Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Метод кластеризації Інтернет-об’єктів з динамічними компонентами

(тема)

Виконав:

студент II курсу, групи СПм-22-5

Штепа Д. С.
(прізвище, ініціали)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва освітньої програми)

Керівник: проф. Можаєв О.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
“____” _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Штепі Дем'яну Сергійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи Метод кластеризації Інтернет-об’єктів з динамічними компонентами

затверджена наказом по університету від “ 01 ” квітня 2024 р. № 257 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи операційна система Windows або Linux, ПК процесор на 1Ггц 1Гб оперативної пам’яті та 100Мб свобідної пам’яті на жорсткому диску, MySQL, технологія JDK.

4. Перелік питань, що потрібно опрацювати у роботі

1) аналітичний огляд;
2) розробка системи кластерізації;
3) програмна реалізація системи та її дослідження;
4) висновки.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 14

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз літературних джерел-	02.04.24-08.04.24	
2	Вибір та обгрунтування методики дослідження	09.04.24-16.04.24	
3	Вибір інструментальних засобів	17.04.24-22.04.24	
4	Розробка моделей протоколів	23.04.24-06.05.24	
5	Проведення експериментів	07.05.24-23.05.24	
6	Оформлення матеріалів кваліфікаційної роботи	24.05.24-03.06.24	
7	Подання кваліфікаційної роботи керівникові та	04.06.24-07.06.24	
	її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	08.06.24-12.06.24	
	вання		

Дата видачі завдання 01 квітня 2024 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Можаєв О.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 80 с., 28 рис., 13 табл., 2 дод., 16 джерел.

ІНТЕРНЕТ-ОБ’ЄКТИ, КЛАСИФІКАЦІЯ, ІНТЕРНЕТ-ПОШУК, КЛАСТЕРНИЙ АНАЛІЗ.

Метою роботи є застосування методів класичного кластерного аналізу для класифікації Інтернет-об’єктів, для персоналізації інформаційного пошуку в Інтернеті. Об'єктом дослідження є процес персоналізації Інтернет пошуку, засновані на вивченні та класифікації Інтернет-об’єктів за допомогою кластерного аналізу. Предметом дослідження є способи математичного опису Інтернет-об’єктів, процедури збору та обробки інформації про ці Інтернет-об'єкти, що дозволяють ефективно застосовувати апарат класичного кластерного аналізу з метою персоніфікації Інтернет-пошуку.

Підвищення рівня персоналізації пошуку, у свою чергу, може бути досягнуто за рахунок розробки перспективних методів класифікації Інтернет-об’єктів, що ґрунтуються на кластерному аналізі, впровадження цих методів у існуючі пошукові системи.

ABSTRACT

Master’s thesis: 80 pages, 28 figures, 13 tables, 2 appendices, 16 sources.

INTERNET OBJECTS, CLASSIFICATION, INTERNET SEARCH, CLUSTER ANALYSIS.

The purpose of the work is the application of classical cluster analysis methods for the classification of Internet objects, for the personalization of information search on the Internet. The object of research is the process of personalization of Internet search, based on the study and classification of Internet objects using cluster analysis. The subject of the research is methods of mathematical description of Internet objects, procedures for collecting and processing information about these Internet objects, which allow to effectively apply the apparatus of classical cluster analysis for the purpose of personalizing Internet search.

Increasing the level of search personalization, in turn, can be achieved through the development of promising methods for classifying Internet objects based on cluster analysis, and the introduction of these methods into existing search systems.Technical and economic calculations of the project are given.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних

ІК – Інтернет-користувач

ІР – Інтернет-ресурс

ПК – персональні комп'ютери

DOM – Document Object Model

XHTML –Xtensible Hyper Text Markup Language

XML – Xtensive Markup Language

ВСТУП

Сьогодні Інтернет є невід'ємною частиною повсякденного життя. Всі сфери людської діяльності тією чи іншою мірою пов'язана з інформаційними технологіями. Величезна кількість ресурсів і інформації, що міститься в мережі перетворило всесвітню павутину на грандіозне сховище погано організованих, неструктурованих даних.

Середня аудиторія пошукової системи Google становить понад 10 млн осіб на добу [4]. Протягом доби ця пошукова система обробляє до 150 млн запитів, видаючи Інтернет-користувачам понад 10000000000000 посилань на Інтернет-ресурси [9]. На жаль, фактом є те, що більшість знайдених ресурсів не містять інформації, що відповідає пошуковим інтересам користувачів.

Інтуїтивно будь-який ІК формує свою систему класифікації та відбору веб-ресурсів для задоволення власних потреб інформації. Користувач Інтернету має свій особистий психологічний портрет і відвідує конкретні «улюблені» ним веб-сторінки. Якщо говорити про поведінку людини в Інтернеті, то можна виділити короткочасні (сесійні) дії ІК, які пов'язані з пошуком конкретної інформації протягом однієї або декількох пошукових сесій. Коли користувач знаходить релевантну інформацію, він припиняє пошук і навіть може вийти з мережі. Крім сесійних дій користувачів можна виділити їхню рутинну поведінку в мережі, наприклад, щоденний ранковий огляд новин про спорт або спілкування в соціальних мережах в обідній час. Великі пошукові системи користуються персональною інформацією та файлами cookie з браузерів для персоналізації результатів пошуку – маркетологи, наприклад, підбирають рекламу залежно від пошукової історії або залежно від статі та віку ІК.

Соціально-демографічна класифікація – основний метод класифікації ІК після їх авторизації на Інтернет-сайтах – забезпечує облік статевих та вікових відмінностей, іншої статичної атрибутивної інформації користувача

[1]. Соціально-демографічна класифікація на сайтах застосовується, наприклад, для цільового напряму рекламних кампаній, але при цьому поведінка користувачів ніяк не застосовується до уваги. Персоналізація користувачів, що проводиться на стороні сайтів, далека від досконалості, оскільки сайти працюють за принципом «клієнт завжди правий», тобто акцент робиться на рекламодавцеві, що вклали великі кошти в просування товару - звідси і кульгають результати пошуку на стороні користувачів. Хороші результати, з допомогою застосування асоціативних методів класифікації [3, 8], досягнуто товарів, реалізованих через Інтернет магазини.

Класифікація дозволяє збільшувати продаж товарів, коли при купівлі одного товару система пропонує придбати супутній товар або набір супутніх аксесуарів. Як показує практика, покупці досить часто купують кілька товарів із однієї класифікаційної групи. Однак, невідомо, наскільки вдало можна застосовувати асоціативні методи для класифікації ІК та ІР з метою персоналізації Інтернет-пошуку? В останні роки в інформаційних джерелах можна зустріти загальні відомості щодо застосування методів кластеризації для класифікації ІК та ІР.

Актуальність. Наведені аргументи свідчать про необхідність подальшого пристосування Інтернету до потреб користувачів та, зокрема, за рахунок персоналізації Інтернет-пошуку. Підвищення рівня персоналізації пошуку, у свою чергу, може бути досягнуто за рахунок розробки перспективних методів класифікації ІК та ІР, що ґрунтуються на кластерному аналізі, впровадження цих методів у існуючі пошукові системи.

Метою роботи є застосування методів класичного кластерного аналізу для класифікації ІК та ІР, для персоналізації інформаційного пошуку в Інтернеті.

Досягнення поставленої мети потрібно вирішити такі основні завдання.

1. Проаналізувати існуючі некластерні методи класифікації ІК та ІР. Проаналізувати існуючі методи кластерного аналізу ІК та ІР, показати їх перевагу порівняно з некластерними методами.

2. Запропонувати адекватний математичний опис Інтернет-об'єктів.

3. Вибрати алгоритм кластеризації Інтернет-об'єктів з числа відомих методів кластерного аналізу, що дозволяє керувати результатом за допомогою вхідних параметрів.

4. Визначити масштаб впливу інформаційної динаміки Інтернет-об'єктів на результати їх кластерного аналізу. Запропонувати методи усунення динамічних факторів при кластеризації ІК та ІР.

5. Розробити та застосувати оригінальний підхід, заснований на принципі узагальнення та одночасної кластерної обробки ІК та ІР.

Об'єктом дослідження є процес персоналізації Інтернет пошуку, засновані на вивченні та класифікації ІК та ІР за допомогою кластерного аналізу.

Предметом дослідження є способи математичного опису ІК та ІР, процедури збору та обробки інформації про ці Інтернет-об'єкти, що дозволяють ефективно застосовувати апарат класичного кластерного аналізу з метою персоніфікації Інтернет-пошуку.

В основі дослідження лежать методи статистичного та кластерного аналізу, теорія графів та web-технології.

1 ДОСЛІДЖЕННЯ АКТИВНОСТІ ІНТЕРНЕТ-КОРИСТУВАЧІВ ДЛЯ ВИРІШЕННЯ ЗАВДАНЬ ПЕРСОНАЛІЗАЦІЇ

1.1 Аналіз активності Інтернет-користувачів у соціальних мережах

В даний час персональна інформація ІК представляє величезний інтерес як для Інтернет-майданчиків, так і для рекламодавців. Справа в тому, що будь-який IP зацікавлений у обробці особистої інформації ІП, які відвідували його сторінки. Це важливо для статистичної обробки відвідуваності з продажу реклами. Можна чітко розділити чоловічі та жіночі сайти, спортивні чи новинні сайти. Наприклад, візьмемо один із великих Інтернет-порталів України – ukr.net. Почнемо із використання персональної інформації користувача компанією ukr.net для персоналізації контенту. Робота користувача починається з реєстрації поштової скриньки (рисунок 1.1) на головній реєстраційній сторінці сайту. На цій сторінці користувач залишає найціннішу інформацію про себе: дату народження, стать, місто та країну проживання.

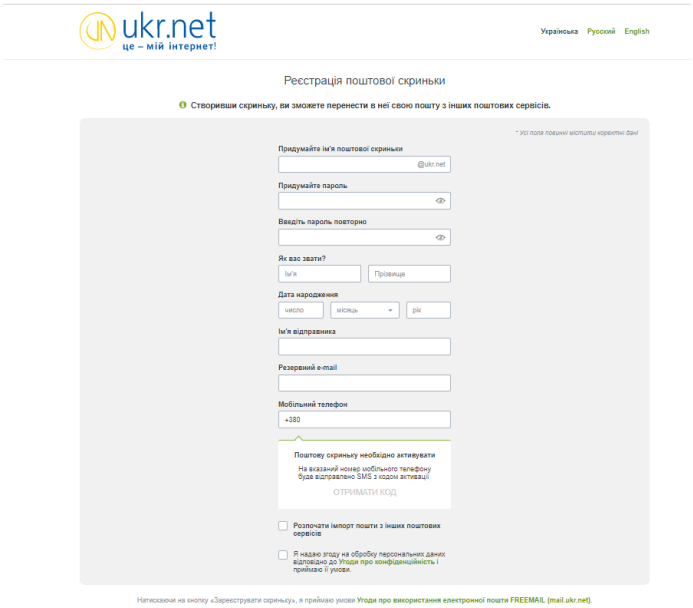

Рисунок 1.1 – Реєстраційна форма для створення поштової скриньки ukr.net

Після того, як ІК залишить персональну інформацію під час реєстрації нової поштової скриньки (рисунок 1.1), ця інформація стає доступною великому числу фахівців (програмістам, маркетологам тощо). З цього моменту починають працювати різні алгоритми для персоналізації інформаційного Інтернет-потоку з порталу ukr.net. Достатньо спостерігати за банерною рекламою на головній сторінці ukr.net після авторизації користувача. Отже, відразу після заповнення реєстраційної форми персональною інформацією ІР починає її використовувати для підбору реклами.

Людина споконвіку жив у соціальному середовищі і тому за своєю поведінкою є соціально залежним суб'єктом. Його активність поширювалася на сім'ю, друзів та соратників, роботу та інші життєві сфери. Зараз, у вік телекомунікацій та комп'ютерних технологій, соціальна активність людини перейшла в кіберпростір. З появою соціальних мереж (Facebook, Instagram), більшість соціально активних людей перейшла до спілкування у віртуальному світі. Люди змогли знайти своїх колег, однокласників та друзів, з якими давно було втрачено зв'язок. Збільшення кількості людей з однаковими інтересами та поглядами призвело до народження ідеї створення спеціалізованих груп.

В рамках однієї і тієї ж спеціалізованої соціальної групи учасники обговорюють конкретні теми та проводять онлайн бесіди з людьми, які мають однаково спрямоване мислення. Практика показує, що в соціальних мережах і блогах інформація поширюється набагато швидше, ніж на сайтах новин. У соціальних мережах людина залишає набагато більше інформації про себе і про свою поведінку, ніж на будь-якому іншому Інтернет-ресурсі: школа, інститут, робота і, звичайно, коментарі. Варто відзначити дуже важливий момент, присутній у соціальних мережах і пов'язаний з так званими like-ами: авторизований користувач може ставити оцінки іншим користувачам, їх статтям, коментарям, фотографіям, тощо, відзначивши like-ом, що сподобалося. Для моментального відображення статусу користувачів

та їх поведінки у межах групи застосовуються нереляційні бази даних, які працюють із високошвидкісної оперативної пам'яттю. Це звані «in memory database» чи бази даних у пам'яті. Основна ідея таких систем – зберігання даних не на дисковому накопичувачі, а у пам'яті. Застосування такого роду БД зменшує час відгуку системи та дозволяє практично моментально перемикатися за групами інтересів користувачів соціальних мереж.

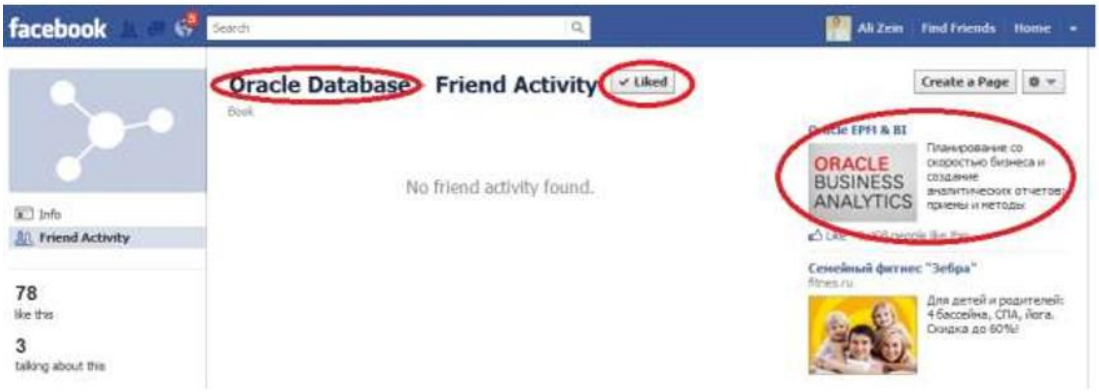

Рисунок 1.2 –Використання лайків для показу реклами у соціальній мережі

Соціальні групи є добрим прикладом для класифікації ІК. Справа в тому, що люди формують групи інтересів, можуть користуватися загальними ресурсами та ділитися досвідом у тому чи іншому напрямі. При цьому будь-яка сформована група буде досить спеціалізованою і може бути легко індексована для швидкого пошуку інтересів користувачів.

1.2 Методи некластерної класифікації інтернет-користувачів та інтернет-ресурсів

Асоціативний метод класифікації широко застосовується в Інтернет магазинах, коли зміст покупних кошиків певної множини покупців аналізується і утворюється певна ймовірнісна закономірність покупок. Проведемо аналіз релевантності між елементами вектора з допомогою асоціативних правил. У БД Інтернет-запитів від 20 вересня 2023 р. випадково

вибрані п'ять ІК виконували пошук товару на сайті market.shop.ua. Представимо таблицю пошуку з векторами пошуку, що складаються з багатьох товарів (ноутбук, планшет, смарт-годинник, смартфон). Шукані множини, представлені в таблиці 1.1 були отримані в умовах, коли час між пошуковими запитами не перевищував чотирьох годин.

Таблиця 1.1 – Таблиця транзакцій пошуку товарів ІК

TID	Транзакції
1	{ноутбук, планшет}
2	{ноутбук, смарт-годинник, смартфон}
3	{смарт-годинник, смартфон}
4	{планшет}
5	{ноутбук, смартфон}

Для початку необхідно розподілити елементи {ноутбук, планшет, смарт-годинник, смартфон} у проміжну таблицю попадань (таблиця 1.2).

Таблиця 1.2 – Таблиця влучень

Набір елементів	Транзакції	Число влучень
{}	{1,2,3,4,5}	5
{ноутбук}	{1,2,5}	3
{планшет}	{1,4}	2
{смарт-годинник}	{2,3}	2
{смартфон }	{2,3,5}	3
{ноутбук, планшет}	{1}	1
{ноутбук, смарт-годинник }	{2}	1
{ноутбук, смартфон}	{2,5}	2
{смарт-годинник, смартфон}	{2,3}	2
{ноутбук, смарт-годинник, смартфон}	{2}	1

Стовпець «Набір елементів» формується за допомогою окремих елементів та можливих комбінацій цих елементів відповідно до реальних результатів таблиці транзакцій (таблиця 1.1). Стовпець «Транзакції» формується за допомогою набору транзакцій, в якому була присутня комбінація елементів в і-му рядку. Значення стовпця «Кількість попадань»

формується на підставі числа елементів стовпця «Транзакції». Тепер можна побудувати асоціативну таблицю елементів (таблиця 1.3).

За асоціативною таблицею елементів проводиться розрахунок ймовірності появи події {смартфон}, якщо подія {ноутбук} мала місце тощо.

Таблиця 1.3 – Асоціативна таблиця елементів

Асоціативний набір	Кількість влучень	Відсоток ймовірності
{ноутбук} → {смартфон}	2	2/3 = 67%
{планшет} → {ноутбук}	1	1/2 = 50%
{смарт-годинник} → {ноутбук}	1	1/2 = 50%
{смарт-годинник} → {смартфон}	2	2/2 = 100%
{смартфон} → {ноутбук}	2	2/3 = 67%
{смартфон} → {смарт-годинник}	2	2/3 = 67%
{ноутбук, смарт-годинник} → {смартфон}	1	1/1 = 100%
{ноутбук, смартфон} → {смарт-годинник}	1	1/2 = 50%
{смарт-годинник, смартфон} → {ноутбук}	1	1/ 2 = 50%
{смарт-годинник} → {ноутбук, смартфон}	1	1/2 = 50%

На прикладі {ноутбук} → {смартфон} у чисельнику буде знаходитись число випадків, коли в транзакції присутні обидва елементи {ноутбук} та {смартфон}: це друга та п'ята транзакція в таблиці 1.1. У знаменнику буде кількість випадків, коли в транзакції тільки є елемент {ноутбук}. Таким чином, в чисельнику буде count({ноутбук, смарт-годинник, смартфон}, {ноутбук, смартфон}) = 2, у знаменнику буде count({ноутбук, планшет}, {ноутбук, смарт-годинник, смартфон}, {ноутбук, смартфон}) = 3. Звідси, ймовірність появи події {смартфон}, якщо настала подія {ноутбук} дорівнюватиме

$$P(\{\text{ноутбук}\} - - > \{\text{пам}'\text{ять}\}) = 2/3 \times 100 = 67\%. \quad (1.1)$$

Метод перетинів ґрунтується на перетині елементів на різних транзакціях. Повернемося до таблиці попадань (таблиця 1.2), і з цієї таблиці виберемо одиночні набори елементів, формуючи таблицю одиночних наборів елементів (таблиця 1.4) {ноутбук}, {планшет}, {смарт-годинник} та {смартфон}.

Таблиця 1.4. – Таблиця одиночних наборів елементів

№	{ноутбук}	{планшет}	{смарт-годинник}	{смартфон}
1	+	+	-	-
2	+	-	+	+
3	-	-	+	+
4	-	+	-	-
5	+	-	-	+

З таблиці одиночних наборів елементів можна скласти таблицю подвійних наборів елементів (таблиця 1.5).

Таблиця 1.5 – Таблиця подвійних наборів елементів

№	{ноутбук, планшет}	{ноутбук, планшет}	{ноутбук, смарт-}	{ноутбук, смартфон}	{планшет, смарт-}	{планшет, смартфон}	{смарт-годинник, смартфон}
1	+	-	-	-	-	-	-
2	-	+	+	-	-	-	+
3	-	-	-	-	-	-	+
4	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-

На етапі формування таблиці подвійних наборів елементів відбувається випадання конкретних комбінацій: випадають {планшет, смарт-годинник} та {планшет, смартфон}. Для розрахунку ймовірності, необхідно спочатку

визначитися з порядком проходження елементів у парах, що формуються - {ноутбук, планшет} або {планшет, ноутбук}, тому що результат розрахунку ймовірностей буде різним:

$$P(\{\text{ноутбук}\} - -> \{\text{планшет}\}) = 1/3 \times 100 = 63\%. \quad (1.2)$$

$$P(\{\text{планшет}\} - -> \{\text{ноутбук}\}) = 1/2 \times 100 = 50\%. \quad (1.3)$$

З таблиці подвійних наборів елементів можна скласти таблицю потрійних наборів елементів (таблиця 1.6).

Таблиця 1.6 – Таблиця потрійних наборів елементів

№	{ноутбук, планшет, смарт-годинник}	{ноутбук, планшет, смартфон}	{ноутбук, смарт-годинник, смартфон}
1	-	-	-
2	-	-	+
3	-	-	-
4	-	-	-
5	-	-	-

На етапі формування таблиці потрійних наборів елементів відбувається випадання конкретних комбінацій: випадають комбінації {ноутбук, планшет, смарт-годинник} та {ноутбук, планшет, смартфон}. Для розрахунку ймовірності необхідно також визначитися з порядком прямування елементів у трійках. Для виконання алгоритму необхідно виконати перерахунок усіх можливих комбінацій (реалізація методу повного перебору).

За допомогою схеми всіх можливих комбінацій можна розрахувати ймовірність появи будь-якої комбінації елементів. Асоціативний метод та метод перетинів широко застосовуються для підбору супутнього товару. Велика кількість покупок груп товарів робить результат розрахунків ймовірностей більш точними, так як ІК показують величезну кількість товарів, яке купувалося одночасно з товаром, що переглядається. Однак ці методи погано масштабуються і як тільки відбувається вихід за рамки транзакції, результати розрахунку ймовірностей псуються. Справа в тому, що

пошукові інтереси ІК неможливо розділити на окремі транзакції і найпопулярніші пошукові терміни зустрічаються у більшості ІК у рамках періоду спостереження $\Delta t = 4$ години. Ці методи не враховують ваги термінів.

Метод частоти термінів широко застосовується для класифікації звичайних текстів. Цей метод можна застосувати і для класифікації ІК з їхньої пошукової історії та ІР з їхнього текстового змісту. Нехай вихідними даними для методу є пошукові терміни ІК у період з 20 по 27 вересня 2023 р. Нехай USR – множина ІП, що спостерігаються, і USR_i – $i_{ий}$ ІП, що спостерігається. Кожен usr_i, $usr_i \in USR$, за залишеними ним персональними даними.

Ці групи використовуються для таргетування реклами ІР, в якій, крім статі, береться до уваги вікова категорія Інтернет-користувача (таблиця 1.7). Після побудови бази даних для структуризації запитів ІК та змісту ІР застосування TF-методу не складає жодних труднощів. З урахуванням наявної обчислювальної потужності, при формуванні статистичних таблиць (таблиці 1.8 і 1.9) було прийнято рішення про обмеження числа термінів nof (V) у глобальному словнику термінів V та їх довжини. У [5] вказано середню довжину українських слів дорівнює 5,28 символів. Однак у пошуковій історії ІК виявилося велике число термінів, довжина яких дорівнює 4, тому необхідно розглядати терміни, довжина яких більша або дорівнює 4.

Таблиця 1.7 – Таблиця соціально-демографічної класифікації

Соц-дем група	стать	вік
USR_1	ЧОЛ	від 12 до 17
USR_2	ЧОЛ	від 18до 24
USR_3	ЧОЛ	від 25 до 34
USR_4	ЧОЛ	від 35 до 44
USR_5	ЧОЛ	від 45 до 55
USR_6	ЖІН	від 12 до 17
USR_7	ЖІН	від 18до 24
USR_8	ЖІН	від 25 до 34
USR_9	ЖІН	від 35 до 44
USR_{10}	ЖІН	від 45 до 55

Таблиця 1.8 – Таблиця терезів шуканих слів за статевою ознакою

Word	fg	f1	f2	f3	f4	f5
наявність	0,110336	0,013158	0,211583	0,182885	0,09277	0,159288
онлайн	0,100315	0,013158	0,11484	0,113593	0,117544	0,067463
скачати	0,091481	0,197368	0,000329	0,090875	0,102195	0,125556
безкоштовно	0,066189	0,013158	0,039816	0,073836	0,074054	0,063247
дивитися	0,064272	0,013158	0,05561	0,06134	0,076612	0,029984
купити	0,059397	0,013158	0,071405	0,105642	0,056954	0,1026
ігри	0,050918	0,013158	0,040803	0,000379	0,086711	0,000234
сайт	0,037042	0,171053	0,046397	0,040515	0,042144	0,038416
відгуки	0,036855	0,013158	0,051662	0,000379	0,042682	0,059733
фото	0,032542	0,013158	0,059559	0,059447	0,000135	0,04029
Facebook	0,031521	0,328947	0,118789	0,000379	0,035142	0,000234
Instagram	0,029917	0,013158	0,000329	0,000379	0,000135	0,046849
відео	0,02923	0,013158	0,050346	0,000379	0,033661	0,033497
фільм	0,01723	0,013158	0,013158	0,000379	0,000379	0,000234
мапа	0,016625	0,016625	0,000329	0,000379	0,021678	0,022722

Таким чином, кожна соціально-демографічна група USR_i може бути представлена числовим вектором $f_{i1} = (f_{i,12}, \ldots, f_{i,2\,j}, \ldots, f_{2i,\,nof\,(Vu)})$ розміром nof (V_u), де $f_{i,\,j}$ – вага j- ого пошукового терміна із глобального словника термінів V_u. Числові координати $f_{i,j}$, $1 \le j \le nof(V_u)$ розташовані у характеристичному векторі f_i, у тому порядку, що й терміни у глобальному словнику V_u. Перехід від вербального до числового представлення результатів дослідження окремо взятої соціально-демографічної групи USR_i у вигляді характеристичного вектора відбувається за рахунок позиційного кодування термінів словника, підрахунку числа їх входжень у запити ІК групи протягом усієї пошукової історії та розрахунку частоти вживання цих термінів (TF значень) групі. Для розрахунку TF-значень застосовується така формула:

$$f_{ij} = \frac{nof(v_{i,j})}{\sum_{j=1}^{nof(Vu)} nof(v_{i,j})}, \quad (1.4)$$

де $nof(v_{i,j})$, $v_{i,j} \in Vu$, – число входження терміну v_j у запити користувачів *i-oї* соц-дем групи протягом усієї пошукової історії. Для порівняння результатів необхідно побудувати глобальний вектор $fg = (fg_1, \ldots, fg_j, \ldots, fg_{nof(Vu)})$, де д ля розрахунку координат fg_j у чисельнику та знаменнику формули 1.1 буде відповідно кількість входження всіх термінів v_j для всіх USR_i, $1 \leq j \leq 10$.

Таблиця 1.9 – Таблиця терезів шуканих слів

Word	fg	f1	f2	f3	f4	f5
наявність	0,110336	0,123077	0,067952	0,103787	0,152243	0,099602
онлайн	0,100315	0,148178	0,16483	0,131092	0,130391	0,09041
скачати	0,091481	0,149798	0,113407	0,090875	0,09145	0,09145
безкоштовно	0,066189	0,013158	0,039816	0,073836	0,074054	0,063247
дивитися	0,064272	0,013158	0,05561	0,06134	0,076612	0,029984
купити	0,059397	0,013158	0,071405	0,105642	0,056954	0,1026
ігри	0,050918	0,013158	0,040803	0,000379	0,086711	0,000234
сайт	0,037042	0,171053	0,046397	0,040515	0,042144	0,038416
відгуки	0,036855	0,013158	0,051662	0,000379	0,042682	0,059733
фото	0,032542	0,013158	0,059559	0,059447	0,000135	0,04029
Facebook	0,031521	0,328947	0,118789	0,000379	0,035142	0,000234
Instagram	0,029917	0,013158	0,000328	0,000379	0,000135	0,046843
відео	0,02923	0,013158	0,050346	0,000379	0,033661	0,033495
фільм	0,01723	0,013158	0,013158	0,000379	0,000379	0,000232
мапа	0,016625	0,016625	0,000329	0,000374	0,021675	0,022721

Статистика отримана, ваги слів розраховані, можна розпочинати формування графіків ваг (рисунок 1.4).

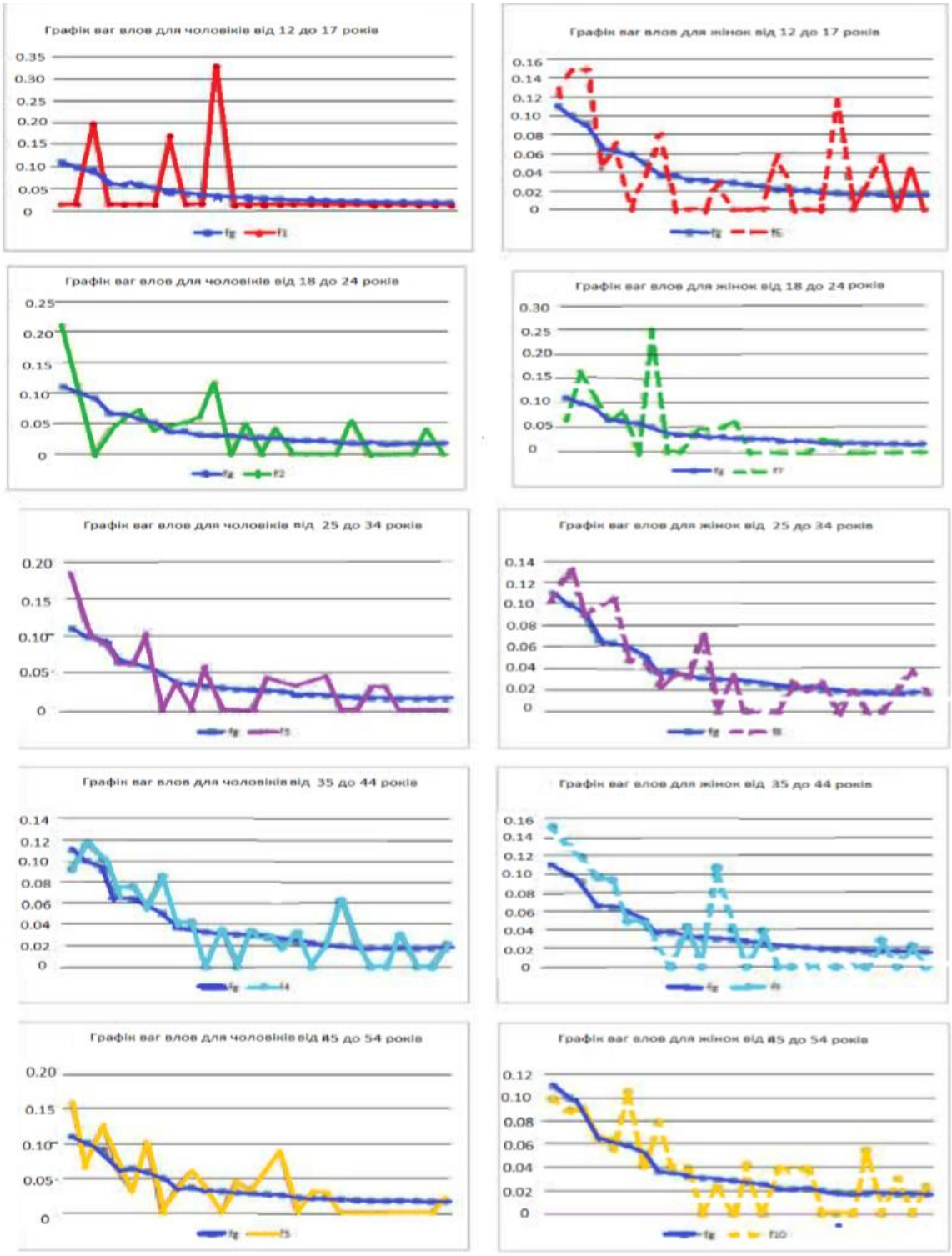

Рисунок1.3 – Графіки ваг слів для ІП, поділених за соціально-демографічними ознаками

На графіках рисунка 1.4 чітко видно, що кожна соц-дем група має свої інтереси та свою поведінку в Інтернеті. Кожен вектор f_i має відхилення щодо

глобального вектора fg. Розглянутий метод TF безпосередньо залежить від якості формування словника Vu та від персональних даних ІП. Метод частоти термінів TF може бути застосований для класифікації ІП, виключно з метою підбору реклами, проте якщо його застосувати для ІР з динамічними елементами, можна зіткнутися з проблемою змін цих показників з кожним завантаженням ІР. Представлені вище методи класифікації можуть бути використані для первинної сегментації Інтернет-об'єктів з обмеженою кількістю характеристик. Це може бути класифікація ІК за соціальними ознаками або ІР за структурою. Для персоналізації пошуку необхідно застосовувати складніші методи з можливістю формування груп об'єктів зі складними характеристиками.

1.3. Кластерні методи класифікації Інтернет-користувачів та Інтернет-ресурсів

Кластеризація – це автоматичне розбиття елементів деякої множини на групи (кластери) залежно від показників їхньої схожості. Елементами множини може бути будь-що: об'єкти з певним набором даних або вектора характеристик. Більшість дослідників вважають, що прабатьком кластерного аналізу є Роберт Коаті Тріон – американський дослідник поведінки тварин, який запропонував систематизувати методи аналізу впливу навколишнього середовища (екологія, соціальний рівень тощо) на поведінку суб'єктів дослідження (тварин, людей) і запропонував групувати суб'єкти дослідження у кластери. Запропонований їм метод дозволив з великою точністю визначати причини та можливі наслідки поведінки людини у стресових ситуаціях, виходячи з її соціального оточення.

У кластеризації існує велика кількість практичних застосувань. Кластеризація дозволяє, наприклад, провести аналіз даних, пошук інформації або групування об'єктів за ознаками та властивостями. Так само

кластеризація сама по собі є важливою формою абстракції даних, і в цій галузі було отримано низку цікавих наукових результатів. Говорячи про кластеризацію Інтернет-об'єктів, необхідно визначити такі базові поняття. Об'єкт – елементарна одиниця, яка може бути представлена за допомогою набору числових характеристик і з якою оперують алгоритми кластеризації. Кожному об'єкту x_i , $i \in I$, зіставляється вектор числових характеристик $z_i = (z_{i,1}, \ldots, z_{i,j}, \ldots, z_{i,n})$. Кардинальність вектора визначає розмірність простору характеристик. Відстань $\rho(x_i,x_k)$ між об'єктами x_i та x_k – результат застосування обраної метрики у просторі характеристик. В даний час існує велика кількість метрик для оцінки відстані між векторами одного і того ж векторного простору. До простих метриків можна віднести евклідову відстань або її квадрат, манхеттенську відстань, статечна відстань та інші [3].

Перенесемо математичний опис абстрактних об'єктів у область дослідження пошукових запитів ІП. Нехай USR – множина спостережуваних ІП, usr_i – перший спостерігається ІП, $usr_i \in$ USR. У довільний момент часу $t_k \in$ T, вказаний ІК можна представити характеристичним вектором такого вигляду:

$$u_i(t_k)=(u_{i,1}(t_{k, \ldots,} u_{i,j}(t_k), \ldots, u_{i,nof(Vu)}(t_k)) \quad (1.5)$$

де $u_{i,j}(t_k)$ – вага j-ого пошукового терміна з глобального словника термінів V_u в момент часу t_k, рівний числу входження цього терміна в запити в пошуковій історії і-го ІП, протягом тимчасового вікна Δt, що спостерігається;

$_{nof}(V)$ – розмір вектора і-го ІП, що дорівнює кількості слів у глобальному словнику термінів V_u.

Числові координати $u_{i,j}(t_k)$, $1 \le j \le {}_{nof}(V_u)$ розташовані в характеристичному векторі в порядку, що відповідає лексикографічному порядку дотримання відповідних термінів у словнику V_u. Перехід від вербального до числового представлення результатів відбувається за рахунок

позиційного кодування термінів та підрахунку числа їх входжень у запити пошукової історії ІП. У наші дні методи кластерного аналізу широко використовуються для вирішення широкого спектру завдань в Інтернеті:

- за способом аналізу даних: точні та нечіткі;
- за кількістю застосувань алгоритмів кластеризації: з одноетапною кластеризацією та з багатоетапною кластеризацією;
- по можливості розширення обсягу оброблюваних даних: масштабовані та не масштабовані;
- за часом виконання кластеризації: потокові (у режимі реального часу) та непотокові (за накопиченням інформації).

Існує ключова різниця між поняттям кластеризація та поняттям класифікація. Кластеризація дозволяє розбити безліч об'єктів на групи (кластери), а класифікація – відносить кожен об'єкт до однієї із заздалегідь визначених груп.

У процесі розв'язання задач кластеризації-класифікації можна виділити чотири групи завдань:

- виділення характеристик об'єктів;
- визначення метрики – для кластеризації об'єктів застосовується метрика близькості об'єктів;
- розбиття об'єктів на групи із застосуванням методів кластерного аналізу;
- класифікація новоствореного об'єкта.

У прикладній статистиці [3] запропоновано 8-ми етапну схему вирішення завдань класифікації. Кожен етап цієї схеми є повноцінним процесом з вхідними і вихідними потоками, можливий і зворотний зв'язок. Кластерні методи зі складними алгоритмами оптимізації застосовуються в пошукових системах, Інтернет-магазинах, системах аналізу контенту сайтів, системах перевірки достовірності текстів дисертацій та ще у багатьох сферах.

Методи кластерного аналізу різноманітні. Їх можна розбити на безліч груп:

- за способом обробки даних: ієрархічні (агломеративні методи та дивізивні методи); неієрархічні методи (ітеративні);

- за способом аналізу даних: точні та нечіткі;

- за кількістю застосувань алгоритмів кластеризації: з одноетапною кластеризацією та з багатоетапною кластеризацією;

- по можливості розширення обсягу оброблюваних даних: масштабовані та не масштабовані;

- за часом виконання кластеризації: потокові (у режимі реального часу) та непотокові (за накопиченням інформації).

Існує ключова різниця між поняттям кластеризація та поняттям класифікація. Кластеризація дозволяє розбити безліч об'єктів на групи (кластери), а класифікація – відносить кожен об'єкт до однієї із заздалегідь визначених груп.

У процесі розв'язання задач кластеризації-класифікації можна виділити чотири групи завдань:

- виділення характеристик об'єктів;

- визначення метрики – для кластеризації об'єктів застосовується метрика близькості об'єктів;

- розбиття об'єктів на групи із застосуванням методів кластерного аналізу;

- класифікація новоствореного об'єкта.

У [3] запропоновано 8-ми етапну схему вирішення завдань класифікації. Кожен етап цієї схеми є повноцінним процесом з вхідними і вихідними потоками, можливий і зворотний зв'язок.

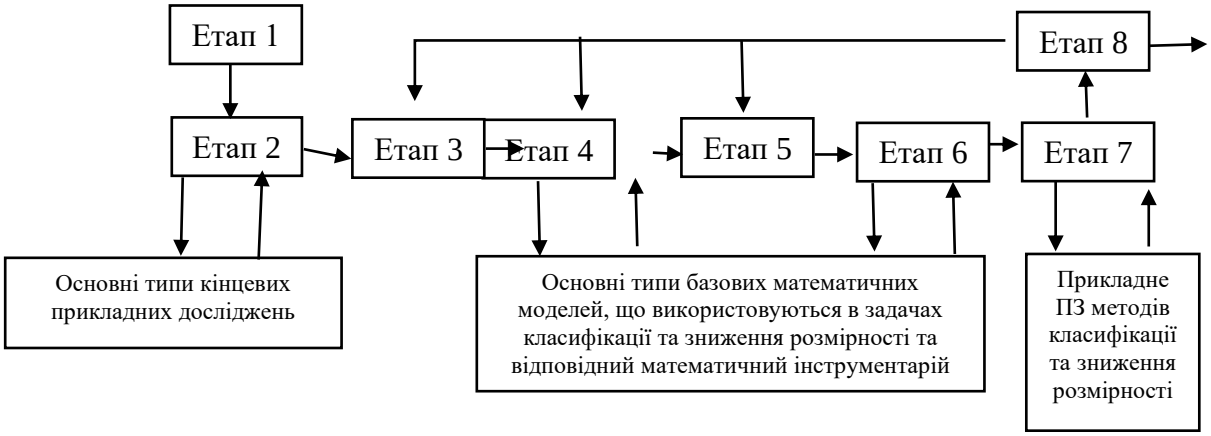

Рисунок 1.4 – Схема поетапного процесу розв'язання задач класифікації

Етап 1. Настановний – на цьому етапі має бути сформульована постановка завдання, що включає характер наукових або практичних висновків, які потрібно отримати на виході.

Етап 2. Постановочний – цьому етапі необхідно сформулювати мети предметно-змістової установки на етапі 1 у термінах основних типів прикладних завдань, які у теорії статистичних методів класифікації.

Етап 3. Інформаційний – полягає у виробленні та реалізації плану збору вихідної статистичної інформації.

Етап 4. Апріорно-математико-постановочний – на підставі висновків та інформації, отриманих в результаті реалізації етапів 1-3, потрібно здійснити попередній вибір базових математичних моделей, які доцільно використовувати в математичній постановці даного конкретного завдання. У цьому чинниками, яких вирішальним чином залежить вибір, є характер кінцевих прикладних цілей дослідження, природа і форма вихідних статистичних даних.

Етап 5. Розвідувальний аналіз – цей етап становлять усілякі методи попередньої статистичної обробки, пропущених вихідних даних з метою виявлення специфіки їхньої ймовірнісної та геометричної природи. На виході етапу мають бути уточнені відомості про фізичний механізм генерування наших вихідних даних, а отже, про базову математичну модель цього механізму.

Етап 6. Апостеріорний математико-постановний – цьому етапі уточнюється математична постановка розв'язуваної завдання з урахуванням висновків.

Етап 7. Обчислювальний – проводиться обчислювальна реалізація наміченого до використання, обраного на попередньому етапі, математичного інструментарію розв'язання задачі.

Етап 8. Підсумковий – аналізуються та інтерпретуються результати виконаної роботи. Залежно від результатів цього аналізу або формуються остаточні наукові або прикладні висновки, або даються уточнення та доповнення до завдання та відбувається повернення до одного з попередніх етапів (зазвичай до етапу 3, 4 або 5). На останньому етапі слід очікувати позитивного результату, що задовольняє поставленим завданням та обраним математичним моделям, інакше необхідно повернутись до одного з попередніх етапів для доведення прийнятих рішень.

2 МЕТОДИ КЛАСТЕРИЗАЦІЇ ІНТЕРНЕТ-ОБ'ЄКТІВ

2.1. Методи аналізу змісту тексту

Лінгвістичний аналіз – метод дослідження тексту, який може бути охарактеризований як лінгвосмисловий аналіз. Вивчення методів, дозволяють автоматизовано «розуміти» текст, тобто. вміти витягувати з нього потрібну інформацію та відповідати на запитання, задані щодо тексту. Лінгвістичний аналіз застосовується, зокрема, до вилучення інформації, машинного перекладу, а також до багатьох областей штучного інтелекту, які стосуються спілкування з користувачем. Існує безліч підходів до лінгвістичного аналізу. Серед них можна виділити статистичний аналіз, аналіз ознак, семантичний аналіз та комбінований підхід.

Статистичний аналіз передбачає вивчення послідовності слів у реченнях тексту, і навіть виведення певних закономірностей виходячи з проведеного вивчення. І тому проводиться підрахунок частоти зустрічі слів у тексті, і навіть ймовірність появи слів друг за другом. З допомогою статистичних методів аналізу тексту вирішується проблема класифікації текстів [18]. Відповідно, аналізуючи зміст слів, можна отримувати ймовірність зустрічі N-грами або частоту термінів (TF або TF-IDF) в художньому або науковому тексті. Після цього знайдені ймовірності перетворюються на ваги і складаються. Текст ставитиметься до того класу, вага якого виявиться більшою.

Аналіз ознак полягає у вивченні морфемних, морфологічних та синтаксичних ознак слів та речень у тексті. Це вивчення необхідно для того, щоб далі можна було будувати модель (структуру) пропозицій і на її підставі отримувати необхідну інформацію. Для побудови такої структури потрібно провести кілька операцій. Перша полягає у визначенні граматичних ознак усіх слів. Для цього в мовах, в яких існують відмінки, відмінювання,

відмінювання і часи, встановлюються всі ці ознаки. Також визначається лема кожного слова – його словникова форма. Для визначення леми слова використовується алгоритм-лематизатор, який або за допомогою порівняння зі словником, або за допомогою послідовного відсікання закінчень та афіксів (префіксів, суфіксів та постфіксів) та додавання нормалізованого закінчення виділяє основу слова. Друга операція зводиться до побудови моделі речень у тексті. З допомогою аналізу граматичних ознак, знайдених раніше, складається залежність слів друг від друга – спочатку у межах речення, та був і межах тексту, якщо необхідний ширший аналіз. Третя операція зводиться до пошуку шуканої інформації. Запит користувача проходить лематизацію, після чого відбувається пошук необхідних лем за всіма пропозиціями. Однак те, що не є великою працею для лінгвіста і просто людини, є великою проблемою для обчислювальної системи.

Описуваний вище аналіз складно чітко сформулювати, тому що, наприклад, слово «історії» можна розглядати не тільки як дальний відмінок, але і як називний / знахідний відмінок множини слова «історія». Окремою перешкодою є омоніми («Мустанг» – це автомобільний бренд, енергетичний напій чи хижак?), тобто. співзвучні слова; якщо в тексті зустрічається кілька слів з однаковими лемами, немає гарантії того, що це одне й те саме слово. Для з'ясування цього необхідно провести складніший кластерний аналіз визначення змісту текстів IP загалом після отримання характеристичних векторів. Без кластерного аналізу (або без застосування методів класифікації текстів) значень слів такі неоднозначності досить важко розв'язати. Ще однією значною проблемою є розбір відсутніх у словнику слів, коли потрібно звертатися до лінгвістичного експерта, який своєю чергою вносить зміни до словника лем. Чітко заданих правил такого аналізу немає; їх можна вивести експериментальним шляхом, проте все одно існуватимуть винятки, відпрацювати які за правилами буде неможливо.

Семантичний аналіз займається розбором тексту щодо значення слів усередині нього. Зазвичай цей вид аналізу застосовується після проведення

граматичного аналізу та доповнює його своїми висновками. Зокрема, семантичний аналіз дозволяє виявляти незв'язність слів і речень усередині тексту, хоча вони можуть бути узгоджені граматично. Також семантичний аналіз дозволяє визначати метафори, переносні значення, справжній зміст співзвучних слів залежно від контексту тощо. На жаль, серйозних успіхів у цій галузі поки що досягти не вдалося, тому цей вид аналізу є найскладнішим і найменш формалізованим, хоч і найбільш затребуваним. Прості способи семантичного аналізу дозволяють класифікувати текст, виділяти емоційне забарвлення тексту (за допомогою виявлення певних слів та аналізу словосполучень на предмет метафор та алегорій) і його тему (за синтаксичними ознаками і кількістю слів, що повторюються в реченнях). Зокрема, за допомогою семантичного аналізу відбувається видача контекстної реклами на багатьох сайтах та пошукових системах [10]. Сторінка, що видається користувачеві, досліджується щодо наявності повторюваних ключових слів, після чого автоматизований генератор реклами видає пов'язану зі знайденими ключовими словами вибірку.

Комбінований підхід має на увазі використання декількох з вищеописаних підходів у зв'язці, послідовній або паралельній, для підвищення точності аналізу. Найчастіше для складного аналізу тексту застосовують аналіз ознак, поєднаний зі статистичним аналізом для ранжування результату пошуку та вирішення неоднозначностей; рідше використовуються вкраплення семантичного аналізу у будь-який з вищеописаних методів. Зокрема, такий підхід використовується в текстових редакторах для виявлення складних помилок (неузгодженість тексту, рекомендації щодо розбиття тексту на абзаци тощо). В рамках даної роботи обробка тексту проводиться за допомогою комбінованого підходу на підставі статистичного методу та методу аналізу ознак.

2.2. Лінгвістична обробка Інтернет запитів

Для застосування методів лінгвістичного аналізу необхідні леми всіх слів тексту (аналіз ознак) та частота зустрічі цих лем у тексті (статистичний аналіз). Слід зазначити, що аналіз моделей зв'язків між словами всередині речень та реченнями всередині тексту в даній роботі не розглядається. Першим кроком до вирішення поставленої проблеми аналізу змісту як ІР, так і запитів ІК є розбиття тексту Інтернет-сторінок та Інтернет-запитів на окремі слова (терміни), тому що текст у формальному визначенні просто набором слів. Тут уже можливі проблеми та неоднозначні трактування: що вважати словом, як ставитись до складних знаків пунктуації тощо. Введемо набір простих правил, що описують більшість випадків, які можуть зустрітися в апріорно правильному тексті. Словом або терміном назвемо послідовність символів букв, обмежена з обох сторін пробілами або розділовими знаками, в якій можуть бути цифри, в тому числі і на першій позиції. Всі розділові знаки та спеціальні символи («+», «-», «/», «=» і т.д.) замінюються пробілами, тим самим, безперервна послідовність символів перетворюється на слова, окремі один від одного пробілами.

Лемою або коренем слова вважатимемо урізану послідовність символів терміну, одержувану допомогою спеціально розробленого алгоритму відсікання закінчень з урахуванням дворівневого словника, що включає глобальний словник термінів і лем словник, який може заповнюватися за допомогою існуючих відкритих словників [5] і динамічно поповнюватися при появі нових термінів. Очевидно, що кільком термінам може відповідати одна лема. Зі сформульованих правил випливає, будь-які тексти дійсно підійдуть під описи, наведені вище. "Правильними" текстами в даному випадку будуть вважатися тексти, як російською, так і англійською мовами, які знаходяться у відкритому доступі в мережі Інтернет і що допускають комп'ютерну обробку. Апріорі вважатимемо, що ІР не містити неприпустимі для заданої мови символи або тексти з синтаксичними друкарськими

помилками. Усі спеціальні символи фільтруються. Виявлення синтаксичних друкарських помилок – окреме серйозне завдання, яке може бути відведене лінгвістичному експерту, який, у свою чергу, може внести виправлення до словника лем. Користуючись довідниками [5], вдалося написати алгоритм заповнення словника термінів і лем, які відповідають наведеним вище правилам.

У роботі застосовуватися комбінований підхід, заснований на статистичному методі, методі аналізу ознак та особливості DOM-моделей IP. Семантичний аналіз вимагає додаткового дослідження і не розглядатиметься в рамках поточної роботи. На рисунку 2.1 подано етапи процесу обробки змісту запитів ІК та текстів IP від первинного «брудного» тексту до лем. Результатом виконання цього процесу є формування статистики лем та нарощування словника за допомогою лінгвістичного експерта. Схема алгоритму лінгвістичної обробки термінів (слів) наведено рисунку 2.1.

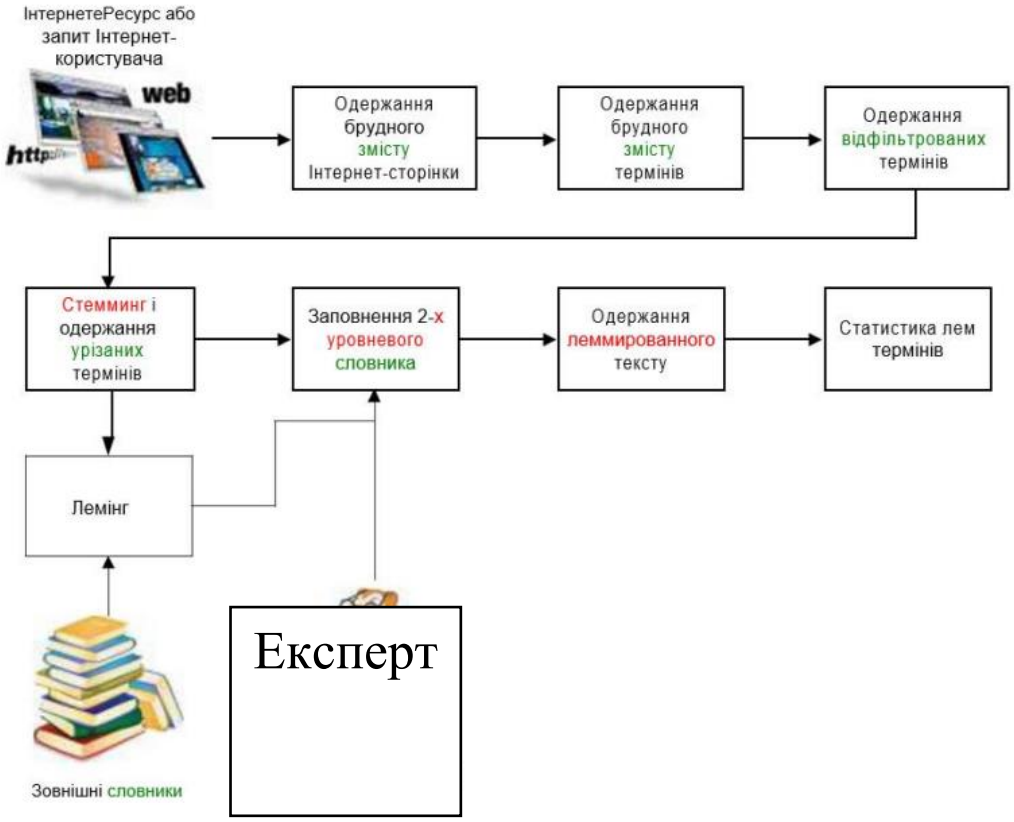

Рисунок 2.1 – Обробка запитів ІК та текстів IP

Необхідно звернути увагу на процес обробки термінів, зробить статистику термінів достовірнішою. Одними з визнаних лідерів серед цього процесу є програми, що реалізують алгоритми усічення закінчень [7]. Алгоритми усічення закінчень можуть бути використані як для російських, так і для англійських термінів, а отримані згодом усічені терміни можуть бути застосовані як леми для формування статистики первісних слів з ліченого тексту. На додаток алгоритм працює з безліччю закінчень, розбитим на кілька підмножин: закінчення герундія («в», «вошi», «вши» – деякі закінчення дієприслівників для російської мови або «ing» для англійської мови), закінчення прикметників («ими», «мі», «ієї»), причетні суфікси («ющ», «вш», «івш»), зворотні постфікси («сь», «ся»), дієслівні закінчення («ла», «чи», «єм»), префікси («до», «за», «над») та закінчення іменників. Процес стеммінгу легко реалізується з використанням БД, для чого достатньо скористатися спеціальним словником масок (маски на T-SQL представляються набором символів, в яких повинні бути символи «%» і/або «_»), який може бути поєднаний зі списком спеціальних символів фільтрації. Усічена форма термінів, що вийшла в результаті обробки, перевіряється по БД за допомогою словника лем. Якщо лему знайдено, то оригінальний термін прив’язується до знайденої леми. Якщо урізаному терміну неможливо було зіставити лему зі словника, він посимвольно порівнює з лемами зі словника. На останніх кроках алгоритму до словника лем включаються урізані терміни, яким неможливо було зіставити вже наявні леми.

Вони зазначаються, щоб лінгвістичний експерт зміг провести перевірку і, за необхідності, виправити чи призначити нові леми. Такий алгоритм сам по собі універсальний і дуже точний, однак і помилки, як показала практика, трапляються досить часто, наприклад, за наявності у слів приставок, що може призвести до отримання несловникового коріння або до надмірного усічення слів. Однак, цю проблему завжди може вирішити кваліфікований лінгвістичний експерт. У запропонованому алгоритмі достатньо

скористатися двома рівневим словником і для статистичної обробки текстів і для формування характеристичних векторів.

Якщо для обробки термінів будуть потрібні інші проміжні результати стеммінгу, то алгоритм потрібно буде переналаштувати на систему словників вищого рівня (наприклад, 3-х або 4-х рівневих словників), додаючи до нього блоки обробки додаткового словника. Резюмуючи все сказане вище, отримаємо наступну послідовність дій, щодо перетворення вихідного запиту ІП/тексту ІР на уявлення, придатне для подальшої обробки алгоритмами кластерного аналізу:

1) виділення всіх термінів із запиту ІП/тексту ІР;

2) стемінг та отримання урізаних термінів після видалення закінчень;

3) перевірка урізаних термінів за словником лем БД. Якщо лему знайдено, перехід до пункту «д»).

4) збереження позначених урізаних термінів, котрим не визначено лема з БД. При необхідності лінгвістичний експерт може підтвердити або змінити позначені урізані терміни, перетворюючи їх на нові леми;

5) формування статистики термінів та характеристичних векторів.

Після виконання перерахованих пунктів здійснюється перехід до кластерного аналізу лінгвістично підготовлених запитів Ікта текстів ІР.

3 РОЗРОБКА ТА РЕАЛІЗАЦІЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ІНТЕРНЕТ-ОБ'ЄКТІВ

3.1. Динамічні зміни в кластерній структурі Інтернет-об'єктів

Сучасні Інтернет ресурси є динамічними об'єктами. Якби вони містили виключно статичні компоненти, то розрахунок центрів відповідних кластерів можна було проводити в дискретні моменти часу, у моменти появи нових IP. Кластеризація Інтернет-користувачів також має динамічний характер, оскільки людська поведінка є динамічним процесом і це відображається в пошукових історіях ІП. У завданнях кластеризації ІК потрібно враховувати не тільки поточні характеристики об'єктів, що кластеризуються, але і їх тимчасові, тобто динамічні зміни за фактом появи нових пошукових запитів. У цьому розділі досліджуються динамічні зміни кластерної структури Інтернет-об'єктів.

Нехай X безліч всіх об'єктів, що спостерігаються $x_i \in X$, $1 \le i \le nof(X)$, віднесених до одного з кластерів $X_l \subseteq X$, $1 \le l \le nof(K)$, де $K = \{X_1, \ldots, X_l, \ldots, X_{nof}(K)\}$ – множина всіх сформованих кластерів. У різні моменти часу $t_k \in T$, k = 0,1,2, ... проводимо спостереження за зміною стану кластерної структури залежно від характеристик об'єктів xi, при цьому стан кожного і-го об'єкта у довільний момент часу tk відображається характеристичним вектором $z_i(T_k)$. Тут необхідно говорити про тимчасову складову як додатковий параметр для всіх елементів вектора, що характеризує об'єкт. Якщо об'єкт дослідження при ієрархічній кластеризації представлений вектором $z_i = (z_{i,1}, \ldots, z_{i,j}, \ldots, z_{i,n})$, який не залежить від часу, то в динамічній системі кластеризації необхідно говорити про вектор $z_i(t_k) = (z_{i,1}(t_k), \ldots, z_{i,j}(t_k), \ldots, z_{i,n}(t_k))$, координати якого прив'язані до моментів часу t_k. У будь-який фіксований момент часу t_k (або інтервал часу Δt_k) можна виділити кілька кластерів, всередині яких об'єкти мають загальними характеристиками. Зміна характеристик об'єкта $u_i \in U$ в

момент часу t_k може призвести до глобальних змін на рівні всієї кластерної структури і тим самим через період часу Δt_k буде необхідно провести нову кластеризацію всіх об'єктів з U. Якщо після закінчення часу Δt_k число кластерів, їх зміст, розміри та положення їх центрів не змінюються, то може йтися про так звану статичну кластерну структуру. Однак зовсім інакша справа в ситуаціях, коли з часом кластерна структура змінюється, коли об'єкти з часом починають володіти деякими новими характеристиками і утворюють групу об'єктів, функціонування яких знаходиться на межі кластерів або навіть за його межами. У такому разі кластерна структура зазнає тимчасових змін і стає динамічною.

Зміни у структурі кластерів: освіта нових кластерів. Для безлічі характеристичних векторів ІК $U_{(tk+1)}$ в момент часу $t_k+1 > t_k$ утворення нових кластерів може бути пов'язане з появою нових об'єктів або з різкою стійкою зміною пошукової активності існуючих об'єктів.

Для першого випадку кожен новий об'єкт дослідження

$$u_{nof(U(tk))+p}(t_{k+1}) \in U(t_{k+1}), \quad p \geq 1,$$

являє собою новий, одиночний ізольований кластер, для якого проводиться розрахунок міри близькості (евклідового) відстані) до решти об'єктів

$$U(t_{k+1}) \setminus \{u_{nof(U(tk))+p}(t_{k+1})\}.$$

За результатами розрахунку евклідової відстані визначається найближчий де

$$U_{nof}(U(t\_k))\} + \mathrm{p}(t_{k+1}) \quad \text{об'єкт} \quad U_{near}(t_{k+1})$$

$$\rho(u_{nof(U(t_k))+p}(t_{k+1}),\ u_{near}(t_{k+1})) = \min_{1 \le i \le nof\ (U(t_k))} \rho\left(u_{nof(U(t_k))+p}(t_{k+1}), u_i(t_{k+1})\right).$$

Визначення найближчого сусіда $U_{near}(t_k+1)$ дозволить ініціалізувати місце розташування нового об'єкта у новій кластерній структурі. Продовжуємо спостерігати за Інтернет-користувачем u_i в період Δt_k. З появою нових об'єктів ІК зростає глобальний словник термінів Vu. Якщо на момент часу t_{k+1} (у попередній інтервал часу Δt_k) ІК $U_{i(tk+1)}$не проводив жодної пошукової діяльності, то збільшення словника термінів V_u ніяк не відбивається на його характеристичному векторі, а лише призводить до появи нових нульових координат: якщо в момент часу t_k характеристичний вектор ІК мав такий вигляд

$$u_i(t_k) = \left(u_{i,1}(t_k), u_{i,2}(t_k), \ldots u_{i,nof(V_u)}(t_k)\right),$$

то в момент часу tk+1 він перетворюється на вигляд:

$$u_i(t_{k+1}) = \left(u_{i,1}(t_{k+1}), u_{i,2}(t_{k+1}), \ldots u_{i,nof(V)}(t_k), 0, 0, \ldots, 0\right).$$

На рисунку 3.1 представлена ілюстрація початкової кластерної структури на момент часу t_k.

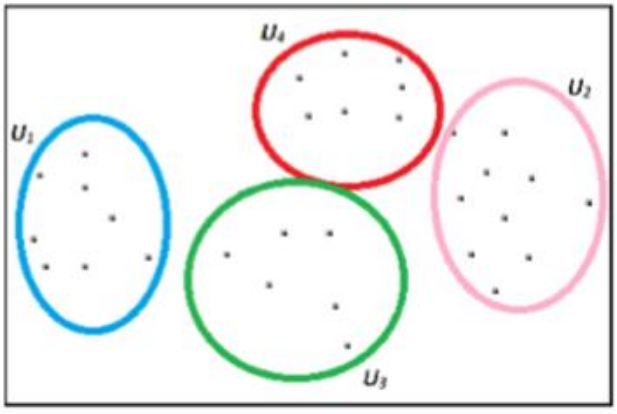

Рисунок 3.1 – Ілюстрація розподілу об'єктів за кластерами у момент часу t_k

Якщо експериментальним шляхом в інтервалі часу Δtk з'явилися нові об'єкти

$$u_{no\beta(U(t_k))+p}(t_{k+1}) \in U(t_{k+1}), \quad p \geq 1,$$

після проведення повторної кластеризації з'являються нові кластери (рисунок 3.2), сформовані за допомогою нових об'єктів.

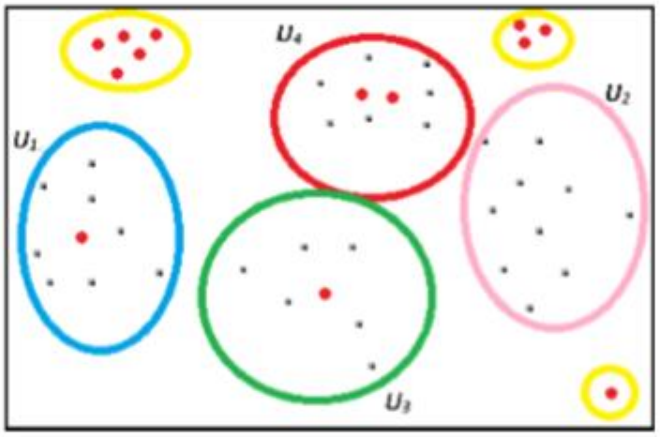

Рисунок 3.2 – Ілюстрація розподілу об'єктів за кластерами у момент часу t_{k+1}

На рисунку 3.2 можна помітити формування нових кластерів з об'єктів, що з'явилися, одиночні кластери, кардинальне число яких дорівнює 1, та насичення (нарощування) вже сформованих кластерів за рахунок надходження нових об'єктів. Якщо на момент часу t_{k+2} пошукова діяльність одного або кількох користувачів різко змінюється, характеристичні вектори цих об'єктів змінюють значення одного або кількох своїх координат і, як наслідок, структура кластера змінюється, що, у свою чергу, призводить до зміни значення міри близькості та відстані між об'єктами одного кластеру Це може спричинити формування нових кластерів (рисунок 3.3).

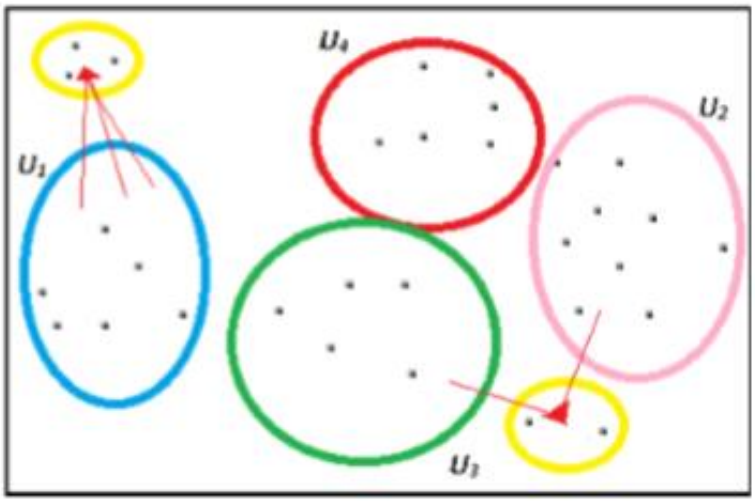

Рисунок 3.3 – Ілюстрація перерозподілу об'єктів у час tk+2

Для вирішення завдання появи нових кластерів можна скористатися коефіцієнтом належності і-ого об'єкта до m-го кластера. Коефіцієнт приналежності $b_{i,m(tk+2)}$ об'єкта $u_{i(tk+2)}$ до кластера $U_{m(tk+2)}$ із кластерної структури $K_{(tk+2)}$ ($Um(tk+2) \in K(tk+2)$) довільний момент часу t_{k+2}:

$$b_{i,m}(t_{k+2}) = \frac{1}{\sum_{l=1}^{nof(K(t_{k+2}))} \left(\frac{(\rho_m(t_{k+2}))^2}{(\rho_l(t_{k+2}))^2} \right)} \quad \text{и} \quad \sum_{m=1}^{nof(K(t_{k+2}))} b_{i,m}(t_{k+2}) = 1, \qquad (3.1)$$

де $nof(K(t_{k+2}))$ – кількість кластерів у кластерній структурі $K(t_{k+2})$;

$$\rho_m(t_{k+2}) = \sqrt{\sum_{j=1}^{nof(V_u(t_{k+2}))} \left(e_{m,j}(t_{k+2}) - u_{i,j}(t_{k+2}) \right)^2} -$$

евклідова відстань між об'єктом u_i та центром e_m m-ого кластера кластерної структури ІП.

Виявлення множини нових кластерів K_{new} починається з виявлення множини нових об'єктів U_{free} з малими ступенями приналежності до всіх

існуючих кластерів. Якщо кількість таких вільних об'єктів $_{nof}(U_{free})$ можна порівняти з розмірами кластерів, вони формують компактну групу об'єктів із загальними властивостями. Компактність об'єктів, визначена нижче (3.6) є ознакою появи нових кластерів. Можливе число нових кластерів $_{nof}(K_{new})$ в момент часу t_k визначається співвідношенням:

$$nof\left(K^{new}\right) = int\left(\frac{nof\left(U^{free}\right)}{d_1 \times N_{min}}\right), \qquad (3.2)$$

де int(...) – ціла частина аргументу;

$_{nof}(U^{free})$ – кількість вільних об'єктів, які не прив'язані до жодного з кластерів;

d1 – задана гранична величина в інтервалі [0,1];

N_{min} = min ($_{nof}$ (U_1^*), …, $_{nof}$ (U_2^*),…, nof ($U_{nof(K)}^*$))– мінімальний розмір кластера, при обчисленні якого враховуються тільки «хороші» об'єкти (U_1^*) ⊆ Ul з досить великими значеннями ступенів належності;

$_{nof}$ (U_1^*), – число добрих об'єктів l-ого кластера, для яких значення ступеня приналежності до зазначеного кластера $b_{i,l} \geq$ d2;

d2 – задана гранична величина в інтервалі [0,1].

Зміни в структурі кластерів це злиття кластерів. За визначенням, злиття кластерів – це формування нового розбиття, коли $_{nof}$(K')-а кількість кластерів перетворюються на $_{nof}$(K'') кількість кластерів, причому $_{nof}$(K'') < $_{nof}$(K')).

Нехай злиття кластерів відбувається у час t_{k+3}. Зі зміною вектора пошуку деякі кластери можуть наблизитися один до одного і злитися в єдину групу (рисунок 3.4).

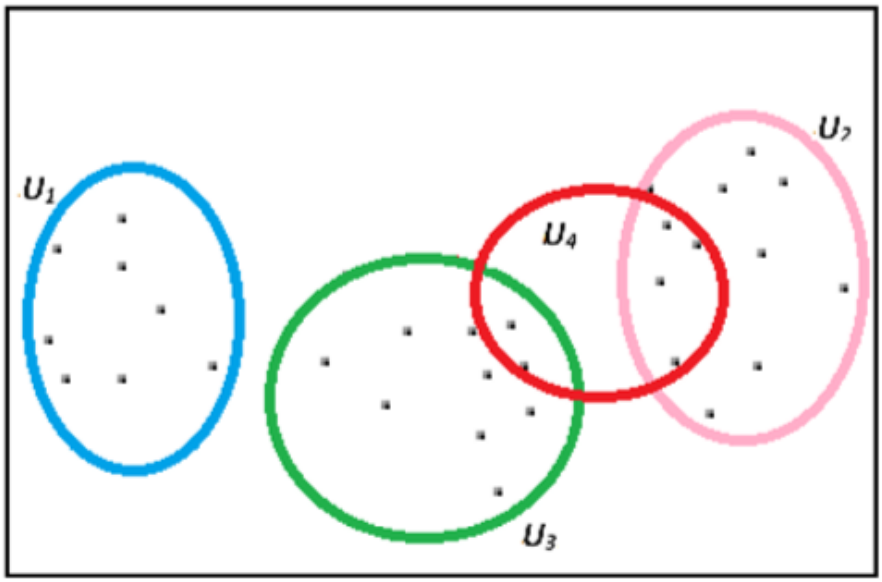

Рисунок 3.4 – Ілюстрація злиття кластера U_4 та перерозподіл його об'єктів та центру між кластерами U_3 та U_2 на t_{k+3}

Виявлення кластерів, що зливаються, починається з виділення об'єктів u_i, що мають високі ступені приналежності (3.1) одночасно для двох кластерів U_l і U_m : $b_{i,l} \approx b_{i,m} \rightarrow 1$.

Якщо число таких об'єктів, що зливаються, досить велике, то це і є ознакою злиття кластерів . Для ітеративних методів кластеризації (методу k-середніх або методу Форель) спостерігається зближення центрів кластерів між кластерами, що зливаються. Кількісним критерієм для оголошення двох кластерів кластерами, що зливаються, може служити міра їх подібності:

$$I_{l,m} = \frac{\sum_{i=1}^{nof(U)} \min(b_{i,l}, b_{i,m})}{\sum_{i=1}^{nof(U)} b_{i,l}}, \qquad (3.3)$$

де $b_{i,l}$ і $b_{i,m}$ – ступеня приналежності i-го об'єкта до кластерів U_l і U_m, відповідно. Однак міра подібності $I_{l,m}$ не є симетричною, оскільки $I_{l,m} \neq c$, тому для виявлення зазначеної подібності краще використовувати міру відповідно до якої кластери будуть вважатися такими, що зливаються, якщо значення $Mc_{l,m}$ перевищує деякий поріг h (при h = 0 всі кластери будуть

вважатися такими, що злилися, при h = 1 кластерів, що злилися, ніколи не буде).

$$Mc_{lm} = \max (I_{l,m}, I_{l,m}), \quad (3.4)$$

Якщо говорити про ієрархічну кластеризацію, то спостерігати за картиною злиття кластерів краще за все на ранніх стадіях їх формування, тому що на пізніших етапах відбувається збільшення кількості об'єктів у кластерах і як наслідок ймовірність повного злиття великих кластерів знижується. Тоді можна буде говорити про розщеплення або дроблення кластерів. Злиття кластерів в ітераційних методах кластеризації має певні недоліки – не враховується форма кластерів та близькість їхніх центрів.

3.2 Зміни у структурі кластерів

Розщеплення або дроблення кластерів можна спостерігати, наприклад, у момент часу t_{k+4}, коли деякі кластери збільшуються у розмірах через велику кількість нових об'єктів, що може призвести до неоднорідності їх внутрішньої структури (рисунок 3.5).

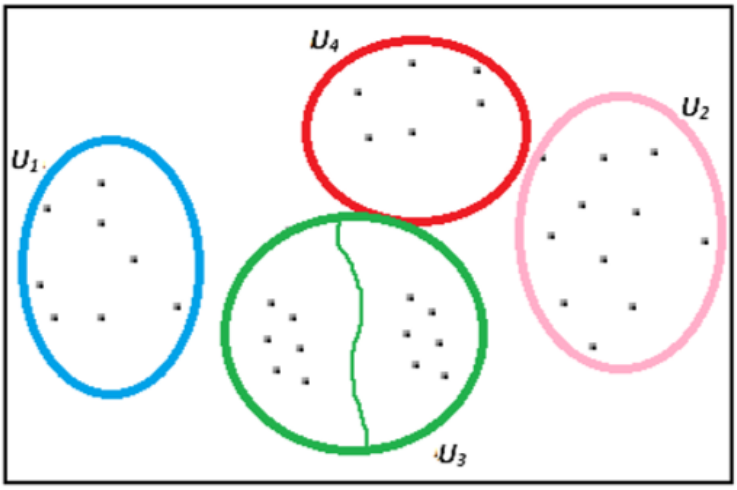

Рисунок 3.5 – Ілюстрація розщеплення кластера U3 та формування всередині нього двох розділених згустків у tk+4

У ієрархічної кластеризації розщеплення чи дроблення кластерів можна виявити більш пізніх етапах, тому що на ранніх стадіях кількість об'єктів та розмірність кластерів не дозволить спостерігати над процесом зміни однорідності малих кластерів. Це не тільки з розмірами кластерів, склад яких може змінитися з часом, але й характером самого алгоритму агломеративної кластеризації. На ранніх стадіях доцільніше говорити про злиття кластерів в ієрархічній системі. Зміни у структурі кластерів: зникнення кластерів. Зникнення кластерів безпосередньо пов'язане зі зникненням об'єктів цих кластерів або з перетворенням їх характеристичних векторів на нульові вектори. Зникнення кластера відбувається, коли спостерігається повний перехід його об'єктів до складу іншого (інших) кластерів (кластерів).

Для дослідження динамічних ефектів у кластерних структурах, було обрано 30 користувачів віком від 18 до 44 років, які проживають на території України (найактивніша група за даними дослідницької компанії TNS [8]) та 100 інформаційних ресурсів, що належать групі сайтів з різною тематикою. Таким чином, застосовано попереднє угруповання об'єктів за статичною інформацією: стать, вік та місце проживання для ІК та тематика для ІР.

За допомогою формули (3.1) експериментально були визначені показники належності, що відображають стан кластерної структури для випадково обраного ІК з обраної групи (таблиця 3.1). Отримані результати графічно інтерпретовані на рисунку 3.7. Графіки підтверджують динамічний характер пошукової активності користувачів: залежно від часу змінюється пошуковий інтерес користувача і, як наслідок, його приналежність до кластерів.

Таблиця 3.1 – Коефіцієнти належності користувачів до кластерів у різні моменти часу

Кластер / Момент часу	U_1	U_2	U_3	U_4	Кластер / Момент часу	U_1	U_2	U_3	U_4
8	0	0	0	0	20	0	0	0	0
9	0	0	0	0	21	0	0	0	0
10	0,5104	0,2713	0,0597	0,1586	22	0,2007	0,5915	0,1928	0,015
11	0,6061	0,1870	0,1559	0,0510	23	0,4092	0,4497	0,1294	0,0117
12	0,1775	0,1297	0,2011	0,4917	24	0,3907	0,3057	0,0937	0,2099
13	0,1896	0,1582	0,3959	0,2563	1	0	0	0	0
14	0,0436	0,1192	0,5991	0,2381	2	0	0	0	0
15	0,0574	0,0857	0,0857	0,3056	3	0	0	0	0
16	0,1601	0,2294	0,3491	0,2614	4	0	0	0	0
17	0,2151	0,16	0,2967	0,3185	5	0	0	0	0
18	0,0653	0,1196	0,4892	0,3259	6	0	0	0	0
19	0,2448	0,0674	0,4691	0,2187	7	0	0	0	0

Результати розрахунку коефіцієнтів належності ресурсу до кластерів (таблиця 3.2) підтверджують динамічність показників ресурсів.

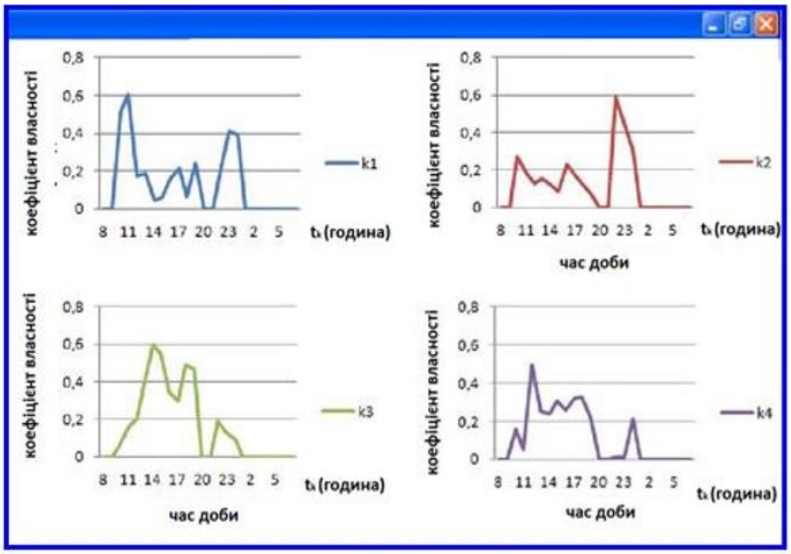

Рисунок 3.7 – Графіки зміни коефіцієнта належності користувача для різних кластерів у різні моменти часу

На рисунку 3.8 побудовані за даними таблиці 3.2 графіки схрещуються, тобто в різні моменти часу Інтернет-ресурс, що спостерігається, належить різним кластерам.

Для усунення цього ефекту пропонується використовувати DOM-

модель ресурсів з метою їхнього «очищення» від динамічних компонентів: інформації партнерських мереж, різноманітних повідомлень, реклами тощо. Поряд з цим DOM-модель може бути використана для розрахунку вагових коефіцієнтів характеристичний вектор ресурсу.

Таблиця 3.2 – Коефіцієнти належності ресурсу до кластерів у різні моменти часу без застосування вагових коефіцієнтів посилення

Кластер Момент часу	W_1	W_2	W_3	Кластер Момент часу	W_1	W_2	W_3
8	0,483	0,208	0,31	20	0,303	0,339	0,358
9	0,382	0,397	0,221	21	0,411	0,289	0,300
10	0,419	0,307	0,0597	22	0,419	0,3	0,281
11	0,362	0,306	0,332	23	0,308	0,352	0,34
12	0,233	0,338	0,429	24	0,15	0,46	0,389
13	0,31	0,363	0,327	1	0,31	0,275	0,416
14	0,309	0,441	0,25	2	0,331	0,343	0,326
15	0,393	0,257	0,351	3	0,377	0,282	0,341
16	0,45	0,313	0,238	4	0,358	0,267	0,375
17	0,401	0,173	0,426	5	0,279	0,451	0,270
18	0,311	0,386	0,303	6	0,352	0,308	0,341
19	0,157	0,449	0,394	7	0,345	0,297	0,358

Враховуючи особливості DOM-моделі Інтернет-сторінок ресурсу, кожному елементу вектора wi(tk) можна порівняти ваговий коефіцієнт, розрахований за формулою

$$\frac{w_p \times k_1}{nW} \times k_2 \times 100, \qquad (3.8)$$

де nW – загальна кількість слів у DOM-моделі сторінки;

w_p – число входжень слова на p-ой позиції у конкретній тезі DOM-моделі сторінки;

k_1 - коефіцієнт посилення ($k_1 > 1$), значення якого розраховується, виходячи з найменування тега, що визначає контекст слова на сторінці;

k_2 – коефіцієнт посилення ($k_2 \geq 1$), значення якого розраховується за формулою відносин площ, які займають слова на сторінці:

$$k_2 = \frac{S_p / cnt_p}{S_{total} / nW},$$

де S_p – Площа області на p-ій позиції;

cnt_p – кількість слів на p-ій позиції;

S_{total} – площа інформаційного тексту;

nW – загальна кількість слів у *DOM*-моделі сторінки.

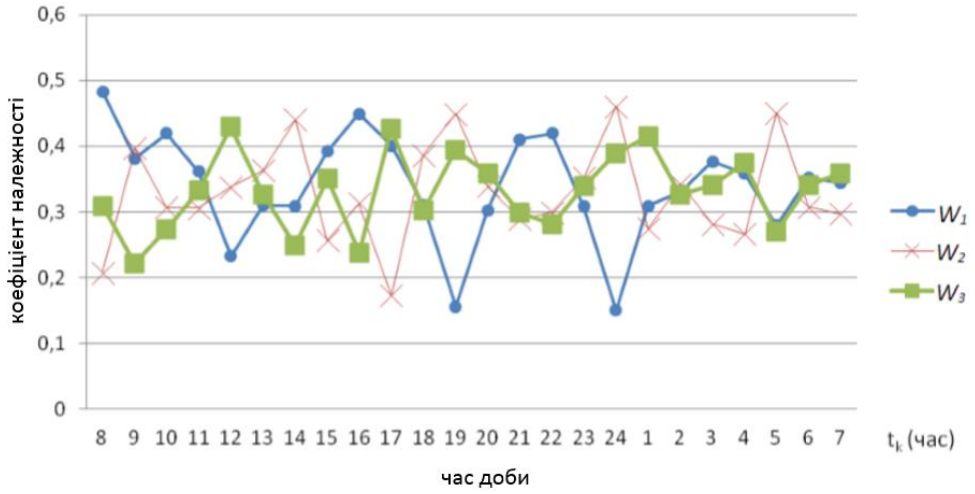

Рисунок 3.8 – Графіки коефіцієнтів належності ресурсу для різних кластерів у різні моменти часу без використання вагових коефіцієнтів посилення

Проведемо повторний розрахунок ступенів належного ресурсу (таблиця 3.3) із застосуванням вагових коефіцієнтів, розрахованих за формулою 3.8 і отримаємо нові графіки (рисунок 3.8).

3.3. Концепція побудови системи персоналізації Інтернет-пошуку

Експериментальні дослідження дозволили визначити тимчасові та обчислювальні витрати на виконання кластерного аналізу реальних масивів ІК та ІР, що дозволило, у свою чергу, оцінити, які програмні модулі КСПП

мають бути встановлені на персональні комп'ютери (ПК) користувачів, а які – на сервери підприємства.

Вже було зазначено, що пошукові запити ІК можна збирати миттєво, саме в ті моменти часу, коли починається їхня пошукова діяльність. Для цього було створено програмний модуль internet_res_search, який дозволяє не тільки відстежувати пошукові запити ІП, а й визначати глибину пошуку, надаючи інформацію про результати пошукової віддачі ситеми. Крім пошукової діяльності, психологічний портрет ІК також формується внаслідок виконання заходів та відвідувань ним ІР.

У зв'язку з цим було реалізовано програмний модуль ie_analyzer, що забезпечує автоматичне стеження та формування Log-файлу з хронологічним списком відвідуваних ІК сторінок. Обробка змісту ІР виявилася досить затратною за часом. Сучасні ІР містять величезну кількість динамічних компонентів в DOM моделі, які можуть постійно змінювати свій зміст, крім того, наявність великих медіа-файлів може сильно збільшити час повного завантаження та читання змісту DOM-моделі ресурсу.

Деякі ресурси можуть містити посилання на неіснуючі елементи на сервері самого ІР, і браузер намагатиметься завантажити їх до моменту timeout. У процесі експериментальних досліджень середній час сканування DOM-моделі однієї сторінки ІР становив 6-7 секунд. Якщо пошукова система, як результат пошукового запиту, видає лише 10 гіперпосилань на знайдені ресурси, то для читання змісту їх DOM-моделей потрібно більше однієї хвилини, що є неприпустимим часом для будь-якого користувача - зайві тимчасові витрати на очікування призводять до серйозних психічних витрат, можуть сильно дратувати людину.

Процес кластерного аналізу ІК та ІР, заснований на виконанні розроблених методів, містить багато етапів (тимчасові вікна для ІП, застосування числових коефіцієнтів посилення з DOM-моделі ІР, боротьба з динамічними елементами ІР та застосування узагальненого характеристичного вектора на основі глобального словника термінів), при

цьому відповідні процедури обробки даних досить час витратні. Так як кластерний аналіз результатів 4-х годинного періоду спостережень займає приблизно 20 хвилин часу (що неприпустимо багато для будь-якого ІП), доцільно виділити спеціальну серверну обчислювальну систему (сервер) під кластерним аналізом Інтернет-об'єктів. З урахуванням вище зазначеного, на рисунку 3.9 представлена узагальнена структура КСПП, що відображає структурну організацію системи. Програмні модулі безпосереднього спостереження internet_res_search та ie_analyzer мають бути встановлені на клієнтських ПК. Програмний модуль internet_res_search реалізовано та підігнано до роботи з пошуковою системою *search*. Справа в тому, що пошукова система *search* вивантажує результати пошуку в HTML-коді самої сторінки, тому немає необхідності перевіряти DOM-модель кожної сторінки, що знижує в рази обчислювальні витрати на читання та аналіз сторінок, отриманих в результаті пошуку.

Усі кластерні розрахунки та вся аналітика мають бути реалізовані програмними модулями, встановленими на корпоративних серверах. Для цієї мети були розроблені спеціальні розрахункові процедури та фільтруючі функції, а також спроектована структура БД InternetDB, яка підтримується системою управління MS SQL Server 2022.

Таким чином, для персоналізації Інтернет-пошуку працівників підприємства, в рамках його корпоративної інформаційно-обчислювальної системи має бути створена триланкова КСВП. На рисунку 3.9 представлені всі три взаємопов'язані ланки корпоративної системи персоналізації пошуку: перша ланка – множина ІП, друга ланка – сервер кластерного аналізу та третя ланка – сервер БД. До мережі Інтернет система підключається через сервер кластерного аналізу. На жаль, у процесі виконання роботи реальну систему, структура якої відповідала б узагальненій структурі рисунка 3.9, створити не вдалося.

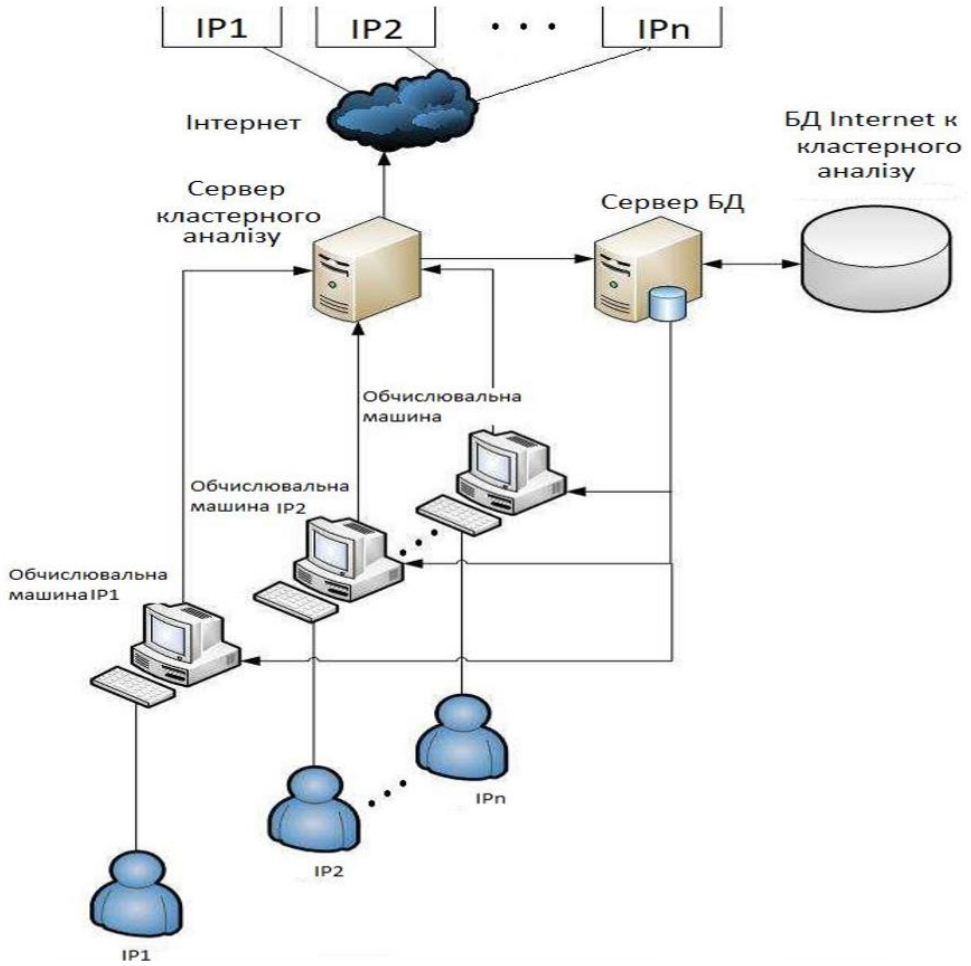

Рисунок 3.9 – Узагальнена структура системи персоналізації

Завдання реалізації та тестування запропонованих методів кластерного аналізу вирішувалися на одній потужній обчислювальній установці, здатній підтримувати безліч віртуальних машин. Структура віртуальної КСВП повною мірою відповідала узагальненій структурі.

3.4. Структуризація даних про пошукову активність Інтернет-користувачів

Великі пошукові системи, такі як *Search* або Google, користуються файлами cookie у браузерах та особистою інформацією, залишеною ІК для персоналізації пошуку. Пошукові системи добре працюють із регіональними запитами при пошуку магазинів та споживчих товарів. Застосування

регіональної та соціально-демографічної інформації за сайтів є статичним методом персоналізації Інтернет контенту, тобто. користувач (його профіль та/або IP-адреса) поміщається в спеціальну базу даних, звідки і починається персоналізація вмісту сайтів. Незважаючи на те, що сучасні пошукові технології повністю використовують статичну інформацію про ІП, результати пошуку залишають бажати кращого. Кожен ІК має свій особистий психологічний портрет та протягом дня відвідує певний, часто фіксований набір веб-сторінок. Будь-який ІК інтуїтивно формує власну систему класифікації та відбору ІР для задоволення своїх потреб в інформації. Незважаючи на те, що ІР проводять свою систему аналізу поведінки ІП, пошукова система не здатна проводити класифікацію ІР для кожного ІК окремо. Користувачі зацікавлені в тому, щоб пошукові системи аналізували їхню пошукову активність, а не тільки статичну інформацію про поле, вік, місцезнаходження тощо.

Для дослідження пошукової активності Інтернет-користувачів у рамках даної роботи необхідно спроектувати БД, що зберігає інформацію про заходи ІК на ІР, а також словник термінів, які містяться у відповідних пошукових запитах. Структура БД заходів Інтернет-користувачів (Internet DB). Пошукова активність ІК відстежується за допомогою спеціального програмного модуля, який записує дані про його дії у Log-файл. Як тільки ІК виконує захід на ІР або залишає ІР, його дії автоматично фіксується та зберігається у спеціальному Log-файлі формат та приклад вмісту якого демонструються таблицею 4.1.

Таблиця 4.1 – Формат файлу заходів ІП

resp_id	comp_id	D	url
1	1	2023-10-20T17:19:57.297	https://www.facebook.com/feed
1	1	2023-10-20T17: 19:59.733	https://www.facebook.com/audio
2	2	2023-10-20T00:03:23.470	https://www.facebook.com/id41028953
2	2	2023-10-20T00:03:58.513	https://www.facebook.com/id89733988
2	3	2023-10-20T00:05:56.807	https://mail.ukr.net/
2	3	2023-10-20T00:05:56.983	https://mail.ukr.net/

Так як формат файлу заходів ІК відомий, можна переступити до проектування БД для обробки та аналізу даних, що містяться в Log-файлі. Зрештою, накопичені в Log-файлах дані трасування заходів ІК мають бути завантажені в БД для подальшої кластеризації ІК щодо їхньої поведінки в Інтернеті. Починаємо з визначення сутностей. На першому кроці (рисунок 4.2) можна виділити дві сутності, що характеризують ІП: власне «Інтернеткористувач» (az_resps) та «місто» (az_cities).

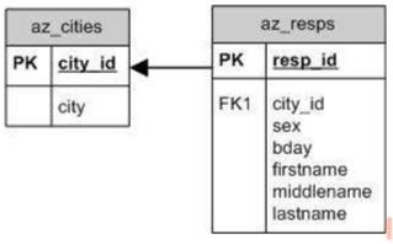

Рисунок 3.10 – Сутності az_resps, az_cities та логічний зв'язок між ними

Завантаживши дані в таблиці az_resps і az_cities, можна порушити створення сутності «захід» (az_visits), показаної рисунок 3.11. Вихідні дані для таблиці заходів відображають заходи ІК на 200 найбільших Інтернет-сайтів, у тому числі на портали google.com, ukr.net, gmail.com, новинні сайти та в різні соціальні мережі.

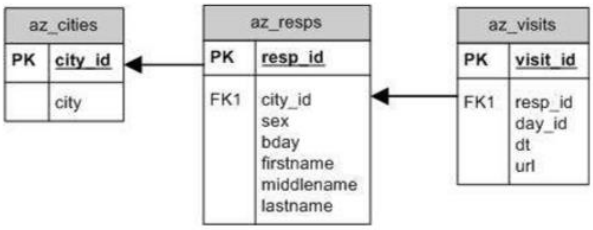

Рисунок 3.11 – Додавання сутності az_visits

Щодня ІК можуть відвідувати понад 2000 сторінок, із них понад 300 сторінок пошукові – запропоновані ІК пошуковими системами. Це зумовлює можливість одразу аналізувати як відвідуваність сайтів, так і пошукову

історію ІП. Таким чином, БД заходів ІК дозволить виявити їхню поведінку, інтереси і, як результат, дозволить згрупувати залежно від інтересів та пошукової спрямованості. Як уже було сказано, IP проводять персоналізацію Інтернет контенту на основі статичної інформації, залишеної відвідувачами сайтів: стать, дата народження, вік та місцезнаходження (легко визначається за IP-адресою). Маркетологи, застосовуючи спеціальні підходи, залежно від статі та віку розбивають ІК на соціальні групи [10]. Тому структуру БД необхідно додати сутність (рисунок 3.12) «соціальна група» az_sd. Щоб зв'язати сутності «Інтернет-користувач» і «соціальна група», створюємо додаткову асоціативну сутність (рисунок 3.13) az_resp_sd. Застосування цього підходу дозволить розподілити обробку заходів з різних баз даних чи серверів.

Алгоритм аналізу та перетворення пошукових рядків, сформованих пошуковими системами, можна розбити на послідовність кроків (рисунок 3.13): отримання доменного імені, декодування, виявлення масок, визначення ключових пошукових слів та поділ пошукового виразу на окремі слова.

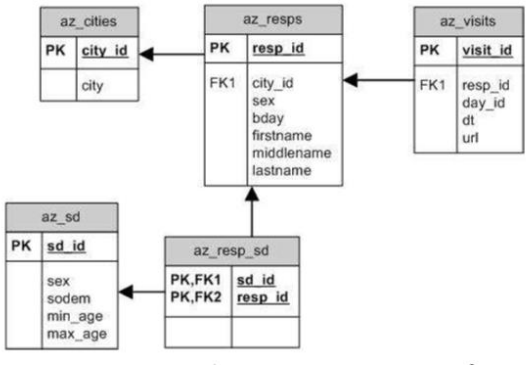

Рисунок 3.12 – Додавання сутності az_resp_sd

На рисунку 3.13 показано п'ять основних кроків обробки пошукової URL-рядку для отримання кінцевих термінів з пошукових рядків. Для кожного кроку окремо розглянемо, яких змін він вимагатиме у структурі БД Internet DB. Перетворення URL та отримання домену сайту. Почнемо з

перетворення даних заходів із обробки URL-сторінок. З цією метою проведемо свого роду реорганізацію отриманих URL. З одного боку, різні ІК можуть відвідувати ту саму сторінку, з іншого боку, URL, на які ІК виконують заходи, мають різні структури та довжини.

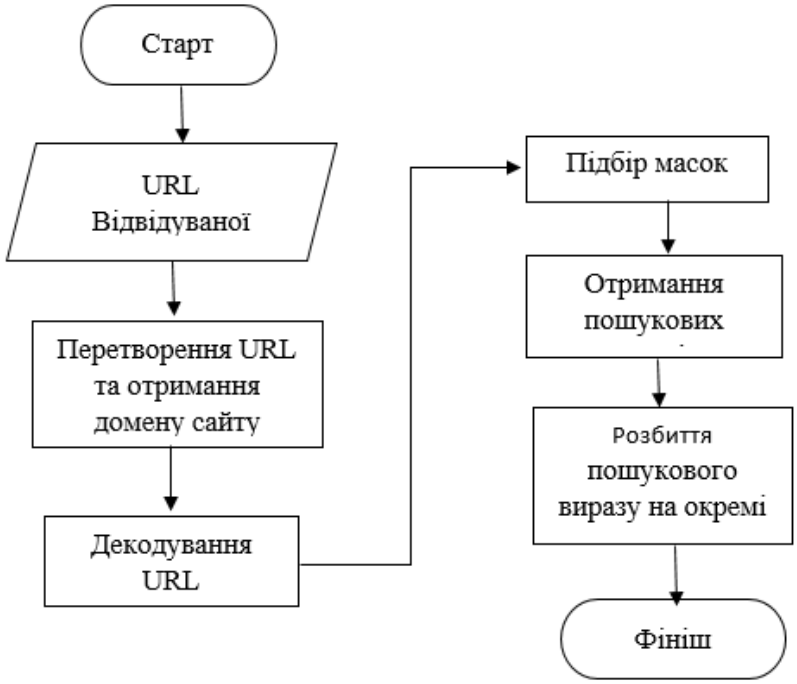

Рисунок 3.13 – Схема алгоритму отримання кінцевих термінів із пошукових рядків

Тому є сенс спочатку визначити унікальні URL, потім знайти їхні доменні імена та згрупувати по сайтах. Тут потрібно розробити спеціальну функцію fnGetDomainFromUrl (вихідний код представлений у додатку Б) для отримання доменного імені другого та третього рівня. Для збереження унікальних URL необхідно створити нову сутність «сторінка» (az_pages), а для збереження доменів – сутність «домен» (az_domain). Вносимо відповідні зміни до структури БД (рисунок 3.14).

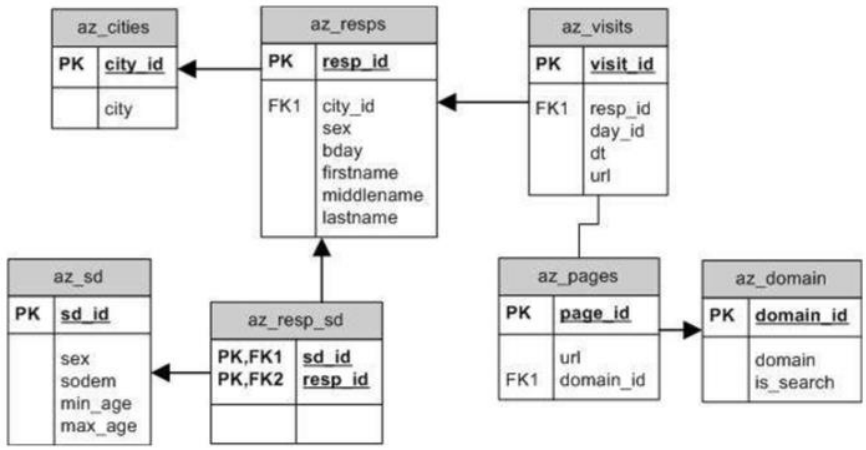

Рисунок 3.14 – Додавання сутностей az_pages та az_domain

Декодування URL у ядрі сервера БД. Працюючи з URL ресурсів, є необхідність виділити пошукові запити, які надають насичену базу дослідження інтересів ИП. Пошукові запити передаються на сервер пошукової системи (search.com.ua) шляхом інкапсуляції пошукового тексту в самій URL-сторінки, із застосуванням спеціального кодування. Зупинимося на способі кодування пошукового рядка для основних пошукових систем у укрнету - search.com.ua, ukr.net тощо. Найбільш поширеними кодуваннями є UTF-8 та ASCII та Windows-1251. Декодування пошукових рядків можна проводити за допомогою web-сервісу «Універсальний декодер – конвертор кирилиці», який є доступним за посиланням http://2cyr.com/decode/?lang=ua (10.10.2023 р.).

Для search.com.ua інкапсульований пошуковий рядок кодований за допомогою Windows-1251 має вигляд:

Лістинг 3.1 – Пошуковий рядок

```
http://search.com.ua/searchsearch?text=%D0%BF%D0%BE%D0%B4%D1%80%
D1%83%D0%B6%D0%BA%D0%B0+%D0%BC%D0%B0%D0%B3%D0%B0%D0%B7%D0%B8%D0%
BD&lr=216&oprn d=1376222967
тобто. пошукові терміни «подружка+магазин» були закодовані як
%D0%BF%D0%BE%D0%B4%D1%80%D1%83%D0%B6%D0%BA%D0%B0+%D0%BC%D0%B0%D0
%B3% D0%B0%D0%B7%D0%B8%D0%BD .
```

Як виявилося, основні пошукові системи укрнета застосовують кодування Windows-1251 для пошукових виразів. Для автоматизації процесу декодування пошукових рядків необхідно написати і потім підключити спеціальну бібліотеку динамічного компонування HttpUtility.dll.

Нова бібліотека динамічного компонування має бути інтегрована в ядро сервера MS SQL Server 2022 для можливості подальшого застосування функцій кодування (Encode(_)) та декодування (Decode(_)). З появою можливості декодування пошукового рядка необхідно внести зміни до структури таблиці az_pages (рисунок 3.7) – додається атрибут декодованого пошукового рядка decoded_url.

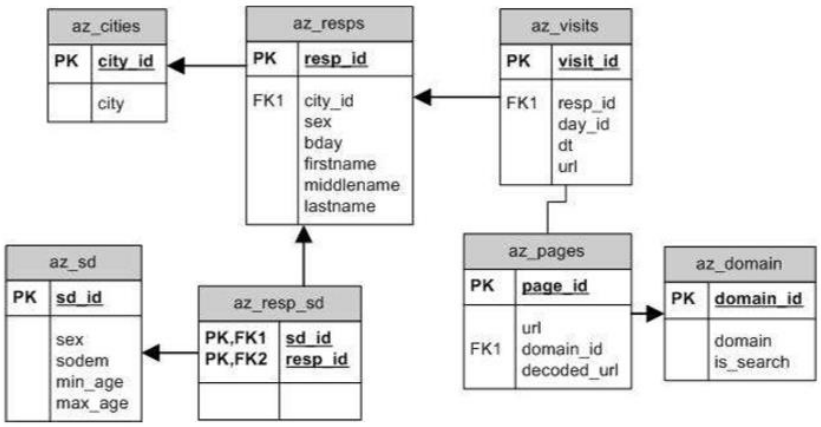

Рисунок 3.15 – Додавання атрибуту decoded_url до сутності az_pages

Вибір масок. Отримавши декодований пошуковий рядок, звертаємо увагу на те, що в додатку до пошукових термінів з'являються допоміжні «сміттєві» терміни, тому необхідно формування пошукових масок-URL з метою усунення «сміття». Проведемо статистичний аналіз рядкових фрагментів пошукового рядка. Після проведеного статистичного аналізу пошукових рядків можна сформувати узагальнені маски для пошукових сайтів, що розглядаються. Маски пошукових сайтів сформовані та з'являється необхідність додавання двох нових сутностей – «маска» (az_mask) та «маска-домен» (az_domain_mask) до структури БД (рисунок 3.16).

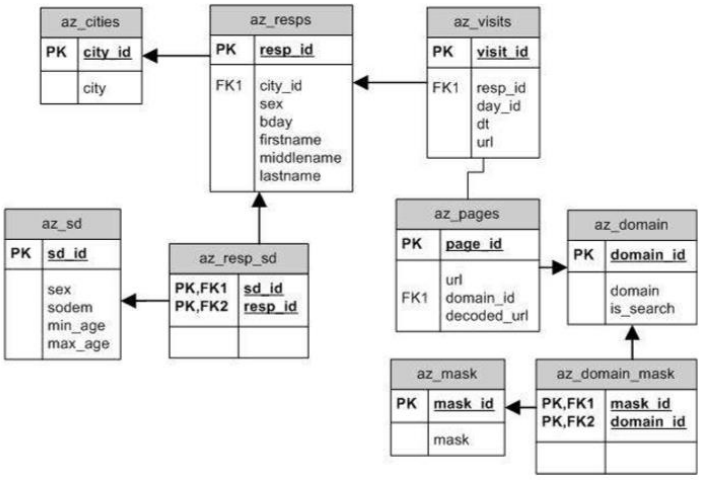

Рисунок 3.16 – Додавання сушіння az_mask та az_domain_mask

Отримання ключових пошукових виразів. Пошукові маски виявлені та завантажені. Можемо розпочинати формування статистики згаданих у запитах термінів. Для цього необхідно перейти до обробки пошукових URL, виділення пошукових термінів, їх лематизації і потім до розрахунку статистики термінів, що зустрічаються. При розгляді URL, визначилися ключові вирази, після яких і розташовуються терміни користувачів. Для кожного сайту свій ключовий вираз. Враховуючи сказане, необхідно додати сутність «ключовий вираз» (az_key_word) до структури БД (рисунок 3.17).

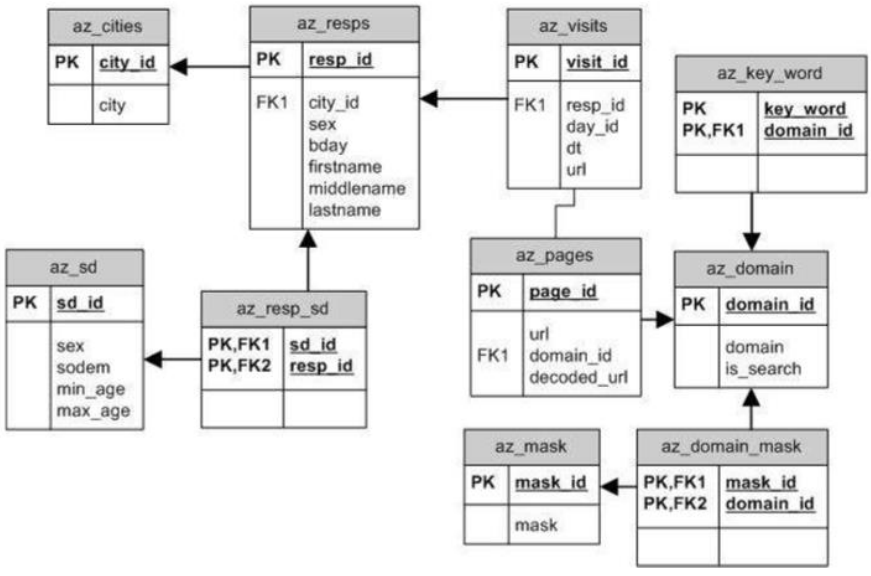

Рисунок 3.17 – Додавання сутності az_key_word

Розбиття пошукового виразу окремі слова. На цьому етапі обробки пошукового рядка значення стовпця decoded_url таблиці az_pages буде складатися зі спеціальних знаків або пробілів і набору слів у пошуковому рядку (#пробіл<слово1> . <слово2> ; <слово3> ,#). Для отримання окремих слів з декодованого та перетвореного URL-рядка (decoded_url) потрібна спеціальна функція az_split() (початковий код у додатку Б), яка розбиває довгий рядок на окремі слова, якщо вірно вказано роздільник між словами. Після того, як пошукові слова виявлені, приступимо до формування таблиці статистики слів. Для цього до БД додамо дві нові сутності: «слово» (az_words) та асоціативну сутність «словопошуковий рядок» (az_pages_words) (рисунок 3.18).

Відсутність унікальних ключів у таблиці az_pages_words пов'язано з можливістю появи одного й того ж слова кілька разів у рядку пошуку.

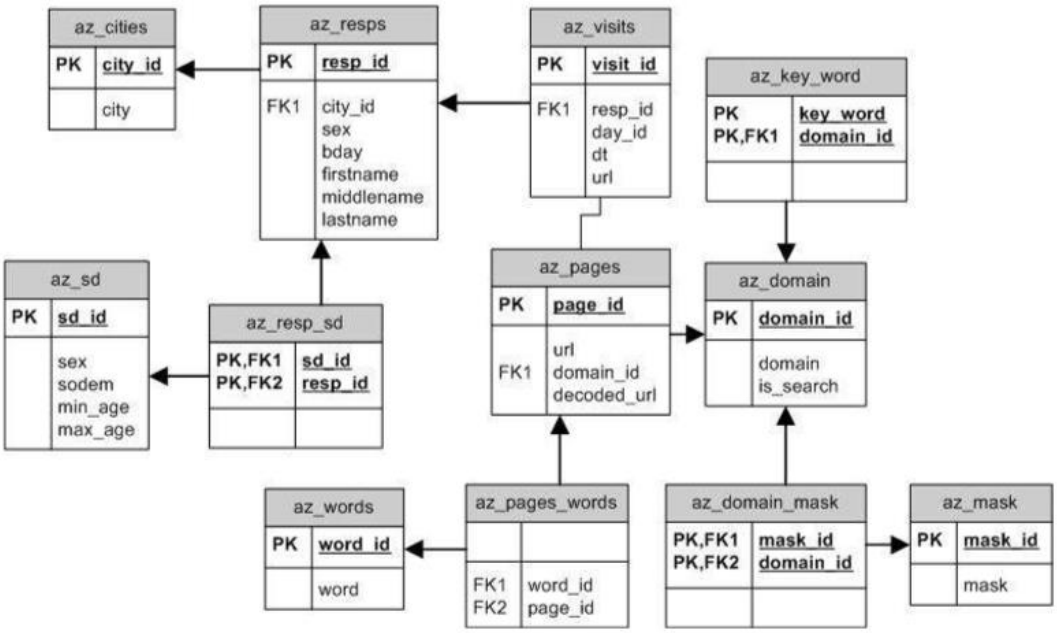

Рисунок 3.18 – Додавання сутностей az_words та az_pages_words

На рисунку 3.18 представлено повну структуру БД для обробки заходів ІКInternetDB. БД пошукової активності ІК містить 12 таблиць, необхідні для

структуризації одержуваних у процесі спостереження даних та подальшого проведення кластерного аналізу ІК з історії пошуку.

3.5. Результати застосування змістової структуризації даних

ІР за своєю структурою є окремо взятою web-сторінкою, або набором web-сторінок з однаковим доменним ім'ям і пов'язаних між собою гіперпосиланнями. Для візуальної інтерпретації ІР необхідно використовувати web-браузер, який перетворює теги DOM-моделі на відповідні візуальні образи. Наприклад, теги та вказують на початок і кінець параграфа, <table> і </table> вказують на таблицю, а <title> та </title> – на заголовок ІР. У свою чергу, кожен тег може містити атрибути, що доповнюють їх візуальні характеристики.

Для *IP HTML* є стандартною мовою розмітки веб-сторінок. У 2000 році була опублікована розширена мова розмітки web-сторінок XHTML. Згодом ІР стають все більш інтерактивними та динамічними завдяки застосуванню динамічних компонентів, інкапсульованих у їх DOM-моделі. Структура БД для зберігання та обробки даних про зміст ІР.

Web-розробники не обмежуються класичним HTML-кодом, використовуючи при створенні ІР різні технології - CSS, JavaScript, Ajax, що призводять до появи динамічних компонентів. У зв'язку з цим необхідно проводити інспекцію ІР, виходячи з DOM-моделей web-сторінок. Використання DOM-моделі дозволяє отримати доступ до будь-яких елементів ІР та їх атрибутів. Це дає можливість маніпулювати web-документами, як об'єктами (object), з усіма їх компонентами, їх атрибутами та властивостями. DOM-модель дозволяє представити ІР у вигляді дерева, кожен вузол якого може бути одночасно як батьківським (parent), так і дочірнім (child) вузлом по відношенню до іншого вузла дерева (рисунок 4.11). Тег HTML є початковою ланкою DOM-моделі, корінням починаючи з якої «розтає дерево» ІР. Для отримання доступу до конкретного

тегу HTML-документа необхідно пройти шлях від кореневого вузла (HTML-тега) до цільового вузла і потім прочитати значення конкретних атрибутів. За допомогою DOM-дерева HTML-документа можна розбити вміст IP на параграфи, списки, розділи, гіперпосилання та інші компоненти структури сторінки.

Для доступу до елементів DOM-моделі є два найбільш поширені методи: – Прямий доступ до HTML-елементу (htmlElement) DOM-моделі за унікальним ідентифікатором. В даному випадку потрібна наявність унікального ідентифікатора необхідного HTML-елемента.

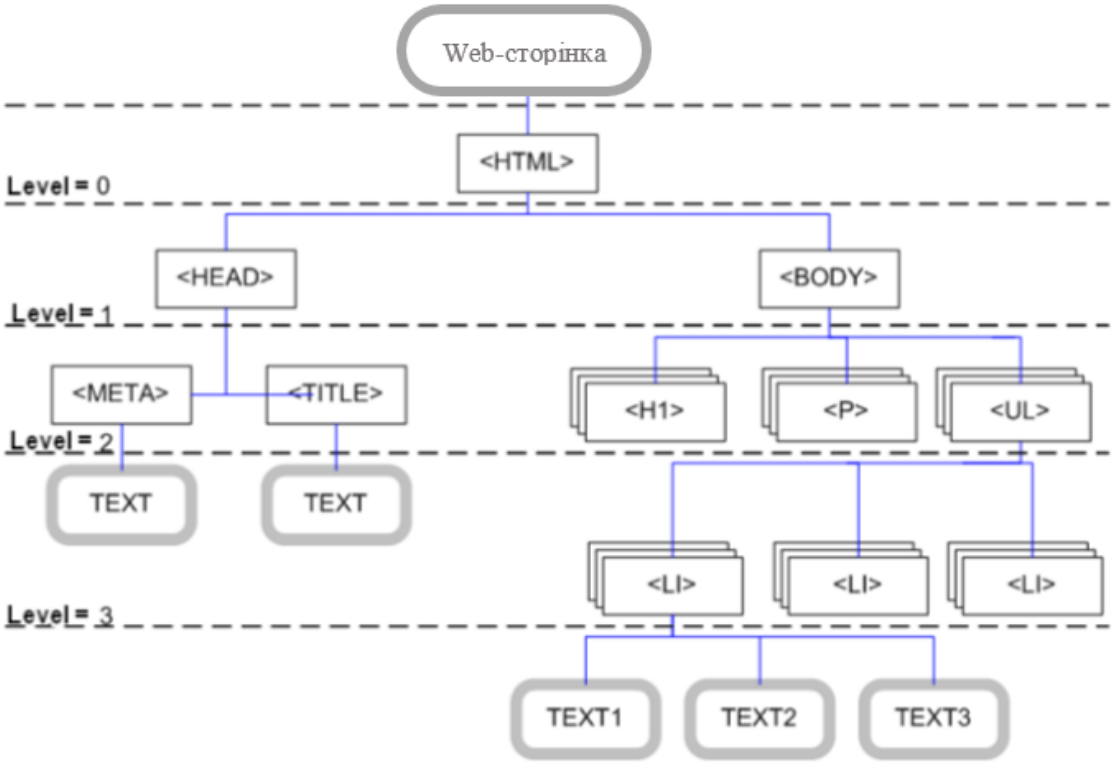

Рисунок 3.20 – Приклад DOM-дерева web-сторінки

Наприклад, на головній сторінці ukr.net маємо: ; - Доступ до HTML-елементу за назвою тега. У цьому випадку необхідно спочатку відібрати набір (htmlElementCollection) з конкретною назвою тега, а потім здійснити пошук

потрібного тега за значеннями атрибутів. Наприклад, на головній сторінці ukr.net, щоб знайти цей HTML-елемент, необхідно спочатку знайти набір елементів, у яких тег називається "a", а потім проводити пошук за атрибутом

Лістинг 3.2 – Пошуковий атрибут

```
"name="clb598679"": <a name="clb598679 " href="http://my.ukr.net " class="social__title__link"><i class="social__title__link__icon icon icon_social icon_social_big icon_social_my"></i><spanclass="social__title__link__text">Мій Світ</span></a> У середовищі Microsoft Visual Studio 2016 достатньо скористатися об'єктом System.Windows.Forms.WebBrowser для доступу до об'єктів DOM-моделі, завантажених URL.
```

Основними функціями для роботи з HTML-елементами є:

- getElementById – функція, що повертає посилання на вузол документа, яку можна використовувати для читання та редагування властивостей та звернення до методів вузла;
- getElementByTagName – функція, що повертає масив із елементів, що мають конкретний тег;
- getAttribute – функція, що повертає значення конкретного атрибуту HTML-елемента.

В алгоритмі доступу до DOM-елементів (рисунок 3.20) getElementById має вищий пріоритет, ніж getElementByTagName. Це актуально, особливо для IP з однаковою структурою, коли різні URL того самого IP мають ідентичну DOM-модель.

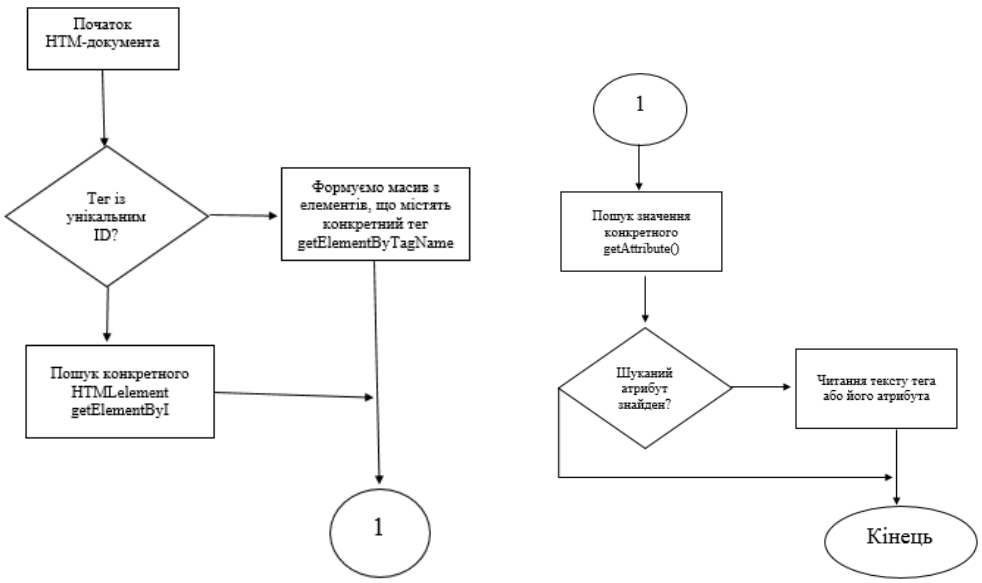

Рисунок 3.20 – Схема алгоритму доступу до DOM-елементів

Для структуризації змісту IP по тэгам досить доповнити раніше розроблену структуру БД для ІК (рисунок 3.21) ще трьома сутностями. Для початку необхідно виділити окрему сутність для всіх URL досліджуваних IP, створивши таблицю Pages. Весь довідник HTML тегів [4,7] DOM-моделі розмістимо у словниковій таблиці HTML_element. Результати читання тегів розташуємо в асоціативній сутності HTML_value. На рисунку 3.21 показана структура з трьох таблиць для зберігання даних про теги та їх значення.

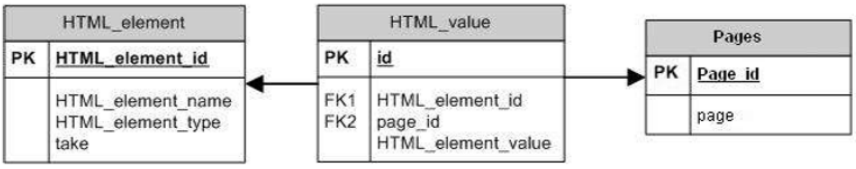

Рисунок 3.21 – Структура БД для зберігання даних про теги та їх значення

Використовуючи експериментальні дані про заходи ІП, отримані з 20 по 27 вереня 2023 року, можемо визначити топ-100 найпопулярніших URL. За вказаний тиждень було всього виконано 6848059 заходів з них 116674 заходи на розділи новин основних Інтернет-порталів - unian.ua, ukr.net і

gmail.com. Заходи виконувались користувачами різної статі, віку та місця проживання України. За результатами експерименту стандартним способом угруповання URL-сторінки сформована таблиця 3.2 топ-100 популярних сторінок новин укрнета. Отримані дані проходили етап структурізації.

3.6. Результати дослідження з оцінки якості персоналізації пошуку

Експеримент проводився з метою порівняння результатів роботи пошукової системи *search* при пошуку багатозначних термінів з результатами класифікації, виконаними після кластерного аналізу КСПП, на тих самих посиланнях. Прикладом омоніма є термін «Мустанг», який має безліч значень – це насамперед хижий ссавець, це марка автомобілів, це художній фільм. Термін «Мустанг» може мати й інші значення – енергетичний напій та назву пісні. Зараз, на жаль, пошукова система *search* не здатна визначити те, в якому сенсі терміном «Мустанг» цікавиться той чи інший ІП. Будь-який ІП, починаючи пошукову діяльність, вводить певний набір термінів у полі пошукового рядка, а потім переходить на його ресурси, що цікавляться, тим самим, виконуючи фільтрацію не потрібних результатів. 20 вересня 2023 року на пошуковий термін «Мустанг», *search* видав понад 3000000 гіперпосилань. Перші 50 гіперпосилань, отримані під час пошуку терміна «Мустанг» в *search* використано при реалізації запропонованої методики кластерного аналізу, що з кількох послідовних етапів. На першому етапі виконувався автоматизований захід за допомогою програми HTMLDocDom за всіма 50 гіперпосиланнями та виконувався збір їхнього текстового змісту. За результатами багаторазового аналізу DOM-моделі сторінок, що відповідають зазначеним гіперпосиланням, застосовувалися числові коефіцієнти посилення та триразова кластеризація IP із зворотним зв'язком. Щойно визначилася множина IP, формувався глобальний словник термінів і далі глобальний словник лем, з урахуванням якого будувалися характеристичні вектора IP.

Для порівняння результатів роботи пошукової системи *search* з результатами застосування запропонованої методики кластерного аналізу було змодельовано та реалізовано ситуацію, коли чотири ІП, зацікавлені різними сутностями, що позначаються терміном «Мустанг», здійснювали Інтернет-пошук. Ці ІК виконували заходи на n число IP, які відповідають їхнім інтересам. Виходячи з пошукової активності ІК та з урахуванням відвідуваних ними IP, формувалися характеристичні вектори ІК і потім були отримані перші середні 4 кластери. За допомогою процедур кластерного аналізу було виконано кластеризацію цих об'єктів, кінцевий результат якої представлений у таблиці 3.4.

Таблиця 3.4 – Кінцевий результат кластерного аналізу.

Кластер	Кількість об’єктів	Список об’єктів кластера
1	28	{2,39,26,21,6,23,35,49,31,10,27,40,13,24,29,38,14,19, 12,50,17,7,37, 28,48,5,1,9}
2	3	{15,33,25}
3	4	{34,11,42,36}
4	15	{45,3,46,41,4,16,22,47,43,20,44,30,8,32,18}

Кластер 1 – кластер з об'єктами, які виявилися ближчими до значення «Мустанг» – автомобіль, кластер 2 – кластер з об'єктами, які виявилися ближчими до значення «Мустанг» – фільм; кластер 3 – кластер з об'єктами, які виявилися ближчими до значення «Мустанг» – тварина; кластер 4 – кластер з об'єктами, що залишилися, з отриманої вибірки. Отримавши кінцевий результат розподілу об'єктів кластерами, можна розрахувати важливі коефіцієнти, що відображають якість отриманого результату. Для оцінки якості можуть застосовуватися такі показники: точність влучення, повнота та випадання. Слід звернути особливу увагу на той факт, що повнота вибірки і випадання не можуть бути представлені як опорні показники умовах обговорюваного експерименту, оскільки у ньому вибірка проводилася дуже обмеженому числі IP. Справа в тому, що при великій

вибірці (тисячі та мільйони IP) щойно збільшується точність результату пошуку, повнота вибірки знижується. Для вибору малого обсягу результат оцінки буде спотвореним. Тим не менш, для повноти картини будуть представлені всі три зазначені вище показники. Точність влучення (precision) – частка релевантної інформації у всьому обсязі інформації, отриманої в результаті пошуку, – обчислюється за формулою:

$$precision = \frac{TR}{TR + FR},$$

де TR – множина релевантних документів, отриманих під час пошуку;

FR – множина нерелевантних документів, отриманих під час пошуку.

Наведемо розрахунки точності потрапляння і порівняємо їх, для пошукової системи *search* (перші 50 гіперпосилань), з одного боку, і, для отриманої кластерної структури, з іншого боку (рисунок 3.22).

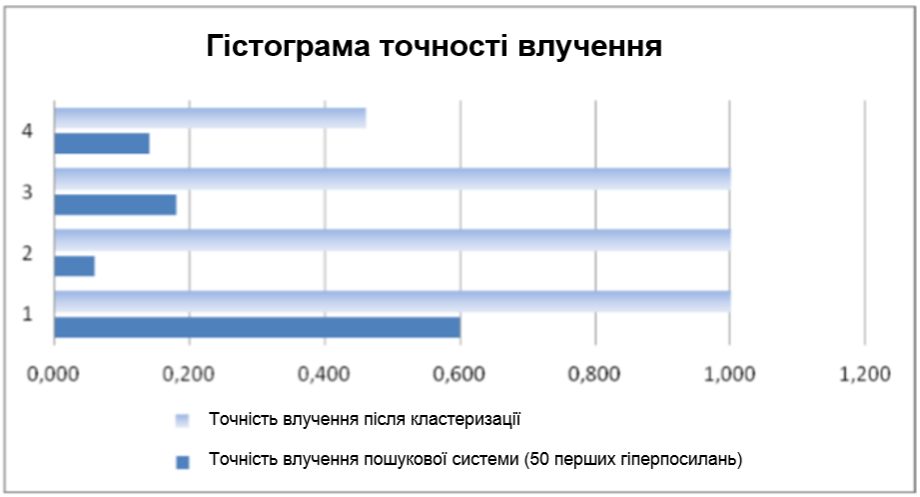

Рисунок 3.22 – Гістограма точності влучення

На рисунку 3.22 видно перевагу застосування запропонованої методики персоналізації пошуку. Точності влучення для 3-х кластерів дорівнює 1 при відвідуванні до 4-х релевантних IP.

$$recal = \frac{TR}{TR+FN}, \quad (3.1)$$

TR – множина релевантних документів, отриманих під час пошуку;

FN – множина релевантних документів, які були отримані під час пошуку, тобто. пошукова система неправдиво визнала цих документів не релевантними.

Наведемо розрахунки повноти вибірки і порівняємо їх для пошукової системи *search* (перші 50 гіперпосилань), з одного боку, і, для отриманої вище кластерної структури, з іншого боку (рисунок 3.23).

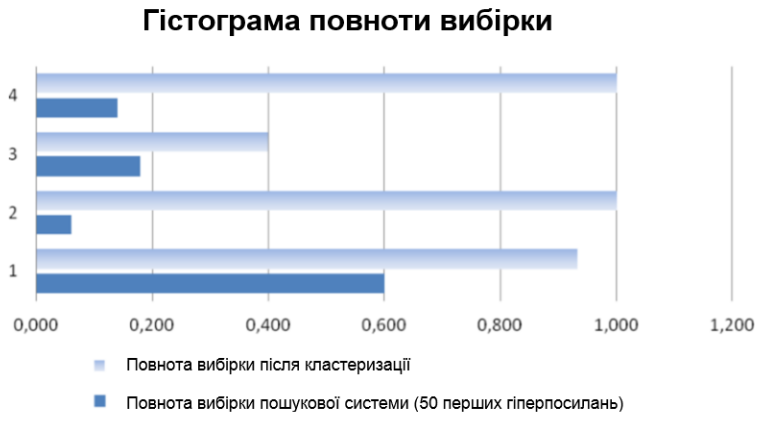

Рисунок 3.23 – Гістограма повноти вибірки

На рисунку 3.23 помітна ще одна перевага застосування запропонованої методики для персоналізації пошуку. Повнота вибірки для 2-х кластерів дорівнює 1. Зрозуміло, що вища точність і повнота, краще. Але в реальному житті максимальна точність і максимальна повнота не можна досягти одночасно, і доводиться шукати певний баланс між ними. Тому, хотілося б мати якусь метрику, яка б поєднувала в собі дані про точність і повноту кластеризації. Випадання (fall-out) або F-мера [3] є гармонійне середнє між точністю та повнотою. Вона прагне нуля, якщо точність чи повнота прагне нуля. Випадання розраховується за такою формулою:

$$F = \frac{2 \times precision \times recall}{precision + recal} \quad (3.2)$$

де F – випадання;

precision – точність влучення;

recall – повнота вибірки.

Випадання F вважається добрим показником для аналізу продуктивності пошукової системи (рисунок 3.24).

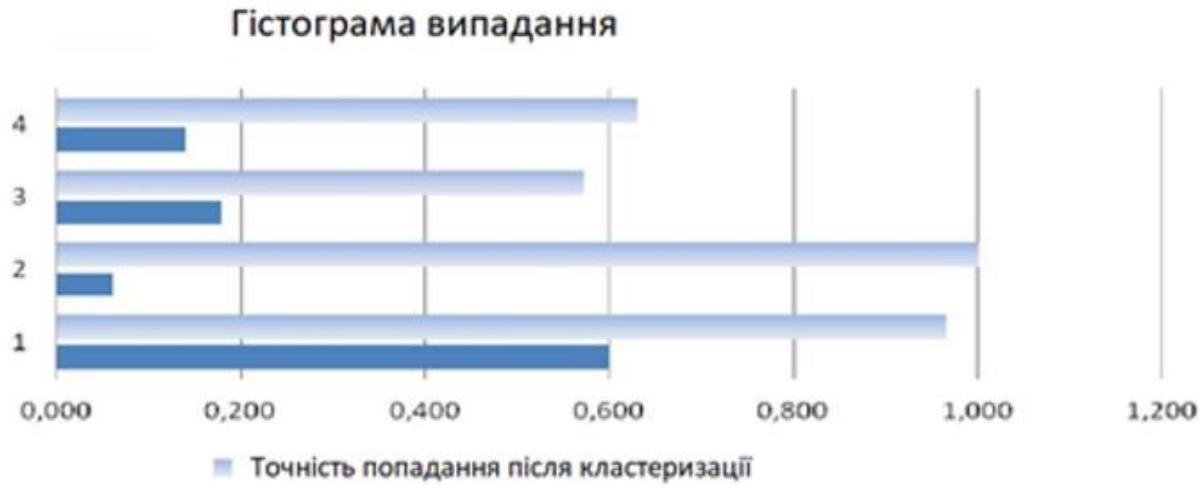

Рисунок 3.24. – Гістограма випадання

Отримані низькі показники для пошукової системи насамперед пов'язані з тим, що дослідження проводилося виключно на перших 50 гіперпосиланнях (немає можливості кластеризувати всі 3000000 IP). Незважаючи на це, в результаті проведених експериментів можна вважати доведеною перевагу застосування кластерного аналізу для персоналізації пошуку. Справа в тому, що 95% всіх ІК користуються гіперпосиланнями, наведеними на перших п'яти сторінках видачі, отриманої під час пошуку в *search*. Якщо застосовувати пропоновану кластерну методику класифікації IP з пошукової історії та інтересам самих ІП, то значною мірою можуть

підвищитися точність потрапляння та повнота вибірки, також збільшиться і випадання. Слід зазначити, що значення зазначених показників збільшуватимуться дедалі більше, прагнучи 1, у разі зростання активності користувачів. При виконанні експерименту було виявлено сайти плагіати, що мають мінімальну відстань між собою. Тим самим було виявлено ще одну перевагу застосування кластерного аналізу при класифікації, а потім персоналізації IP.

ВИСНОВКИ

В результаті проведеного дослідження було запропоновано комплексний підхід до кластеризації Інтернет-об’єктів.

У роботі проведено аналіз існуючих методів класифікації об'єктів, досліджено можливість їх застосування до Інтернет-об'єктів.

Запропонована й реалізована процедура лінгвістичної обробки тексту, що базується на використанні дворівневого словника термінів, можливістю застосування відкритих словників. При необхідності передбачена можливість звертання до «лінгвістичного експерта» для лематизації нових або нестандартних термінів.

Запропоновано схему кластеризації Інтернет-об’єктів зі зворотним зв'язком. Реалізація схеми дозволяє перетворювати динамічні Інтернет-об’єкти у статичні та застосовувати до останніх стандартні алгоритми кластерного аналізу. Розроблена математична модель представлення й процедури формування характеристичних векторів Інтернет-обєктів, числові координати яких, розташовані в порядку, відповідному до лексикографічного порядку проходження термінів в глобальному словнику. Перехід від вербального до числового показу координат відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у текст пошукових запитів або текстовий контент статичних компонентів Dom-моделі IP

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Mozhaiev O.,Kuchuk N.,Shtepa D., Sorobei B. Study of the Internet of Things network construction tasks. // Системи управління, навігації та зв'язку. Збірник наукових праць. – Полтава: ПНТУ, 2024. – Т. 1 (75). – С. 137-141. doi: 10.26906/SUNZ.2024. Zhovinsky E.Ya., Kryuchenko N.O., Paparyha P.S. Geochemistry of Environmental Objects of the Carpathian Biosphere Reserve. Kyiv, 2013. 100 p.

2. Kurimo M. Unsupervised segmentation of words into morphemes / M. Kurimo, M. Creutz, E. Arsoy. – Morpho challenge, 2015. – 95 p.

3. Russell M. Application to automatic speech recognition / M. Russell // Proc. INTERSPEECH-2016. – Pittsburgh, USA, 2016. – P. 1021-1024.

4. Schlippe T. Grapheme-to-Phoneme Model Generation for Indo-European Languages / T. Schlippe, S. Ochs, T. Schultz // Proc. ICASSP-2012. – Kyoto, Japan, 2022. – P. 45-51/

5. Huang C. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition / C. Huang, E. Chang, J. Zhou, K. Lee // Proc. INTERSPEECH-2000. – Beijing, China, 2020. – P. 818-821.

6. Hannemann M. Combinations of Confidence Measures for the Detection of Out-of-Vocabulary Segments in Large Vocabulary Continuous Speech Using Differently Constrained Recognizers / M. Hannemann. – Magdeburg : Otto-von-Guericke-Universitat, 2018. – 252 p.

7. Bourlard H. Links between Markov models and multilayer perceptron's / H. Bourlard, C.J. Wellekens // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – Vol. 12, No. 12. – 2020. – P. 1167-1178.

8. Bourlard H. Towards increasing speech recognition error rates / H. Bourlard, H. Hermansky, N. Morgan // Speech Communication. – 2016. – Vol. 18. – P. 205-231.

9. Hinton G. Deep neural networks for acoustic modeling in speech

recognition: The shared views of four research groups / G. Hinton, L. Deng, G. Dahl, A. Mohamed // IEEE Signal Process. Mag. – 2019. – Vol. 29, No. 6. – P. 82–97.

10. Dong Yu. Automatic Speech Recognition: A Deep Learning Approach / Yu. Dong, Li. Deng. – London : Springer Verlag, 2015. – 321 p.

11. Hermansky H. Tandem connectionist feature extraction for conventional HMM systems / H. Hermansky, D. Ellis, S. Sharma // Proc. ICASSP-2020. – Istanbul, 2020. – Vol. 3. – P. 1635-1638.

12. Shelly G., Woods D. HTML, XHTML, AND CSS. – Boston: Course Technology, Cengage Learning. – 2021.

13. Singh A. Web Content Extraction to Facilitate Web Mining // International Journal of Electronics and Computer Science Engineering. 2022. № 1 // Електроний ресурс // URL: http://www.ijecse.org/wp-content/uploads/2022/06/Volume-1Number3PP-1292-1299.pdf.

14. Soumen C. Mining the Web Discovering Knowledge from Hypertext Data. – San Francisco: Morgan Kauffman Publishers. – 2020.

15. Sundar G., Narmadha D., Haran A. Combinational Scheme for Efficient Content Extraction from Web Pages // Australian Journal of Basic and Applied Sciences. 2024. №1.

16. Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density // Lab of High Volume language Information Processing & Cloud Computing Beijing Lab of Intelligent Information Technology, Beijing Institute of Technology. 2021 // Электронный ресурс // URL: http://disnet.cs.bit.edu.cn/DOM%20Based%20Content%20Extraction%20via%20 Text %20Density.pdf