplots.
11. Додати на графіки назву та діапазон значень осей.
12. Аналогічно побудувати 3D графік за трьома змінними (Surface Plots).
13. Змінити властивості графіка – кут нахилу та повороту.
14. Створити нову таблицю з чотирьома змінними: *номер залікової книжки, прізвище, курс, успішність (середній бал)*. Заповнюти таблицю довільними значеннями.
15. Відсортувати значення в заліжності від типу даних змінних: текст – в алфавітному порядку, числові дані – за зростанням.
16. Побудувати графік залежності середнього балу від курсу навчання.
17. Змінити властивості графіка, функцію кривої з лінійної на довільну нелінійну.
18. Проаналізувати графік та описати залежності.
19. Побудувати гістограму для змінної *успішність*.
20. Змінити функцію кривої Normal на будь-яку іншу, використовуючи вікно властивостей графіка.
21. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Система «STATISTICA», розроблена компанією Statsoft, є однією з найбільш популярних статистичних програм для пошуку закономірностей, прогнозування, класифікації, візуалізації даних. Може застосовуватися в

економіці, промисловості, медицині, наукових дослідженнях і інших сферах людської діяльності. Клієнтами Statsoft є найбільші компанії зі світовим ім'ям.

STATISTICA – це інтегрована система аналізу та управління даними. STATISTICA - це інструмент розробки користувальницьких додатків в бізнесі, економіці, фінансах, промисловості, медицині, страхуванні та інших областях. STATISTICA легка в освоєнні і використанні.

У системі існує можливість застосовувати класичні й новітні методи проведення аналізу даних: кластерний, факторний, кореляційний, дисперсійний аналіз, лінійну й нелінійну регресії, нейронні мережі й ін. Візуалізація вихідних, проміжних, вихідних даних може бути здійснена вибором з великої кількості різних графіків, піктографіків і діаграм.

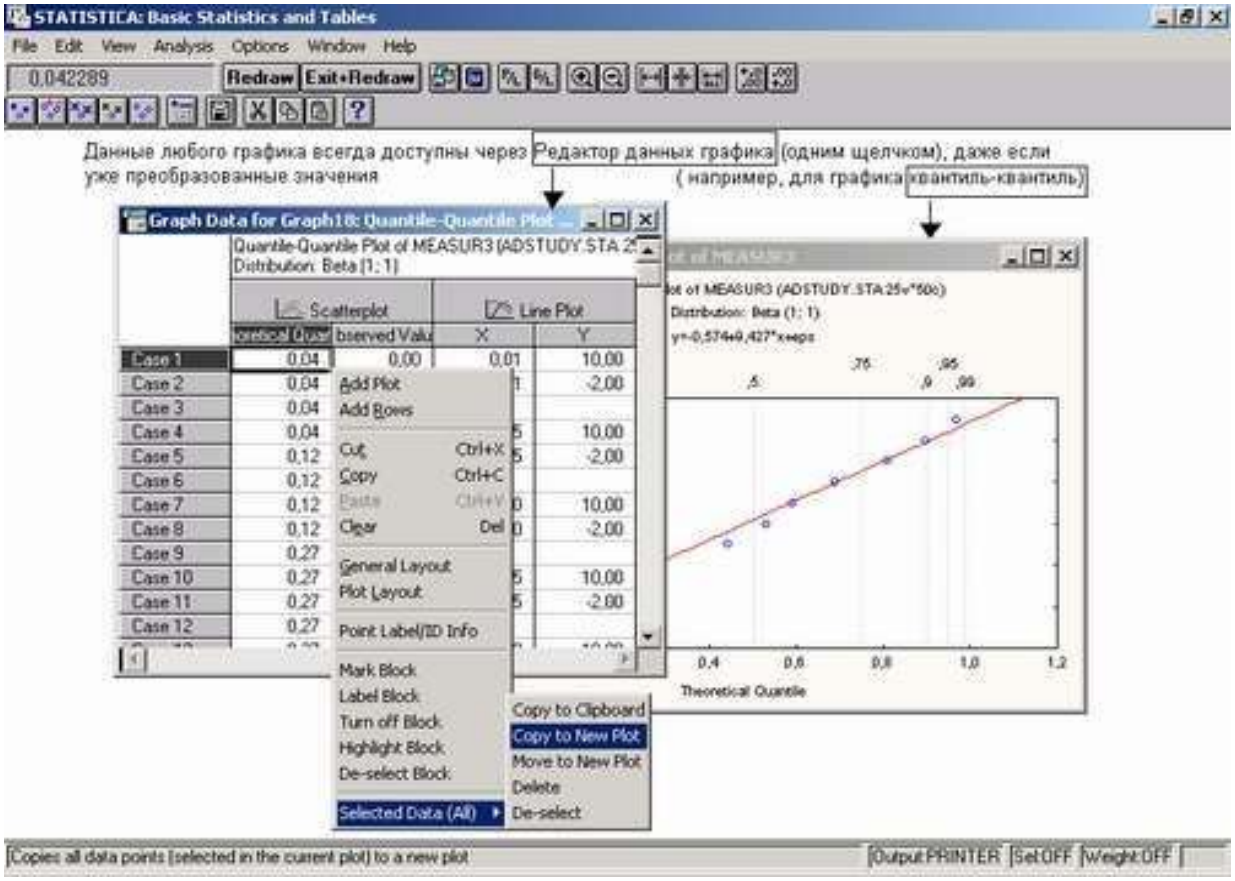

Рис. 5. Вікно системи STATISTICA

STATISTICA складається з набору модулів, в кожному з яких зібрані тематично зв'язкові групи процедур. При перемиканні модулів можна або

залишати відкритим тільки одне вікно програми STATISTICA, або всі викликані раніше модулі, оскільки кожен з них може виконуватися в окремому вікні.

Контрольні запитання

1. Що таке система «STATISTICA» і для чого вона потрібна?
2. Охарактеризуйте особливості графічного формату *.stg.
3. Як експортувати графік в інший застосунок?
4. Як імпортувати на графік системи STATISTICA графічний об’єкт з іншої програми?
5. Як помістити текст на графік STATISTICA (звіти, таблиці тощо)?

Лабораторна робота № 5.
РОЗГОРТАННЯ ПЛАТФОРМИ РОЗПОДІЛЕНИХ ОБЧИСЛЕНЬ HADOOP

Навчальні питання:

1. Основні властивості платформи побудови розподіленних застосунків для масово-паралельної обробки Hadoop.
2. Модулі платформи Hadoop.
3. Використання моделі програмування MapReduce для аналізу великих даних.

Завдання

1. Ознайомитися з архитекторою платформи Hadoop.
2. Проаналізувати модулі, які додаються до плаформи Hadoop.
3. Встановити окрему операційну систему Ubuntu Server 16.04.3 LTS на віртуальну машину VirtualBox.
4. Встановити Java – OpenJDK.

5. Створити окрему облікову запис користувача для запуску платформи Hadoop.
6. Налаштувати IP-адресу серверу та доменне ім’я вузлу.
7. Налаштувати доступ за ssh та відключити використання IPv6.
8. Встановити та налаштувати Apache Hadoop.
9. Створити на головному вузлі файл подкачки.
10. Запустити міні-кластер Hadoop.
11. Виконати задачу Word Count – завантажити в HDFS декілька текстових файлів.
12. За допомогою консолі або використовуючи веб-інтерфейс ResourceManager'a по адресу http://master:8088/cluster/apps/ відслежувати роботу платформи.
13. По завершенні виконання аналізу, відкрити папку /out в HDFS та ознайомитис з результатом.
14. Зберегти результат аналізу в локальну файлову систему.
15. Виконати додаткові команди:
    - Видалення діректорії з HDFS: /usr/local/hadoop/bin/hdfs dfs –rm -r /out
    - Відміна режиму HDFS тільки для читання, який виникає із за збою: /usr/local/hadoop/bin/hdfs dfsadmin -safemode leave
    - Остановлення Yarn: /usr/local/hadoop/sbin/stop-yarn.sh
    - Остановлення DFS: /usr/local/hadoop/sbin/stop-dfs.sh
16. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Hadoop – популярна програмна платформа (software framework) побудови розподілених додатків для масово-паралельної обробки (massive parallel processing, MPP) даних в рамках обчислювальної парадигми *MapReduce*.

Hadoop складається з чотирьох модулів:

– сполучна програмне забезпечення Hadoop Common;

– розподілена файлова система HDFS;

– система для планування завдань і управління кластером;

– платформа програмування і виконання розподілених MapReduce обчислень Hadoop MapReduce.

У Hadoop Common входять бібліотеки управління файловими системами, підтримуваними Hadoop, і сценарії створення необхідної інфраструктури управління розподіленою обробкою.

HDFS (Hadoop Distributed File System) файлова система, призначена для зберігання файлів великих розмірів, по блоках розподілених між вузлами обчислювального кластера.

YARN (англ. Yet Another Resource Negotiator) – модуль, що відповідає за управління ресурсами кластерів і планування завдань. YARN забезпечує можливість паралельного виконання декількох різних завдань в рамках кластера та їх ізоляцію (за принципами мультіарендності).

Hadoop MapReduce – програмний комплекс для програмування розподілених обчислень в рамках парадигми MapReduce.

Комплекс передає на вхід згортки відсортовані висновки від базових оброблювачів, зведення складається з трьох фаз – shuffle (тасування, виділення потрібної секції виведення), sort (сортування, угруповання по ключам висновків від розподільників – досортування, що вимагається в разі, коли різні атомарні обробники повертають набори з однаковими ключами, при цьому правила сортування на цій фазі можуть бути задані програмно і використовувати будь-які особливості внутрішньої структури ключів) і власне reduce (згортка списку) – отримання результуючого набору. Для деяких видів обробки згортка не потрібно, і комплекс повертає в цьому випадку набір відсортованих пар, отриманих базовими обработчиками.

MapReduce – це інфраструктурне рішення, здатне ефективно використовувати сьогодні кластерні, а в майбутньому багатоядерні архітектури.

Більшість сучасних паралельних комп'ютерів належить до категорії «багато потоків команд, багато потоків даних» (Multiple Program Multiple Data, MPMD). Кожен вузол виконує фрагмент загального завдання, працюючи зі своїми власними даними, а на додаток існує складна система обміну повідомленнями, що забезпечує узгодження спільної роботи.

Контрольні запитання

1. Основне призначення платформи.
2. З яких компонентів складається платформа Hadoop?
3. Для чього використовується парадигма MapReduce?
4. Принцип фази map.
5. Принцип фази reduce.
6. Яки переваги надає використання парадигми MapReduce при обробці «великих даних»?

Лабораторна робота № 6.
НАЛАШТУВАННЯ РЕПЛІКАЦІЇ НА СКБД MYSQL

Навчальні питання:

1. Основне призначення репликації даних.
2. Види репликацій даних.
3. Використання репликацій в MySQL.

Завдання

1. Ознайомитися з поняттями реплікації даних.
2. Розглянути типи реплікацій: Master-slave; Master-slave для деякількох slave-серверов; Master-master.

3. Визначити, коли виконується затримання репликації та як отримати актуальні дані при аналізі.
4. Визначити проблеми репликації в MySQL та як їх вирішувати.
5. Переваги асинхронної репликації.
6. Встановити окрему операційну систему Ubuntu Server 16.04.3 LTS на віртуальну машину VirtualBox.
7. Налаштувати статичний IP-адрес та доменне ім’я вузлу.
8. Налаштувати доступ за ssh та відключити використання IPv6.
9. Встановити на віртуальну машину MySQL-сервер.
10. Налаштувати master-slave реплікації:
    - налаштувати провідний сервер;
    - створити базу даних, в якої створити таблицу з наступними даними:

      create database clusterdb;

      create table clusterdb.simples (id int not null primary key);

      insert into clusterdb.simples values (999),(1),(2),(3);

    - переглянути зміст таблиці: *select * from clusterdb.simples*
11. Налаштувати права на реплікацію:
    - створити профіль користувача: *mysql -u root -p --prompt='master> '*
    - створити та призначити права користувачу для репліки: *grant replication slave on *.* to 'slave_user'@'%' identified by 'slavepass';*
    - обновити права доступу: *flush privileges*;
12. Заблокувати всі таблиці в створенної базі даних:

      use clusterdb;

      flush tables with read lock;

13. Перевірити статус провідного серверу: *show master status*;
14. Вийти з коснолі MySQL.
15. Створити дамп бази даних: *mysqldump -u root -p clusterdb > /home/user/clusterdb.sql*
16. Зайти в консольта розблокувати таблиці.
17. Створити базу на провідному сервері.
18. Налаштувати провідній сервер.
19. Активувати репликації на провідному сервері.
20. Перевірити роботу репликації.
21. Налаштувати права на реплікацію.
22. Активувати репликації на первому мастер-сервері.
23. Перевірити роботу репликації.
24. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Реплікація (англ. Replication) – механізм синхронізації вмісту декількох копій об'єкта (наприклад, вмісту бази даних). Під репликацией також розуміють процес копіювання даних з одного джерела на інший (або на безліч інших).

Реплікація є однією з технік масштабування даних в розподілених системах. Реплікація діляться на синхронні і асинхронні.

У разі синхронної реплікації, якщо одна репліка (копія) оновлюється, всі інші репліки того ж фрагмента даних також повинні бути оновлені в одній і тій же транзакції. Це означає, що всі репліки залишаються повинні суперечити одна одній. Недоліком даного методу є високі накладні витрати на синхронізацію реплік.

У разі асинхронної реплікації оновлення однієї репліки поширюється на інші через деякий час, а не в тій же транзакції. Таким чином, при асинхронної реплікації вводиться затримка, або час очікування, протягом якого окремі репліки можуть бути фактично неідентичних. У той же час накладні витрати на реплікацію істотно зменшуються.

Master-slave реплікація працює з одним основним сервером бази даних, який називається ведучим. На ньому відбуваються всі зміни в даних (будь-які запити MySQL). Ведений сервер постійно копіює всі зміни з майстра. З додатка на сервер відправляються запити читання даних (запити SELECT). Отже, провідний сервер відповідає за зміни даних, а ведений – за читання.

Master-slave реплікація на кілька slave серверів – перевага цього типу реплікації в тому, що можна використовувати більш одного відомого сервера. Зазвичай слід використовувати не більше 20 відомих серверів при роботі з одним ведучим.

MySQL реалізує асинхронну реплікацію. Це означає, що дані на відомому сервері можуть з'явитися з невеликою затримкою.

Контрольні запитання

1. Дати визначення реплікації даних.
2. В чому відмінність синхронної та асинхронної реплікації?
3. В чому полягає master-slave реплікація?
4. В чому полягає master-master реплікація?
5. Недоліки асинхронної реплікації.
6. Де застосовується асинхронная реплікація.

Лабораторна робота № 7.

ВІЗУАЛІЗАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ ІНСТРУМЕНТУ GEPHI

Навчальні питання:

1. Основне призначення та принцип роботи інструменту візуалізації даних Gephi.
2. Генерація даних.
3. Формати графов Gephi.

Завдання

1. Завантажити інструмент візуалізації даних Gephi (http://gephi.org/) і встановити його на своєму комп'ютері.
2. Ознайомитися з матеріалами, що дають уявлення про основи роботи з інструментом Gephi.
3. Сформулювати ідею мережевого графа, присвяченого активності у соціальних медіа.
4. Зібрати дані для візуалізації (наприклад, в соціальній мережі Facebook за допомогою додатка Netvizz), проаналізувати їх та агрегувати в форматі .gdf.
5. Побудувати граф, використовуючи інструментарій Gephi.
6. Додати до графу назви вершин, змінити розмір вершин відповідно до ступеня їх важливості.
7. Підготувати мережевий граф до друку.
8. Зберегти результат візуалізації у форматах .svg та .pdf.
9. Створити звіт з лабораторної роботи, в якому стисло описати виконані дії та навести необхідні ілюстрації.

Методичні рекомендації

Gephi – це пакет програмного забезпечення з відкритим кодом для мережевого аналізу. Він дозволяє доволі чітко розрізняти усі можливі зв'язки між вершинами та добре взаємодіє з бібліотекою NetworkX, завдяки цьому можна задавати колір та розмір вершин, а сам Gephi надає можливість задавати необхідні примітки для кожної вершини, регулювати розміри міток для вузлів графа, а також має функцію затінення інших вузлів графа при наведенні на конкретну вершину.

Gephi є дуже потужним засобом візуалізації різних даних. Використання баз даних дозволяє швидко і легко отримати необхідні вхідні дані для візуалізації, але крім цього Gephi підтримує багато інших форматів таких як .gexf, .gephi, .gml, .graphml та ін.

Gephi містить в собі набір основних укладок, а також безліч інших інструментів для аналізу графів.

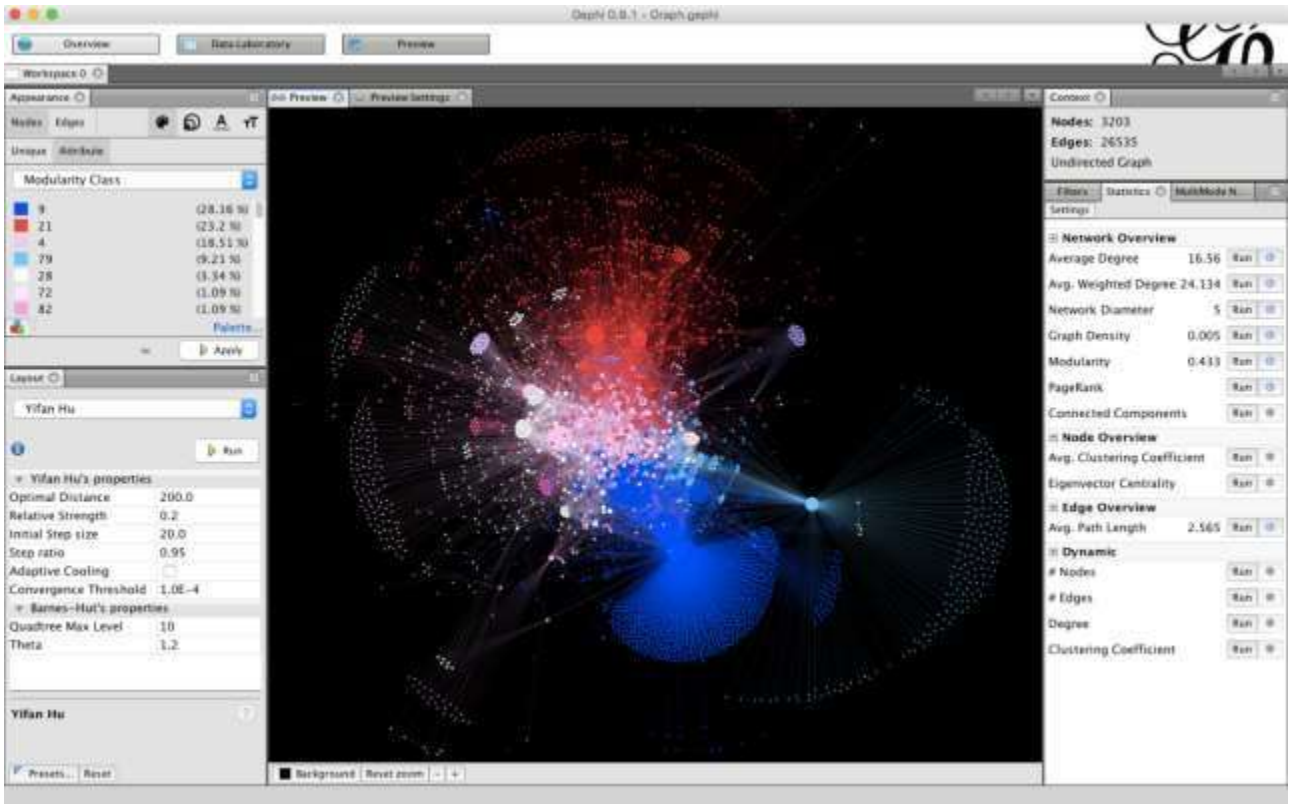

Рис. 6. Вікно інструменту Gephi

Контрольні запитання

1. Основне призначення інструменту Gephi.
2. Які алгоритми використовують в Gephi для аналізу великих даних?
3. Як експортувати граф?
4. Яки типи візуалізації використовуються в інструменті Gephi?
5. Які формати підтримує інструмент Gephi?
6. Як здійснити модіфікацію даних в в інструменті Gephi?

Теми для самостійної роботи

1. Методи і техніка аналізу великих даних.
2. Основні поняття та положення концепції Big Data.
3. Основні технології та інструменти роботи з великими даними.
4. Аналіз даних (Rapid Miner, Weka).
5. Екосистема Apache Hadoop.
6. Основи машинного обчислення засобами Spark (основні концепції).
7. Технологій обробки великих масивів даних (запити до даних з Hive).
8. Аналіз великих даних за допомогою Microsoft R.
9. Інструменти Hive.
10. Основні концепції Map Reduce.
11. Графові бази даних
12. Графова база даних Neo4j. Мова запитів до графів Cypher.
13. Створення та оцінки моделей розподілу (модель Azure Machine Learning).
14. Аналітика великих даних: принципи, напрямки і задачі.
15. Порівняння використання реляційних й нереляційних способів збереження даних.
16. Візуалізація даних Gephi.
17. Мова Python та R, стек бібліотек аналізу даних.
18. Консолідація даних.
19. Основні конструкції мови R, консолідація даних, візуалізація.
20. Аналітика потокових даних в платформі Storm.
21. Виконання програм в Hadoop.
22. Архітектура та організація даних в Apache Spark.
23. Обробка даних в GraphX.
24. Застосування технологій великих даних для задач управління в реальному часі.
25. Алгоритми класифікації та кластеризації.
26. Інструменти Sqoop.

27. Основні характеристики великих даних.
28. Консолідація даних.
29. HDFS – основи організації.
30. Виконання програм в Hadoop.
31. Алгоритми класифікації.
32. Алгоритми кластеризації.
33. Нейронні мережі як реалізація алгоритмів машинного навчання
34. Інтелектуальні алгоритми
35. Застосування технологій великих даних для задач управління в реальному часі.
36. Візуалізація даних. Веб-додатки створені для візуалізації.

ПЕРЕЛІК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ ТА ІНФОРМАЦІЙНИХ ДЖЕРЕЛ

Базова література

1. Zgurovsky M.Z., Zaychenko Y.P. Big Data: Conceptual Analysis and Applications. Springer, 2020. – 298 p.
2. Вайгенд Андреас. BIG DATA. Вся технология в одной книге. М.: Эксмо, 2018. – 384 с
3. Волкова Светлана (ред.). Просто BIG DATA. СПб.: Страта, 2019. – 148 с.
4. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.
5. Дэви С. Основы Data Science и Big Data. Python и наука о даннях // С. Дэви, М. Арно, А. Мохамед. – СПб.: Питер, 2017. – 336 с.: ил.
6. Плас Дж.В. Python для сложных задач. Наука о данных и машинное обучение. – СПб.: Питер, 2018. – 576 с.
7. Свейгарт Э. Автоматизация рутинных задач с помощью Python: практическое руководство для начинающих. – М.: Вильямс, 2017. – 573 с.

Допоміжна література

1. Дэвенпорт Том, Хо Ким Джин. О чем говорят цифры. Как понимать и использовать данные Манн, Иванов и Фербер, 2014.
2. Маккинни У. Python и анализ данных М.: ДМК Пресс, 2015. – 482 с.
3. Фрэнк Билл. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики М.: Альпина Паблишер, 2014. – 430 с.
4. Фрэнкс Билл. Укрощение больших данных М.: Манн, Иванов и Фербер, 2014. – 352 с.
5. Технологии Big Data: как использовать большие данные в маркетинге https://www.uplab.ru/blog/big-data-technologies/

Інформаційні ресурси

1. Шаховська Н. Б., Болюбаш Ю. Я. Модель великих даних “сутність характеристика”. Режим доступу: http://ena.lp.edu.ua:8080/bitstream/ntb/29775/1/20_186- 196.pdf
2. Intern. Journal of Data Science and Analytics. Special issue on Data Science in Europe. 2018. Vol. 6, Issue 3. P. 163–269.
3. Применение графовых баз данных [Електронний ресурс] // NitrosData – Режим доступу до ресурсу: http://nitrosdata.ru/2019/02/20/primenenie-grafovyh-baz-dannyh/
4. Документація Neo4j. [Електроний ресурс] Режим доступу https://neo4j.com/docs/
5. OrientDB: Матеріал з Вікіпедії–вільної енциклоредії. [Електроний ресурс] Режим доступу: https://ru.wikipedia.org/wiki/OrientDB
6. OrientDBCommunity. [Електроний ресурс] Режим доступу https:// orientdb. org/docs/
7. ArangoDB: Матеріал з Вікіпедії–вільної енциклопедії [Електроний ресурс] Режим доступу: https://ru.wikipedia.org/wiki/ArangoDB
8. ArangoDB v3.5.0 Documentation. [Електроний ресурс] Режим доступу: https://download.arangodb.com/nightly/doc/ArangoDB_Drivers_3.5.0-devel.pdf

ЗМІСТ

ЕЛЕКТРОННЕ НАВЧАЛЬНЕ ВИДАННЯ

Сергій Вікторович Кухаренко

АНАЛІЗ ВЕЛИКИХ ДАНИХ

Методичні вказівки
до виконання лабораторних та самостійних робіт
з дисципліни
«Аналіз великих даних»

для підготовки здобувачів вищої освіти
галузі знань 12 «Інформаційні технології»