**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ**

**Інститут (факультет )**__Автоматизації і комп'ютерних_____
**Кафедра** _____Інформаційних систем_____

| **«До захисту в ЕК»** | **«До захисту допущено»** |
|---|---|
| Директор інституту(декан факультету) | Завідувач кафедри |
| _____ __Андрій Форсюк__ | _____ _Сергій_Чумаченко_ |
| (підпис) (ім'я та прізвище) | (підпис) (ім'я та прізвище) |
| «___» _____ 20__р. | «___» _____ 20__р. |

**КВАЛІФІКАЦІЙНА РОБОТА**
**НА ЗДОБУТТЯ ОСВІТНЬОГО СТУПЕНЯ МАГІСТРА**

зі спеціальності_122 «Комп'ютерні науки та інформаційні технології»_____
(код та назва спеціальності)
освітньо-професійної програми __Інформаційні управляючі системи та технології_____
на тему:__Дослідження та застосування технології WebUsagemining для аналізу використання Web-ресурсів у мережі Internet

Виконав: здобувач_2_ курсу, групи_4_

___Базь Валентин Романович_____
(прізвище, ім'я, по батькові повністю) (підпис)

Керівник __М'якшило Олена Михайлівна_____ _____
(прізвище , ім'я та по батькові повністю) (підпис)

Консультанти_Олена М'якшило _____
(ім'я та прізвище) (підпис)

_____ _____
(ім'я та прізвище) (підпис)

_____ _____
(ім'я та прізвище) (підпис)

Рецензент __Олег Клименко _ _____
(ім'я та прізвище) (підпис)

Я як здобувач(ка) Національного університету харчових технологій розумію і підтримую політику університету з академічної доброчесності. Я не надавав(-ла) і не одержував(-ла) недозволеної допомоги під час підготовки цієї роботи. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Здобувач_____
(підпис)

Київ - 2022_р.

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

Інститут (факультет) __Автоматизації і комп'ютерних систем_____
Кафедра __Інформаційних систем_____
Освітній ступінь__магістр_____
**Спеціальність**__122 «Комп'ютерні науки»_____
**(код і назва)**
**Освітньо-професійна програма** Інформаційні управляючі системи та технології
_____
**(назва)**

**ЗАТВЕРДЖУЮ**
**Завідувач**
**кафедри** Інформаційних систем

_____Чумаченко С.В._____
**"** 11 **"** __листопада__**2021року**

# З А В Д А Н Н Я

## НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧА

_____Базь Валентин Романович_____
(прізвище, ім'я, по батькові)

1. Тема _Дослідження та застосування технології WebUsagemining для аналізу використання Web-ресурсів у мережі Internet _____
_____
керівник роботи__М'якшило Олена Михайлівна, доцент, к.т.н._____,
( прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом закладу вищої освіти від "11" _11_2021 року №884-кс__
2. Строк подання здобувачем роботи__2.02.2022 року_____
3. Вихідні дані до роботи Інформація про веб-сайт кафедри ІС, log-файли з веб-серверу_____
_____
_____
_____
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити) _Опис технології web mining, опис особливостей веб-сайту кафедри інформаційних систем, постановка завдання дослідження, дослідження особливостей методів попередньої обробки, дослідження методами Data Science, створення рекомендацій на основі отриманих результатів_
_____
_____
_____
5. Перелік графічного матеріалу
_Схеми та результати процессів web usage mining, діаграми результатів досліджень,_____

6. Консультанти розділів роботи

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | завдання прийняв |
| 1 | М’якшило О.М. | | |
| 2 | М’якшило О.М. | | |
| 3 | М’якшило О.М. | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

7. Дата видачі завдання__11 листопада 2021 р._____

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів виконання кваліфікаційної роботи | Строк виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Огляд задач та інструментів аналізу даних в мережі Інтернет. Постановка задачі дослідження. | 15.11 – 28.11.21 | Виконано |
| 2 | Дослідження та обґрунтування вибору методів Web Usage Mining | 29.11 – 19.12.21 | Виконано |
| 3 | Дослідження реальних даних з сайту кафедри ІС | 20.12 – 23.01.2022 | Виконано |
| 4 | Створення автореферату та підготовка презентації | 24.01 – 06.02.2022 | Виконано |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

**Здобувач** _____ ____Базь В.Р._____
(підпис) (прізвище та ініціали)

**Керівник роботи** _____ ____М’якшило О.М.____
( підпис ) (прізвище та ініціали)

Зміст

## Вступ

**Актуальність теми.** Користуючись веб-сайтом відвідувачі самі того не підозрюючи залишають багато інформації. Ці дані зберігаються у вигляді log-файлів. Це об’ємні, текстові файли які практично неможливо обробити в ручну.

Технологія Web usage mining вже давно використовується для дослідження використання веб-ресурсів, хоч і в переважній більшості, комерційних сайтів. З’явилось багато методів що використовуються в тій чи іншій ситуації. Навіть для проведення деяких невеликих, в масштабі всієї технології Web usage mining, процесів існує не тощо методи, а й цілі групи методів. В той же час, оскільки методів багато і різниця між деякими з практичної точки зору неочевидна, вибір методів часто залежить від дослідника, або можливостей програмного рішення для дослідження.

За допомогою Web usage mining можна досліджувати поведінку користувача на сайті, які сторінки та в якому порядку відвідує, як потрапляє на сайт і як його покидає і т. п. Таким чином можна краще розуміти користувача і використати це для виявлення та виправлення дуже не очевидних проблем веб-сайту.

**Зв’язок роботи з науковими програмами, планами, темами.** Дана робота виконувалась згідно з планом та програмою наукових досліджень на кафедрі інформаційних систем Національного університету харчових технологій за тематикою «Дослідження та впровадження інформаційних технологій у галузях харчової промисловості та освіті, № держреєстрації 0117U003475».

**Об'єкт дослідження.** Об'єктом дослідження є інформація накопичена в журналі використання веб-ресурсів сайту кафедри інформаційних систем НУХТ.

**Предмет дослідження.** Предметом досліджень є процеси та методи технології Web usage mining.

**Мета і завдання дослідження.** Мета цього дослідження – виявлення проблем та формування пропозицій щодо підвищення ефективності використання сайту кафедри інформаційних систем НУХТ на основі

інтелектуального аналізу використання веб-ресурсів сайту кафедри. З цього випливають наступні завдання:

1. Дослідження особливостей функціонування сайту кафедри інформаційних систем та визначення джерел інформації для використання інструментів Web usage mining.
2. Дослідження процесу технології Web usage mining та обґрунтування вибору методів для інтелектуального аналізу функціонування веб-сайту кафедри інформаційних систем
3. Проведення етапу попередньої підготовки даних;
4. Дослідження особливостей використання методів ідентифікації користувача та сеансу
5. Дослідження використання веб-ресурсів методом асоціації;
6. Дослідження використання веб-ресурсів методом кластеризації.
7. Дослідження використання веб-ресурсів методом класифікації;
8. Формування рекомендацій щодо покращання роботи сайту.

**Методи дослідження.** В даній роботі були використані такі методи:

Евристичні методи – в рамках цієї роботи використовуються для ідентифікації користувачів та ідентифікації сеансів. Методи інтелектуального аналізу даних. В рамках цього дослідження використовується три типи алгоритмів: асоціація, класифікація, кластеризація.

**Наукова новизна одержаних результатів.**

- Визначено особливості використання технології Web usage mining для дослідження використання веб-ресурсів не комерційного сайту.
- Виявлено вплив обмеженої кількості даних на результат використання технології Web usage mining.
- Запропоновано рекомендації що до виправлення виявлених недоліків сайту кафедри.

**Практичне значення отриманих результатів.** Отримані результати дозволяють краще зрозуміти поведінку користувачів на сайті. Також отримані результати дослідження можуть бути використані як відправна точка для виправлення недоліків сайту кафедри інформаційних систем.

**Апробація результатів магістерської роботи.** Основні положення та результати наукової роботи доповідались на III міжнародній науково-практичній конференції «Сучасні тенденції розвитку інформаційних систем і телекомунікаційних технологій».

**Публікації.** Ключові положення наукової роботи опубліковані в тезах III міжнародної науково-практичної конференції «Сучасні тенденції розвитку інформаційних систем і телекомунікаційних технологій» у 2022 році, а також на VIII Міжнародній науково-технічній Internet-конференції «Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами» у 2021 році.

**Структура роботи.** Дана робота складається з вступу, трьох розділів, висновку, двох додатків, окрім цього 33 рисунки та 2 таблиці.

# Розділ 1. Огляд задач та інструментів аналізу даних в мережі Інтернет. Постановка задачі дослідження.

## 1.1 Web Mining. Основні поняття

Web mining – це технологія, суть якої полягає у використанні алгоритмів та методів Data mining для пошуку залежностей та знань в мережі інтернет. Виділяють три категорії, так би мовити напрямки, у Web mining:

- Вилучення веб контенту, або Web content mining
- Вилучення структур, або Web structure mining
- Аналіз використання веб ресурсів, або Web usage mining

Напрямок Web content mining використовується для вирішення задач, пов'язаних з пошуком знань в мережі інтернет. Пошук та дослідження веб контенту вимагає використання інформаційного пошуку, машинного навчання та методів Data mining, таких як класифікація і кластеризація. В даному напрямі досліджується саме наповнення сторінок.

Web structure mining досліджує взаємозв'язки між сторінками основуючись на зв'язках між ними, тобто досліджується сама структура веб сайту. Під час дослідження, зазвичай моделюється структура сайту з певним рівнем деталізації. В самому простому випадку гіперпосилання представляють у вигляді графа G = (D, L), де D – набір сторінок, L – набір посилань. Web structure mining може використовуватись як підготовка до Web content mining.

Напрямок Web usage mining, у якості даних для дослідження використовує записи з log-файлів веб-сервера, журналу дій користуача. Цей процес ще називають "дослідження потоків кліків", оскільки в журнал заносяться всі дії користувача. Ціллю Web usage mining, зазвичай є виявлення уподобань користувачів при використанні ресурсів в мережі інтернет. На основі такого дослідження можна виявити неочевидні недоліки в структурі та наповненні сайту, та виправити їх.

Залежно від цілей та напрямку дослідження, можуть використовуватись різні методи Data mining. На рис.1.1 показано зв’язок між категоріями Web mining та методами Data mining [4].

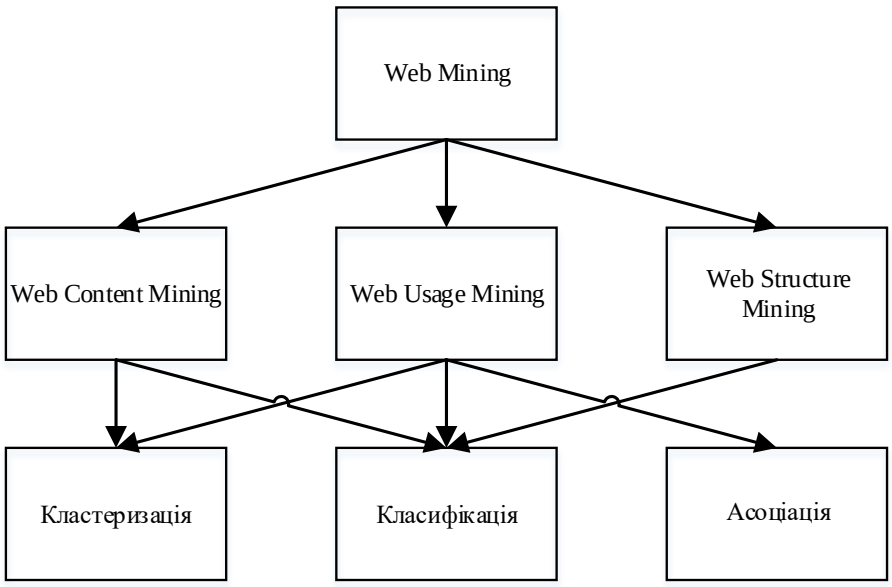



Рис.1.1 Зв’язок між Web mining та Data mining

## 1.2 Журнал дій користувача у Web Usage Mining

Як було сказано в попередньому пункті, основним джерелом даних для дослідження у технології Web usage mining, є журнал дій користувача. Цей журнал представляє собою текстовий файл, зазвичай дуже об’ємний, в який послідовно заносяться дані про всі дії користувача на сайті. В журналі дій завжди є поля: “Віддалений хост ”, “Дата/Час”, “HTTP запит”, “Код стану”, “Кількість переданих даних”. Що до іншої інформації, то її наявність залежить від формату ведення журналу і самого веб серверу. Розглянемо найбільш популярні.

Common log format, або CLF це найбільш простий формат. Він має такі поля: “Віддалений хост ”, “Ідентифікація”, “Аутентифікація”, “Дата/Час”, “HTTP запит”, “Код стану”, “Кількість переданих даних”.

Extended common log format, або ECLF це трохи модифікований формат CLF. До вже наявних полів додали це поля “Напрям ” та “User Agent”. Зараз це

найбільш популярний формат. Приклад запису: 109.248.148.245 - - [25/Oct/2021:07:01:30 +0300] "GET /form/ HTTP/1.0" 200 11684 "http://is.nuft.edu.ua/form/" "Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/90.0.4430.93 Safari/537.36"

Формат MS IIS, розроблений компанією Microsoft включає такі поля: “IP адреса клієнта”, “Ім’я користувача”, “Дата”, “Час”, “Сервіс”, “Ім’я сервера”, “IP адреса сервера”, “Пройдений час”, “Кількість даних відправлених клієнтом”, “ Кількість даний відправлених сервером”, “Код стану сервісу”, “Код стану Windows”, “Тип запиту”, “Ціль операції”, “Параметри”. Формат MS IIS містить найбільше полів і надає найбільше корисної інформації. Проте і для зберігання такої кількості інформації потрібно більше місця [2].

## 1.3 Процес Web Usage Mining

Процес Web Usage mining означає виявлення та аналіз моделі в потоках кліків, транзакціях користувачів та інших пов’язаних даних, зібраних або створених в результаті взаємодії користувача з веб-ресурсами на одному чи кількох веб-сайтах. Метою процесу є виявлення, моделювання та аналіз моделі поведінки користувачів, які взаємодіють з а веб-сайтом. Виявлені закономірності зазвичай представляють у вигляді колекцій сторінок, об’єктів або ресурсів, до яких часто звертаються або які використовуються групами користувачів із спільними потребами чи інтересами.

В цілому процес Web Usage mining можна розділити на три взаємозалежні етапи:

1. Збір даних і попередня обробка;
2. Виявлення шаблонів;
3. Аналіз шаблонів;

На етапі попередньої обробки дані очищаються та розподіляються на набори транзакцій користувачів, що представляють діяльність кожного окремого користувача під час різних сеансів відвідування сайту.

Інші джерела знань, такі як вміст або структура сайту, знання семантичної області з сайту, онтології (наприклад, каталоги продуктів або ієрархії концепцій), використовуються для попередньої обробки або для покращення даних про транзакцій користувача.

На стадії виявлення шаблонів, статистичні операції, бази даних та операції машинного навчання використовуються для отримання прихованих закономірностей, що відображають типову поведінку користувачів, а також підсумкову статистику по веб-ресурсам, сеансам і користувачам.

На останньому етапі процесу виявлені закономірності та статистичні дані додатково обробляються, фільтруються, що, можливо, призводить до сукупних моделей користувачів, які можна використовувати як вхідні дані для:

- Рекомендацій;
- Інструменти візуалізації;
- Інструменти веб-аналітики;
- Створення звітів;

Загальний процес зображено на рис. 1.2.

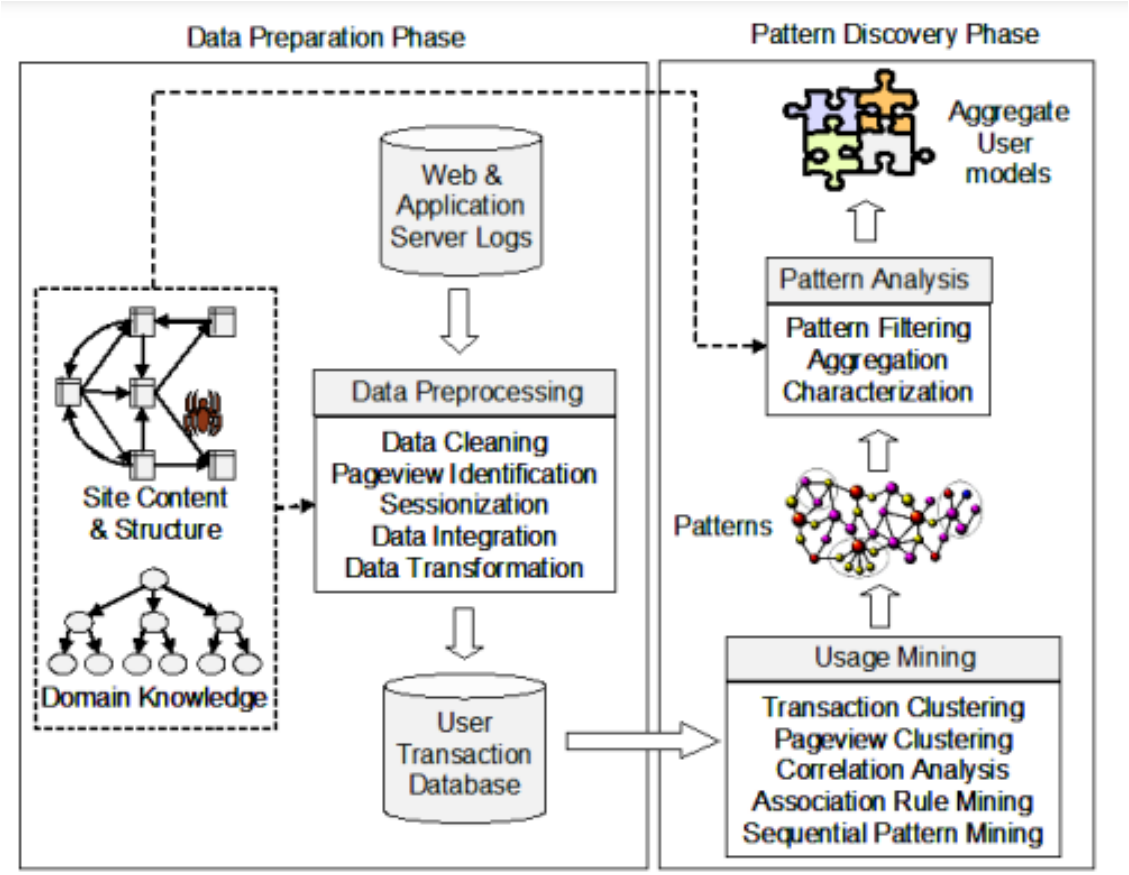



Рис.1.2. Загальний процес Web Usage mining

## 1.4 Інформація про веб-сайт кафедри інформаційних систем НУХТ

Веб-сайт — це сукупність програмних, інформаційних, медійних засобів, пов'язаних між собою, оформлених у вигляді окремих сторінок і доступних в мережі Інтернет.

Виділяють такі типи веб-сайтів:

- Рекламні веб-сайти. Такі сайти створюються спеціально для рекламних цілей, хоча напряму й не займаються продажом товарів або послуг. Для них характерними є використання на сторінках великої кількості графіки та анімацій. Для залучення користувачів використовують розважальні методи.
- Веб-сайти продавці. Ці веб-сайти створюються, як магазини.

- Веб-сайти альтруїсти. Це інформаційні веб-сайти що надають безкоштовні послуги. Зазвичай такі сайти займаються збором якої небуть інформації про користувача.
- Веб-сайт підтримки. На такого типу веб-сайтах розміщуються наприклад оновлення для ПЗ та інструкція, актуальні новини що до компанії або продукту.

Хоча в цілому і виділяють такі типи, більшість сайтів вдало поєднує характеристики з кількох типів. Це призвело до появи сайтів - “офісів”, де в рамках одного сайту надаються різні послуги, наприклад продаж та підтримка, та скоріш за все ще і збір інформації.

Сайт кафедри інформаційних систем повністю підходить під визначення сайту-альтруїста:

- Сайт наповнено інформацією про кафедру, викладачів і т. п.
- На сайті є можливість заповнити анкету та залишити відгук.

Потрапляючи на сайт через головну сторінку (як і планується), користувач бачить актуальні новини про події на кафедрі, а також меню для доступу до інших розділів сайту. Також слід додати що меню доступне на будь якій сторінці сайту, тобто з будь якої сторінки (крім адміністративної панелі) можна потрапити на будь яку іншу сторінку. На рис. 1.3 зображено головну сторінку сайту.

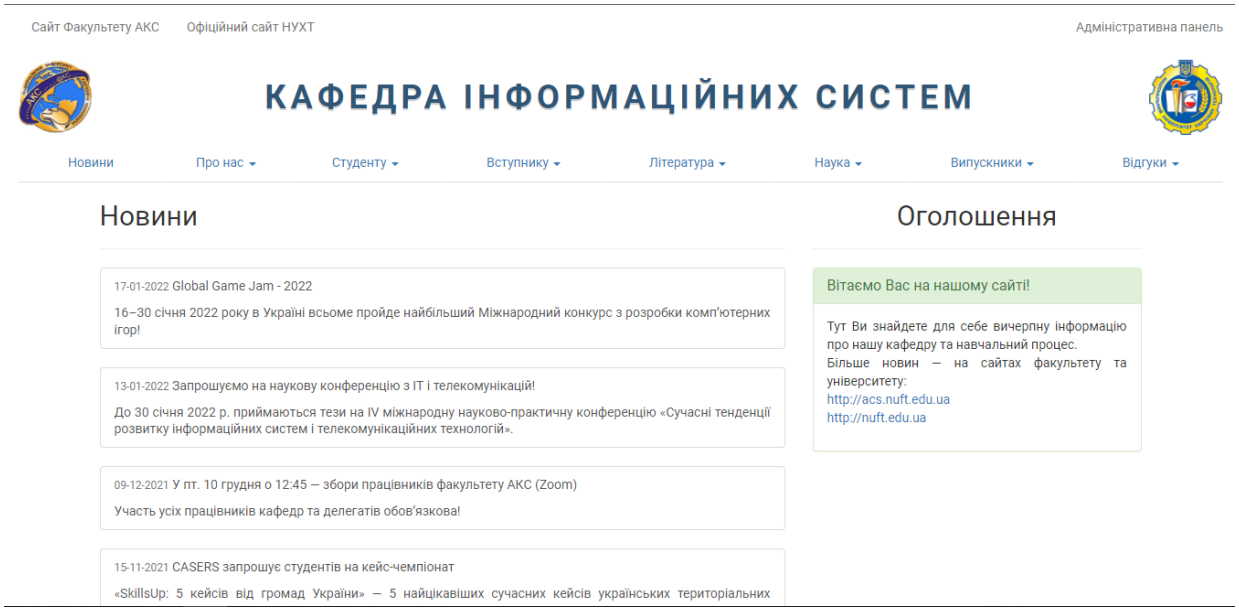



Рис. 1.3 Головна сторінка сайту

Майже всі пункти меню зроблені у вигляді випадного списку і при натиску на них мають вигляд як на рис. 1.4.

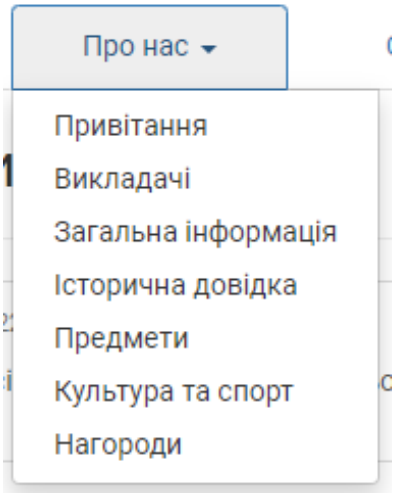



Рис. 1.4 Вкладка “Про нас” у вигляді випадного списку

На сайті можна знайти інформацію про навчальний процес та в цілому про кафедру. Інформація на сайті досить зручно поділена по розділах. Наприклад в розділі “Студенту” можна знайти інформацію що є актуальною для студентів, в розділі “Про нас” інформацію про кафедру, викладачів, предмети і т. д.. Інформація на сторінках подана коротко і по темі.

На рис.1.5 показана карта сайту по якій можна зрозуміти з яких розділів складається сайт.


Рис.1.5 Карта сайту кафедри ІС

В цілому можна сказати що це досить простий інформаційний сайт, цільовою аудиторією якого є користувачі, що якось пов’язані з кафедрою ІС. В основному це вступники, студенти, викладачі та випускники цієї кафедри.

Сайт кафедри інформаційних систем розміщено на веб-сервері Nginx. Це досить популярний веб-сервер. За даними з статистики Netcraft, станом на кінець 2021 року, 22.23% сайтів використовувало Nginx.

З причини налаштувань, як я думаю пов’язаних з економією місця, веб-сервер зберігає дані про активність користувачів лише за останні п’ять днів. Тому для дослідження доступно не так багато даних як хотілось би. Але, в цілому, навіть аналіз такого невеликого об’єму даних має дати якийсь результат.

## 1.5 Постановка задачі

Основним завданням було виявлення неявних закономірностей у використанні сторінок сайту кафедри. Проте для вирішення цього завдання слід виконати декілька попередніх кроків:

1. Синхронізувати дані з декількох журналів
2. Попередньо обробити дані. Видалити, трансформувати і т. д.
3. Ідентифікувати користувачів
4. Ідентифікувати сеанси
5. Дослідити використання веб-ресурсів за допомогою Статистики
6. Дослідити використання веб-ресурсів за допомогою Кластеризації
7. Дослідити використання веб-ресурсів за допомогою Класифікації
8. Дослідити використання веб-ресурсів за допомогою Асоціації
9. Зробити висновки щодо наявності або відсутності проблем сайту

Для вирішення деяких завдань існує багато можливостей та методів. Опираючись на знання про веб-сайт та веб-сервер слід обрати декілька методів, та порівняти результат їх використання на прикладі дослідження журналу сайту кафедри.

В результаті виконання цих задач будуть виявлені не тільки деякі закономірності в моделі поведінки користувача, а й розглянуто можливість або неможливість, доцільність або недоцільність використання деяких методів та рекомендацій що до проведення етапів Web usage mining.

## 1.6 Висновки до розділу 1

1. Технологія Web usage mining є одним з напрямків більш широкого поняття Web mining, та використовує деякі методи з Data Science
2. В основі технології Web usage mining лежить використання даних із журналу дій користувача.
3. Web usage mining складається з двох послідовних фаз: попередня обробка та дослідження шаблонів.

4. Веб-сайт кафедри інформаційних систем це простий, інформаційний сайт що розміщено на популярному веб-сервері Nginx.
5. В результаті постановки задачі, було сформовано дев’ять кроків, послідовних завдань, виконання яких призведе до виявлення закономірностей в поведінці користувача.

# Розділ 2. Дослідження та обґрунтування вибору методів Web Mining

## 2.1 Метрики

У веб-аналітиці можуть використовуватись багато метрик. В різних джерелах можна знайти статті та теми “32 ключові метрики у веб-аналітиці” або “5 ключових типів метрик сайту” і тому подібні. Тобто єдиного, правильного списку немає. Метрики обираються для конкретного дослідження конкретного сайту. В таблиці 1 показано основні метрики та параметри що використовуються найбільш популярними засобами [5]:

| Google Analytics | Яндекс.Метрика | SimilarWeb | Liveinternet |
|---|---|---|---|
| Перегляд сторінки (Pageview) | Перегляд сторінки | — | Перегляд |
| Сеанс (Session) | Сеанс | Візит (Total Visits) | Сесія |
| Користувач (User) | Відвідувач | Відвідувач (Unique Visitors) | Відвідувач |

Таблиця 1. Метрики та параметри

Використання саме цих програмних рішень (Google Analytics, Яндекс.Метрика, SimilarWeb, Liveinternet) неможливе, так як потрібно додавати програмний код на сторінку сайту. Проте можна побачити що їх основні

метрики по суті не відрізняються. Тому за основу можна взяти такий список параметрів:

- Користувач
- Сеанс
- Перегляд

Оскільки веб-аналітика використовується переважно маркетологами та для дослідження комерційних веб-сайтів, використовуються такі метрики як: середня ціна покупки, коефіцієнт покинутих корзин, та взагалі метрики що базуються на прибутку, покупці, вартості і тому подібні. Використання даних метрик для дослідження сайту кафедри взагалі не можливе.

Використання деяких метрик значно підвищило б якість дослідження, проте, через невеликий об’єм досліджуваних даних, їх використання хоч і можливе, але не принесе хоч скільки небуть корисний результат. До таких метрик можна віднести: кількість нових користувачів, періодичність візитів, коефіцієнт відтоку, та інші метрики для використання яких потрібні дані за великий проміжок часу.

Із параметрів які доцільно використовувати, можна виділити: час проведений на сторінці, тип сторінки, платформа користувача. Тут слід пояснити чому саме ці параметри.

Платформу користувача можна визначити дуже точно так як це відображається в журналі, майже зі 100% точністю. Розділивши користувачів, сеанси або перегляди саме таким чином, можна виявити проблеми що притаманні для якоїсь конкретної платформи.

Час проведений на сторінці також легко визначається і очевидно що це корисні дані. Але, є деякі нюанси. Для останньої сторінки в сеансі, час визначити неможливо. Так як час проведений на сторінці це різниця між часом завантаження однієї сторінки і часом завантаження наступної, то для сторінки виходу час буде дорівнювати нулю. З іншого боку, таким чином можна визначати точку виходу.

Тип сторінки. Даний параметр використовується у багатьох методах дослідження, проте у трохи іншому вигляді. Часто використовують такі параметри як, точка входу або точка виходу. При точному моделюванні поведінки користувача також виділяють конкретні кроки, послідовність сторінок. Проте точне моделювання є досить ресурсовитратним та виходить за рамки саме цього дослідження. Під типом сторінки я розумію те, що сторінка може бути входом, виходом, проміжним етапом та вхід/вихід якщо ввесь сеанс це перегляд однієї сторінки.

Таким чином досліджувати використання веб-ресурсів сайту кафедри інформаційних систем я буду використовуючи такі вхідні параметри: Перегляд, Користувач, Сеанс, Час, Тип сторінки, Платформа.

## 2.2 Етап збору та попередньої обробки даних

Важливим завданням для будь якого аналізу даних є створення відповідного цільового набору даних, до якого можуть використовуватися алгоритми аналізу даних і статистичні методи. Це особливо важливо для Web Usage mining, де через характеристики даних кліків та їх зв’язок з іншими пов’язаними даними зібраних з кількох джерел і через декілька каналів. Процес підготовки даних часто займає найбільше часу та обчислень і вважається найбільш інтенсивним етапом у процесі Web Usage mining, і часто вимагає використання спеціальних алгоритмів та евристичних методів, які зазвичай не використовуються в інших ситуаціях. Цей процес має вирішальне значення для успішного виявлення корисних шаблонів з даних. Процес може включати попередню обробку вихідних даних, інтегрування даних з кількох джерел і трансформування інтегрованих даних у форму, придатну для введення та проведення конкретного аналізу даних. У сукупності це називається підготовкою даних.

Підготовка даних про використання створює ряд унікальних проблем, які привели до створення різноманітних алгоритмів та евристичних методів для завдань попередньої обробки, таких як злиття та очищення даних, ідентифікація користувача та сеансу.

Успішне застосування методів аналізу даних дуже залежить від правильного проведення попередньої обробки.

На рис.2.1 подано короткий огляд основних завдань та елементів при попередній обробці даних [1].

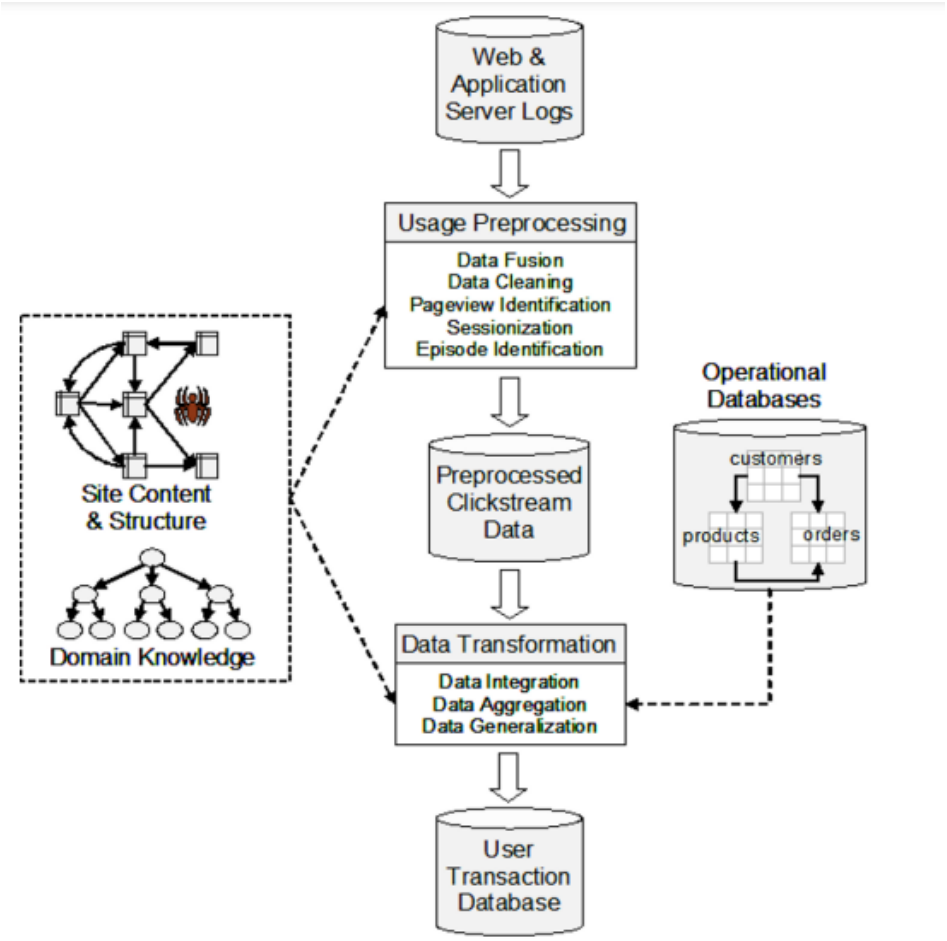



Рис.2.1 Етапи підготовки даних для Web Usage mining.

## 2.3 Джерела та типи даних

Основними джерелами даних, які використовуються при аналізі використання ресурсів веб-сайтів, є журнал сервера файли, які включають

журнали доступу до веб-сервера та журнали сервера програм. Додаткові джерела даних, які також необхідні як для підготовки даних, так і для виявлення шаблонів включає файли сайту та метадані, операційні бази даних, шаблони додатків і знання предметної області. У деяких випадках і в деяких користувачів, додаткові дані можуть бути доступні на рівні клієнта або проксі (Інтернет-провайдера), а також із зовнішніх джерел.

Дані, отримані з різних джерел, можна розділити на чотири категорії.

**Дані про використання:** дані журналу, які автоматично збираються з сайту та серверів додатків і показують дрібнозернисту навігаційну поведінку відвідувачів. Це основне джерело даних для Web Usage mining. Кожне звертання до сервера, що відповідає HTTP-запиту, генерує одиночний

запис у журналах доступу до сервера. Кожен запис журналу (залежно від формату журналу) може містити поля, що визначають час і дату запиту,

IP-адреса клієнта, запитуваний ресурс, можливі параметри, які використовуються при виклику веб-програми, статус запиту, використаний метод HTTP, агент користувача (тип і версія браузера та операційної системи), посилання на Веб-ресурс, якщо доступні, файли cookie на стороні клієнта, які однозначно дозволяють визначити повторного відвідувача. Типовим прикладом журналу доступу до сервера зображено на рис.2.2, на якому показано кілька часткових записів журналу. Дані в записах журналу були змінені для захисту конфіденційності.

| | |
|---|---|
| 1 | 2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://dataminingresources.blogspot.com/ |
| 2 | 2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727)<br>http://maya.cs.depaul.edu/~classes/cs589/papers.html |
| 3 | 2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200<br>318814 HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1)<br>http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey |
| 4 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/ |
| 5 | 2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636<br>HTTP/1.1 maya.cs.depaul.edu<br>Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)<br>http://maya.cs.depaul.edu/~classes/cs480/announce.html |

Рис.2.2. Приклад журналу дій користувача

Залежно від цілей аналізу ці дані повинні бути трансформовані та агреговані на різних рівнях абстракції. У веб-використанні найпростіший рівень абстракції даних – це перегляд сторінки. А перегляд сторінки – це сукупне представлення колекції веб-об’єктів та відображення їх в браузері в результаті дії одного користувача. Концептуально кожен перегляд сторінки може бути розглядається як набір веб-об’єктів або ресурсів, що представляють конкретну «подію користувача», наприклад, читання статті, перегляд сторінки продукту або додавання товару в кошик. На рівні користувача найпростіший рівень поведінкової абстракції — це абстракція сеансу. Сеанс, або сесія - це послідовність переглядів сторінок одним користувачем за одне відвідування. Поняття сесії можна додатково абстрагувати, вибравши підмножину переглядів сторінок у сеансі які є значущими або релевантними для поставлених завдань аналізу.

**Дані про вміст:** Дані вмісту сайту – це сукупність об’єктів і відносин, які передаються користувачеві. Здебільшого ці дані складається з комбінацій текстових матеріалів та зображень. Дані джерела, що використовуються для доставки або створення цих даних, включають статичні HTML/XML сторінки,

мультимедійні файли, динамічно згенеровані сегменти сторінки, скрипти та колекції записів з баз даних.

**Структурні дані:** Дані структури представляють погляд дизайнера на організацію контенту на сайті. Ця організація представлена через структуру зав'язків між сторінками, що відображається через гіперпосилання. Дані структури також включають внутрішню структуру вмісту в межах сторінки. Наприклад, можуть бути як документи HTML, так і XML представлені у вигляді деревоподібних структур на просторі тегів сторінки. Структури гіперпосилань для сайту зазвичай фіксуються в автоматично згенерованій «карті сайту». Інструмент картографування сайту повинен мати можливість захоплення і представлення взаємозв'язків між переглядами й усередині сторінок. Для динамічно створених сторінок, інструменти відображення сайту повинні або включати знання внутрішніх основних програм і скриптів, які генерують HTML вміст, або повинні мати можливість генерувати сегменти вмісту за допомогою вибірки параметрів, що передаються таким додаткам або скриптам.

**Дані користувача:** Оперативна база даних для сайту може містити додаткову інформацію про профіль користувача. Такі дані можуть включати демографічну інформацію про зареєстрованих користувачів, оцінки користувачів щодо різних об'єктів, таких як продукти або фільми, минулі покупки чи історії відвідувань користувачів, а також інші явні або неявні представлення інтересів користувачів. Деякі з цих даних можуть бути зафіксовані анонімно, доки можна розрізнити різних користувачів. Наприклад, анонімна інформація, що міститься в клієнтських файлах cookie, може вважатися частиною інформації профілю користувачів і використовуватися для ідентифікації повторних відвідувачів сайту. Багато програм персоналізації вимагають зберігання попередньої інформації профілю користувача [1].

## 2.4 Ключові етапи попередньої обробки даних

Як зазначено на рис.2.1, завдання для попередньої обробки даних включають злиття та синхронізацію даних із кількох файлів журналів, очищення даних, ідентифікація перегляду сторінок, ідентифікація користувача, ідентифікація сеансу (або сеансування), ідентифікація епізоду та інтеграція даних кліків із іншими джерелами даних, наприклад вмістом, а також інформацією про користувачів і продукт з оперативних баз даних. Тепер розглянемо деякі основні завдання попередньої обробки [1].

### 2.4.1 Злиття та очищення даних

Для великомасштабних веб-сайтів типовим є те, що вміст, який подається користувачам, надходить з кількох веб-серверів або серверів додатків. У деяких випадках кілька серверів з надлишковим вмістом використовуються для зниження навантаження на будь-який конкретний сервер. Злиття даних відноситься до об’єднання файлів журналів з кількох веб- та серверів додатків. Для цього може знадобитися глобальна синхронізація між ними. За відсутності спільних вбудованих ідентифікаторів сеансів, евристичні методи на основі поля «referrer» у журналах сервера разом із різними сеансами і методи ідентифікації користувача (див. нижче) можна використовувати для виконання злиття. Цей крок дуже важливий для «міжсайтового» аналізу веб-використання, де аналіз поведінки користувачів виконується над файлами журналів кількох пов'язаних веб-сайтів.

Очищення даних зазвичай залежить від конкретного сайту та включає такі завдання, як видалення сторонніх посилань на вбудовані об’єкти, які можуть не бути важливими для цілей аналізу, включаючи посилання на файли стилів, графічні або звукові файли. Процес очищення також може включати видалення принаймні деяких полів даних (наприклад, кількість байтів передано

або використана версія протоколу HTTP тощо), які можуть не надавати корисної інформації в задачах аналізу даних.

Очищення даних також тягне за собою видалення посилань через сканер навігації. Нерідкі випадки, коли типовий файл журналу містить значний відсоток посилань (іноді більше 50%), отриманих у результаті діяльності ботів або павуків. Відомі сканери пошукових систем зазвичай ідентифікуються та видаляються шляхом підтримки списку відомих сканерів.

Інші «доброхідні» сканери, які дотримуються стандартного виключення протоколів роботів, починають сканування свого сайту, спочатку намагаючись отримати доступ до файлу виключення «robots.txt» у кореневому каталозі сервера. Таким чином, такі сканери можуть бути визначеними шляхом пошуку всіх сеансів, які починаються зі (спроби) доступу до цього файлу. Однак значна частина посилань сканерів походить від тих, що або не ідентифікують себе явно (наприклад, у полі «агент»), або неявно; або від тих сканерів, які свідомо маскуються під законних користувачів. У цьому випадку для ідентифікації посилань сканера може знадобитися використання евристичних методів, які відрізняють типову поведінку веб-сканерів від фактичних користувачів. Була проведена певна робота щодо використання класифікації алгоритмів побудови моделей сканерів і веб-роботів навігації, але такі підходи мали лише обмежений успіх і потрібно більше роботи в цій сфері [1].

### 2.4.2 Ідентифікація перегляду сторінки

Ідентифікація переглядів сторінок дуже залежить від структуру сайту, а також вмісту сторінки та тематики сайту. Концептуально кожен перегляд сторінки може розглядатись як набір веб-об’єктів або ресурсів, що представляють конкретну «подію користувача», наприклад, натискання на посилання або кнопку, перегляд сторінки. Для статичного сайту, з одним кадром, кожен файл HTML може мати особистий зв'язок з переглядом сторінки.

Однак для сайтів з кількома фреймами, кілька файлів складають певний перегляд сторінки. Для динамічних сайтів, перегляд сторінки може представляти собою комбінацію статичних шаблонів і вмісту, створеного серверами додатків на основі набору параметрів [1].

### 2.4.3 Ідентифікація користувача

Використання технології Web usage mining не вимагає знань про ідентичність користувача. Однак слід розрізняти різних користувачів один від одного. Оскільки користувач може відвідати сайт кілька разів, сервер реєструє записи кількох сеансів для кожного користувача. Ми використовуємо фразу “запис активності користувача” для позначення послідовності зареєстрованих дій, що належать одному користувачеві [1].

За відсутності механізмів аутентифікації, а це найбільше поширено, підхід до розрізнення унікальних відвідувачів полягає у використанні cookie. Однак не всі сайти використовують файли cookie, і через проблеми конфіденційності, файли cookie на стороні клієнта іноді відключаються користувачами. Однієї лише IP-адреси, зазвичай, недостатньо для відображення записів журналу на набір унікальних відвідувачів, хоча використання того методу також може бути. В основному це пов’язано з поширенням проксі-серверів ISP, які призначають чергування IP-адрес клієнтам, коли вони переглядають веб сторінку. Це не рідкість, щоб знайти багато записів журналу, що відповідають обмеженій кількості IP-адрес проксі-сервера, це адреси від великих інтернет-провайдерів. Отже, два входження однієї IP-адреси (навіть розділені достатнім проміжком часу), насправді може відповідати двом різним користувачам. Без аутентифікації користувача або файлів cookie на стороні клієнта, все ще можна точно ідентифікувати унікальних користувачів за допомогою комбінації IP-адрес та іншої інформації, такої як дані з поля User Agent.

Розглянемо, наприклад, рис.2.3. На ньому зображено частину частково попередньо обробленого файлу журналу (мітки часу вказано лише у вигляді

годин і хвилин). Використовуючи комбінації поля IP і поля User agent у файлі журналу, ми можемо розділити журнал на записи діяльності для трьох окремих користувачів (зображено на рис.2.4).

| Час | IP | URL | Referer | User Agent |
|---|---|---|---|---|
| 0:01:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 0:09:00 | 1.2.3.4 | B | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 0:10:00 | 2.3.4.5 | C | - | Mozilla/5.0 (Linux; Android 6.0.1; ... |
| 0:12:00 | 2.3.4.5 | B | C | Mozilla/5.0 (Linux; Android 6.0.1; ... |
| 0:15:00 | 2.3.4.5 | E | C | Mozilla/5.0 (Linux; Android 6.0.1; ... |
| 0:19:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 0:22:00 | 2.3.4.5 | D | B | Mozilla/5.0 (Linux; Android 6.0.1; ... |
| 0:22:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 0:25:00 | 1.2.3.4 | E | C | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 0:25:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 0:33:00 | 1.2.3.4 | B | C | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 0:58:00 | 1.2.3.4 | D | B | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 1:10:00 | 1.2.3.4 | E | D | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 1:15:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 1:16:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 1:17:00 | 1.2.3.4 | F | C | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| 1:26:00 | 1.2.3.4 | F | C | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 1:30:00 | 1.2.3.4 | B | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| 1:36:00 | 1.2.3.4 | D | B | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |

Рис.2.3. Журнал

| | Час | IP | URL | Referer | User Agent |
|---|---|---|---|---|---|
| User1 | 0:01:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 0:09:00 | 1.2.3.4 | B | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 0:19:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 0:25:00 | 1.2.3.4 | E | C | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 1:15:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 1:26:00 | 1.2.3.4 | F | C | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 1:30:00 | 1.2.3.4 | B | A | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| | 1:36:00 | 1.2.3.4 | D | B | Mozilla/5.0 (Linux; Android 11; M2003J15SC)... |
| User2 | 0:22:00 | 1.2.3.4 | A | - | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 1:16:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 0:25:00 | 1.2.3.4 | C | A | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 0:33:00 | 1.2.3.4 | B | C | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 0:58:00 | 1.2.3.4 | D | B | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 1:10:00 | 1.2.3.4 | E | D | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 1:17:00 | 1.2.3.4 | F | C | Mozilla/5.0 (Windows NT 10.0; Win64; x64)... |
| | 0:10:00 | 2.3.4.5 | C | - | Mozilla/5.0 (Linux; Android 6.0.1; ... |
| | 0:12:00 | 2.3.4.5 | B | C | Mozilla/5.0 (Linux; Android 6.0.1; |

Рис.2.4. Журнал з ідентифікованими користувачами

### 2.4.4 Ідентифікація сесій

Ідентифікація сесії це процес сегментації запису активності кожного з користувачів на сеанси, кожен сеанс представляє одне відвідування сайту. Веб-сайти без додаткової інформації про аутентифікацію від користувачів і без таких механізмів, як вбудовані ідентифікатори сеансів, повинні покладатися на евристичні методи виявлення сесії. Метою таких методів є відновлення, на основі даних кліків, фактичної послідовності дій що виконується одним користувачем під час одного сеансу відвідування сайту. В ідеальному випадку це дасть можливість повторно побудувати точну послідовність навігації користувача під час сеансу [1].

Як правило, методи виявлення сесії поділяються на дві основні категорії: орієнтовані на час та орієнтовані на структуру. Орієнтовані на час застосовуються для глобальної оцінки локального тайм-ауту, щоб розрізняти послідовні сеанси, тоді як структурно-орієнтовані методи використовують або статичну структуру сайту, або неявну структуру зв'язків, записану в полях рефератів журналів сервера.

Як приклад, два варіанти методів орієнтованих на час  та базовий варіант навігаційно-орієнтованого наведено нижче. Кожен наведений метод сканує журнали активності користувачів та виводить набір створених сеансів:

- Метод №1. Загальна тривалість сеансу не може перевищувати поріг, який дорівнює $\theta$. Враховуючи що $t_0$, мітка часу для першого запиту в створеному сеансі, запит з міткою часу t присвоюється цьому сеансу, якщо тільки $t - t_0 \leq \theta$.
- Метод №2. Загальний час, проведений на сторінці, не може перевищувати поріг, який дорівнює $\delta$. Враховуючи що $t_1$, мітка часу для запиту, призначеного створеному сеансу, наступний запит із міткою часу $t_2$ присвоюється даному сеансу, якщо і тільки $t_2 - t_1 \leq \delta$.

- Метод №3. Запит додається до створеного сеансу, якщо реферер для цього запиту було раніше викликано в цьому сеансі. В іншому випадку запит використовується як початок нової сесії. Слід зауважити, що при використанні цього методу можливо, що запит потенційно може належати більш ніж одному «відкритому» створеному сеансу, запит міг бути здійснений раніше під час кількох сеансів. У цьому випадку додаткова інформація може бути використана для тлумачення.

Приклад застосування методів виявлення сесії наведено в Рис.2.5 і Рис.2.6. На рис.2.5 Метод №1, описаний вище, $\theta = 30$ хвилин було використано для розділення запису активності користувача (з прикладу на рис.2.4) на два окремих сеансу.

| | | Час | IP | URL | Referer |
|---|---|---|---|---|---|
| User1 | Сеанс1 | 0:01:00 | 1.2.3.4 | A | - |
| | | 0:09:00 | 1.2.3.4 | B | A |
| | | 0:19:00 | 1.2.3.4 | C | A |
| | | 0:25:00 | 1.2.3.4 | E | C |
| | Сеанс2 | 1:15:00 | 1.2.3.4 | A | - |
| | | 1:26:00 | 1.2.3.4 | F | C |
| | | 1:30:00 | 1.2.3.4 | B | A |
| | | 1:36:00 | 1.2.3.4 | D | B |

Рис.2.5. Метод №1

| | | Час | IP | URL | Referer |
|---|---|---|---|---|---|
| User1 | Сеанс1 | 0:01:00 | 1.2.3.4 | A | - |
| | | 0:09:00 | 1.2.3.4 | B | A |
| | | 0:19:00 | 1.2.3.4 | C | A |
| | | 0:25:00 | 1.2.3.4 | E | C |
| | | 1:26:00 | 1.2.3.4 | F | C |
| | Сеанс2 | 1:15:00 | 1.2.3.4 | A | - |
| | | 1:30:00 | 1.2.3.4 | B | A |
| | | 1:36:00 | 1.2.3.4 | D | B |

Рис.2.6. Метод №3

Якщо застосувати метод №2 з порогом 10 хвилин, запис користувача буде розглядатися як три сеанси, а саме: A→B→C→E, A і F→B→D. З іншого боку, на рис. 7 зображено приклад використання методу №2 на той самий запис

активності користувача. У цьому випадку один раз запит на F (з часом штампа 1:26:00) досягнуто,  і є дві відкриті сесії, а саме A→B→C→E і A. Але F додається до першого, оскільки його посилання C було викликано в сесії 1. Запит на B (з відміткою часу 1:30:00) потенційно може належати до обох відкритих сеансів, оскільки його реферер, A, викликається як у сеансі 1, так і у сеансі 2. У цьому випадку він додається до другої сесії, оскільки вона є останньою відкритою сесією [1].

### 2.4.5 Ідентифікація епізоду

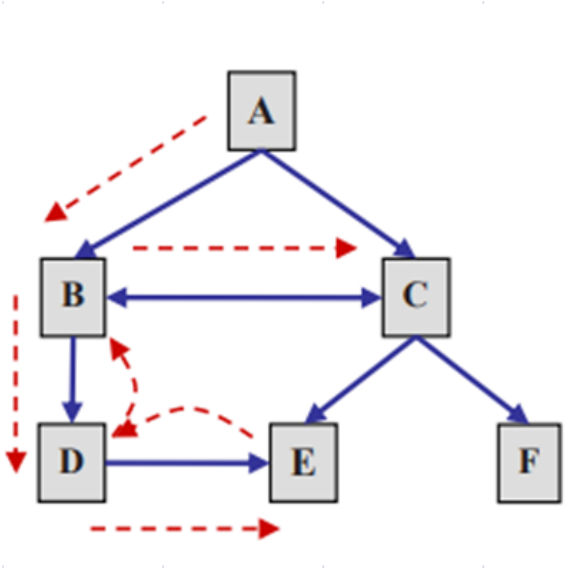
Ідентифікація епізоду може бути виконана як останній крок попередньої обробки даних, щоб зосередитися на відповідних підмножинах переглядів сторінок у кожному сеансі користувача. Епізод — це підмножина або послідовність сеансів, що складається із семантично або функціонально пов’язаних переглядів сторінок. Це завдання може вимагати автоматичної або напівавтоматичної класифікації переглядів сторінок на різні функціональні типи або на класи концепцій відповідно до онтології предметної області або ієрархії концепцій.

### 2.4.6 Завершення шляху

Ще одне потенційно важливе завдання попередньої обробки, яке зазвичай є виконується після ідентифікації сесії, є пошук кінцевої точки. На стороні клієнта або проксі кешування часто може призвести до відсутності посилань на доступ до цих сторінок або до об’єктів, які були кешовані. Наприклад, якщо користувач повертається на сторінку A під час того ж сеансу, то другий доступ до сторінки A, ймовірно, призведе до перегляду попередньо завантаженої версії A, яка була кешована на  стороні клієнта, і тому запити до сервера не надходять. Це призводить до того що друге посилання на A не записується в

журналах сервера. Відсутня посилання через кешування можна евристично вивести через шлях завершення, яке покладається на знання структури сайту та реферера інформація з журналів сервера. У разі динамічно генерованої сторінки, додатки на основі форм, які використовують метод HTTP POST, призводять до того що ввесь або частина параметра не додається до URL-адреси, доступ до якої здійснюється користувачем.

Простий приклад відсутніх посилань наведено на рис.2.7. Також наведено граф, що представляє структуру зв'язків сайту. Стрілки представляють навігаційний шлях, за яким слідує гіпотетичний користувач. Перейшовши на сторінку E, користувач повернувся назад (наприклад, за допомогою кнопки «Назад» у браузері) на сторінку D, а потім B, з якої було здійснено перехід на сторінку C. Зворотні посилання на D і B не відображаються в файлі журналу, оскільки ці сторінки кешуються на стороні клієнта (тому для цих сторінок не було зроблено явний запит до сервера). Файл журналу показує це після запиту на E, наступний запит користувача на сторінку C з реферера B. Іншими словами, у записі діяльності є пробіл відповідно до навігації користувача зі сторінки E на сторінку B. Враховуючи граф сайту можна зробити висновок про два відсутні посилання (тобто E → D і D →B) [1].




**Реальний шлях користувача:**
**A>B>D>E>D>B>C**

| URL | Referer |
|---|---|
| A | - |
| B | A |
| D | B |
| E | D |
| C | B |

Рис.2.7. Структура зв'язків сайту

Враховуючи вище сказане, слід зазначити, що варіантів кінцевої точки, загалом, багато (можливо нескінченно).

## 2.5 Дослідження сесій та користувачів

Статистичний аналіз попередньо оброблених даних сесії становить найбільш поширену форму аналізу. У цьому випадку агрегуються заздалегідь такі дані, як дні, сеанси, відвідувачі або домени. Метод стандартного статистичного аналізу, на основі цих даних, можна використовувати, щоб отримати знання про поведінку користувача.

Такий підхід використовують багато доступних комерційних інструментів для аналізу веб-журналу. Звіти на основі цього типу аналізу можуть включати інформацію про найчастіше відвідувані сторінки, середній час перегляду сторінки, середня довжина шляху на сайті, точку входу і виходу та інші сукупні показники. Незважаючи на недостатню глибину, при цьому типі аналізу, отримані знання можуть бути потенційно корисними для покращення продуктивності системи та забезпечення підтримки рішень. Крім того, все частіше з’являються нові комерційні інструменти веб-аналітики що включають різноманітні алгоритми аналізу даних, що призводить до появи більш складних показників сайту та клієнтів.

Іншою формою аналізу інтегрованих даних про використання є Online Analytical Processing (OLAP). OLAP забезпечує більш інтегровану структуру для аналізу з вищим ступенем гнучкості. Джерелом даних для аналізу OLAP є зазвичай багатовимірне сховище даних, яке об'єднує дані про використання, вміст, і дані електронної комерції на різних рівнях агрегації для кожного параметра. Інструменти OLAP дозволяють змінювати рівні агрегації по кожному виміру протягом аналізу. Масштаби аналізу в такій структурі можуть базуватися на різних полях, доступних у файлах журналів, і можуть включати тривалість часу, домен, запитуваний ресурс, агент користувача та реферери. Це

дозволяє проводити дослідження частин журналу, пов’язаних із певним інтервалом часу, або на більш високому рівні абстракції щодо структури шляху URL. Вихідні дані запитів OLAP також можна використовувати як вхідні дані для різноманітних інструментів Data mining або візуалізації даних [1].

## 2.6 Кластерний аналіз та сегментація

Кластеризація — це метод аналізу даних, який об’єднує елементи, що мають схожі характеристики у групи. У сфері використання веб ресурсів, існує два види цікавих кластерів, які можна виявити: кластери користувачів і кластери сторінок.

Кластеризація записів користувачів (сеансів або транзакцій) є однією з найбільш частою задачею аналізу в Web usage mining та веб-аналітиці. Кластеризація користувачів, як правило, створює групи, які демонструють подібні моделі перегляду. Такі знання особливо корисні для визначення демографічних даних користувачів, щоб наприклад виконати сегментацію ринку в програмах електронної комерції або надати персоналізований веб-контент користувачам зі схожими інтересами. Подальший аналіз груп користувачів на основі їхніх демографічних атрибутів (наприклад, вік, стать, рівень доходу тощо) може привести до відкриття цінної інформації.

Відображаючи транзакцій користувача у багатовимірний простір, як вектори переглядів сторінок, стандартні алгоритми кластеризації, наприклад k-середніх, може розділити цей простір на групи транзакцій, які близькі один до одного на основі міри відстані або подібності між векторами. Отримані таким чином кластери транзакцій представляють сегменти відвідувачів на основі їхньої навігаційної поведінки або інших атрибутів, які були записані у файлі транзакції. Однак кластери транзакцій самі по собі не є ефективним засобом виявлення типових шаблонів поведінки користувачів. Кожен кластер транзакцій, потенційно може містити тисячі транзакцій користувачів із сотнями

посилань на перегляди сторінок. Кінцева мета в кластеризації транзакцій користувачів – це надати можливість для аналізу кожного сегмента [1].

## 2.7 Використання правил асоціації

Використовуючи правила асоціації та статистичний аналіз кореляції можна виявити групи елементів або сторінок, до яких зазвичай звертаються разом. Це, у свою чергу, дає можливість веб-сайтам краще організувати вміст сайту, або надавати ефективні рекомендації.

Найбільш поширені підходи до виявлення асоціацій засновані на апріорному алгоритмі. Цей алгоритм знаходить групи елементів (перегляди сторінок відображаються в попередньо обробленому журналі), які відбуваються часто разом у багатьох транзакціях (тобто, задовольняючи заданий дослідником мінімум поріг підтримки). Такі групи називають частими наборами предметів. Правила асоціації, які задовольняють мінімальний поріг довіри потім генеруються з частих наборів елементів.

Виявлення правил асоціації із даних веб-транзакцій надає багато переваг. Наприклад, правилами також можна скористатися щоб оптимізувати структуру сайту. Якщо сайт не надає прямий зв'язок між двома сторінками A і B, відкриття правила, $A \rightarrow B$, вказує на те, що надання прямого гіперпосилання від A до B може допомогти користувачеві у пошуку необхідної інформації [1].

## 2.8 Використання методу класифікації

Класифікація — це процес відношення елемента даних в один із кількох попередньо визначених класів. Можна працювати з даними про користувачів або сторінками, які належать до певного класу чи категорії. Це вимагає виділення та вибору ознак, які найкраще описують властивості даного класу чи категорії. Класифікацію можна здійснити шляхом використання алгоритмів

навчання під наглядом, таких як дерева рішень, наївні Байєсівські класифікатори, класифікатори k-найближчих сусідів і опорний вектор. Також можливе використання раніше виявлених кластерів та правил асоціації для класифікації нових користувачів.

Методи класифікації відіграють важливу роль у програмах веб-аналітики для моделювання поведінки користувачів відповідно до різних попередньо визначених метрик [1].

## Розділ 3. Дослідження даних з журналу сайту кафедри ІС

Для виконання практичних задач в даному розділі буде використано такі рішення:

1. Мова програмування Python, середовище PyCharm Community;
2. MS Excel 2007
3. Analog 6.0
4. MS Analysis Services

### 3.1 Злиття даних

І так для дослідження мені дано п’ять окремих файлів. Це означає що дані за одну добу записуються в один файл журналу.

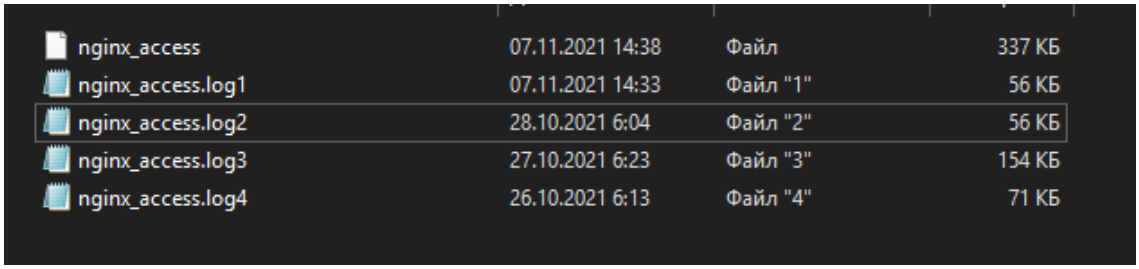



Рис. 3.1. Журнали для дослідження

На Рис. 3.1 показано журнали дій відразу після завантаження з веб-серверу. Їх можна дослідити кожен окремо, але результат буде менш точним. Тому ці файли слід об’єднати в один, при цьому не можна допустити зміну даних в записах та порядок їх розміщення. Можливо це було б хоч трохи складне завдання для роботи з журналами з дуже великого сайту, де одночасно,

паралельно ведеться декілька журналів, які потрібно синхронізувати. Але тут записи додаються послідовно, тому можна просто створити текстовий файл і додавати дані з кожного журналу (по одному журналу), в правильній послідовності. Взагалі такі прості, але об'ємні та "рутинні", дії з текстовими файлами можна автоматизувати. Для цього я буду використовувати мову Python. Код програм наведено в додатку 2. І в результаті у мене має з'явитися один файл що містить дані з п'яти журналів.

Навіть такі "сирі" дані можна спробувати дослідити, хоча б для того щоб побачити всю важливість процесу очищення та взагалі етапу попередньої обробки. Щоб не витрачати багато часу, я скористаюсь уже готовим рішенням Analog 6.0.

Analog це безкоштовний програмний засіб для аналізу журналів веб-серверів. Це досить старий засіб, перша версія вийшла ще в 1995 році, проте останнє оновлення було в середині 2021 року. Використовувати цей засіб можна з Windows, macOS, Linux, та багатьох Unix-подібних систем. На даний момент можливе, в деякій мірі, автоматичне використання. Тобто дані кожної доби аналізуються, потрібно лише увімкнути режим представлення log-файлів. Що до використання в ручну, тут все просто але дуже примітивно, на рівні 1995 року. Всі налаштування відбуваються через текстовий файл "analog.txt". Загалом більшість параметрів уже налаштовано і їх можна не змінювати, проте обов'язково слід вказати ім'я та шлях досліджуваного файлу, в іншому випадку буде досліджуватись файл-приклад. В цілому налаштування виглядає так:

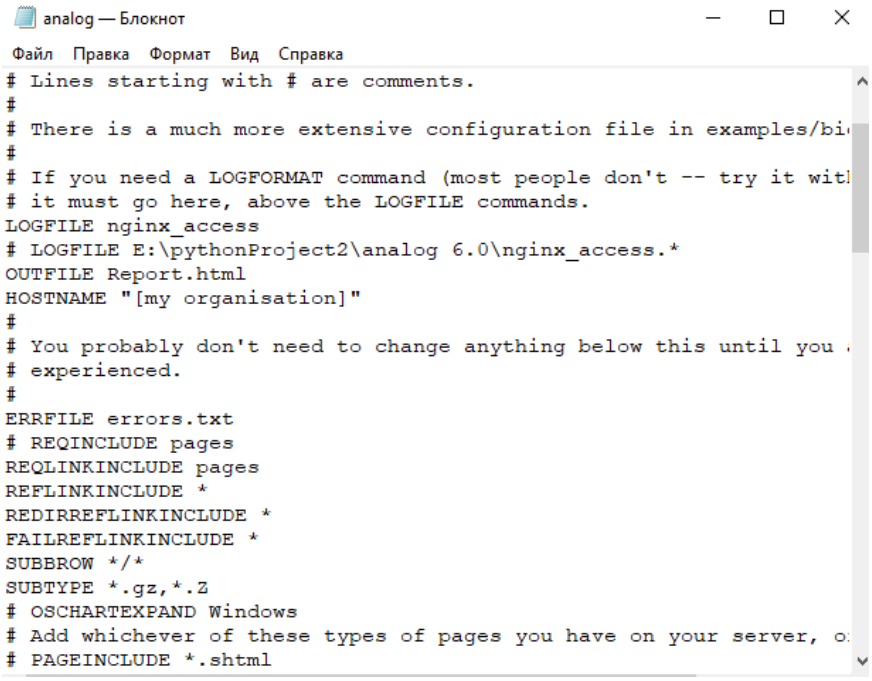



Рис.3.2. Відкритий файл “analog.txt”

Після встановлення шляху до log-файлу, налаштуванні інших параметрів та зберігання файлу “analog.txt”, запускаємо файл “analog.cfg”. В результаті виконання цього файлу автоматично буде побудовано діаграми, які можна подивитись у браузері, відкривши файл “Report.html”.

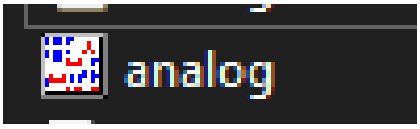



Рис.3.3 Файл “analog.cfg”

В результаті дослідження за допомогою Analog 6.0 я отримав такі корисні результати: (Додаток А). Як можна побачити, деякі діаграми дійсно мають практичну користь, наприклад як:

- Статистика активності по часу та дню тижня
- Статистика ОС користувачів
- Статистика успішних та неуспішних запиті

Більшість діаграм хоч і надають якусь інформацію але вона не несе практичної користі і іноді людському оку не зовсім зрозуміло що взагалі аналізується. Це відбувається тому що з точки зору будь якої системи для обробки, дані не розрізняються на важливі і неважливі, так зване сміття.

Взагалі саме  цьому випадку, для журналів сайту кафедри ІС, очищення даних важливо тільки для того, щоб результати дослідження були точнішими. У випадку, якщо це були б журнали якогось великого або дуже популярного сайту, або цього ж сайту кафедри але за значно більший проміжок часу, очищення ще б значно зменшило об’єм досліджуваного файлу.

Саме тому на етапі попередньої обробки слід виконувати ще й очищення.

## 3.2 Очищення даних

В результаті дій що були проведені в п.3.1 в мене з’явився файл, технічно придатний для дослідження, але практичної користі в цьому поки що мало. В залежності від кінцевої мети процес очистки даних, можливо, треба буде виконати декілька разів, але по різному. Наприклад: для дослідження сторінок що цікавлять живу аудиторію, слід видалити всі записи що містять інформацію про ботів, павуків і т. п..

Очищення деяких параметрів, для сайту кафедри ІС, буде актуально для будь-якої фінальної мети. В полях “Ідентифікація” та “Авторизація” в усіх записах журналу стоїть “-”, це означає що дані поля можна видалити. Або поле протоколу передачі даних, воно не використовується в жодному аналізі і його також можна видалити. Що до інших полів та параметрів, видалення записів з ними буде проводитись окремо та буде залежать від конкретної мети [3].

Щоб ще раз переконатись у важливості даного процесу, перед початком очищення зафіксуємо що в журнал містить 1703 записи и має об’єм 390 Кб. Тепер я видалю всі записи що створені не людьми. Зазвичай записи ботів, пошукових агентів, роботів відрізняються полем User-Agent. Для їх очистки до достатньо видалити записи що містять в полі User-Agent такі слова як “bot”, “robot”, “spider”, “bit”. Тепер файл містить 1073 записи і має об’єм 244 Кб.

Таким чином було відсіяно 37% записів. Можна продовжити очистку видаливши записи що також не несуть корисної інформації. До таких можна віднести скрипти, стилі, шрифти, зображення. Після очистки файл містить 561 запис і має об’єм 124 Кб. Якщо після цих дій проаналізувати отриманий файл за

допомогою Analog 6.0, можна побачити як змінились діаграми. В цілому вони стали більш зрозумілими.

## 3.3 Ідентифікація користувача та сесії

Для дослідження поведінки користувача цього користувача очевидно слід ідентифікувати. Під ідентифікацією розуміється те, що я, як дослідник, маю відрізняти записи що належать одному користувачеві від записів іншого користувача.

Як і в більшості сайтів в інтернеті, на сайті кафедри ІС відсутня ідентифікація користувачів. Це можна зрозуміти, поглянувши на журнал, а саме на поле №2, де у всіх записах стоїть “-”. В розділі №1 було розглянуто декілька методів для ідентифікації користувача в таких випадках. Мною було обрано два методи:

1. Метод ідентифікації користувача по одній лише IP-адресі
2. Метод ідентифікації користувача по User-Agent та IP-адресі разом.

Для початку слід додати в журнал нове поле Time Stamp. Значення в ньому, в деякій мірі, відповідають часу. Тільки от час там лише в секундах і від якоїсь початкової дати. Ця дата може бути будь яка, на вибір дослідника. Я обрав "25/10/2021 00:00:00", так як ця дата найближча до початкової дати в журналі, і мені не доведеться працювати з дуже великими числами. Для цього різницю днів, дня запису і початкової дати, треба помножити на 86400 і ще додати кількість секунд що залишились [3].

Для зручності я перейду в MS Excel 2007. Я просто імпортую текстові дані з журналу. Тепер треба відсортувати записи спочатку по IP-адресі а потім по Time Stamp (стовпчик D). Результат на рис.3.4

| D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|
| 85073 | 1.55.205.101 | - | - | 25.10.2021 | 23:37:53 | GET | /form/ |
| 85074 | 1.55.205.101 | - | - | 25.10.2021 | 23:37:54 | POST | /form/ |
| 111260 | 103.246.144.45 | - | - | 26.10.2021 | 6:54:20 | GET | /form/ |
| 111261 | 103.246.144.45 | - | - | 26.10.2021 | 6:54:21 | POST | /form/ |
| 63917 | 103.250.166.12 | - | - | 25.10.2021 | 17:45:17 | GET | /form/ |
| 63935 | 103.250.166.12 | - | - | 25.10.2021 | 17:45:35 | GET | /form/ |
| 63943 | 103.250.166.12 | - | - | 25.10.2021 | 17:45:43 | GET | /pages/ |

Рис.3.4. Імпортовані та вже відсортовані дані

Слід сказати що, стандартно, MS Excel не правильно сортує IP-адреси. Сортування відбувається як сортування звичайного тексту, але для задачі ідентифікації користувача, такого цілком достатньо.

Приступлю до реалізації методів. При реалізації обох методів першому запису присвоюється Користувач №1, а далі реалізую методи через формули (представлені формули використовую по усьому стовпцю B і C):

- Метод №1 =ЕСЛИ(E2<>E1;МАКС(C$1:C1)+1;"")
- Метод №2 =ЕСЛИ(ИЛИ(E2<>E1; P2<>P1);МАКС(B$1:B1)+1;"")

Після цього можна спостерігати що в деяких місцях методи не збігаються (рис 3.5)

| | | | |
|---|---|---|---|
| | | 63943 | 103.250.166.12 |
| 4 | 4 | 55546 | 109.191.131.206 |
| | | 55547 | 109.191.131.206 |
| 5 | | 252336 | 109.191.131.206 |
| | | 252337 | 109.191.131.206 |
| 6 | 5 | 25290 | 109.248.148.245 |

Рис 3.5. Розбіжності в результатах

Хоча IP-адреса для користувача №4 і №5(за методом User-Agent та IP-адреси) одна, скоріш за все це різні користувачі. Окрім User-Agent в них ще значно відрізняється час входу. Слідуючи з опису методу у Розділі №1, слід вважати що це два різні користувачі.

В результаті (рис 3.6.) було виявлено що методом User-Agent та IP-адреси було виявлено більше користувачів, і сам спосіб вважається більш точним.

| | | | |
|---|---|---|---|
| 560 | | 197 | 169 |
| 561 | | 198 | 170 |

Рис 3.6. Результат порівняння роботи методів виявлення користувачів

В багатьох джерелах [3] трапляється інформація що ідентифікацію сесій, в деяких випадках, можна не проводити. Це рішення залежить кількості даних, отриманих з журналів. Тобто якщо даних багато – ідентифікацію сесій необхідно проводити, даних мало – можна не проводити. У випадку, коли окремі сесії не виділяють, прийнято вважати що: 1 користувач = 1 сесія. Але все ж таки, в такому випадку рішення залишається за дослідником.

Як можна було побачити на сайті за 5 днів було 198 користувачів, що недуже багато і по дням і по користувачам. З цього слідує що в даному випадку проведення ідентифікації сесії необов’язкове.

Для деяких наступних процесів, із-за специфіки сайту, мені знадобляться дані саме про окремі сесії. Тому я приступлю до їх виявлення.

Насправді це не складний процес, хоча сама точність виявлення повністю за лежить від рішень дослідника. Суть методу, що я буду використовувати, полягає в розділенні записів активності одного користувача на сесії основуючись на полі Time Stamp, створеному ще на етапі попередньої обробки. Оскільки один користувач міг декілька разів виходити та заходити на сайт на протязі дня, або тижня (або більшого проміжку часу), очевидно що це різні сесії. Із-за різниці в полях дата та час ці сесії можна виявити. Проте поле Time Stamp це теж час, але у більш зручному форматі, тому саме його я і буду використовувати. Суб’єктивність полягає в такому питанні: скільки часу, між послідовними діями користувача, маже пройти, щоб це можна було вважати новим, окремим візитом на сайт?

Порівняю як такий вибір вплине на результат ідентифікації. Вважаю що перерва між сесіями може бути 1/2/6/24 годин або наприклад 5 днів. Результати занесу в таблицю 2.

Табл. 2. Кількість окремих сесій

| | Користувачів | Окремих сесій |
|---|---|---|
| 3600 секунд | 198 | 216 |
| 7200 секунд | | 215 |
| 21600 секунд | | 212 |
| 86400 секунд | | 206 |
| 432000 секунд | | 198 |

Очевидно що із збільшенням часу, кількість ідентифікованих сесій буде зменшуватись. Але ця кількість не може бути меншою за число ідентифікованих користувачів. Знаючи тематику сайту, журнал якого досліджую, я буду вважати що число окремих сесій саме 216.

Що до самого методу ідентифікації цих сесій, його я реалізував за допомогою Excel за формулою: =ЕСЛИ(И(A2=A1; D2-D1<3600);B1;B1+1)

### 3.3.1 Точка входу

Після того, як користувачів ідентифіковано, можна приступити до вирішення ще однієї задачі, результат якої може бути неочевидним. Це пошук початкової сторінки через яку користувач потрапляє на сайт. Окрім того що це є обов’язковим для побудови моделі поведінки користувача, отримані результати можуть бути корисними для розуміння проблем сайту.

І так після, проведення попереднього етапу, було виявлено 198 користувачів сайту. Оскільки для виконання попереднього етапу було обрано MS Excel, ID користувача розміщено якраз напроти URL першої сторінки, на яку він потрапив, можна просто підвести статистику. Для більшого розуміння можна побудувати діаграму (рис 3.7).

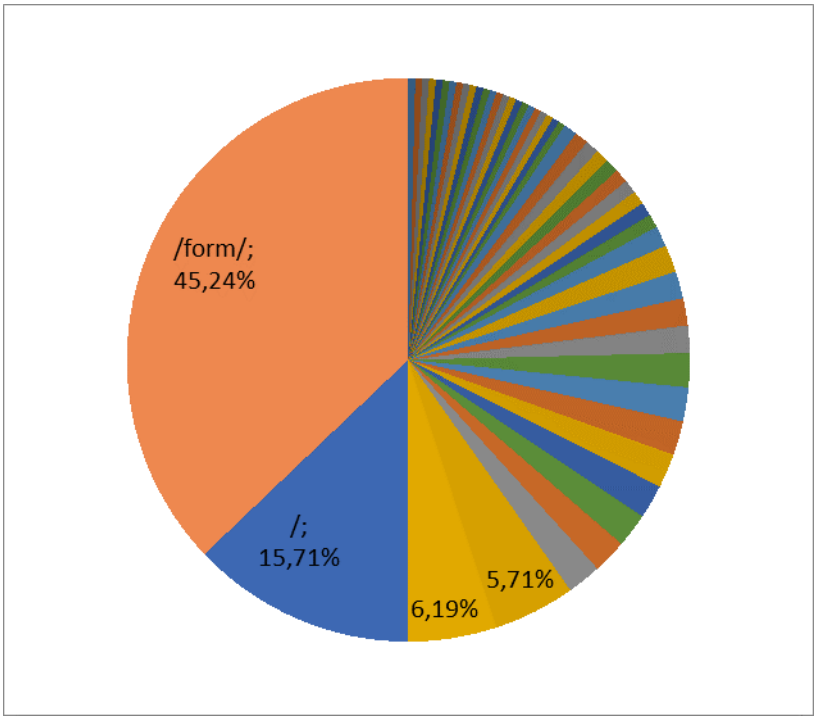



Рис 3.7 Діаграма точки входу на сайт

Таким чином було виявлено що для користувачів початковою сторінкою на сайті є:

1. У 45,24% це - http://is.nuft.edu.ua/form
2. У 15,71% це - http://is.nuft.edu.ua/
3. У 6,19% це - http://is.nuft.edu.ua/pages/view/rozklad
4. У 5,71% це - http://is.nuft.edu.ua/news/view/sparta_camp

Пояснення. 45,24% користувачів потрапляє на сайт починаючи з сторінки з анкетою для випускників, в той час як головну сторінку використовує лише 15,71%. Для 6,19% сторінкою входу є розклад дзвінків (можливо ці 6,19% першокурсники?). Користувачі, що входять до 5,71%, потрапляли на сайт через

одну новину, що на час написання цієї роботи вже втратила актуальність. Інші сторінки використовуються у менш ніж 2,5% кожна.

Уже тут очевидна проблема сайту (можливо в наповненні, можливо в структурі): головна сторінка не є основною точкою входу на сайт.

## 3.4 Використання правил асоціації

Серед трьох напрямків Web Mining тільки у процесі Web Usage Mining можна використати правила асоціації. Оскільки за допомогою Web Usage Mining досліджують поведінку користувача на сайті, використання правил асоціації дозволяє з якоюсь вірогідністю, показати його частковий маршрут. Це дасть змогу використати отриману інформацію для практичних цілей а саме для пошуку недоліків у структурі, у наповнені сайту, та інших.

Серед багатьох програмних засобів для вирішення цієї задачі, я обрав MS Excel 2007 з надбудовою для проведення інтелектуального аналізу даних (для її роботи потрібен MS Analysis Services).

Оскільки дані для дослідження уже в таблицях Excel, і вже було проведено деяку попередню обробку в попередніх пунктах, то можна приступити до виявлення правил асоціації.

Даний процес буде відтворено двічі. Перший раз у якості ідентифікатора транзакції буде ID користувача, другий раз - ID сесії. У якості елементу транзакції, в обох випадках буде URL сторінки. Результати можна порівняти і зробити висновок що до доцільності етапу ідентифікації сесій.

### 3.4.1 Ідентифікатор транзакції – ID користувача

В результаті обробки за допомогою методу інтелектуального аналізу даних, а саме пошуку асоціативних правил, було виявлено такі залежності (рис 3.8) і такі правила (рис 3.9), а також підтримку (рис 3.10).

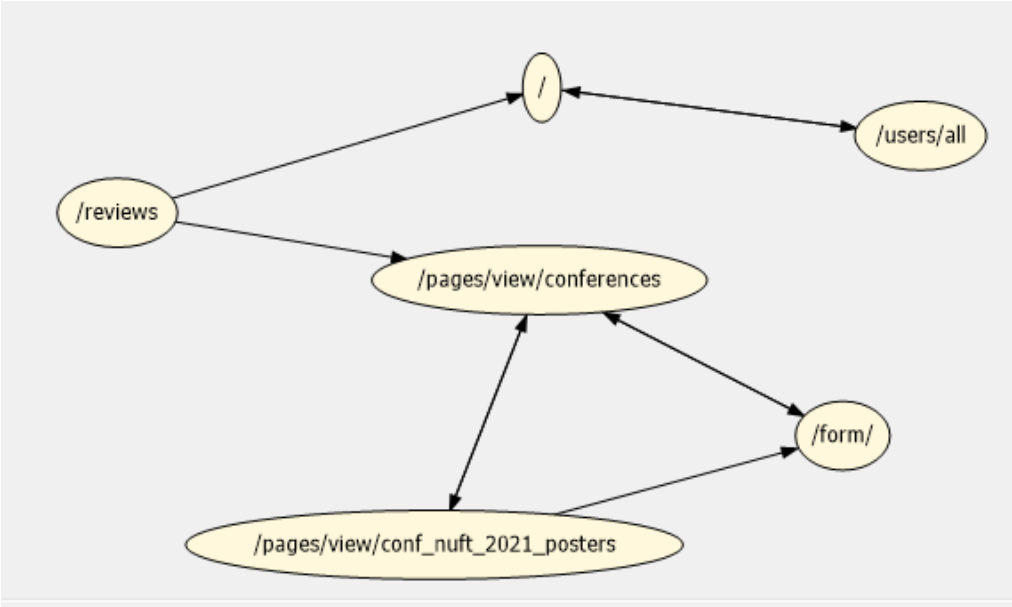



Рис. 3.8 Залежності

| Ве... | Важность | Правило |
|---|---|---|
| 1,000 | 0,880 | /pages/view/conf_nuft_2021_posters -> /pages/view/conferences |
| 1,000 | 0,827 | /pages/view/conf_nuft_2021_posters, /form/ -> /pages/view/conferences |
| 0,833 | 0,307 | /pages/view/conferences -> /form/ |
| 0,786 | 0,215 | /pages/view/conf_nuft_2021_posters -> /form/ |
| 0,786 | 0,215 | /pages/view/conf_nuft_2021_posters, /pages/view/conferences -> /form/ |
| 0,714 | 0,591 | /reviews -> / |
| 0,714 | 0,604 | /reviews -> /pages/view/conferences |
| 0,615 | 0,573 | /users/all -> / |
| 0,389 | 1,811 | /pages/view/conferences -> /pages/view/conf_nuft_2021_posters |
| 0,367 | 1,202 | /pages/view/conferences, /form/ -> /pages/view/conf_nuft_2021_posters |
| 0,316 | 0,681 | /form/ -> /pages/view/conferences |
| 0,216 | 0,797 | / -> /users/all |

Рис. 3.9 Правила

| Подде... | Разм... | Набор элементов |
|---|---|---|
| 95 | 1 | /form/ |
| 37 | 1 | / |
| 36 | 1 | /pages/view/conferences |
| 30 | 2 | /pages/view/conferences, /form/ |
| 14 | 1 | /pages/view/rozklad |
| 14 | 1 | /pages/view/conf_nuft_2021_posters |
| 14 | 2 | /pages/view/conf_nuft_2021_posters, /pages/vie... |
| 13 | 1 | /news/view/sparta_camp |
| 13 | 1 | /users/all |
| 11 | 3 | /pages/view/conf_nuft_2021_posters, /pages/vie... |
| 11 | 2 | /pages/view/conf_nuft_2021_posters, /form/ |
| 10 | 1 | /users/login |
| 8 | 2 | /users/all, / |
| 7 | 1 | /form |
| 7 | 1 | /reviews |
| 5 | 2 | /reviews, / |
| 5 | 2 | /reviews, /pages/view/conferences |

Рис. 3.10 Підтримка

Виходячи з отриманих правил уже можна зробити деякі висновки про поведінку користувача та певну закономірність у порядку перегляду сторінок.

Розберу декілька правил. Наприклад із першого правила слідує що, якщо користувач перейшов на сторінку конференцій то також він перегляне з постери до цієї конференції. Насправді, хоча закономірність вірна, слід зробити поправку на те що матеріал для дослідження актуальний станом на осінь 2021. Тобто вірним висновком буде: якщо користувач відвідав сторінку конференцій то скоріш за все він передивиться матеріли до найближчої конференції.

Із восьмого та останнього правила слідує що перегляди головної сторінки і сторінки з викладачами кафедри взаємопов’язані.

### 3.4.2 Ідентифікатор транзакції – ID сесії

В результаті точно таких же дій, як і в попередньому підпункті було виявлено такі залежності (рис 3.11) і такі правила (рис 3.12), а також підтримку (рис 3.13).

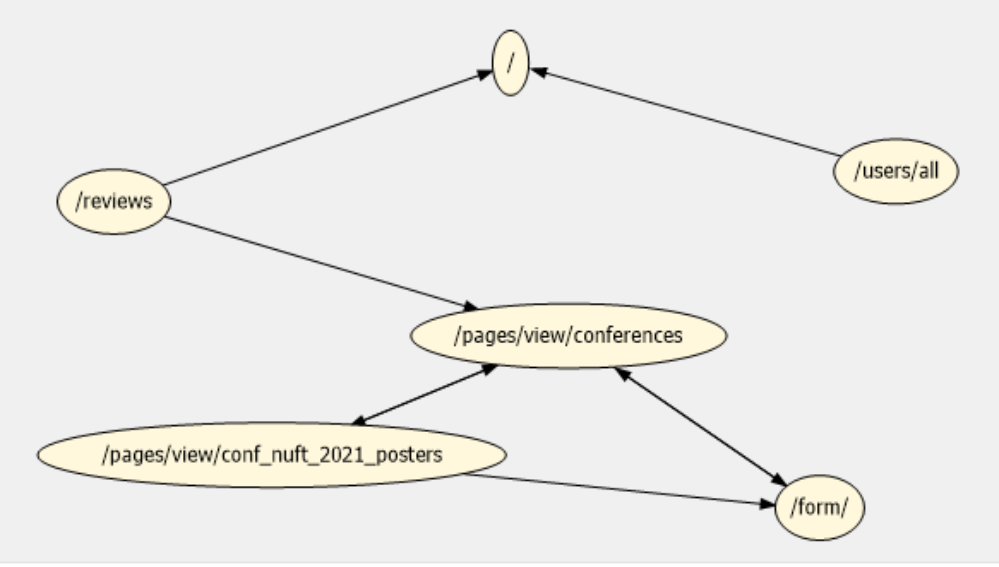



Рис. 3.11 Залежності

| Веро... | Важность | Правило |
|---|---|---|
| 1,000 | 0,920 | /pages/view/conf_nuft_2021_posters -> /pages/view/conferences |
| 1,000 | 0,866 | /pages/view/conf_nuft_2021_posters, /form/ -> /pages/view/conferences |
| 0,833 | 0,333 | /pages/view/conferences -> /form/ |
| 0,786 | 0,240 | /pages/view/conf_nuft_2021_posters -> /form/ |
| 0,786 | 0,240 | /pages/view/conf_nuft_2021_posters, /pages/view/conferences -> /form/ |
| 0,714 | 0,580 | /reviews -> / |
| 0,714 | 0,643 | /reviews -> /pages/view/conferences |
| 0,571 | 0,528 | /users/all -> / |
| 0,389 | 1,856 | /pages/view/conferences -> /pages/view/conf_nuft_2021_posters |
| 0,367 | 1,246 | /pages/view/conferences, /form/ -> /pages/view/conf_nuft_2021_posters |
| 0,306 | 0,725 | /form/ -> /pages/view/conferences |

Рис. 3.12 Правила

| Поддержка | Разм... | Набор элементов |
|---|---|---|
| 98 | 1 | /form/ |
| 41 | 1 | / |
| 36 | 1 | /pages/view/conferences |
| 30 | 2 | /pages/view/conferences, /form/ |
| 15 | 1 | /pages/view/rozklad |
| 14 | 1 | /users/all |
| 14 | 1 | /pages/view/conf_nuft_2021_posters |
| 14 | 2 | /pages/view/conf_nuft_2021_posters, /pages/vie... |
| 13 | 1 | /news/view/sparta_camp |
| 11 | 1 | /users/login |
| 11 | 3 | /pages/view/conf_nuft_2021_posters, /pages/vie... |
| 11 | 2 | /pages/view/conf_nuft_2021_posters, /form/ |
| 8 | 2 | /users/all, / |
| 7 | 1 | /form |
| 7 | 1 | /reviews |
| 6 | 1 | /users/view/samsonov |
| 5 | 2 | /reviews, / |
| 5 | 2 | /reviews, /pages/view/conferences |

Рис. 3.13 Підтримка

Можна побачити заміна ID користувача на ID сесії, як і передбачалось, на залежності та правила суттєво не вплинула. Правил дійсно стало менше (11 проти 12), а самі вони – більш важливі. Однак це ніяк не означає що у всіх випадках можна знехтувати ідентифікацією сесії. Це лише підтверджує що при досліджені невеликого об’єму даних, цей етап можна пропустити.

## 3.5 Дослідження методом кластерного аналізу

Для знаходження найбільш популярних груп що відображають використання веб-ресурсів, побудуємо модель використання методом кластеризації. Вхідними даними для створення кластерів буде витрачений на перегляд сторінки час.

Позначимо через J множину переглядів сторінок групи А:

$J = \{j_1, j_2, \dots j_i, \dots, j_n\}$, де $j_i$ – перегляд сторінки.

Необхідно побудувати множину кластерів K та відображення E множини J на множину K, тобто $E: J \rightarrow K$.

Відображення E задає модель даних, що є рішенням задачі. Якість рішення задачі визначається кількістю вірно кваліфікованих даних.

Кожен перегляд сторінки визначається набором своїх характеристик та часом: $j_i = \{z_1, z_{2, \dots,} z_h, \dots, z_m\}$, де $z_m$ – характеристика перегляду.

Кожна змінна $z_h$ може приймати значення з деякої множини значень характеристик перегляду: $z_h = \{\nu_h^1, \nu_h^2, \dots\}$

Задача кластеризації полягає у побудові множини:

$K = \{k_1, k_2, \dots, k_l, \dots, k_g\}$, де $k_l$ – кластер, який містить схожі перегляди сторінок з множини J: $k_l = \{j_i, j_p \mid j_i \in J, j_p \in J$ та $d(j_i, j_p) < \sigma\}$, де $\sigma$ - величина, яка визначає міру близькості для включення об’єктів в один кластер; $d(j_i, j_p)$ – відстань між об’єктами.

Невід'ємне значення $d(j_i, j_p)$ є відстанню між елементами $j_i$ та $j_p$, якщо виконуються наступні умови:

- $d(j_i, j_p) \geq 0$, для всіх $j_i$ та $j_p$;
- $d(j_i, j_p) = 0$, тоді і тільки тоді, коли $j_i = j_p$;
- $d(j_i, j_p) = d(j_i, j_p)$;
- $d(j_i, j_p) \leq d(j_i, j_r) + d(j_r, j_p)$

Якщо відстань $d(j_i, j_p)$ менше деякого значення $\sigma$, то елементи, які характеризують перегляди сторінок, близькі і розміщуються в одному кластері. В іншому випадку елементи відмінні один від одного та розміщуються у різні кластери.

Дослідження структур кластерів надає можливість визначити сукупність значень характеристик перегляду сторінки.

Для проведення кластерного аналізу я обрав такі критерії як: URL сторінки; час, проведений користувачем на сторінці; та поле де визначено чим

являється сторінка для сесії(входом, виходом, проміжним етапом, або одночасно і входом і виходом). В результаті виділилось 11 кластерів (рис.3.14).

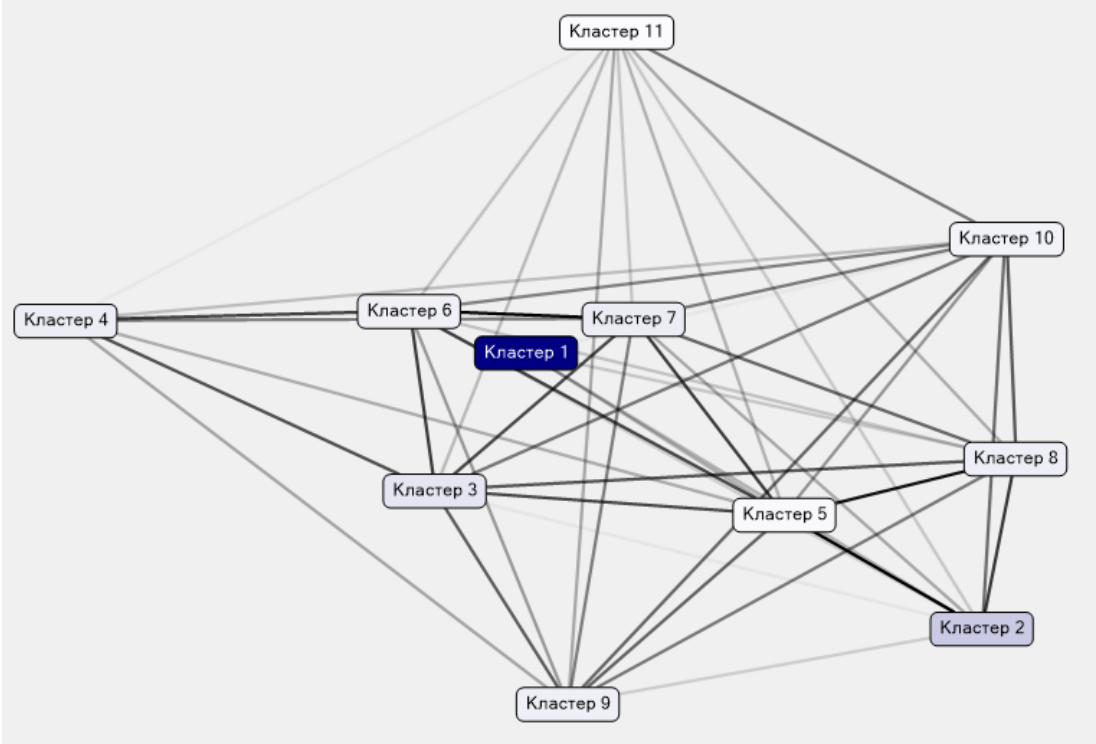



Рис.3.14 Кластери

Найбільш цікавими є кластери з найбільш міцним зв'язком, це кластер №2 та кластер №5. Профілі цих кластерів зображено на рис.3.15

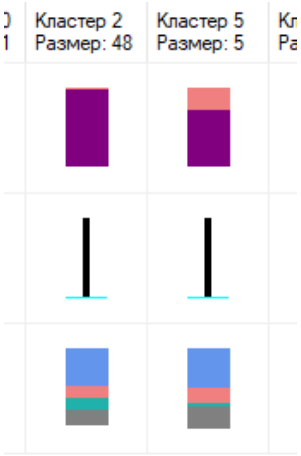



Рис.3.15 Профілі кластерів №2 та №5

Цікавими ці кластери є тому що підтверджують результат пункту 3.3.1 Точка входу.

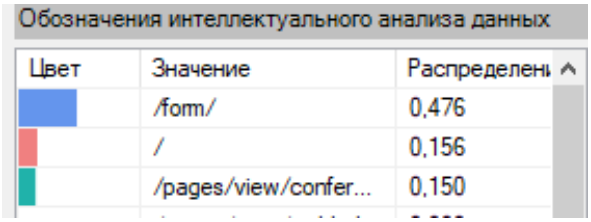



Рис.3.16 Розподілення в кластері №2

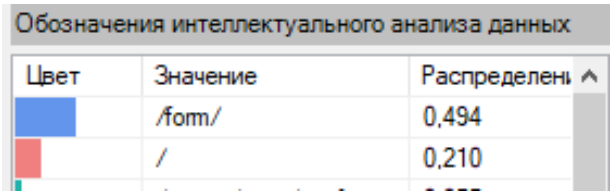



Рис.3.17 Розподілення в кластері №5

Розподілення в цих кластерах майже співпадає із результатом отриманим статистичним методом. Але якщо результат статистичного методу досить загальний і охоплює всі сеанси без виключень, то в кластерах №2 та №5 об’єднані лише ті сеанси що містять біль як один перегляд сторінки. При чому час перебування на сторінці входу становить 1-1,5 секунд. Тобто очевидно що для користувача це не та сторінка, яку він шукав.

Якщо з цього зробити практично  корисний висновок то він виглядатиме так: Навіть користувачі, які потрапили на сайт не випадково, все одно використовують для входу сторінку з анкетою а не головну.

## 3.6 Дослідження методом класифікації

Для використання даного методу слід спочатку визначитись із класами. Із поля User Agent можно визначити платформу з якої було здійснено запит. Нехай класами будуть найбільш популярні платформи: Windows, Android, iPhone, Linux, Mac OS. Вхідними даними буде поле URL адреса сторінки та поле Платформа. В результаті я отримую таке дерево рішень (рис.3.18).

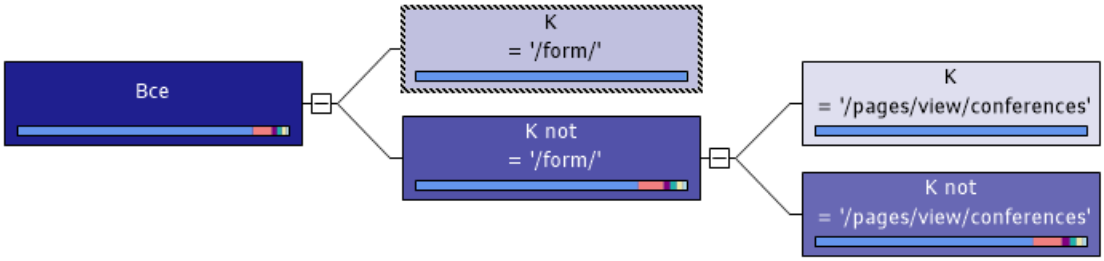



рис.3.18 Дерево рішень.

З цього дерева рішень слідує що майже 99% переглядів сторінки з анкетою (/form/), відбулося з Windows, і ще майже 1% з Linux. Сторінку з інформацією про конференції з Windows переглядають у 100% випадків. Що до інших сторінок - результат на рис.3.19.

| Значение | Вар... | Вероят... | Гистограм... |
|---|---|---|---|
| Android | 27 | 11,15% | |
| iPhone | 6 | 3,24% | |
| Linux | 5 | 2,86% | |
| Mac OS | 5 | 2,86% | |
| Windows | 204 | 77,78% | |
| Інша ОС | 3 | 2,11% | |
| Отсутствует | 0 | 0,00% | |

рис.3.19 Статистика по іншим сторінкам

З даного аналізу слідує що найбільш популярна платформа для використання сайту кафедри ІС – це Windows. А також, що до сторінки анкети, що вона не користується популярністю серед мобільних платформ, можливо що дана сторінка погано оптимізована під смартфони.

## 3.7 Розгляд проблем та рекомендації для їх вирішення

Проблема непопулярності деяких сторінок серед мобільних платформ, на мою думку полягає не у тому що конкретно з цими сторінками щось не так, або з ними не зручно працювати. Проблема криється у навігації сайту. Якщо поглянути на меню сайту використовуючи повноекранний режим комп'ютера, то все здається зручно, хоч і якось мінімалістично. Проте, якщо використовувати сайт зі смартфону, або просто зменшити вікно на комп'ютері,

можна зрозуміти що меню дуже не зручне. І взагалі хочеться як найшвидше покинути сайт.

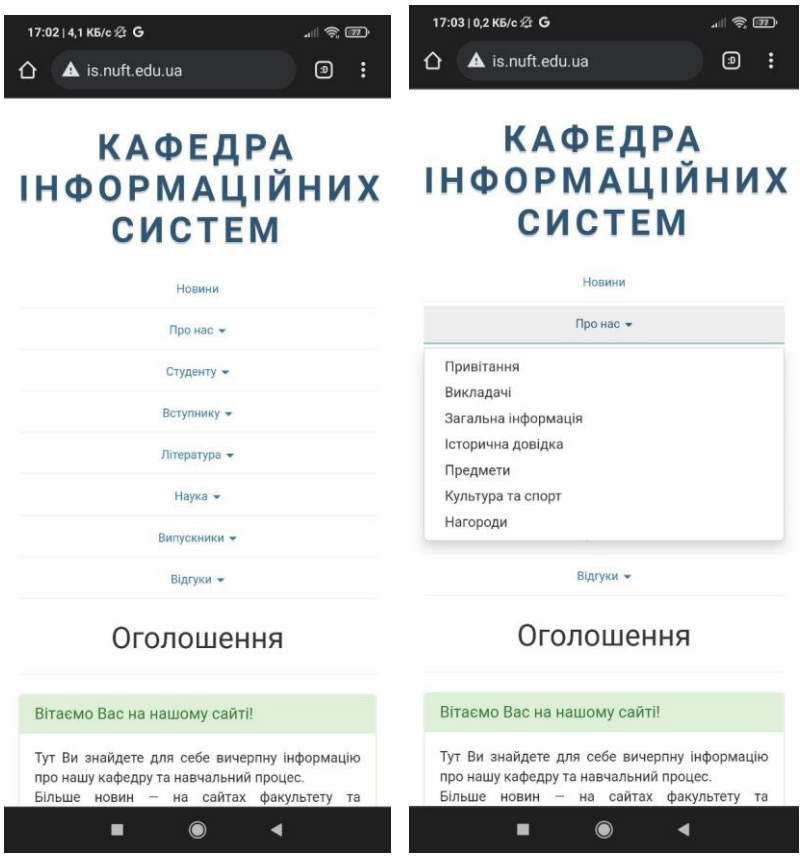



Рис.3.20-3.21 Мобільна версія сайту

По-перше, меню в мобільній версії сайту займає багато місця на екрані, і до того ж його не можна згорнути. При переходах між сторінками сайту користувач кожен раз спочатку бачить повне меню а не інформацію що шукає.

По-друге, пункти меню розкриваються як списки. Це не проблема для повної версії сайту. Але на смартфоні, при розгортанні однин список перекриває все меню. Видно що цьому не було приділено достатньо уваги при розробці.

Повертаючись саме до сторінок анкетування та конференцій, оскільки саме на їх прикладі найбільш помітне небажання користуватись сайтом із смартфону, скоріш за все користувач просто не може знайти ці сторінки (або не бажає шукати) використовуючи це меню. І дійсно, можливість переходу до цих

сторінок знаходиться в останніх пунктах меню що можуть бути перекриті будь яким іншим відкритим пунктом меню.

Скоріш за все даний сайт розроблювався в основному для використання з комп'ютера, і оптимізації під мобільні екрани не було приділено достатньо уваги.

В ідеалі, слід цілком переробити мобільну версію сайту для більш зручного використання на смартфонах. Якщо менш радикально, то хоча б переробити меню. Слід додати можливість його згорнути/розгорнути, виправити перекривання одних пунктів меню розгорнутим іншим пунктом, можливо змінити шрифт та розмір заголовків меню.

Щодо проблеми з точкою входу на сайт. Користувачі що заходять через сторінку з анкетуванням, а таких більшість, скоріш за все переходять за прямим посиланням (через Email, Telegram, Viber тощо). Скоріш за все тут також проблема криється в навігації. Проте, тут проблема більш глобальна і торкається загалом сайту. Оновлення меню може не вистачити. Оскільки головна сторінка зовсім не "головна", я думаю, слід попрацювати над наповненням та структурою саме цієї сторінки.

## 3.8 Висновки до розділу 3

1. Хоча використання статистичних методів на багато простіше з точки зору реалізації, зазвичай отриманої таким способом інформації достатньо лише для виявлення поверхневих проблем.

2. Використання методів з технології Data mining, а саме кластеризації, класифікації та асоціації, дозволяє глибше дослідити дані і виявити більше неочевидних закономірностей.

3. В результаті п.3.5 та п.3.6 я отримав дані близькі до тих що і при використанні статистичних методів. Але це більш практичні дані, а не "суха" статистика. Результати більш точно вказують на слабкі місця, та проблеми сайту.

4. Отже використання методів інтелектуального аналізу (кластеризації, класифікації та асоціації) даних дозволяє провести більш якісне дослідження журналу подій користувача, ніж при використанні статистичних методів.

# Висновки

В процесі виконання роботи було:

- Обґрунтовано необхідність проведення очищення даних для проведення більш точного дослідження.
- Досліджено методи ідентифікації користувача для отримання найбільш точних даних про кількість відвідувачів сайту.
- Досліджено методи ідентифікації сеансу для виявлення найбільш доцільного методу групування переглядів сторінок в набори, що називають сеансами.
- Досліджено використання веб-ресурсів методом асоціації, що показало залежність у використанні деяких сторінок, у рамках одного сеансу.
- Доведено що при невеликому об’ємі даних для дослідження, можна не проводити ідентифікацію сеансів.
- Досліджено використання веб-ресурсів методом кластеризації, що дало змогу підтвердити та поглибити знання про неправильне використання деяких сторінок. Конкретно в якості точки входу, найчастіше використовується не головна сторінка.
- Досліджено використання веб-ресурсів методом класифікації, що показало що, в цілому сайтом мало користуються з мобільних платформ. На деякі важливі сторінки користувачі взагалі не заходять з мобільних платформ.
- На основі отриманих результатів виявлено деякі неочевидні недоліки сайту кафедри інформаційних систем, а саме проблема в

незручній навігації, а також проблема в оптимізації веб-сайту під мобільні платформи.

- Також сформовано рекомендації що до виправлення виявлених проблем.

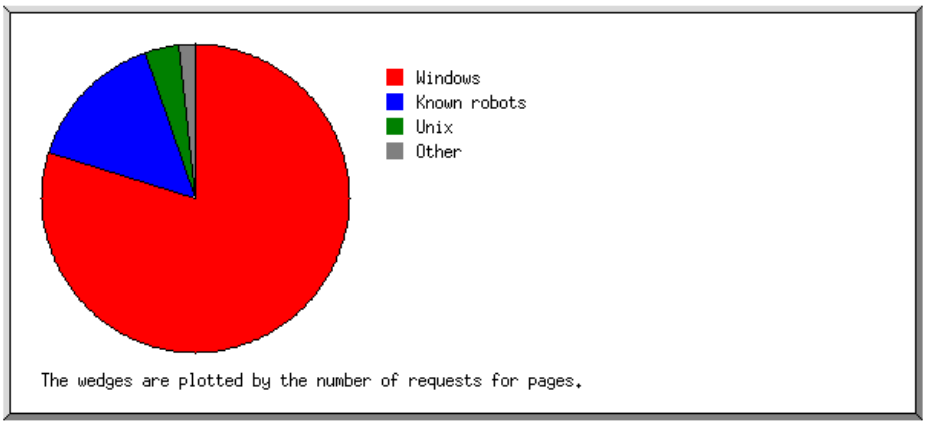
# Список використаних джерел


1. Bing Liu (2011). *Web data mining : exploring hyperlinks, contents, and usage data*. Heidelberg ; New York: Springer.
2. BaseGroup Labs. (n.d.). *Web Mining: анализ использования веб-ресурсов, обработка веб-лога*. [online] Available at: https://basegroup.ru/community/articles/web-usage-mining-part1 [Accessed 10 Nov. 2021].
3. BaseGroup Labs. (n.d.). *Web Mining: анализ использования веб-ресурсов, обработка веб-лога*. [online] Available at: https://basegroup.ru/community/articles/web-usage-mining-part2 [Accessed 10 Nov. 2021].
4. BaseGroup Labs. (n.d.). *Web Mining: основные понятия*. [online] Available at: https://basegroup.ru/community/articles/basic-conceptions [Accessed 01 Nov. 2021].
5. Ключевые метрики веб-аналитики 2021: терминология и различия [онлайн]. (без дати). *Completo*. [Дата звернення 4 лютого 2022]. Режим доступу: https://blog.completo.ru/metriki-2021/
6. Web Mining: интеллектуальный анализ данных в сети Internet | Портал знань, портал знаний, дистанційне навчання [онлайн]. (без дати). *Портал знань — Знання повинні бути доступними | Портал знань, портал знаний, дистанційне навчання*. [Дата звернення 4 лютого 2022]. Режим доступу: http://www.znannya.org/?view=technologies-km-3-1
7. Базь В. Р. Дослідження та застосування технології Web usage mining для аналізу сайту кафедри ІС / Базь В. Р., М'якшио О. М. // Наукові праці Четвертої міжнар. наук.-практ. конф. «Сучасні тенденції розвитку інформаційних систем і телекомунікаційних технологій», 1–2 лютого 2022 р. (Київ, Україна). – К. : НУХТ, 2022. – (у друці).

8. Базь В. Р. Дослідження та застосування технології WebUsageMining для аналізу сайту кафедри ІС / Базь В. Р. // Наукові праці Восьмої міжнар. наук.-техн. Internet-конф. «Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційнотехнічними та технологічними комплексами», 26 листопада 2021 р. (Київ, Україна). – К. : НУХТ, 2022. – 71 с.

# Додаток А

*This report lists the operating systems used by visitors.*

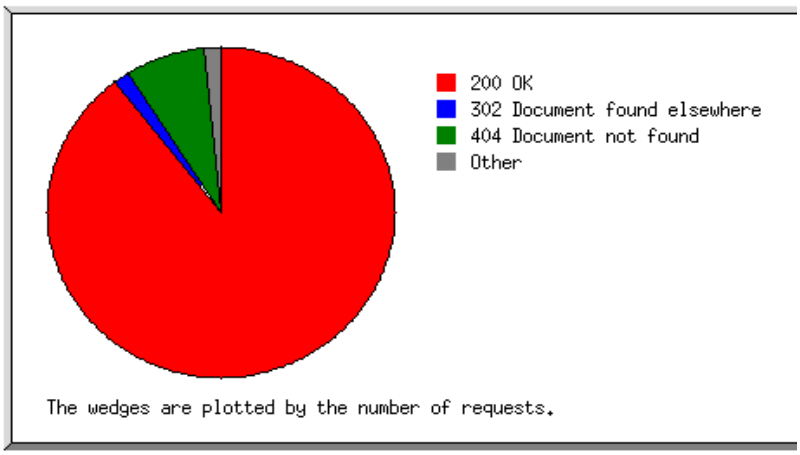



Listing operating systems, sorted by the number of requests for pages.

| no. | reqs | pages | OS |
|---|---|---|---|
| 1 | 760 | 240 | Windows |
| | *279* | *112* | *Unknown Windows* |
| | *449* | *111* | *Windows NT* |
| | *32* | *17* | *Windows XP* |
| 2 | 595 | 45 | Known robots |
| 3 | 117 | 11 | Unix |
| | *117* | *11* | *Linux* |
| 4 | 58 | 3 | Macintosh |
| 5 | 9 | 2 | OS unknown |

Рис А.1 Статистика по ОС користувачів

*This report lists the HTTP status codes of all requests.*




Listing status codes, sorted numerically.

| reqs | status code |
|---|---|
| 1525 | 200 OK |
| 24 | 302 Document found elsewhere |
| 16 | 304 Not modified since last retrieval |
| 127 | 404 Document not found |
| 11 | 4xx [Miscellaneous client/user errors] |

Рис А.2 Статистика по статусу запитів

Each unit (▬) represents 1 request for a page.

| hour | reqs | pages |
|---|---|---|
| 0 | 107 | 22 |
| 1 | 28 | 14 |
| 2 | 228 | 26 |
| 3 | 24 | 4 |
| 4 | 28 | 1 |
| 5 | 13 | 5 |
| 6 | 22 | 10 |
| 7 | 42 | 17 |
| 8 | 22 | 2 |
| 9 | 53 | 12 |
| 10 | 75 | 11 |
| 11 | 71 | 5 |
| 12 | 46 | 10 |
| 13 | 114 | 17 |
| 14 | 86 | 31 |
| 15 | 59 | 18 |
| 16 | 57 | 22 |
| 17 | 77 | 20 |
| 18 | 48 | 8 |
| 19 | 53 | 3 |
| 20 | 95 | 13 |
| 21 | 42 | 9 |
| 22 | 123 | 16 |
| 23 | 28 | 7 |

Рис А.3 Статистика запитів по часу доби

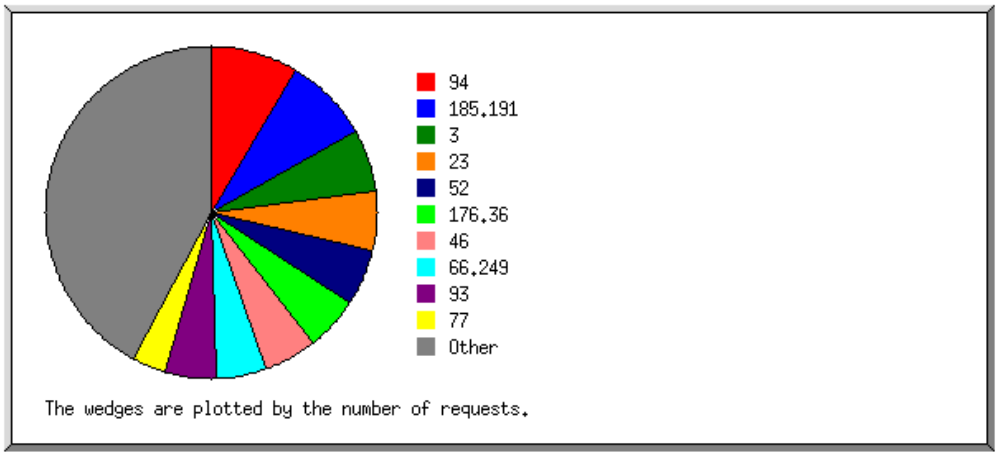
*This report lists the total activity for each day of the week, summed over all the weeks in the report.*

Each unit (▬) represents 2 requests for pages or part thereof.

| day | reqs | pages |
|---|---|---|
| Sun | 0 | 0 |
| Mon | 250 | 46 |
| Tue | 266 | 57 |
| Wed | 535 | 82 |
| Thu | 232 | 62 |
| Fri | 258 | 56 |
| Sat | 0 | 0 |

Рис А.4 Статистика запитів по дню тижня (Ця статистика була б корисна при наявності даних за більший проміжок часу)

*This report lists the organisations of the computers which requested files.*

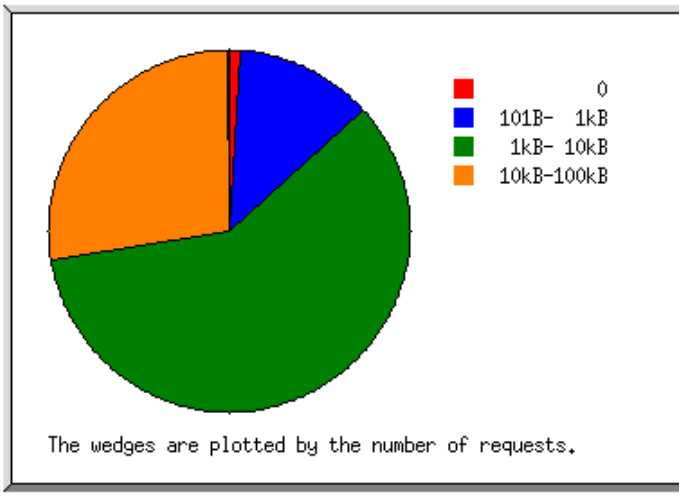



Listing the top 20 organisations by the number of requests, sorted by the number of requests.

| reqs | %bytes | organisation |
|---|---|---|
| 132 | 6.30% | 94 |
| 129 | 3.68% | 185.191 |
| 93 | 2.70% | 3 |
| 89 | 3.08% | 23 |
| 85 | 2.54% | 52 |
| 80 | 5.21% | 176.36 |
| 78 | 6.52% | 46 |
| 77 | 2.65% | 66.249 |
| 76 | 6.38% | 93 |
| 51 | 4.37% | 77 |
| 50 | 5.16% | 91 |
| 40 | 3.67% | 109 |
| 31 | 2.04% | 95 |

Рис А.5 Статистика по IP-адресах запитів (Приклад малокорисного результату)

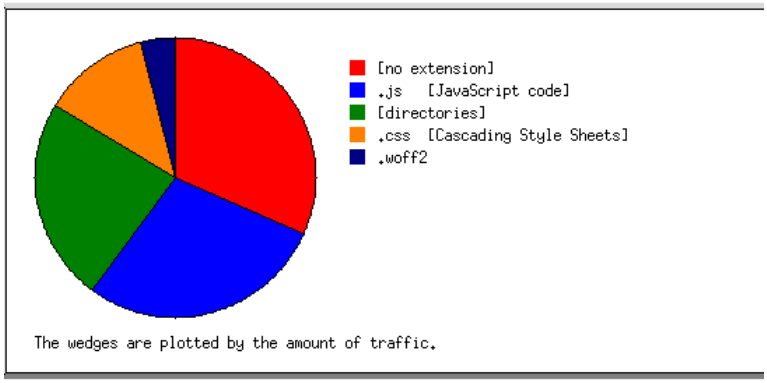
*This report lists the sizes of files.*




| | size | reqs | %bytes |
|---|---|---|---|
| | 0 | 17 | |
| 1B- | 10B | 0 | |
| 11B- | 100B | 0 | |
| 101B- | 1kB | 189 | 0.98% |
| 1kB- | 10kB | 910 | 31.59% |
| 10kB- | 100kB | 423 | 65.18% |
| 100kB- | 1MB | 2 | 2.25% |

Рис А.6 Статистика по об’єму запитуваних файлів (Приклад малокорисного результату)

*This report lists the extensions of files.*

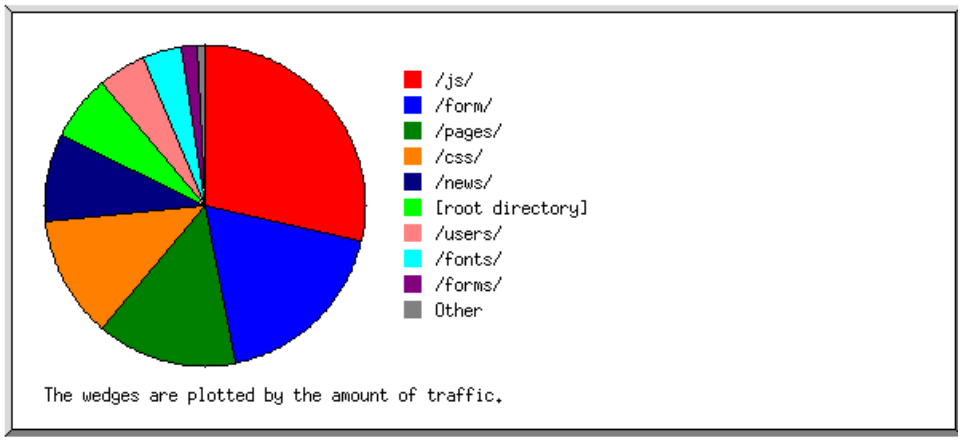



Listing extensions with at least 0.1% of the traffic, sorted by the amount of traffic.

| reqs | %bytes | extension |
|---|---|---|
| 816 | 31.57% | [no extension] |
| 235 | 28.46% | .js [JavaScript code] |
| 303 | 23.66% | [directories] |
| 161 | 12.36% | .css [Cascading Style Sheets] |
| 26 | 3.95% | .woff2 |

Рис А.7 Статистика по найбільш запитуємим форматам файлів (Приклад малокорисного результату)

*This report lists the directories from which files were requested. (The figures for each directo*




Listing directories with at least 0.01% of the traffic, sorted by the amount of traffic.

| reqs | %bytes | directory |
|---|---|---|
| 236 | 28.59% | /js/ |
| 180 | 18.32% | /form/ |
| 247 | 14.12% | /pages/ |
| 153 | 12.22% | /css/ |
| 301 | 8.95% | /news/ |
| 122 | 6.69% | [root directory] |
| 174 | 4.83% | /users/ |
| 26 | 3.95% | /fonts/ |
| 59 | 1.56% | /forms/ |
| 30 | 0.61% | /admin/ |
| 5 | 0.10% | /reviews/ |
| 6 | 0.05% | /user/ |
| 2 | 0.01% | [not listed: 1 directory] |

Рис А.8 Статистика по найбільш запитуємим директоріям (Приклад малокорисного результату)

# Додаток Б

```
import shutil

f = open("output_file.txt","r")
lines = f.readlines()
f.close()
f = open("output_file_redakt_dla_tez.txt","w")

with open('output_file.txt','wb') as wfd:
    for f in ['nginx_access.log4.txt', 'nginx_access.log3.txt',
'nginx_access.log2.txt', 'nginx_access.log1.txt', 'nginx_access.log0.log']:
        with open(f, 'rb') as fd:
            shutil.copyfileobj(fd, wfd)
```

Рис Б.1 Код програми для об’єднання журналів

```
      import datetime

f = open("output_file_redakt_dla_tez.txt ","r")
lines = f.readlines()
f.close()
f = open("Timestamp.txt","w")

char1='['
char2=']'
      data_start= "25/10/2021 00:00:00"
for line in lines:
  if line.find("bot")==-1:
    if line.find("spider") == -1:
      if line.find("bit") == -1:
        if line.find("css") == -1:
          if line.find("js") == -1:
            if line.find(".php") == -1:
              original_date = line[line.find(char1) + 1: line.find(char2)]
              d1 = datetime.datetime.strptime(data_start, '%d/%m/%Y %H:%M:%S')
              d2 = datetime.datetime.strptime(original_date, '%d/%m/%Y
%H:%M:%S')
              stamp = ((d2 - d1).days * 86400) + ((d2 - d1).seconds)
              line = line.replace('[', '', 1)
              line = line.replace(']', '', 1)
              line = line.replace('"', '|', 2)
              line = line.replace(' ', '|', 4)
              line = line.replace(' ', '', 1)
              line = line.replace(' ', '|', 1)
              line = line.replace(' ', '|', 1)
              line = line.replace(' ', '', 1)
              line = line.replace(' ', '|', 3)
              f.write("|" + str(stamp) + "|" + line)
f.close()
```

Рис Б.2 Код програми для попередньої обробки