Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Післядипломної освіти_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження методів кластерізації ресурсів інформаційних мереж
(тема)

Виконав: студент 2 курсу, групи ІПЗмзд-17-1__
спеціальності _121- Інженерія програмного забезпечення_____
(код і повна назва спеціальності)
освітньо-професійної програми _Інженерія програмного_забезпечення_____
(повна назва освітньої програми)

_____Баткова Т.В._____
(прізвище, ініціали)

Керівник _____проф. Шубін І.Ю._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2019 р.

Харківський національний університет радіоелектроніки

Факультет післядипломної освіти

Кафедра програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121– Інженерія програмного забезпечення
(код і повна назва)

Освітньо-професійна програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові Батковій Таїсії Володимирівні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів кластерізації ресурсів інформаційних мереж

затверджена наказом по університету від « »_____2019 р № __

2. Термін подання студентом роботи до екзаменаційної комісії «15» червня 2019 р.

3. Вихідні дані до роботи *Алгоритми обробки великих обсягів даних, алгоритми захисту даних, методи стримінгу великих даних та пояснювальна записка. Використовувати ОС Windows, середовище об'єктно-орієнтованого проектування.*

4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, методи пошуку корисних даних, опис об'єктних моделей, використовувані. методи та алгоритми, архітектура програмної системи, опис розробленої програмної системи. , результати тестування програмної системи*

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) _____*Титульний аркуш, мета завдання, обґрунтування доцільності розроблення, постановка задачі, об'єктна модель системи, базові моделі, методи й алгоритми, структура бази даних, структурно-логічна схема взаємодії даних, інтерфейс програмної системи, результати функціонування програмної системи, демонстраційні матеріали*_____
_____

6. Консультанти розділів роботи

Найменування розділу	Консультант (посаду, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Шубін І.Ю.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	19 квітня 2019 р.	
2.	Огляд існуючих методів	27 квітня 2019 р.	
3.	Методи швидкого детектування відрізків	15 травня 2019 р.	
4.	Підготовка пояснювальної записки	20 травня 2019 р.	
5.	Спецчастина	28 травня 2019 р.	
6.	Підготовка презентації та доповіді	05 червня 2019 р.	
7.	Попередній захист	05 червня 2019 р.	
8.	Нормоконтроль, рецензування	06 червня 2019 р.	
9.	Занесення диплома в електронний архів	11 червня 2019 р.	
10.	Допуск до захисту в зав. кафедри	11 червня 2019 р.	

Дата видачі завдання _ «____»_ _____2019 р.

Студент _____
(підпис)

Керівник роботи _____ __проф. Шубін І.Ю._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 106 с., 43 рис., 8 табл., 4 додатки, 26 джерел.

КЛАСТЕРНИЙ АНАЛІЗ, ІНФОРМАЦІЙНИЙ ПОШУК, АНАЛІЗ ДАНИХ, СЕМАНТИЧНИЙ АНАЛІЗ.

Об'єктом дослідження є методи персоналізації Інтернет-пошуку, засновані на вивченні й класифікації інтернет-користувачів та інтернет-ресурсів за допомогою кластерного аналізу.

Метою є застосування методів класичного кластерного аналізу для класифікації інтернет-користувачів та інтернет-ресурсів, для персоналізації інформаційного пошуку в Інтернеті

Методи розробки базуються на методах математичного моделювання.

Як результат обчислювального експерименту, проведений вибір алгоритму кластеризації, що забезпечує показники кластерної структури.

CLUSTER ANALYSIS, INFORMATION SEARCH, ANALYSIS OF DATA, SEMANTIC ANALYSIS.

The object of the research is the methods of personalization of Internet search, based on the study and classification of Internet users and Internet resources with the help of cluster analysis.

The purpose is to apply methods of classical cluster analysis for the classification of Internet users and Internet resources, to personalize the information search on the Internet

Methods of development are based on methods of mathematical modeling.

As a result of the computational experiment, a clustering algorithm is chosen, which provides cluster structure indices.

ЗМІСТ

ВСТУП

Інтернет в 21-ому столітті став невід'ємною частиною повсякденного життя. Економічна, соціальна й наукова діяльність людства тією чи іншою мірою пов'язана з Інтернет-технологіями. У наші дні можна проводити переговори з партнерами по бізнесу, грошові перекази, онлайн-консультації, навчатися й багато чого іншого, не виходячи з будинку. Мобільний Інтернет прив'язав людину до віртуального світу – у будь-який час й у будь-якому місці на земній кулі, маючи доступ до Інтернету, можна бути в курсі всього, що відбувається на планеті.

Величезна кількість ресурсів інформації, що утримується в них, перетворило всесвітню павутину в грандіозне сховище погано організованих, неструктурованих даних. Пошук інформації в мережі Інтернет став долею людства. Величезна кількість «сміття», що видається пошуковими системами, робить актуальною проблему персоналізації пошуку в Інтернеті, адаптації пошукових систем до запитів окремих користувачів або їх груп..

Інтуїтивно будь-який ІК формує свою систему класифікації й відбору веб-ресурсів для задоволення власних потреб в інформації. Користувач Інтернету має свій особистий психологічний портрет і відвідує конкретні, «улюблені» їм веб-сторінки. Якщо говорити про поведінку людини в мережі Інтернет, то можна виділити короткочасні (сесійні) дії ІК, які пов'язані з пошуком конкретної інформації протягом однієї або кілька пошукових сесій. Коли користувач знаходить релевантну інформацію, він припиняє свій пошук і навіть може вийти з мережі. Крім сесійних дій користувачів можна виділити їх рутинну поведінку в мережі, наприклад, щоденний ранковий огляд новин про спорт або спілкування в соціальних мережах під час обідньої перерви.

Великі пошукові системи, такі як Google, Bing і т.ін. користуються персональною інформацією й файлами cookie із браузерів для персоналізації результатів пошуку – маркетологи, наприклад, підбирають рекламу залежно від

пошукової історії або залежно від статі й віку ІК. Вдаліше всього застосовується регіональний або географічний таргетинг – люди думають, що Яндекс дійсно порозумнішав і сказати що, це не так, не можна. Насправді, Яндекс добре працює з регіональними запитами при пошуку магазинів/товарів місцевого користування/споживання.

Програмісти працюють над алгоритмами, що підвищують релевантність документів запитам за допомогою розрахунку ваг пошукових термінів, що дозволяє відбирати релевантні результати й переваги користувачів. У компанії Яндекс крім лінгвістичного аналізу контенту, індексу цитування, функції DCG (Discounted cumulative gain) [1], системи машинного навчання Матрикснет [2] і фільтрів негативних ознак у число таких методів входять і різні процедури обліку й обробки первинної персональної інформації. Коли користувачі видають запити Google, приблизно в 20% випадків вони формулюють запити неоднозначно. Технологія компанії,» уміє враховувати множина неявних цілей користувачів і показувати потрібні відповіді. В основі роботи лежить статистика пошукових запитів ІК.

Соціально-демографічна (далі СД) класифікація – основний метод класифікації ІК після їхньої авторизації на Інтернет-сайтах – забезпечує облік статевих і вікових відмінностей, іншої статичної атрибутивної інформації користувача. СД класифікація на сайтах застосовується, наприклад, для таргетування рекламних кампаній, але при цьому поведінка користувачів ніяк не береться до уваги. Проведена на сторінках сайтів персоналізація користувачів далека від досконалості, тому що сайти працюють за принципом «клієнт завжди правий», тобто акцент робиться на рекламодавцях, що вклали великі кошти в просування товару – тому й кульгають результати пошуку на сторінках користувачів.

Наведені аргументи свідчать про необхідність подальшого пристосування Інтернету до потреб користувачів і, зокрема, за рахунок персоналізації Інтернет пошуку. Підвищення рівня персоналізації пошуку, у свою чергу, може бути досягнуте за рахунок розробки перспективних методів класифікації ІК та ІР,

заснованих на кластерному аналізі, впровадженню цих методів в існуючі пошукові системи.

Проблема: відсутність ефективних методів і засобів, що забезпечують персоналізацію пошуку інформації в Інтернеті.

Про персоналізацію пошуку жаркі дискусії ведуться вже майже 20 років – усі зацікавлені в тому, щоб результати пошуку в Інтернеті були як можна більш релевантними запитам користувачів. Однак недостатня наукова пропрацьованість проблеми, закритість більшості практично реалізованих рішень провідними компаніями постачальниками Інтернет послуг обумовила необхідність дослідження теоретичних і практичних питань застосування методів кластерного аналізу для персоналізації пошуку.

По темі кластерного аналізу існує багато літератури. Вона охоплює загальні питання математичного опису об'єктів і алгоритми їх кластеризації. Кластеризація об'єктів із статичними властивостями широко застосовується повсякденно переважно в аналітичній діяльності. Однак, методи кластеризації динамічних об'єктів, таких як IP, недостатньо розроблені й, крім того, мало хто з дослідників розглядав ідею узагальненого показу об'єктів різної природи, що мають подібні властивості.

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

Проаналізовано методи, спрямовані на рішення проблеми персоналізації й підвищення якості результатів пошуку в Інтернеті. Ці методи дозволяють використовувати існуючі класичні алгоритми кластерного аналізу для Інтернет-Об'єктів – ІК та IP – з урахуванням особливостей їх математичного опису.

Як результат обчислювального експерименту, проведений вибір алгоритму кластеризації ІК та IP, що забезпечує показники кластерної структури.

Розроблені й програмно реалізовані методи, що забезпечують виконання кластерного аналізу для персонализації пошуку в Інтернеті. Програмна реалізація зазначених методів здійснена у вигляді віртуальної системи персоналізації пошуку. аналітична частина проекту, виконується кластеризація Інтернет-об'єктів.

Використовуючи зазначені інструменти, експерт-аналітик на основі результатів кластерного аналізу одержує чітку картину про розподіл ІК та ІР по кластерах залежно від декількох вхідних параметрів: тривалості періоду спостереження за активністю ІК, числа кластерів, значень коефіцієнтів підсилення й мінімальної довжини термінів.

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

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

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 ІР.

Розроблений набір програмних модулів (програмна система) для спостереження за активністю ІК і одержання текстового змісту ІР з їх обліком .

У середовищі MS SQL Server розроблені спеціальні збережені процедури, що виконують усі необхідні розрахунки – від формування словників термінів до кінцевого розподілу об'єктів по кластерах.

1 АНАЛІЗ СТАНУ РОЗВ'ЯЗАННЯ ПРОБЛЕМИ ТА ОБҐРУНТУВАННЯ ЦІЛЕЙ ДОСЛІДЖЕННЯ

1.1 Алгоритми використання інформації для рішення завдань персоналізації

При спробі одержання знань із web-а ми не можемо орієнтуватися на суворі структури й компоненти, тому що в Інтернеті присутня величезна кількість розподіленої, гетерогенної, неструктурованої й динамічно мінливої інформації. Незважаючи на це, IP навчилися бути ближче до ІК, перестали бути ізольованими від них. Як тільки ІК заходить на IP, він одразу залишає свій слід: стають відомі його місце розташування (географія), персональні дані (стать, вік і т.д.), його історія пошуку. Наразі персональна інформація ІК становить величезний інтерес як для Інтернет-майданчиків, так і для рекламодавців. Будь-який IP зацікавлений в обробці особистої інформації ІК та його сторінки, що відвідували. Це важливо для статистичної обробки відвідуваності з метою продажу реклами. Можна чітко розділити чоловічі й жіночі сайти, спортивні або сайти новин. Для прикладу, виберемо один з великих Інтернет-порталів – gmail.com.

Робота користувача починається з реєстрації поштової скриньки на головній реєстраційній сторінці сайту. Практика показує, що в соціальних мережах і блогах інформація поширюється набагато швидше, ніж на сайтах новин. У соціальних мережах люди залишають набагато більше інформації про себе й про свою поведінку, ніж на якому-небудь іншому Інтернет-ресурсі: про школу, інститут, роботу й, звичайно, коментарі. Варто відзначити дуже важливий момент, що зареєстрований у соціальних мережах пов'язаний з так званими лайками: авторизований користувач може ставити оцінки іншим користувачам, їх статтям, коментарям, фотографіям і т.д., відзначивши, що сподобалося лайком (рис. 1.1).

Для моментального відображення статусу користувачів і їх поведінки в рамках групи застосовуються нереляційні бази даних, що працюють із високошвидкісною оперативною пам'яттю. Це так звані in memory database або бази даних у пам'яті. Основна ідея таких систем – зберігання даних не на

дисковому накопичувачі, а прямо в пам'яті. Застосування такого роду БД зменшує час відгуку системи й дозволяє практично моментально перемикатися по групах інтересів користувачів соціальних мереж.

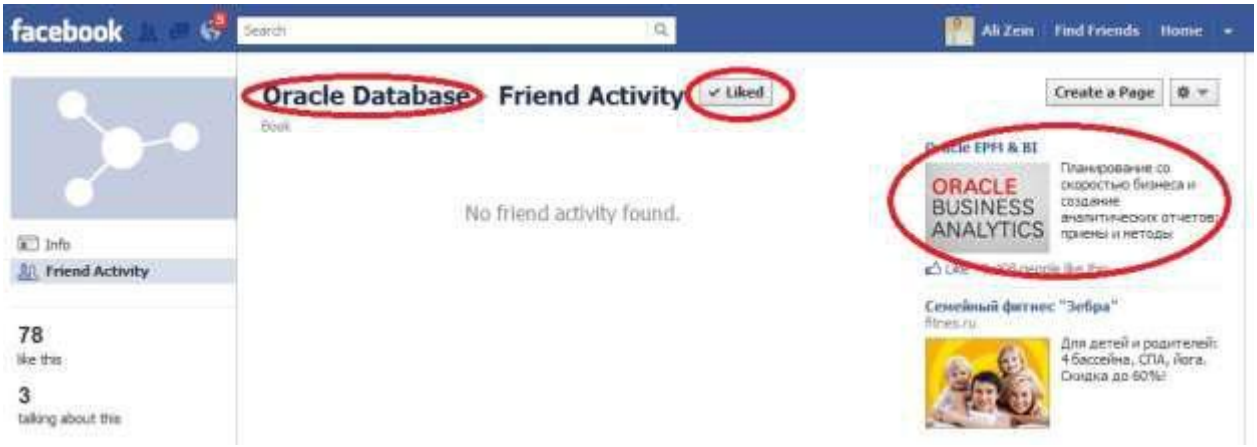

Рисунок 1.1 – Використання лайків для показу реклами в соціальній мережі.

Соціальні групи являють гарний приклад для класифікації ІК. Справа в тому, що люди формують групи інтересів, можуть користуватися загальними ресурсами й ділитися досвідом у тій чи іншій галузі. При цьому будь-яка сформована група буде досить спеціалізованою й може бути легко індексована для швидкого пошуку в області інтересів користувачів.

1.2 Методи некластерної класифікації користувачів і ресурсів

Асоціативний метод класифікації широко застосовується в Інтернет-магазинах, коли зміст кошиків множина і покупців аналізується й утворюється якась імовірна закономірність покупок. Проведено аналіз релевантності між елементами вектора за допомогою асоціативних правил. У БД Інтернет-запитів від 20 грудня 2018 р. випадково обрані п'ять ІК, що виконували пошук товару на сайті rozetka.ua. Таблиця пошуку з векторами пошуку, що складаються із товарів

{фотоапарат, реєстратор, навігатор, пам'ять}, та пошукові дані, представлені в табл. 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	{фотоапарат, реєстратор}
2	{фотоапарат, навігатор, пам'ять}
3	{навігатор, пам'ять}
4	{реєстратор}
5	{фотоапарат, пам'ять}

Для початку необхідно розподілити елементи {фотоапарат, реєстратор, навігатор, пам'ять} у проміжну таблицю влучень (табл. 1.2).

Таблиця 1.2 – Таблиця влучень

Набір елементів	Транзакції	Число влучень
{}	{1,2,3,4,5}	5
{фотоапарат}	{1,2,5}	3
{реєстратор}	{1,4}	2
{навігатор}	{2,3}	2
{пам'ять}	{2,3,5}	3
{фотоапарат, реєстратор}	{1}	1
{фотоапарат, навігатор }	{2}	1
{фотоапарат, пам'ять}	{2,5}	2
{навігатор, пам'ять}	{2,3}	2
{фотоапарат, навігатор, пам'ять}	{2}	1

Стовпець «Набір елементів» формується за допомогою окремих елементів та можливих комбінацій цих елементів, відповідно до реальних результатів

таблиці транзакцій. Стовпець «Транзакції» формується за допомогою набору транзакцій, у якому була присутня комбінація елементів в рядку. Значення стовпця «Число влучень» формується на підставі числа елементів стовпця «Транзакції».

Тепер можна побудувати асоціативну таблицю елементів (табл. 1.3):

Таблиця 1.3 – Асоціативна таблиця елементів

Асоціативний набір	Число влучень	Відсоток імовірності
{фотоапарат} → {пам'ять}	2	2/3 = 67%
{реєстратор} → {фотоапарат}	1	1/2 = 50%
{навігатор} → {пам'ять}	2	2/2 = 100%
{пам'ять} → {фотоапарат}	2	2/3 = 67%
{пам'ять} → {навігатор}	2	2/3 = 67%
{фотоапарат, навігатор} → {пам'ять}	1	1/1 = 100%
{фотоапарат, пам'ять} → {навігатор}	1	1/2 = 50%
{навігатор, пам'ять} → {фотоапарат}	1	1/2 = 50%
{навігатор} → {фотоапарат, пам'ять}	1	1/2 = 50%

По асоціативній таблиці елементів проводиться розрахунок імовірності появи події {пам'ять}, якщо подія {фотоапарат} мала місце і т.д. На прикладі {фотоапарат} → {пам'ять} у чисельнику буде перебувати число випадків, коли в транзакції присутні обидва елемента {фотоапарат} і {пам'ять}: це друга й п'ята транзакція в таблиці 1.1. У знаменнику буде число випадків, коли в транзакції тільки присутній елемент {фотоапарат}.Таким образом, в чисельнику буде count({фотоапарат, навігатор,пам'ять}, {фотоапарат, пам'ять}) = 2, в знаменнику Буде count({фотоапарат,реєстратор}, {фотоапарат, навігатор, пам'ять}, {фотоапарат, пам'ять}) = 3. Звідси, імовірність появи події {пам'ять}, якщо настала подія реєстратор}, {фотоапарат, навігатор, пам'ять}, {фотоапарат,

пам'ять}) = 3. Звідси, імовірність появи події {пам'ять}, якщо настала подія {фотоапарат} буде рівна P({фотоаппарат}− − > {пам’ять}) =2/3 ∗100 = 67%

Метод перетинань ґрунтується на перетинанні елементів на різних транзакціях. Повернемося до таблиці влучень (табл. 1.2), і з цієї таблиці вибрано одиночні набори елементів, формуючи таблицю одиночних наборів елементів (табл. 1.4) {фотоапарат}, {реєстратор}, {навігатор} і {пам'ять}.

Таблиця 1.4. – Таблиця одиночних наборів елементів

№№	{фотоапарат}	{реєстратор}	{навігатор}	{пам'ять}
1	+	+	-	-
2	+	-	+	+
3	-	-	+	+
4	-	+	-	-
5	+	-	-	+

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

Таблиця 1.5 – Таблиця подвійних наборів елементів

№№	{фотоапарат реєстратор}	{фотоаппарат навигатор}	{фотоаппарат память}	{реєстратор, навігатор}	{реєстратор, пам'ять}	{навігатор, пам'ять}
1	+	-	-	-	-	-
2	-	+	+	-	-	+
3	-	-	-	-	-	+
4	-	-	-	-	-	-
5	-	-	-	-	-	-

На етапі формування таблиці подвійних наборів елементів відбувається випадання конкретних комбінацій: випадають {реєстратор, навігатор} і {реєстратор, пам'ять}. Для розрахунку ймовірності необхідно спочатку визначитися з порядком проходження елементів у формованих парах – {фотоапарат, реєстратор} або {реєстратор, фотоапарат}, тому що результат розрахунку ймовірностей буде різним: –

$$P(\{\text{фотоапарат}\}-->\{\text{реєстратор}\})= \frac{1}{3} \times 100 = 63\%$$

$$P(\{\text{реєстратор}\}-->\{\text{фотоапарат}\})= 12 \times 100 = 50\%$$

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

Таблиця 1.6 – Таблиця потрійних наборів елементів

№№	{фотоапарат, реєстратор, навігатор}	{фотоапарат, реєстратор, пам'ять}	{фотоапарат, навігатор, пам'ять}
1	-	-	-
2	-	-	+
3	-	-	-
4	-	-	-
5	-	-	-

На етапі формування таблиці потрійних наборів елементів відбувається випадання конкретних комбінацій: випадають комбінації {фотоапарат, реєстратор, навігатор} і {фотоапарат, реєстратор, пам'ять}. Для розрахунку ймовірності необхідно також визначитися з порядком проходження елементів у трійках.

Для виконання алгоритму необхідно виконати перерахунок усіх можливих комбінацій (рис. 1.2)

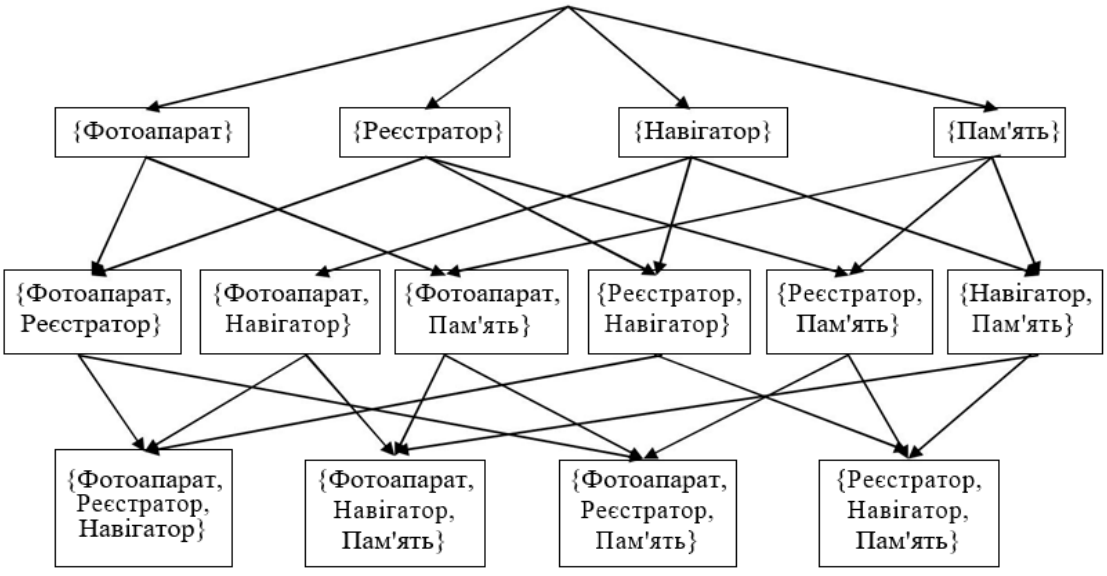

Рисунок 1.2 – Схема всіх можливих комбінацій.

За допомогою схеми всіх можливих комбінацій можна розрахувати ймовірності появи будь-якої комбінації елементів.

Асоціативний метод і метод перетинань широко застосовуються для добору супутнього товару. Велика кількість покупок груп товарів робить результат розрахунків ймовірностей більш точними, тому що ІК показує величезну кількість товарів, яка купувалася одночасно з товаром, що проглядається.

Однак ці методи погано масштабуються й, як тільки відбувається вихід за рамки транзакції, результати розрахунку ймовірностей псуються. Справа в тому, що пошукові інтереси ІК неможливо розділити на окремі транзакції й самі популярні пошукові терміни зустрічаються в більшості ІК у рамках періоду спостереження $\Delta t = 4$ години. Ці методи не враховують важливості термінів.

Метод частоти термінів широко застосовується для класифікації звичайних текстів. Цей метод можна застосувати й для класифікації ІК по їхній пошуковій історії й ІР по їхньому текстовому змісту.

Нехай вихідними даними для методу є пошукові терміни ІК у період з 13 по 19 травня 2019 р. Нехай USR – множина спостережуваних ІК і usri – інший ІК, за яким ведеться спостереження. Кожний usri, usri∩usr, по залишених ним персональним даним може бути розподілено в одну з десяти соціально-демографічних груп Usri $\in \{USR_1, ..., USR_{10}\}$ таких, що $\cup \upsilon\sigma\rho\iota$ = USR і Usri ∩ Usrk = ∅, i ≠ k. Ці групи використовуються для таргетування реклами ІР, у якій крім статі береться до уваги вікова категорія Інтернет-користувача (табл. 1.7).

Після побудови бази даних для структуризації запитів ІК та змісту ІР застосування Tf-Методу не вимагає ніякої праці. Обліком наявної обчислювальної потужності, при формуванні статистичних таблиць було ухвалене рішення про обмеження числа термінів nof(V) у глобальному словнику термінів V і їх довжини. В [5] зазначена середня довжина російських слів рівна 5,28 символів. Однак у пошуковій історії ІК виявилася велика кількість термінів, довжина яких рівна 4, тому необхідно розглядати терміни, довжина яких більше або рівна 4.

Таблиця 1.7 – Таблиця СД класифікації

Соц-дем	Стать	Вік
USR_1	ЧОЛОВІК	від 12 до 17
USR_2	ЧОЛОВІК	від 18 до 24
USR_3	ЧОЛОВІК	від 25 до 34
USR_4	ЧОЛОВІК	від 35 до 44
USR_5	ЧОЛОВІК	від 45 до 55
USR_6	ДРУЖИН	від 12 до 17
USR_7	ДРУЖИН	від 18 до 24
USR_8	ДРУЖИН	від 25 до 34
USR_9	ДРУЖИН	від 35 до 44
USR_{10}	ДРУЖИН	від 45 до 55

Таким чином, кожна соціально-демографічна група Usri може бути представлена числовим вектором $fi = (fi_{,1}, \ldots, fi_{,j}, \ldots, fi_{,nof(Vu)})$ розміром *nof(Vu),* де $f_{i,j}$ – вага *j*-ого пошукового терміна із глобального словника термінів Vu. Числові координати $fi_{,j}$, $1 \le j \le nof(Vu)$ розташовані в характеристичному векторі fi, у тому ж порядку, що й терміни в глобальному словнику *Vu*.

Перехід від вербального до числової вистави результатів дослідження окремо взятої соціально-демографічної групи Usri у вигляді характеристичного вектора відбувається за рахунок позиційного кодування термінів словника, підрахунку числа їх входжень у запити ІК групи протягом усієї пошукової історії й розрахунку частоти вживання цих термінів (Tf-значень) у групі.

Для розрахунку Tf-значень застосовується наступна формула:

$$f_{ij} = \frac{nof(v_{i,j})}{\sum_{j=1}^{nof(Vu)} nof(v_{i,j})}, \tag{1.1}$$

де $nof(v_{i,j})$,

$v_{i,j} \in Vu$, – число входжень терміна vj у запити користувачів *i*-тої СД групи протягом усієї пошукової історії.

Для порівняння результатів необхідно побудувати глобальний вектор

$$fg = (fg_1, \ldots, fgj, \ldots, fgnof_{(Vu)}),$$

де для розрахунку координат fgj у чисельнику й знаменнику формули (1.1) буде відповідно число входжень усіх термінів vj для всіх
Usri, $1 \le j \le 10$.

Таблиця 1.8 – Таблиця ваг шуканих слів для чоловіків різного віку

Word	fg	f_1	f_2	f_3	f_4	f_5
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
Гри	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
відкликання	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
контакті	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
Москві	0,027396	0,013158	0,000329	0,045816	0,028948	0,06231
Ціна	0,025855	0,013158	0,042119	0,039	0,018715	0,08714
офіційний	0,022355	0,013158	0,000329	0,034457	0,031911	0,000234
магазин	0,021917	0,013158	0,000329	0,038243	0,000135	0,02811
Фільм	0,01723	0,013158	0,000329	0,000379	0,030295	0,000234
інтернет	0,016834	0,013158	0,000329	0,000379	0,000135	0,000234
Гра	0,01673	0,013158	0,040803	0,000379	0,000135	0,000234
Карта	0,016625	0,013158	0,000329	0,000379	0,021678	0,022722

Статистика отримана, ваги слів розраховані, наступним кроком є формування графіків ваг (рис. 1.3). На графіках видно, що кожна СД група має свої інтереси й свою поведінку в мережі Інтернет. Кожний вектор fi має своє відхилення щодо глобального вектора fg.

Розглянутий метод прямо залежить від якості формування словника Vu і від персональних даних ІК. Метод частоти термінів може бути застосований для класифікації ІК, винятково з метою добору реклами, однак якщо його застосувати для ІР з динамічними елементами, то можна зіштовхнутися із 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 по СД ознаках або ІР за структурою. Для персоналізації пошуку необхідно застосовувати більш складні методи з можливістю формування груп об'єктів зі складними характеристиками.

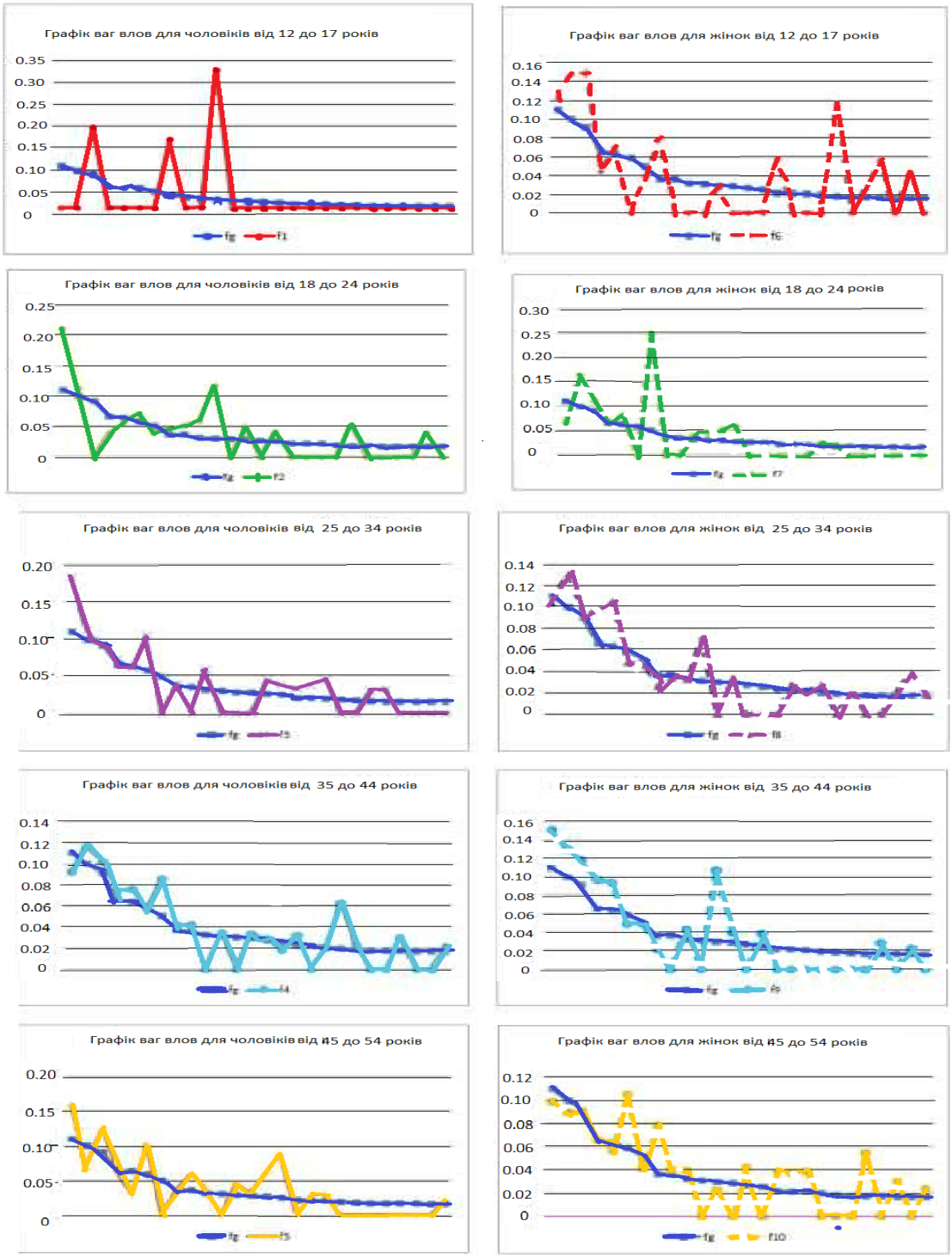

Рисунок 1.3 – Графіки ваг слів для ІК, розділених по СД ознаках [6]

1.3 Кластерні методи класифікації користувачів ресурсів

Кластеризація – це автоматична розбивка елементів множини на групи (кластери) залежно від показників їх схожості. Елементами множини можуть бути що завгодно: об'єкти з певним набором даних або напрямом характеристик. Більшість дослідників вважають, що прабатьком кластерного аналізу є Тріон Роберт Чоут (Robert Choat Tryon) – американський дослідник поведінки тварин, який запропонував систематизувати методи аналізу впливу навколишнього середовища (екологія, соціальний рівень і т.д.) на поведінку суб'єктів дослідження (тварин, людей) і запропонував групувати суб'єкти дослідження в кластери. Запропонований ним метод дозволив з великою точністю визначати причини й можливі наслідки поведінки людини в стресових ситуаціях, виходячи з її соціального оточення.

У кластеризації існує велика кількість практичних застосувань. Кластеризація дозволяє, наприклад, провести аналіз даних, пошук інформації або групи об'єктів за ознаками й властивостями. Так само кластеризація сама по собі є важливою формою абстракції даних, і в цій області був отриманий ряд цікавих наукових результатів.

Об'єкт – елементарна одиниця, яка може бути представлена за допомогою набору числових характеристик, і в якій оперують алгоритми кластеризації. Кожному об'єкту x_i, $i \in I$, ставиться у відповідність вектор числових характеристик $z_i = (z_{i,1}, \ldots, z_{i,j}, \ldots, z_{i,n})$. Кардинальність вектора n визначає розмірність простору характеристик. Відстань $\rho(xi,xk)$ між об'єктами xi і xk – результат застосування обраної метрики в просторі характеристик. У цей час, існує велика кількість метрик для оцінки відстані між векторами того самого векторного простору. До простих метрик можна віднести евклідову відстань або її квадрат, манхеттенську відстань, степенна відстань і інші [3]. До складних метрик можна віднести відстань між центрами ваги, групову середню відстань, відстань Чебишева й інші.

Перенесемо математичний опис абстрактних об'єктів в область дослідження пошукових запитів ІК. Нехай USR – множина ІК, за якими спостерігаємо, usri – і-ий ІК, за яким спостерігаємо, usri $\in$ USR. У довільний момент часу $t_k \in T$, зазначений ІК можна представити характеристичним вектором наступного виду:

$$u_i(t_k) = (u_{i,1}(t_k), \ldots, u_{i,j}(t_k), \ldots, u_{i,nof(Vu)}(t_k)), \quad (1.2)$$

де $u_{i,j}(t_k)$ – вага *j*-ого пошукового терміна із глобального словника термінів Vu у момент часу tk, що дорівнює числу входжень цього терміна в запити в пошуковій історії і-го ІК, протягом спостережуваного тимчасового вікна $\Delta\tau$;

nof(V) – розмір вектора і-го ІК, дорівнює числу слів у глобальному словнику термінів Vu.

Числові координати $u_{i,j}(tk)$, $1 \le j \le nof(Vu)$ розташовані в характеристичному векторі в порядку, відповідному до лексикографічного порядку проходження відповідних термінів у словнику Vu. Перехід від вербального до числового представлення результатів відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у запити пошукової історії ІК.

Методи кластерного аналізу широко застосовуються для вирішення широкого спектра завдань в Інтернеті. Кластерні методи зі складними алгоритмами оптимізації застосовуються в пошукових системах, Інтернет-магазинах, системах аналізу контенту сайтів, системах перевірки текстів і ще в багатьох сферах. Методи кластерного аналізу можна розбити на декілька груп: по способу обробки даних: ієрархічні (агломеративні методи й дивизивні методи); неієрархічні методи (ітеративні); по способу аналізу даних: чіткі й нечіткі; по числу застосувань алгоритмів кластеризації: з одноетапною кластеризацією і з багатоетапною кластеризацією; по можливості розширення обсягу оброблюваних даних: масштабовані й немасштабовані; за часом виконання кластеризації: потокові (у режимі реального часу) і непотокові (по нагромадженню інформації).

Існує ключова різниця між поняттям кластеризація й поняттям класифікація. Кластеризація дозволяє розбити множину об'єктів на групи

(кластери), а класифікація – відносить кожний об'єкт до однієї із заздалегідь певних груп. У процесі виконання завдань кластеризації-класифікації можна виділити чотири групи завдань:

– виділення характеристик об'єктів;

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

– розбивка об'єктів на групи із застосуванням методів кластерного аналізу; класифікація об'єкта, що знову з'явився.

У [6] запропонована 8-мі етапна схема виконання завдань класифікації. Кожний етап цієї схеми являє собою повноцінний процес із вхідними й вихідними потоками, можливий і зворотний зв'язок (рис. 1.4).

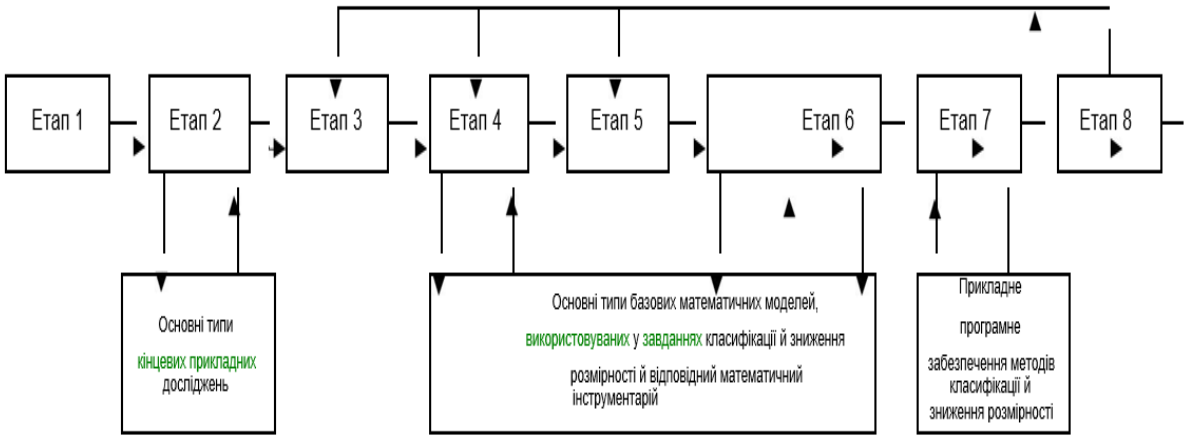

Рисунок 1.4 – Схема поетапного процесу виконання завдань класифікації [3]

Етап 1. Настановний – на цьому етапі повинна бути сформульована постановка завдання, що включає в себе характер наукових або практичних висновків, які потрібно одержати на виході.

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

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

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

Етап 5. Розвідницький аналіз – цей етап становлять усілякі методи попередньої статистичної обробки, пропущених вихідних даних з метою виявлення специфіки їх імовірнісної й геометричної природи. На виході етапу повинні бути уточнені відомості про фізичний механізм генерування наших вихідних даних, а виходить, про базову математичну модель цього механізму.

Етап 6. Апостеріорний математико-постановочний – на цьому етапі уточнюється математична постановка розв'язуваного завдання з урахуванням висновків.

Етап 7. Обчислювальний – проводиться обчислювальна реалізація наміченого до використання, обраного на попередньому етапі, математичного інструментарію рішення завдання.

Етап 8. Підсумковий – аналізуються й інтерпретуються результати проробленої роботи. Залежно від результатів цього аналізу, або формуються остаточні наукові чи прикладні висновки, або даються уточнення й доповнення до завдання й 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 і обраним математичним моделям, а якщо ні, то необхідно повернутися до одного з попередніх етапів для доведення ухвалених рішень.

1.4 Математичні моделі кластерних методів

Ієрархічні алгоритми можна віднести до нечітких алгоритмів кластерізації, тому що число кластерів заздалегідь невідоме. У чистому виді ієрархічна кластерізація є одноетапною, немасштабованою й непотоковою, тому що застосовується єдиний алгоритм кластерізації за принципом віддаленості (найближчого або далекого) сусіда на основі статистики конкретних досліджуваних об'єктів.

В якості вихідних даних виступають пошукові запити ІК. Широко застосовувані асоціативні й статичні методи можуть бути використані для попереднього сортування об'єктів. Зокрема, ці методи можуть бути використані для визначення тривалості пошукового інтересу в мережі або для добору схожих товарів в Інтернет-магазинах. Для роботи зі складними об'єктами такими, як ІК та ІР можуть застосовуватися ієрархічні алгоритми кластеризації, тому що агломеративні й дивизивні ієрархічні алгоритми відмінно справляються із завданням згрупування об'єктів з більшою кількістю ознак – властивостей або характеристик. Отже, якщо досліджувати пошукову діяльність ІК за якийсь період часу Δt, то наприкінці періоду спостереження кожному користувачеві можна буде співставити характеризуючий вектор, координати якого будуть використані для побудови ієрархій [7].

Дослідження структури ієрархій зручно вести в термінах теорії графів.У графові ієрархії вершина може бути кінцем декількох стрілок, але вона є початком тільки однієї стрілки. Ієрархія є бінарною, тоді й тільки тоді, коли в її графові кожна вершина, відповідна до нескінченності, що містить більш одного елемента, є кінцем двох стрілок .

Ієрархічною класифікацією множина і об'єктів $X = \{X_1,\ldots, Xnof_{(X)}\}$ називається побудова ієрархії на X, що відбиває наявність однорідних, у певному змісті, класів X і взаємозв'язки між класами (рис. 1.5). Алгоритми ієрархічної

класифікації бувають дивизивні, у яких початкова нескінченність X поступово розділяється на

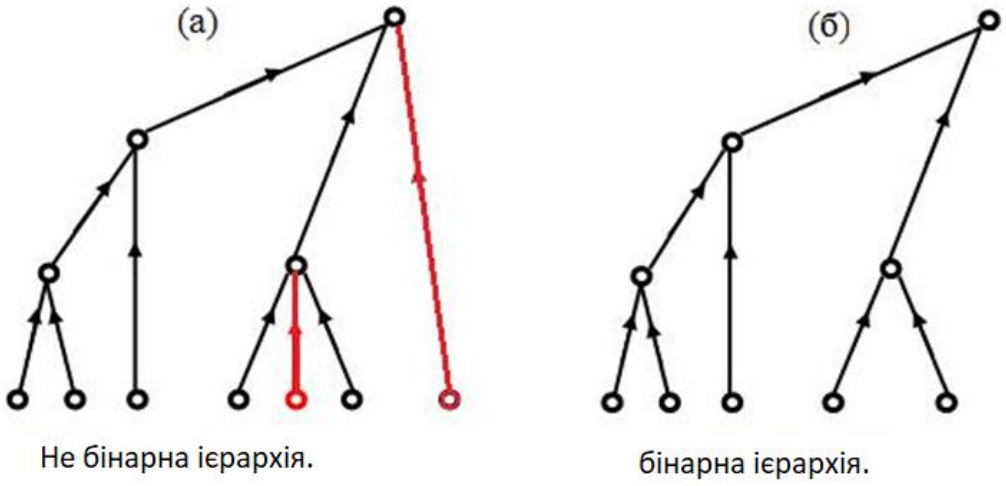

Рисунок 1.5 – Бінарна (б) і не бінарна (а) ієрархії

усе більш дрібні підмножини, і агломеративні – у яких крапки нескінченності X поступово поєднуються в усе більші підмножини. Отримані графи ієрархій за допомогою цих алгоритмів називаються, відповідно, дивизивними й агломеративними.

Відмінними від ієрархічних (агломеративних і дивизивних) методів кластеризації, є ітераційні методи, наприклад метод *k*-середніх, який використовує в якості вхідного параметра число класів, на які проводиться розбивка нескінченності X, або метод Форель, для якого необхідно вказати радіус куль, якими покривається вибірка X. Враховуючи сказане,
ітераційні методи слід віднести до точних методів кластеризації.

1.5 Постановка задач дослідження

Показано, що статичні методи класифікації, засновані на персональній інформації ІК (стать, вік, місце проживання й ін.), широко застосовуються в соціальних мережах для цільового добору Інтернет-реклами, надання ІК іншої

адресної інформації. Некластерні методи класифікації (асоціативний метод, метод перетинань і ін.) широко застосовуються в Інтернеті для добору «схожого» або «супутнього» товару в Інтернет-магазинах. Незважаючи на простоту й поширеність статичних і некластерних методів, вони не можуть бути використані для персоналізації пошуку в широкому розумінні цього слова: ці методи можуть бути спрямовані лише на досягнення локальної мети – одержання комерційної вигоди.

За аналізом методу частоти термінів доведено, що кожна соціально-демографічна група ІК може бути охарактеризована особистим графіком ваг слів, що відбивають інтереси користувачів, що входять у групу. Як виявилося, кожна СД група унікальна у своїй пошуковій поведінці, звідси випливає, що ІК можуть бути виділені в окремі групи за статевою й віковою приналежністю й це автоматично приведе до поділу за інтересами, відповідними до цих груп. Незважаючи на індивідуальність кожного ІК, їх можна й потрібно групувати, як за допомогою статичної інформації (стать, вік і т.д.), так і на основі аналізу їх пошукової діяльності в Інтернеті.

Досліджені основні методи класичного кластерного аналізу, коли об'єкти дослідження представлені за допомогою числових характеристичних векторів, після попереднього формування глобального словника термінів і застосування методу позиційного кодування. Кластеризація дозволяє розбити велику кількість об'єктів на групи (кластери), а класифікація – відносить кожний об'єкт до однієї із заздалегідь встановлених груп. Для виконання завдань кластеризації й класифікації була обрана схема рішення відповідних завдань, що полягає з восьми ключових етапів.

В роботі поставлено наступні задачи.

– розробити алгоритми використання результатів кластеризації для персоналізації пошуку – результати аналізу пошукової активності ІК у поточному інтервалі часу можуть бути застосовані для прогнозу його інформаційних потреб у наступних інтервалах часу;

– розробити набір програмних модулів для спостереження за активністю ІК і одержання текстового змісту IP з обліком їх Dom-моделей;

– розробити спеціальні збережені процедури, що виконують усі необхідні розрахунки – від формування словників термінів до кінцевого розподілу об'єктів по кластерах;

– зазначені модулі й збережені процедури мають утворити єдину програмну систему, яка, бувша встановленою на сервери локальної мережі, що дозволить організувати на підприємстві корпоративну систему персоналізації пошуку

2 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

2.1 Методи аналізу змісту тексту

Лінгвістичний аналіз – метод дослідження тексту, який може бути охарактеризований як лінгвосмисловий аналіз. Він являє собою вивчення методів, що дозволяють автоматизовано «розуміти» текст, тобто вміти витягати з нього потрібну інформацію й відповідати на запитання, задані щодо тексту. Лінгвістичний аналіз застосовується, зокрема, до витягу інформації, машинного перекладу, а також до багатьох областей штучного інтелекту, що стосуються спілкування з користувачем.

Існує множина підходів до лінгвістичного аналізу. Серед них можна виділити статистичний аналіз, аналіз ознак, семантичний аналіз і комбінований підхід.

Статистичний аналіз має на увазі вивчення послідовності слів тексту, а також робить висновок певних закономірностей на підставі проведеного вивчення. Для цього проводиться підрахунок частоти зустрічі слів у тексті, а також імовірність появи слів один за одним. За допомогою статистичних методів аналізу тексту вирішується проблема класифікації текстів [7]. Відповідно, аналізуючи зміст слів, можна отримати ймовірності зустрічі N-грами або частоту термінів у науковому тексті. Після цього знайдені ймовірності переводяться у ваги й складаються. Текст буде належати до того класу, вага якого виявиться більше. Аналіз ознак полягає у вивченні морфемних, морфологічних синтаксичних ознак слів і пропозицій у тексті. Це вивчення необхідне для того, щоб далі можна було будувати модель (структуру) пропозицій і на її підставі витягати необхідну інформацію.Для побудови такої структури необхідно зробити кілька операцій. Перша полягає у визначенні граматичних ознак усіх слів. Для цього в мовах, у яких існують відмінки, дієвідміни, відмінювання й часи, встановлюються всі ці ознаки. Також визначається лема кожного слова – його словникова форма. Для визначення леми слова використовується алгоритм-

млематизатор, який або за допомогою порівняння зі словником, або за допомогою послідовного відсікання закінчень і афіксів (префіксів, суфіксів і постфіксів) і додавання нормалізованого закінчення виділяє основу слова. Друга операція зводиться до побудови моделі пропозицій у тексті. За допомогою аналізу граматичних ознак, знайдених раніше, складається залежність слів одне від одного – спочатку в межах пропозиції, а потім і в межах тексту, якщо необхідний більш широкий аналіз. Третя операція зводиться до пошуку потрібної інформації. Користувацький запит проходить лематизацію, після чого відбувається пошук необхідних лем по всім пропозиціям. Однак те, що не представляє великої праці для лінгвіста й звичайної людини, є великою проблемою для обчислювальної системи. Описуваний вище аналіз складно чітко сформулювати, тому що, наприклад, слово «історії» можна розглядати не тільки як давальний відмінок, але і як називний/ знахідний відмінок множини слова «історія». Окремою завадою є омоніми («ягуар» – це автомобільний бренд, енергетичний напій або хижак?), тобто співзвучні слова; якщо в тексті зустрічається кілька слів з однаковими лемами, немає гарантії того, що це – те саме слово. Для з'ясування цього необхідно провести більш складний кластерний аналіз для визначення змісту текстів ІР у цілому після одержання характеристичних векторів. Без кластерного аналізу (або без застосування методів класифікації текстів) значень слів такі неоднозначності досить складно пояснити. Ще однієї значною проблемою є розбір відсутніх у словнику слів, тоді потрібно звертатися до лінгвістичного експерта, який у свою чергу вносить зміни в словнику лем. Чітко заданих правил для такого аналізу не існує; їх можна вивести експериментальним шляхом, однак завжди будуть існувати виключення, опрацювати які за правилами буде неможливо.

Семантичний аналіз дозволяє виявляти незв'язність слів і речень у тексті, хоча вони й можуть бути погоджені граматично. Також семантичний аналіз дозволяє визначати метафори, переносні значення, дійсний зміст співзвучних слів залежно від контексту, і т.д. Прості способи семантичного аналізу дозволяють класифікувати текст, виділяти емоційне забарвлення тексту (за допомогою

виявлення певних слів і аналізу словосполучень, що містять метафори і алегорії) і його тему (по синтаксичних ознаках і кількості повторюваних слів у реченнях). Зокрема, за допомогою семантичного аналізу відбувається видача контекстної реклами на багатьох сайтах і в пошукових системах [10]. Сторінка, запропонована користувачеві, досліджується на предмет наявності повторюваних ключових слів, після чого автоматизований генератор реклами видає пов'язану зі знайденими ключовими словами вибірку.

Комбінований підхід має на меті використання декількох з вищеописаних підходів у поєднанні, послідовному або паралельному, для підвищення точності аналізу. Найчастіше для складного аналізу тексту застосовують аналіз ознак, сполучений зі статистичним аналізом для ранжирування результату пошуку й визначенню неоднозначностей; рідше використовуються вкраплення семантичного аналізу в кожному з вищеописаних методів. Зокрема, такий підхід використовується в текстових редакторах для виявлення складних помилок (неузгодженість тексту, рекомендації з розбивки тексту на абзаци, і т.д.).

У рамках даної роботи, обробка тексту проводиться за допомогою комбінованого підходу на підставі статистичного методу й методу аналізу ознак.

2.2 Алгоритми лінгвістичної обробки запитів і текстів Інтернет-ресурсів

Для застосування методів лінгвістичного аналізу необхідні леми всіх слів тексту (аналіз ознак) і частота зустрічі цих лем у тексті (статистичний аналіз). Слід зазначити, що аналіз моделей зв'язків між словами усередині речень і реченнями усередині тексту в даній роботі не розглядається.

Першим кроком до рішення поставленої проблеми аналізу змісту, як ІР, так і запитів ІК є розбивка тексту Інтернет-сторінок і Інтернет-запитів на окремі слова (терміни), тому що текст у формальному визначенні є просто набором слів. Тут уже можливі проблеми й неоднозначні трактування: що вважати словом, як

ставитися до складних знаків пунктуації і т.д. Уведемо набір простих правил, що описують більшість випадків, які можуть зустрітися в апріорно правильному тексті.

Словом або терміном назвемо послідовність символів-букв, обмежених по обидва боки пробілами або розділовими знаками, у якому можуть бути цифри, у тому числі й на першій позиції. Усі розділові знаки й спеціальні символи («+», «-», «/», «=» і т.д.) заміняються пробілами, тим самим, безперервна послідовність символів перетворюється в слова, відділені одне від одного пробілами.

Лемою або коренем слова будемо вважати урізану послідовність символів терміна, одержану за допомогою спеціально розробленого алгоритму відсікання закінчень із урахуванням дворівневого словника, що включає глобальний словник термінів і словник лем, який може заповнюватися за допомогою існуючих відкритих словників [11] і динамічно поповнюватися з появою нових термінів. Очевидно, що декільком термінам може відповідати одна лема.

Зі сформульованих правил випливає, що будь-які тексти дійсно підійдуть під описи, наведені вище. «Правильними» текстами в цьому випадку будуть вважатися тексти, як на російській, так і на англійській мовах, що перебувають у відкритому доступі в мережі Інтернет, й допускають комп'ютерну обробку. Апріорі будемо вважати, що ІР не містить неприпустимі для заданої мови символи або тексти із синтаксичними друкарськими помилками. Усі спеціальні символи фільтруються. Виявлення синтаксичних друкарських помилок – окреме серйозне завдання, яке може бути відведене лінгвістичному експертові, який, у свою чергу, може внести виправлення в словник лем. Користуючись довідниками [12], вдалося написати алгоритм заповнення словника термінів і лем, що відповідають наведеним вище правилам.

Може застосовуватися комбінований підхід, заснований на статистичному методі, методі аналізу ознак і особливостях Dom-моделей ІР. Семантичний аналіз вимагає додаткового дослідження, і не буде розглядатися в рамках поточної роботи. На рис. 2.1 представлені етапи процесу обробки змісту запитів ІК і текстів ІР від первинного «брудного» тексту до лем. Результатом виконання цього

процесу є формування статистики лем і нарощування словника за допомогою лінгвістичного експерта. Схема алгоритму лінгвістичної обробки термінів (слів) наведено на рис. 2.2.

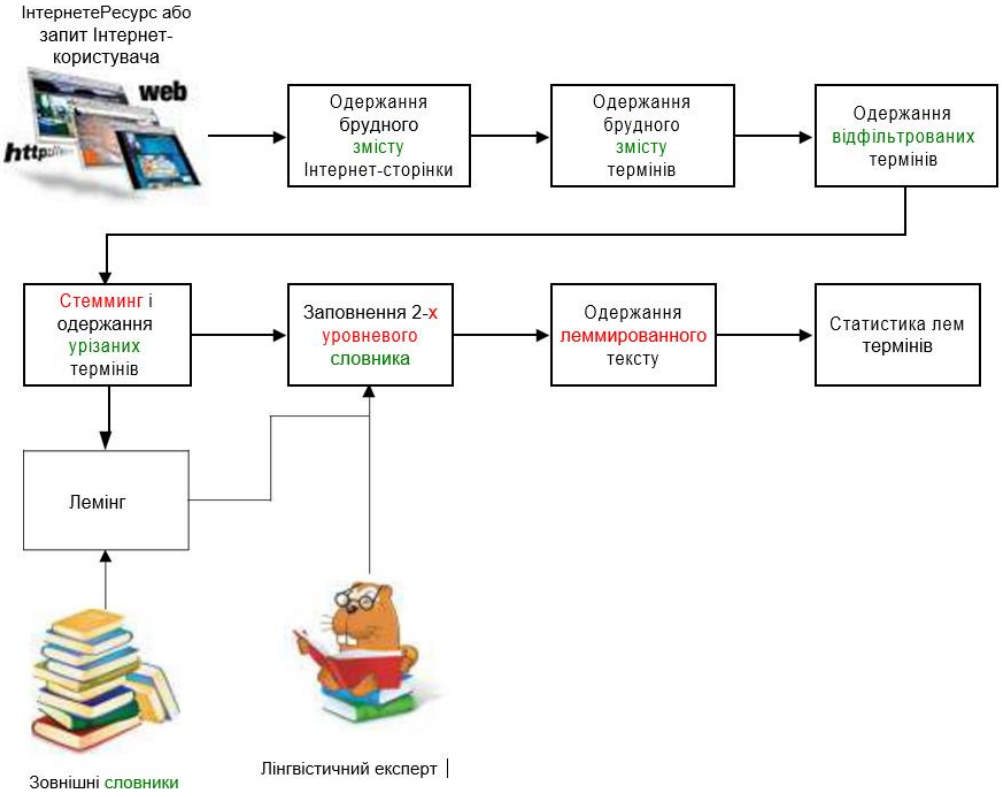

Рисунок 2.1 – Процес лінгвістичної обробки запитів ІК та текстів ІР

Через те, що на етапі роботи розглянутого алгоритму відбувається повна перевірка слів з тексту ІР або пошукового запиту ІК, доцільно проводити тут же підрахунок зустрічі слів у тексті, застосовуючи відомі статистичні методи. Потім, на підставі отриманої інформації будуть сформовані характеристичні вектори класифікованих Інтернет-об'єктів і глобальний словник термінів, з якими будуть далі працювати алгоритми кластерного аналізу.

В алгоритмі рис. 2.2, можна скористатися особливостями Dom-моделі ІР, що дозволяє локалізувати позиції термінів і тим самим виділяти особливо важливі терміни (наприклад, заголовки й найменування Інтернет-сторінок), з метою підвищення значень числових характеристик у ключових термінах або навіть відсортування динамічних елементів, «що прослизнули» при читанні змісту текстів ІР.

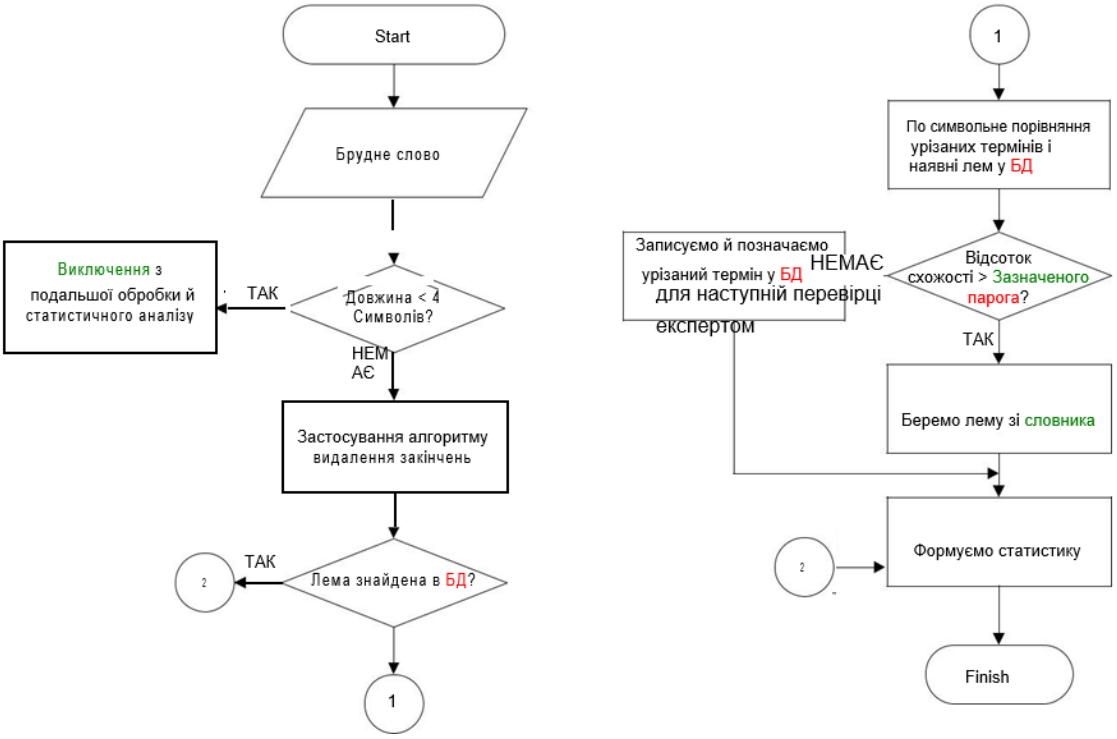

Рисунок 2.2 – Схема алгоритму лінгвістичної обробки термінів

За допомогою описаного вище алгоритму, текст IP розбивається на терміни, які вже можна обробляти. Необхідно звернути увагу на процес обробки термінів, а точніше на стемінг (stemming) і лематизацію термінів для приведення слів до початкової форми й тим самим згрупувати їх, що, у свою чергу, зробить статистику термінів більш достовірною.

Початковою стадією роботи майже всіх лематизаторів є стемінг. Цей процес має на увазі виділення кореня слова, тоді як лематизатор на підставі кореня підбирає базову форму, що підходить для нього. Одними з визнаних фаворитів серед стемерів є програми, що реалізують алгоритми усікання закінчень [13]. Алгоритми усікання закінчень можуть бути використані як для російських, так і для англійських термінів, а отримані згодом усічені терміни можуть бути застосовані в якості лем для формування статистики первісних слів з потрібного тексту. У добавок алгоритм працює з великою кількістю закінчень, розбитих на кілька підмножин: закінчення герундія («в», «возі» – деякі закінчення

дієприслівників для української мови або «ing» для англійської мови), закінчення прикметників («ий», «ими», «є»), дієприслівникові суфікси («ючи», «вши», «ши»), зворотні постфікси («сь», «ся»), дієслівні закінчення («ла», «чи», «їм»), префікси (« до», « за», « над») і закінчення іменників.

Процес стемінгу легко реалізується з використанням БД, для чого досить скористатися спеціальним словником масок (маски на T-SQL представлені набором символів, у яких повинні бути присутні символи «%» і/або «_»), що можуть бути сполученими зі списком спеціальних символів для фільтрації. Усічена форма, що вийшла у результаті обробки термінів, перевіряється по БД за допомогою словника лем. Якщо лема знайдена, то оригінальний термін прив’язується до знайденої леми (рис. 2.2). Якщо урізаний термін неможливо зпівставити з лемою зі словника, то він посимвольно порівнюється з лемами зі словника. На останніх кроках алгоритму в словник лем включаються урізані терміни, які неможливо було зіставити з вже наявними лемами. Вони позначаються, щоб лінгвістичний експерт зміг провести перевірку й, при необхідності, поправити або призначити нові леми. Такий алгоритм сам по собі універсальний і досить точний, однак і помилки, як показала практика, трапляються досить часто, наприклад, при наявності в словах префіксів, що може привести до одержання несловникового кореня або до надмірного усікання слів. Однак цю проблему завжди може розв'язати кваліфікований лінгвістичний експерт.

У запропонованому алгоритмі достатньо скористатися дворівневим словником і для статистичної обробки тексту, і для формування характеристичних векторів. Якщо для обробки термінів будуть потрібно інші проміжні результати стемінга, то алгоритм потрібно буде переналаштувати на систему словників більш високого рівня (наприклад, 3- або 4- рівневих словників), додаючи до нього блоки обробки додаткового словника.

Таким чином, наступна послідовність дій по перетворенню вихідного запиту ІК/тексту ІР у об’єкт, придатний для подальшої обробки алгоритмами кластерного аналізу:

– виділення всіх термінів із запиту ІК/тексту ІР;

– стемінг і отримання урізаних термінів після видалення закінчень;

– перевірка урізаних термінів по словникові лем БД. Якщо лема знайдена, перехід до пункту д).

– збереження позначених урізаних термінів, для яких не визначена лема із БД. При необхідності лінгвістичний експерт може підтвердити або змінити позначені урізані терміни, перетворюючи їх у нові леми;

– формування статистики термінів і характеристичних векторів.

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

2.3 Методи динамічної зміни в кластерній структурі Інтернет-об'єктів

Сучасні Інтернет-ресурси є динамічними об'єктами. Якби вони містили винятково статичні компоненти, то розрахунок центрів відповідних кластерів можна було б проводити в дискретні моменти часу, у моменти появи нових ІР. Кластеризація Інтернет-користувачів також носить динамічний характер, тому що людська поведінка є динамічним процесом і це відбивається в пошукових історіях ІК. У завданнях кластеризації ІК потрібно враховувати не тільки поточні характеристики об'єктів, що кластерізуються, але і їх тимчасові, тобто динамічні зміни за фактом появи нових пошукових запитів. У цій главі досліджуються динамічні зміни кластерної структури Інтернет-об'єктів.

Нехай X нескінчена кількість усіх об'єктів для спостереження $x_i \in X$, $1 \le i \le nof(X)$, віднесених до одного із кластерів $X_l \subseteq X$, $1 \le l \le nof(K)$, де $K = \{X_1, \ldots, Xl, \ldots, Xnof_{(K)}\}$ – нескінчена кількість усіх сформованих кластерів. У різні моменти часу $t_k \in T$, $k = 0,1,2, \ldots$ проводимо спостереження за зміною стану кластерної структури залежно від характеристик об'єктів xi, при цьому стан кожного *i*-го об'єкта в довільний момент часу tk відображається

характеристичним вектором $z_i(t_k)$. Тут необхідно говорити про часову складову як додаткового параметра для всіх елементів вектора, що характеризує об'єкт. Якщо об'єкт дослідження при ієрархічній кластеризації представлений вектором zi = ($z_{i,1}$, …, $z_{i,j}$, …, $z_{i,n}$), який не залежить від часу, то в динамічній системі кластеризації необхідно говорити про вектор $z_i(t_k)$ = ($z_{i,1}(t_k)$, …, $z_{i,j}(t_k)$, …, $z_{i,n}(t_k)$), 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, усередині яких об'єкти мають загальні характеристики. Зміна характеристик об'єкта $u_i \in U$ у момент часу t_k може привести до глобальних змін на рівні всієї кластерної структури й тим самим через період часу Δt_k буде необхідно провести нову кластеризацію всіх об'єктів з U. Якщо після витікання часу Δtk число кластерів, їх зміст, розміри й положення їх центрів не змінюються, то мова може йти про так звану статичну кластерну структуру. Однак зовсім інакше виглядає справа в ситуаціях, коли із часом кластерна структура змінюється, коли об'єкти із часом починають отримувати деякі нові характеристики й утворюють групу об'єктів, функціонування яких перебуває на межі кластерів або навіть за його межами. У такому випадку кластерна структура переносить тимчасові зміни й стає динамічною.

Кластеризація розглядається, як на завдання моніторингу сукупності Інтернет-об'єктів з n-мірним характеристичним вектором числових ознак $z_i(tk)$, де індекс і відповідає номеру об'єкта. Вимір характеристик даних об'єктів здійснюється в дискретні, не обов'язково рівновіддалені моменти часу t_k. Через інтервал часу Δt_k проводиться перевірка стабільності кластерної структури й при необхідності, тобто при наявності динамічних змін, її корекція. Наприклад, для кластеризації ІК, з метою виключення впливу занадто старих спостережень, моніторинг стану кластерів доцільно організувати за принципом тимчасового вікна, тобто в nof(Vu)-мірному просторі при аналізі кластерної структури повинні враховувася тільки об'єкти, зафіксовані в останньому тимчасовому вікні Δt_k. Якщо говорити про тимчасове вікно для обліку нових Інтернет-об'єктів можна припустити наступні динамічні зміни в структурі кластерів: утворення нових

кластерів, злиття кластерів, розщеплення або дроблення кластерів, зникнення кластерів, переміщення центрів кластерів [15].

Для характеристичних векторів ІК $U(tk_{+1})$ у момент часу $tk_{+1} > tk$ 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 $u_{nof}(U(tk))+p(tk_{+1}) \in Y(tk_{+1})$, $p \geq 1$, являє собою новий, одиночний ізольований кластер, для якого проводиться розрахунок заходу близькості (евклідової відстані) до всіх інших об'єктів $U(tk_{+1}) \backslash \{u_{nof}(U(tk))+p(tk_{+1})\}$. За результатами розрахунку евклідової відстані визначається найближчий до $u_{nof}(U(tk))+p(tk+1)$ об'єкт $u_{near}(tk+1)$:

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

Визначення найближчого сусіда $u_{near}(t_{k+1})$ дозволить ініціалізувати місце положення нового об'єкта в новій кластерній структурі.

Продовжується спостереження за Інтернет-користувачем u_i у період часу Δt_k. З появою нових об'єктів ІК збільшується глобальний словник термінів Vu. Якщо на момент часу t_{k+1} (у попередній інтервал часу Δt_k) ІК $ui(t_{k+1})$ не проводив ніяку пошукову діяльність, то збільшення словника термінів Vu ніяк не відображається на його характеристичному векторі, а лише призводить до появи нових нульових координат: якщо в момент часу t_k характеристичний вектор ІК мав такий вид

$$ui\ (tk\ ) = (\upsilon\iota,1(\tau\kappa\ )^{,ui},2\ (\tau\kappa\ )^{,\ldots,ui},nof\ (Vu\ )\ (\tau\kappa\ )),$$

то в момент часу tk_{+1} він перетворився у вид:

$$u_i\ (t_{k\ +1}) = (\upsilon\iota,_1(\tau\kappa\ +1\ )^{,\ ui},2\ (\tau\kappa\ +\mathbf{1}\ )^{,\ldots,ui},nof\ (V\ )\ (\tau\kappa\ )^{,0,0,\ldots,0}).$$

На рис. 2.1 представлена ілюстрація первісної кластерної структури в момент часу t_k.

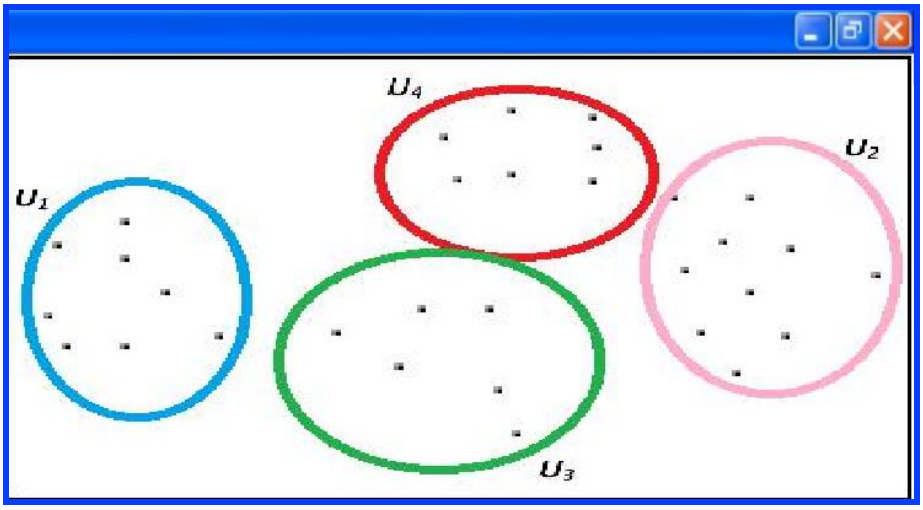

Рисунок 2.3 – Ілюстрація розподілу об'єктів по кластерах у момент часу t_k

Якщо експериментальним шляхом в інтервалі часу Δt_k з'явилися нові об'єкти $unof(U(t_k))+p(t_{k+1}) \in Y(t_{k+1})$, $p \geq 1$, то після проведення повторної кластеризації з'являються нові кластери (рис. 2.4), сформовані за допомогою нових об'єктів.

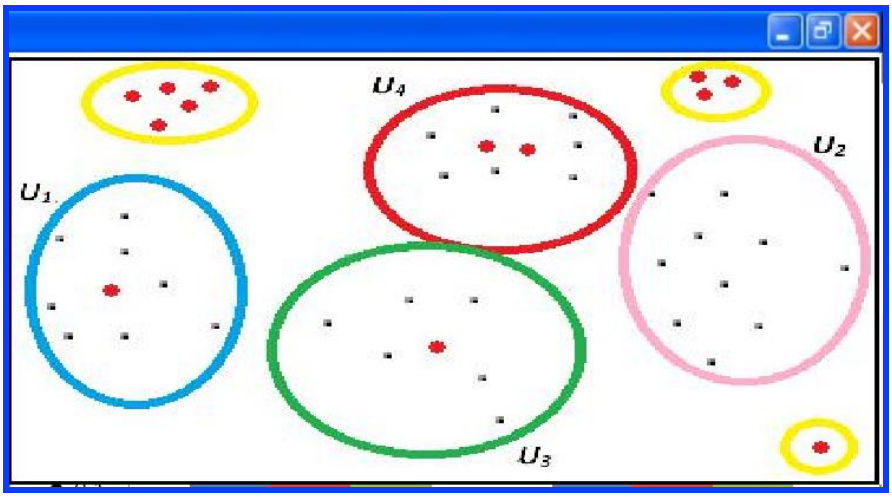

Рисунок 2.4 – Ілюстрація розподілу об'єктів по кластерах у момент часу t_{k+1}

На рис. 2.4 показано формування нових кластерів з, об'єктів, що знову з'явилися, одиночні кластери, кардинальне число яких рівно 1, і насичення (нарощування) уже сформованих кластерів за рахунок появи нових об'єктів.

Якщо на момент часу t_{k+2} пошукова діяльність одного або декількох користувачів різко змінюється, характеристичні вектори цих об'єктів міняють значення одного або декількох своїх координат і, як наслідок, структура кластера змінюється, що, у свою чергу, приводить до зміни значення точок близькості й відстані між об'єктами одного кластера. Це може призвести до формування нових кластерів (рис. 2.5).

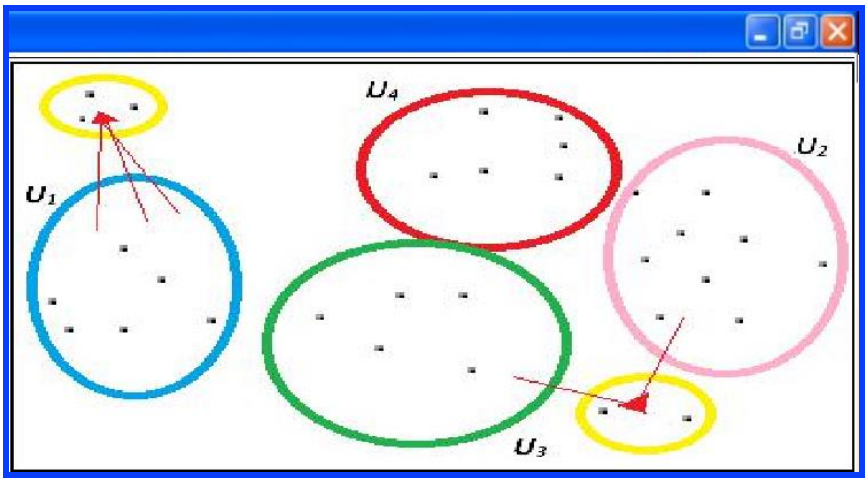

Рисунок 2.5 – Ілюстрація перерозподілу об'єктів у момент часу t_{k+2}

Для вирішення завдання появи нових кластерів можна скористатися коефіцієнтом приналежності *i*-ого об'єкта до *m*-ого кластера.

Коефіцієнт приналежності $bi_{,m}(tk_{+2})$ об'єкта $ui(tk_{+2})$ до кластера $Um(tk_{+2})$ із кластерної структури $K(tk_{+2})$ ($Um(tk_{+2}) \in K(tk_{+2})$) у довільний момент часу t_{k+2}:

$$b_{i,m}(t_{k+2}) = \frac{1}{\sum_{l=1}^{nof(K(t_{k+2}))} \left(\frac{(\rho_m(t_{k+2}))^2}{(\rho_l(t_{k+2}))^2} \right)} \quad \text{и} \quad \sum_{m=1}^{nof(K(t_{k+2}))} b_{i,m}(t_{k+2}) = 1, \tag{2.1}$$

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

$\rho_m(t_{k+2}) = \sqrt{\sum_{j=1}^{nof(V_m(t_{k+2}))} \left(e_{m,j}(t_{k+2}) - u_{i,j}(t_{k+2})\right)^2}$ – евклидова відстань між об'єктом u_i і центром e_m m-ого кластера кластерної структури ІК. Виявлення нових кластерів Knew починається з виявлення нових об'єктів U_{free} з малими ступенями приналежності

до всіх існуючих кластерів. Якщо число таких вільних об'єктів nof(U_{free}) порівнюють з розмірами кластерів, вони формують компактну групу об'єктів із загальними властивостями. Компактність об'єктів, є ознакою появи нових кластерів. Можливе число нових кластерів nof (k_{new}) в момент часу t_k визначається співвідношенням:

$$nof\left(K^{new}\right) = int\left(\frac{nof\left(U^{free}\right)}{d_1 \times N_{min}}\right), \tag{2.2}$$

де int(...) – ціла частина аргументу;

nof(U^{free}) – число вільних об'єктів, які не прив'язані до жодного із кластерів;

d_1 – задана гранична величина в інтервалі [0,1];

Nmin = min (νοϕ (Y_1^*), ..., nof (υλ*), ...,nof (υνοϕ* $_{(K\ )}$)) мінімальний розмір кластера, при обчисленні якого враховуються тільки «гарні» об'єкти Ul* з досить великими значеннями ступенів приналежності;

nof (υλ*) – число гарних об'єктів l-ого кластера, для яких значення ступені приналежності до зазначеного кластера $b_{i,l} \geq d_2$;

d_2 – задана гранична величина в інтервалі [0,1].

3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ

3.1 Розробка алгоритмів зміни в структурі кластерів

По визначенню, злиття кластерів – це формування нової розбивки, коли nof(K’) кількість кластерів перетворюються в nof(K’’) кількість кластерів, причому nof(K”) < nof(K’).

Нехай злиття кластерів відбувається в момент часу t_{k+3}. Зі зміною пошукового вектора деякі кластери можуть наблизитися одне до одного й злитися в єдину групу (рис. 3.1).

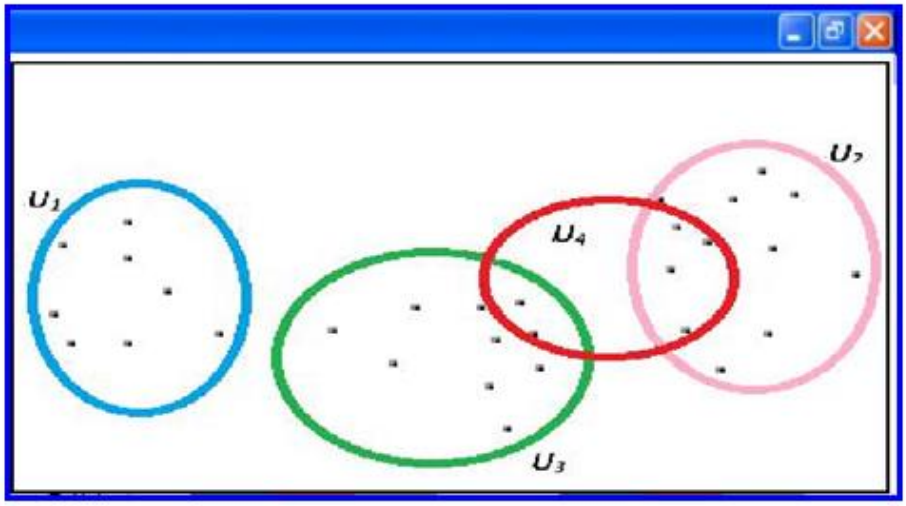

Рисунок 3.1 – Ілюстрація злиття кластера U_4 і перерозподіл його об'єктів і центру між кластерами U_3 і U_2 в tk_{+3}

Виявлення кластерів, що зливаються, починається з виділення об'єктів ui, що мають високі ступені приналежності (3.1) одночасно для двох кластерів U_l і U_m : $bi_{,l} \approx bi_{,m} \rightarrow 1$. Якщо число таких об'єктів, що зливаються, досить велике, то це і є ознакою злиття кластерів. Для ітеративних методів кластеризації (методу *k*-середніх) спостерігається зближення центрів кластерів між кластерами, що зливаються. Кількісним критерієм для оголошення двох кластерів кластерами, що зливаються, може бути їх подібності:

$$I_{l,m} = \frac{\sum_{i=1}^{nof(U)} \min(b_{i,l}, b_{i,m})}{\sum_{i=1}^{nof(U)} b_{i,l}}, \quad (3.1)$$

де $b_{i,l}$ і $b_{i,m}$ – ступені приналежності *i*-го об'єкта до кластерів Ul і Um, відповідно.

Однак захід подібності $Il_{,m}$ не є симетричною, тому що $Il_{,m} \neq Im_{,l}$, тому для виявлення зазначеної подібності краще використовувати формулу

$$Mc_{l,m} = \max\left(I_{l,m}, I_{m,l}\right), \quad (3.2)$$

відповідно до якої кластери будуть вважатися такими, що зливаються, якщо значення $Mc_{l,m}$ перевищує деякий поріг h (при h = 0 усі кластери будуть вважатися злитими, при h = 1 кластерів, що злилися, ніколи не буде).

Якщо говорити про ієрархічну кластеризацію, то спостерігати за картиною злиття кластерів найкраще на ранніх стадіях їх формування, тому що на більш пізніх етапах відбувається збільшення кількості об'єктів у кластерах і як наслідок імовірність 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 і близькість їх центрів.

3.2 Алгоритм зміни в структурі кластерів

Розщеплення або дроблення кластерів можна спостерігати, наприклад, у момент часу t_{k+4}, коли деякі кластери збільшуються в розмірах через велику кількість нових об'єктів, що може привести до неоднорідності їх внутрішньої структури (рис. 3.2).

Багатоекстремальності гістограм ознак – в [13] запропоновано розглядати кластер, як розщеплений, якщо

$$R = \frac{f_{\max} - f_{\min}}{f_{\max}} \geq g\,, \tag{3.3}$$

де fmax і fmin – значення глобальних максимуму й мінімуму гістограми [13] хоча б для одного з компонентів характеристичного вектора, а – граничне значення.

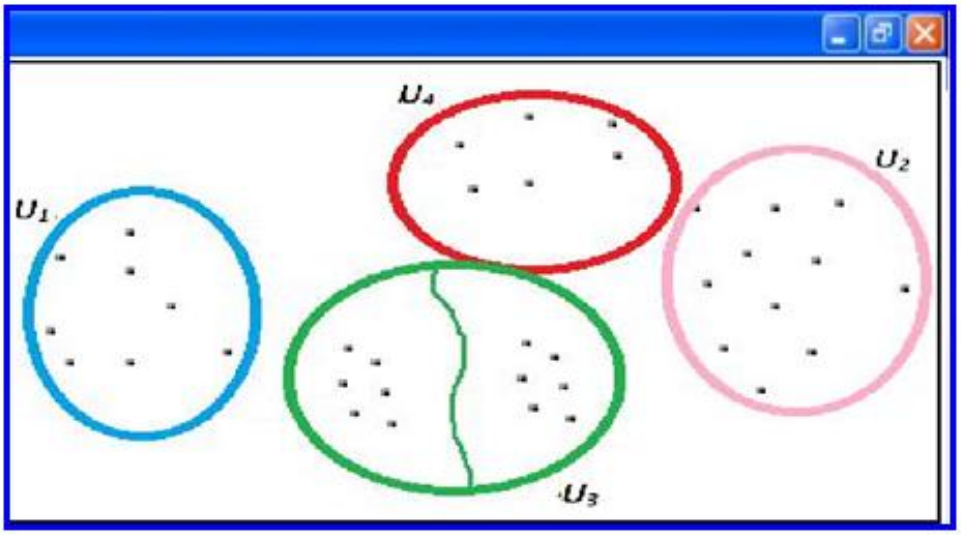

Рисунок 3.2 – Ілюстрація розщеплення кластера U3 і формування всередині нього двох розділених згустків в t_{k+4}

Ієрархічну кластеризацію розщеплення або дроблення кластерів можна виявити на більш пізніх етапах, тому що на більш ранніх стадіях кількість об'єктів і розмірність кластерів не дозволить спостерігати над процесом зміни однорідності малих кластерів. Це пов'язане не тільки з розмірами кластерів, склад яких може змінитися із часом, але й з характером самого алгоритму агломеративної кластеризації. На більш ранніх стадіях доцільніше говорити про злиття кластерів в ієрархічній системі.

Зникнення кластерів прямо пов'язане зі зникненням об'єктів цих кластерів або з перетворенням їх характеристичних векторів у нульові вектора.

Зникнення кластера відбувається, коли спостерігається повний перехід його об'єктів до складу іншого (інших) кластера (кластерів). Якщо в момент часу tk_{+4}

кластер $Ul(tk_{+4})$ містив $nof(Ul(tk_{+4}))$ елементів, а кластер $Um(tk_{+4})$ містив $nof(Um(tk_{+4}))$ елементів і $\overrightarrow{Ul(tk_{+4})} \cap \upsilon\mu(tk_{+4}) = \varnothing$, то зникнення кластера Um у момент часу tk_{+5} буде оголошено, якщо $Ul(tk_{+5}) = Ul(tk_{+4}) \cup \upsilon\mu(tk_{+4})$ і $Um(tk_{+5}) = \varnothing$.

Інша причина зникнення кластера пов'язана з перетворенням характеристичних векторів його об'єктів у нульові вектора. Нехай у момент часу tk_{+4} $Um(tk_{+4}) = \{um_1, ..., umi, ..., umnof_{(Um(tk+3))}\}$, де $umi \neq (0, ..., 0)$. Якщо в момент часу tk_{+5} $Um(tk_{+5}) = \{ 01, 02,...,0_{nof\ (Y\ m(t\ k\ +3\ ))} \}$, то кластер Um перестає існувати й можна сказати, що відбулося його зникнення.

В ієрархічних алгоритмах кластеризації немає поняття «центр кластера».Частково в агломеративному алгоритмі на більш високих рівнях ієрархії приналежність до кластера визначається мінімальною відстанню від будь-якого елемента цього кластера: розрахунок центрів кластерів відсутня. У методі k-середніх на кожній ітерації проводиться перерахунок координат центрів і в цих випадках стрибкоподібні переміщення центрів кластерів призводять до глобальних змін на рівні кластерів у цілому й на рівні об'єктів зокрема. Із часом поведінка ІП змінюється, їхні пошукові запити, які формують пошукові вектори, можуть викликати повільні зміни положень центрів кластерів – дрейф центрів кластерів (рис. 3.3).

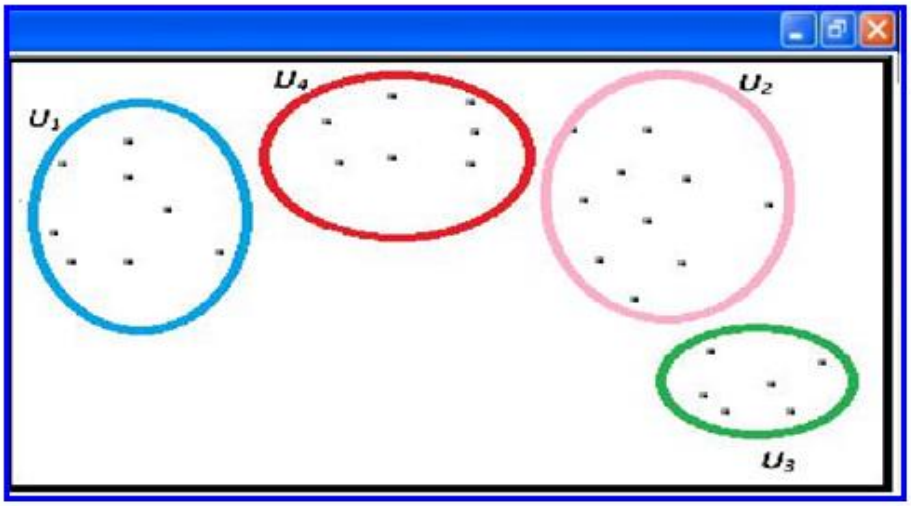

Рисунок 3.3 – Ілюстрація дрейфу кластерів у момент часу t_{k+6}

Такі зміни можуть бути досить незначними й носити безперервний характер, але їх доцільно виявляти й відслідковувати, оскільки вони, зрештою,

можуть стати причиною стрибкоподібних змін кластерної структури. Виявити ступінь дрейфу кластерів [50] можна з використанням формули компактності Pl кластера Ul стосовно віднесених до нього об'єктів:

$$P_l = \frac{\sum_{i=1}^{nof(U_l^*)} b_{i,l}^2 \times \rho(u_i^*, e_l)^2}{\sum_{i=1}^{nof(U_l^*)} b_{i,l}}, \tag{3.4}$$

де nof (Y_l^*)– число гарних об'єктів lгого кластера, для яких значення ступеня приналежності до зазначеного кластера $0 \le bi_{,l} \le 1$;

$b_{i,l}$ – коефіцієнт приналежності об'єкта ui кластеру Ul, $0 \le bi_{,l} \le 1$;

ρ(ui*,el) – евклидова відстань між гарним об'єктом ui* $\in U_l^*$ і центром el кластера Ul;

Дрейф має місце, якщо компактність кластера, розрахована для двох останніх тимчасових вікон, зменшується, тобто якщо слушна нерівність

$$P_l(t_{k+6}) = P_l(t_{k+6}) - P_l(t_{k+5}) < q, \tag{3.5}$$

де $q < 0$ – задана гранична величина.

Якщо умова (3.5) виконується, то необхідно проводити перерахунок координат центру кластера U_l.

3.3 Перехід від динамічної до статичної кластеризації

Моделі із застосуванням числових коефіцієнтів підсилення Інтернет-користувачів за їхньою поведінкою ніяк не можна віднести до статичних об'єктів – протягом доби вони характеризуються різною пошуковою активністю в мережі. Динамічна активність ІП пов'язана, у першу чергу, з їх географічним місцем, статтю, віком, соціальним станом і, звичайно, з їх розпорядком дня. Проте, розрізняють статичну й динамічну кластеризацію. Статична кластеризація має місце, якщо число кластерів, їх розміри й центри із часом не змінюються. Однак

якщо із часом кластерна структура змінюється, об'єкти кластеризації починають змінювати свої властивості, утворюючи нові кластери, або переходячи з одного кластера в інший, тоді в цьому випадку потрібна динамічна кластеризація. Статичні методи кластеризації — ієрархічні або ітераційні широко застосовуються для класифікації різного роду об'єктів. Методи динамічної кластеризації застосовуються рідше, тому що вимагають суттєво більших обчислювальних витрат, обумовлених періодичним спостереженням за великою кількістю об'єктів, що класифікуються, обробкою отриманих даних і формуванням кластерних структур.

Зрозуміло, що персоналізація Інтернет-пошуку повинна базуватися на кластеризації як користувачів, так і інформаційних ресурсів. Необхідно відслідковувати пошукову історію користувачів, поєднуючи їх у кластери за інтересами і змістом ресурсів, а також за темами. Питання полягає в тому, чи можна застосовувати статичні методи кластерного аналізу динамічних об'єктів? Як можна знизити вплив динаміки досліджуваних об'єктів на якість кластеризації є завданням роботи.

3.4 Метод дослідження динаміки кластерних структур

Дослідження динаміки кластерних структур припускає паралельне спостереження, як за користувачами, так і за ресурсами. У дискретні моменти часу $t_k \in T$, $k = 0,1,\ldots$, здійснюється формування числових векторів, що характеризують поточну пошукову активність користувачів і поточний зміст ресурсів. Експеримент починається з побудови зазначених векторів у момент часу t_0, розрахунку евклідової відстані між об'єктами, формування кластерів користувачів і ресурсів. Через інтервали часу $\Delta\tau$ $t_{k+1} - t_k$ здійснюється перевірка стабільності раніше сформованої кластерної структури.

Спостереження проводиться протягом 24 годин з інтервалом $\Delta\tau$ рівним 1 годину. При аналізі кластерної структури враховуються тільки результати пошукової активності за останні 4 години (період в 4 години підтверджується результатами експериментів таблиці 3.1 – це максимальний період часу протягом якого користувач може цікавитися конкретною темою), включаючи останню вибірку в момент спостереження. Це дозволяє організувати свого роду так зване ковзне тимчасове вікно спостереження й виключити вплив занадто старих спостережень – дані старих спостережень віддаляються. На підставі отриманої інформації й математичного аналізу динаміки кластерної структури робиться висновок про доцільність застосувань статичної або динамічної кластеризації, як для користувачів, так і для ресурсів.

Аналіз пошукової активності користувачів залежно від географії й часу доби дозволяє на неформальному рівні зробити висновок про необхідність застосування динамічної кластеризації для їхньої класифікації. Динамічна кластеризація дозволяє проводити класифікацію користувачів у будь-який момент часу й на будь-якій території. Динамічна кластеризація має бути застосована й для них, тому що зміст ресурсів, що включає досить велику кількість динамічних компонентів, постійно змінюється. Нехай U – велика кількість користувачів, за якими спостерігають, ui – *i*-й спостережуваний користувач, ui $\in$ U. Якщо проводиться спостереження за пошуковою активністю користувача ui, то в довільний момент часу tk $\in$ T можна сформувати характеристичний (пошуковий) вектор ігго користувача, що має наступний вид:

$$u_i(t_k) = (u_{i,1}(t_k), \ldots, u_{i,j}(tk), \ldots, u_{i,nof(Vu)}(t_k)),$$

де $u_{i,j}(tk)$ – числові координати j-го пошукового терміна із глобального словника термінів Vu у момент часу tk, дорівнює числу входжень цього терміна;

запити і-го користувача, виконані протягом відповідного тимчасового вікна;

nof(Vu) – розмір пошукового вектора ігго користувача, дорівнює числу слів у глобальному словнику термінів.

Числові координати $u_{i,j}(t_k)$, $1 \le \varphi \le \nu o \phi(Vu)$, розташовані в характеристичному векторі в тому ж лексикографічному порядку, що й терміни в словнику Vu.

Перехід від вербального до числової презентації результатів пошукової активності відбувається за рахунок позиційного кодування термінів і підрахунку числа їх входжень у запити користувача. Інший, відомий спосіб презентації вербальної інформації, що використовує покажчик був виключений з розгляду, тому що загальне число слів, застосовуване для розрахунків, не є постійним. Крім того, пропонований спосіб добре узгодиться з методами реляційної презентації й обробки інформації, заснованими на застосуванні Sql-Мови.

Нехай W – множина ресурсів, за якими спостерігаємо, w_i – *i*-ий ресурс, за яким ведеться спостереження, $w_i \in W$. У довільний момент часу $t_k \in T$ можна представити *i*-й ресурс характеристичним вектором наступного виду:

$$w_i(tk) = (w_{i,1}(tk), \ldots, w_{i,j}(tk), \ldots, w_{i,nof(Vw)}(tk)),$$

де $w_{i,j}(tk)$ – числові координати j-го пошукового терміна із глобального словника термінів Vw у момент часу t_k, дорівнює числу входжень цього терміна текст і-го ресурсу, протягом часового вікна для спостережень.

Для роботи з текстовим змістом ресурсів використовано DOM-моделі їх сторіноктобто об'єктні моделі Інтернет-документа, що дозволяє одержати доступ до текстового вмісту як статичних, так і динамічних (утримуючих динамічні компоненти) сторінок ресурсів, за якими спостерігаємо. Слід зазначити, що навіть при статичному HTML-коді сторінок, їх об'єктні моделі можуть мати динамічний характер. Тому кластеризація ресурсів, заснована на прямому аналізі HTML-коду сторінок, не є доцільною.

Проблема впливу динамічних компонентів на поведінку ресурсів у кластерних структурах була виявлена після тривалого спостереження за змістом їх Dom-моделей – ресурси можуть увесь час переміщатися між кластерами. Більше того, спостереження показують, що для Dom-моделей ресурсів характерна періодичність, тобто через n(tk) моментів часу текстовий зміст цих ресурсів повторюється.

У науково-технічній літературі розглянуті різні математичні методи для визначення стану кластерів у кластерних структурах. В одних джерелах пропонується використовувати ступінь приналежності об'єкта до кластера на

підставі матриці розбивки, в інших – рекомендується розраховувати формулу компактності об'єктів усередині кластера. Є множина інших чисельних показників для аналізу стану кластерів, таких як, наприклад, передбачуване число нових кластерів і ступінь дрейфу об'єктів усередині них. У рамках даної глави використовується формула розрахунку ступені приналежності об'єктів до різних кластерів.

Для дослідження динамічних ефектів у кластерних структурах, були обрано 100 інформаційних ресурсів, що належать групі сайтів новин з різною тематикою. Таким чином, застосована попередня група об'єктів за статичною інформацією: стать, вік і місце проживання для ІК і тематика для ІР.

За допомогою формули (3.1) експериментально були визначені показники приналежності, що відбивають стани кластерної структури для випадково обраного ІК з обраної групи. Отримані результати графічно інтерпретовано на рис. 3.4. Графіки підтверджують динамічний характер пошукової активності користувачів: залежно від часу змінюється пошуковий інтерес користувача й, як наслідок, його приналежність до кластерів.

Результати розрахунку коефіцієнтів приналежності ресурсу до кластерів (табл. 3.2) підтверджують динамічність характеристик ресурсів. На рис. 3.5 побудовані за даними таблиці 3.2 графіки схрещуються, тобто в різні моменти часу Інтернет-ресурс, за яким ми спостерігаємо, належить різним кластерам.

Для усунення цього ефекту пропонується використовувати Dom-модель ресурсів з метою їх «очищення» від динамічних компонентів: інформації партнерських мереж, різного роду повідомлень, реклами і т.д. Поруч із цим Dom-модель може бути використана для розрахунку вагових коефіцієнтів характеристичного вектора ресурсу.

Враховуючи особливості Dom-моделі Інтернет-сторінок ресурсу, кожному елементу вектора wi(tk) може бути зіставлений ваговий коефіцієнт, розрахований по формулі:

$$\frac{w_p \times k_1}{nW} \times k_2 \times 100, \tag{3.6}$$

де nW – загальне число слів в Dom-моделі сторінки;

w_p – число входжень слова на *p*-ой позиції в конкретному тегу Dom-моделі сторінки;

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

k_2 – коефіцієнт підсилення ($k_2 \geq 1$), значення якого розраховується по формулі відносин площ, що займають слова на сторінці:

$$k_2 = \frac{S_p / cnt_p}{S_{total} / nW}, \tag{3.7}$$

де Sp – площа області на *p*-ой позиції; cntp – число слів на *p*-ой позиції; Stotal– площа інформаційного тексту; nW – загальне число слів в Dom-моделі сторінки.

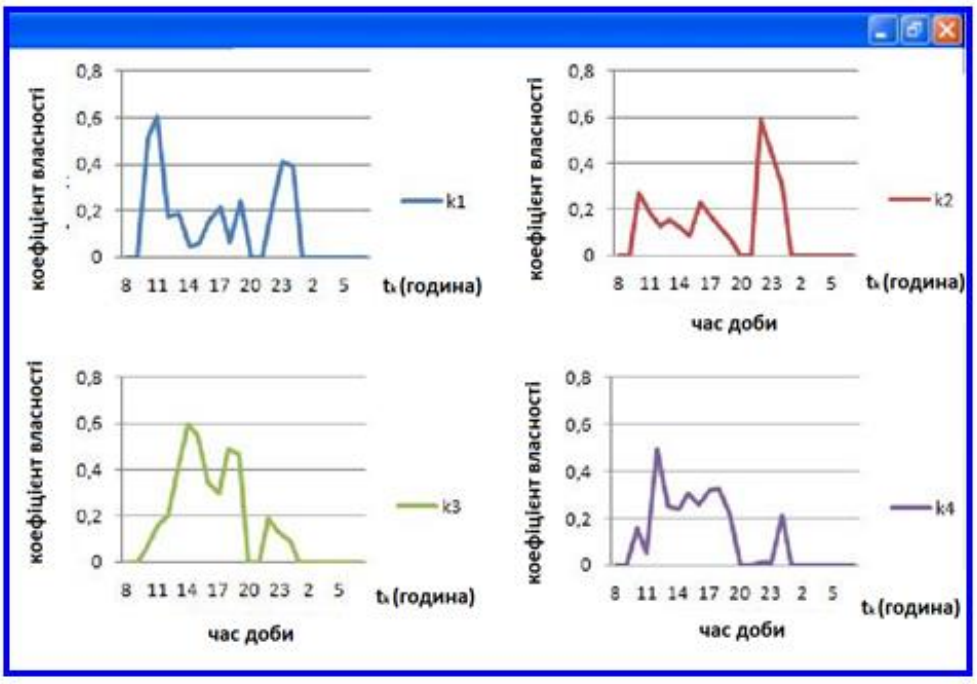

Рисунок 3.4 – Графіки зміни коефіцієнта приналежності користувача до різних кластерів у різні моменти часу

Таблиця 3.2 – Коефіцієнти приналежності ресурсу до кластерів у різні моменти часу без застосування вагових коефіцієнтів підсилення

Кластер / Момент часу (tk)	W1	W2	W3	Кластер / Момент часу (tk)	W1	W2	W3
t_0=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	0,274	**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

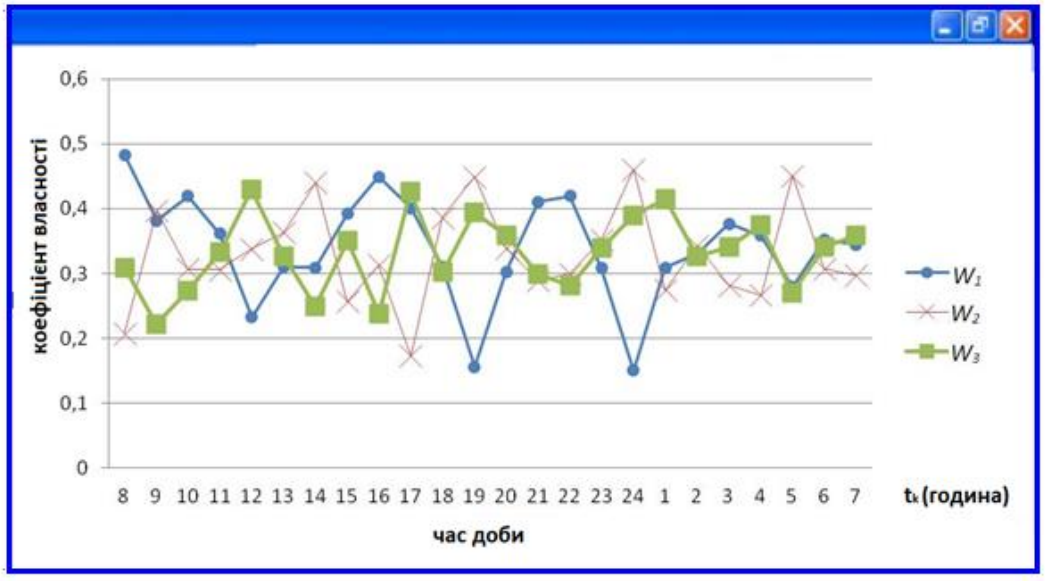

Рисунок 3.5 – Графіки коефіцієнтів приналежності ресурсу для різних кластерів у різні моменти часу без використання вагових коефіцієнтів підсилення.

Проведено повторний розрахунок ступенів приналежності ресурсу, за яким проводиться спостереження, (рис 3.6) із застосуванням вагових коефіцієнтів, розрахованих по формулі (3.7) і буде одержано нові графіки (рис. 3.7).

Момент часу (tk) \ Кластер	W_1	W_2	W_3	Момент часу (tk) \ Кластер	W_1	W_2	W_3
t_0=8	0,207	0,533	0,260	20	0,089	0,559	0,352
9	0,167	0,602	0,231	21	0,350	0,592	0,059
10	0,270	0,481	0,249	22	0,071	0,586	0,343
11	0,307	0,560	0,134	23	0,255	0,487	0,259
12	0,329	0,624	0,047	24	0,214	0,628	0,158
13	0,13	0,613	0,257	1	0,190	0,605	0,204
14	0,222	0,581	0,197	2	0,316	0,530	0,154
15	0,320	0,547	0,133	3	0,222	0,481	0,298
16	0,219	0,575	0,206	4	0,264	0,528	0,208
17	0,119	0,500	0,381	5	0,360	0,604	0,035
18	0,249	0,635	0,116	6	0,239	0,502	0,259
19	0,298	0,567	0,135	7	0,087	0,528	0,385

Рисунок 3.6 – Приналежність ресурсу до кластерів у різні моменти часу із застосуванням вагових коефіцієнтів підсилення

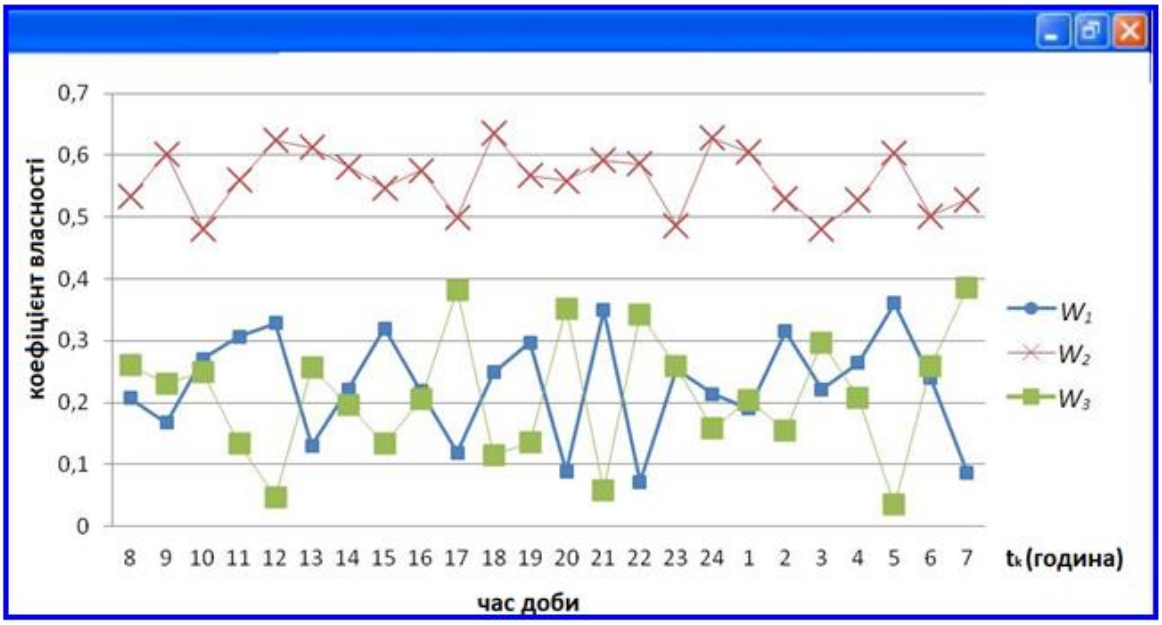

Рисунок 3.7 – Графіки коефіцієнтів приналежності ресурсу для різних кластерів у різний час доби із застосуванням вагових коефіцієнтів.

Використання коефіцієнтів підсилення, побудованих на підставі Dom-моделі, приводить до зрушення розрахункового коефіцієнта приналежності убік одного із кластерів (графік W_2). Інші графіки зміщуються вниз. До застосування коефіцієнтів підсилення середні значення коефіцієнтів приналежності до

кластерів W_1 і W_3 рівні 0.340 і 0.331, а після їхнього застосування знизилися до 0.229 і 0.211, відповідно.

З метою персоналізації Інтернет-пошуку був проведений кластерний аналіз користувачів і ресурсів. Якщо для кластеризації користувачів доцільно застосовувати динамічний підхід, то динамічна кластеризация ресурсів приводить до невизначеностей: у різні, але близькі моменти часу, інформаційні ресурси можуть належати різним кластерам. Причиною цьому є наявність інтерактивних динамічних компонентів, які періодично обновляють свій зміст. З рис. 3.7 видно, що при застосуванні динамічної кластеризації коефіцієнти приналежності одного й того ж самого IP до кластерів перебувають у постійній варіації, отже, підтвердити його приналежність до конкретного кластера неможливо. Використання коефіцієнтів підсилення, побудованих на підставі DOM-моделі, приводить до зрушення розрахункового коефіцієнта приналежності убік одного із кластерів. Таким чином, IP, які на перший погляд були динамічними, втрачають цю властивість. Запропонований метод кластеризації ресурсів із застосуванням коефіцієнтів підсилення можна застосовувати для будь-яких сучасних IP. Це дозволяє відмовитися від динамічної кластеризації ресурсів, зберігаючи на високому рівні ступінь їх приналежності до тих або інших кластерів і, як наслідок, забезпечити стабільність кластерної структури в цілому.

4 ОПИС РОЗРОБЛЕНОЇ ПРОГРАМНОЇ СИСТЕМИ

4.1 Концепція побудови інформаційної системи

Експериментальні дослідження дозволили визначити тимчасові й обчислювальні витрати на виконання кластерного аналізу реальних масивів ІК і ІР, що дозволило, у свою чергу, оцінити, які програмні модулі КСПП повинні бути встановлені на персональні комп'ютери (ПК) користувачів, а які – на сервери підприємства. Зазначені дослідження проводилися на персональному комп'ютері із центральним процесором AMD FX-6100 Six-Core Processor з тактовою чистотою 3.30 ГГц, що мають оперативну пам'ять ємністю 16 ГБ. На машині в різний час був установлений браузер Internet Explorer 10-ої і 11-ої версії. Максимальна пропускна здатність Інтернетет-каналу становила 5 Мб/с.

Пошукові запити ІК можна збирати одномоментно саме в ті моменти часу, коли починається їхня пошукова діяльність. Для цієї мети був створений програмний модуль `internet_res_search`, який дозволяє не тільки відслідковувати пошукові запити ІК, але й визначати глибину пошуку, тим самим, надаючи інформацію про результати пошукової видачі. Крім пошукової діяльності, портрет ІК також формується в результаті виконання заходів і відвідувань ним ІР. У зв'язку із цим був реалізований програмний модуль `ie_analyzer` , що забезпечує автоматичне спостереження й формування LogaФайлу із хронологічним списком відвідуваних ІК сторінок.

Сучасні ІР містять величезне число динамічних компонентів в Dom-моделі, які можуть постійно змінювати свій зміст, крім того, наявність більших медіафайлів (відео або картинки високої роздільної здатності) може сильно збільшити час повного завантаження й читання змісту Dom-моделі ресурсу. Деякі ресурси можуть містити посилання на неіснуючі елементи, що перебувають на сервері самого ІР, і браузер буде намагатися завантажити їх до моменту timeout. У процесі експериментальних досліджень середній час сканування Dom-моделі однієї сторінки ІР становило 6-7 секунд. Якщо пошукова система, як результат

пошукового запиту, видає 10 гіперпосилань на знайдені ресурси, то для читання змісту їх Dom-моделей буде потрібно більше однієї хвилини, що є неприпустимим часом для будь-якого користувача – зайві часові витрати на очікування призводять до серйозних витрат, можуть сильно дратувати людину. Процес кластерного аналізу ІК і ІР, заснований на виконанні розроблених методів, містить багато етапів (часові вікна для ІК, застосування числових коефіцієнтів підсилення з Dom-моделі ІР, боротьба з динамічними елементами ІР і застосування узагальненого характеристичного вектора на основі глобального словника термінів), при цьому відповідні процедури обробки даних потребують досить багато часу. Тому що кластерний аналіз результатів 4-го періоду спостережень займає приблизно 20 хвилин часу (що неприпустимо багато для будь-якого ІК), доцільно виділити спеціальну серверну обчислювальну систему (сервер) під кластерний аналіз Інтернет-об'єктів.

Прикладом розробленої інформаційної технології є представлена на рис. 4.1 узагальнена структура СПП, що відображає структурну організацію системи. Програмні модулі безпосереднього спостереження internet_res_search і ie_analyzer повинні бути встановлені на клієнтських ПК. Програмний модуль internet_res_search реалізований до роботи з пошуковою системою. Пошукова система вивантажує результати пошуку в HTML-коді самої сторінки, тому немає ніякої необхідності інспектувати Dom-модель кожної сторінки, що знижує в рази обчислювальні витрати на читання й аналіз сторінок, отриманих у результаті пошуку. Усі кластерні розрахунки й уся аналітика повинні бути реалізовані програмними модулями, встановленими на корпоративних серверах. Для цієї мети були розроблені спеціальні розрахункові процедури й фільтруючі функції, а також спроектована структура БД, підтримувана системою керування MS SQL Server.

Таким чином, для персоналізації Інтернетет-пошуку працівників підприємства, у рамках його інформаційно-обчислювальної системи повинна бути створена трьохланкова СПП. На рис 4.1 представлені всі три взаємозалежні ланки корпоративної системи персоналізації пошуку: перша ланка – множина ІК, друга

ланка – сервер кластерного аналізу й третя ланка – сервер БД. До мережі Інтернет система підключається через сервер кластерного аналізу.

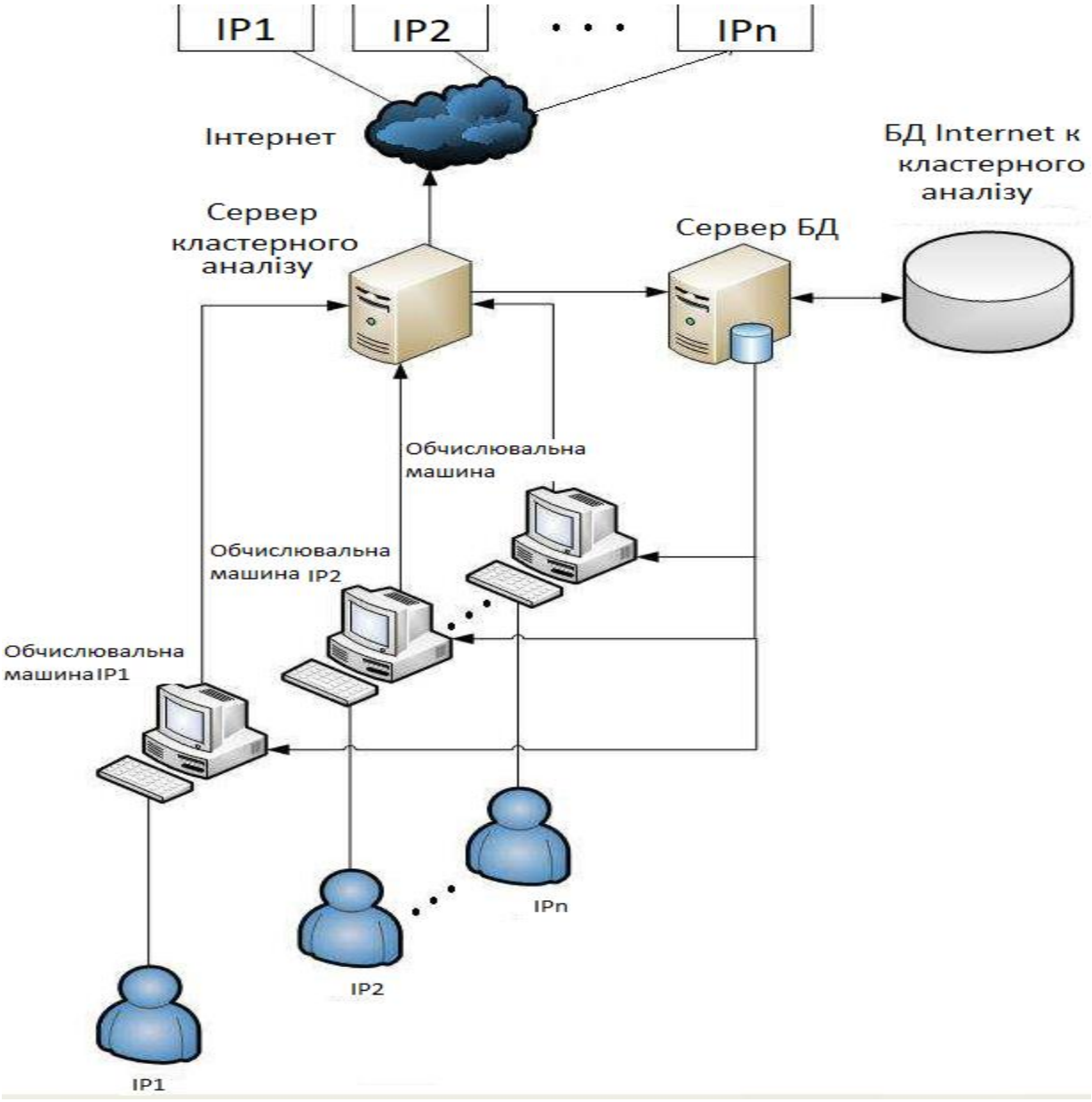

Рисунок 4.1 – Узагальнена структура корпоративної системи персоналізації пошуку

Завдання реалізації й тестування пропонованих методів кластерного аналізу зважувалися на одній потужній обчислювальній установці, здатній підтримувати множину віртуальних машин. Структура віртуальної КСПП повною мірою відповідала узагальненій структурі, представленій на рисунку 4.1

На рис. 4.2 представлена повна структура БД для обробки заходів ІК Internetdb. БД пошукової активності ІК містить 12 таблиць, необхідних для структуризації одержуваних у процесі спостереження даних і наступного проведення кластерного аналізу ІК по історії пошуку.

Для цього в БД додано дві нові сутності: «слово» (az_words) і асоціативну сутність «словопошуковий рядок» (az_pages_words). Відсутність унікальних ключів таблиці az_pages_words пов'язане з можливістю появи одного й того самого слова кілька разів у пошуковому рядку.

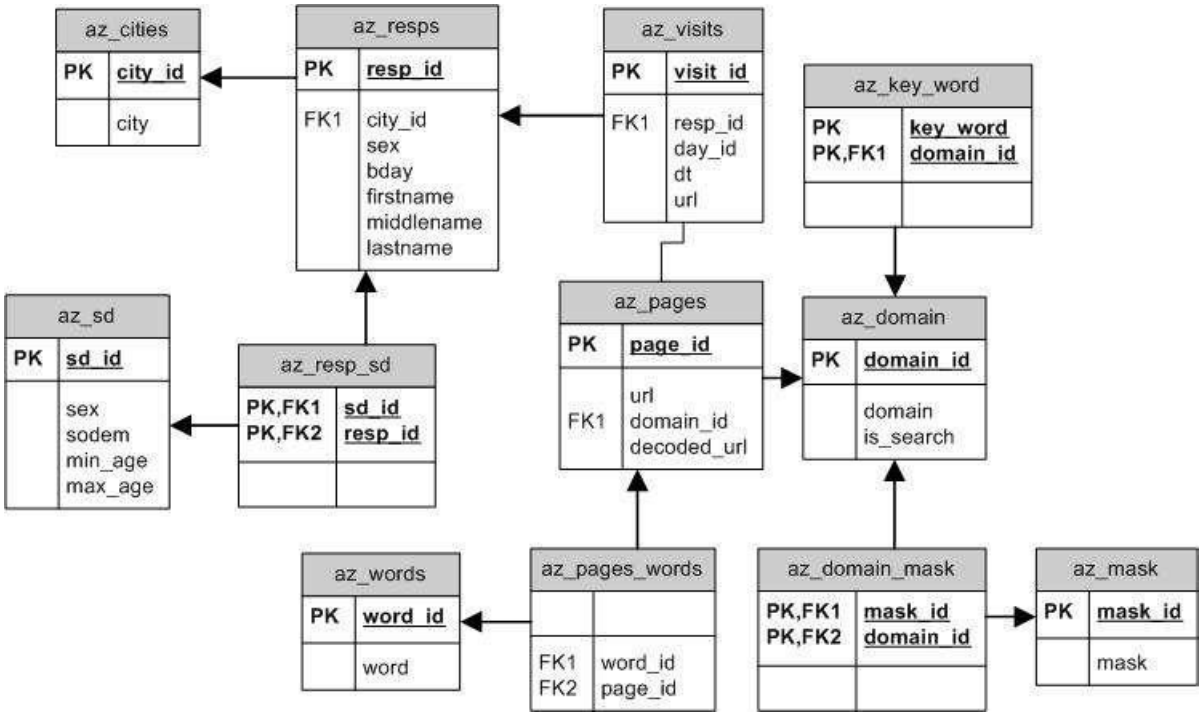

Рисунок 4.2 – Додавання сутностей az_words і az_pages_words

4.2 Програмний засіб структуризації даних про зміст Інтернет-ресурсів

IP по своїй структурі є або окремо взятою web-сторінкою, або набором web-сторінок з однаковим доменним іменем, що пов'язані між собою гіперпосиланнями. Для візуальної інтерпретації IP необхідно використовувати

web-браузер, який перетворить теги Dom-моделі у відповідні візуальні образи. Наприклад, теги і вказують на початок і кінець параграфа, <table> і </table> вказують на таблицю, а <title> і </title> – на заголовок IP. У свою чергу кожен тег може містити атрибути, що доповнюють їхні візуальні характеристики.

Для IP HTML є стандартною мовою розмітки web-сторінок. Згодом IP стають усе більше інтерактивними й динамічними завдяки застосуванню динамічних компонентів, інкапсульованих у них Dom-моделі.

Структура БД для зберігання й обробки даних про зміст IP. Web-розроблювачі не обмежуються класичним HTML-кодом, використовуючи при створенні IP технології, – CSS, Javascript, Ajax, що призводять до появи динамічних компонентів. У зв'язку із цим необхідно проводити інспекцію IP виходячи з Dom-моделей web-сторінок. Використання Dom-моделі дозволяє отримати доступ до будь-яких елементів IP та їх атрибутів. Це дає можливість маніпулювати web-документами, як об'єктами («`object`»), з усіма їхніми компонентами, їх атрибутами й властивостями. Dom-модель дозволяє представити IP у вигляді дерева, кожний вузол якого може бути одночасно як батьківським (parent), так і дочірнім (child) вузлом стосовно іншого вузла дерева (рисунок 4.3). Тег HTML є початковою ланкою Dom-моделі, коренем, з якого «виросте дерево» IP.

Для одержання доступу до конкретного тегу HTML-документа, необхідно пройти шлях від кореневого вузла (HTML-тега) до цільового вузла й потім прочитати значення конкретних атрибутів. За допомогою Dom-дерева HTML-документа можна розбити зміст IP на параграфи, списки, розділи, гіперпосилання й інші компоненти структури сторінки.

Для доступу до елементів Dom-моделі є два методи:

– прямий доступ до HTML-елементу (HTMLelement) Dom-моделі по унікальному ідентифікатору. У цьому випадку, необхідна наявність унікального ідентифікатора, необхідного HTML-елементу. Наприклад, на головній сторінці google.com:

```
<div class="portal-headline__projects" id="portal-headline__box">;
```

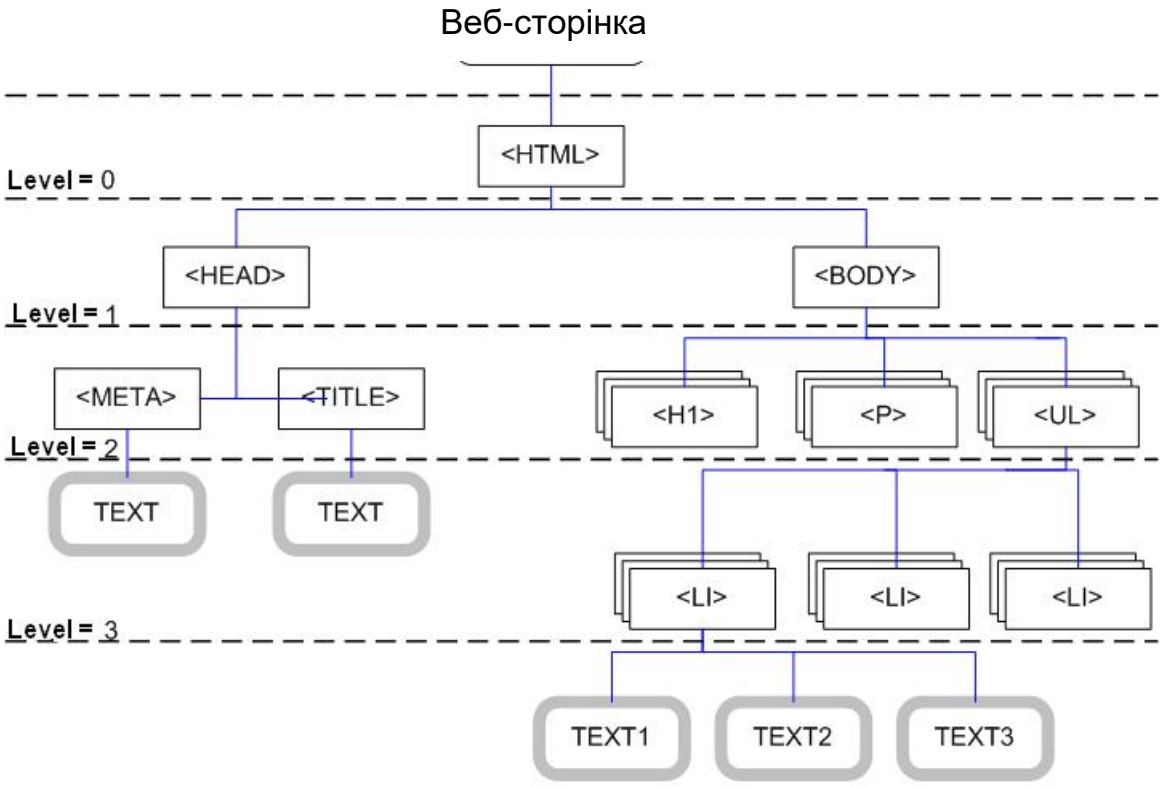

Рисунок 4.3 – Приклад Dom-дерева web-сторінки

– доступ до HTML-елементу за назвою тега – у цьому випадку необхідно спочатку відібрати набір (HTMLelementcollection) з конкретною назвою тега, а потім зробити пошук потрібного тега за значеннями атрибутів. Наприклад, на головній сторінці gmail.com, щоб знайти цей HTML-елемент, необхідно спочатку знайти набір елементів, у яких тег називається «a», а потім проводити пошук по атрибуту «`name = "clb598679"`»:

```
<a name="clb598679" href="http://my.gmail.com "
class="social__title__link"><i class="social__title__link__icon icon
icon_social icon_social_big icon_social_my"></i><span
class="social__title__link__text"> </span></a>
```

Основними функціями для роботи з HTML-елементами є:

`getelementbyid` – функція, що повертає посилання на вузол документа, яку можна використовувати для читання й редагування властивостей і звертання до методів вузла;

`getelementbytagname` – функція, що повертає масив з елементів, що мають конкретний тег;

`getattribute` – функція, що повертає значення конкретного атрибута.

В алгоритмі доступу до Dom-елементів (рисунок 4.3) `getelementbyid` має більш високий пріоритет, ніж `getelementbytagname`. Це актуально, особливо для IP з однаковою структурою, коли різні URL того самого IP мають ідентичну Dom-модель.

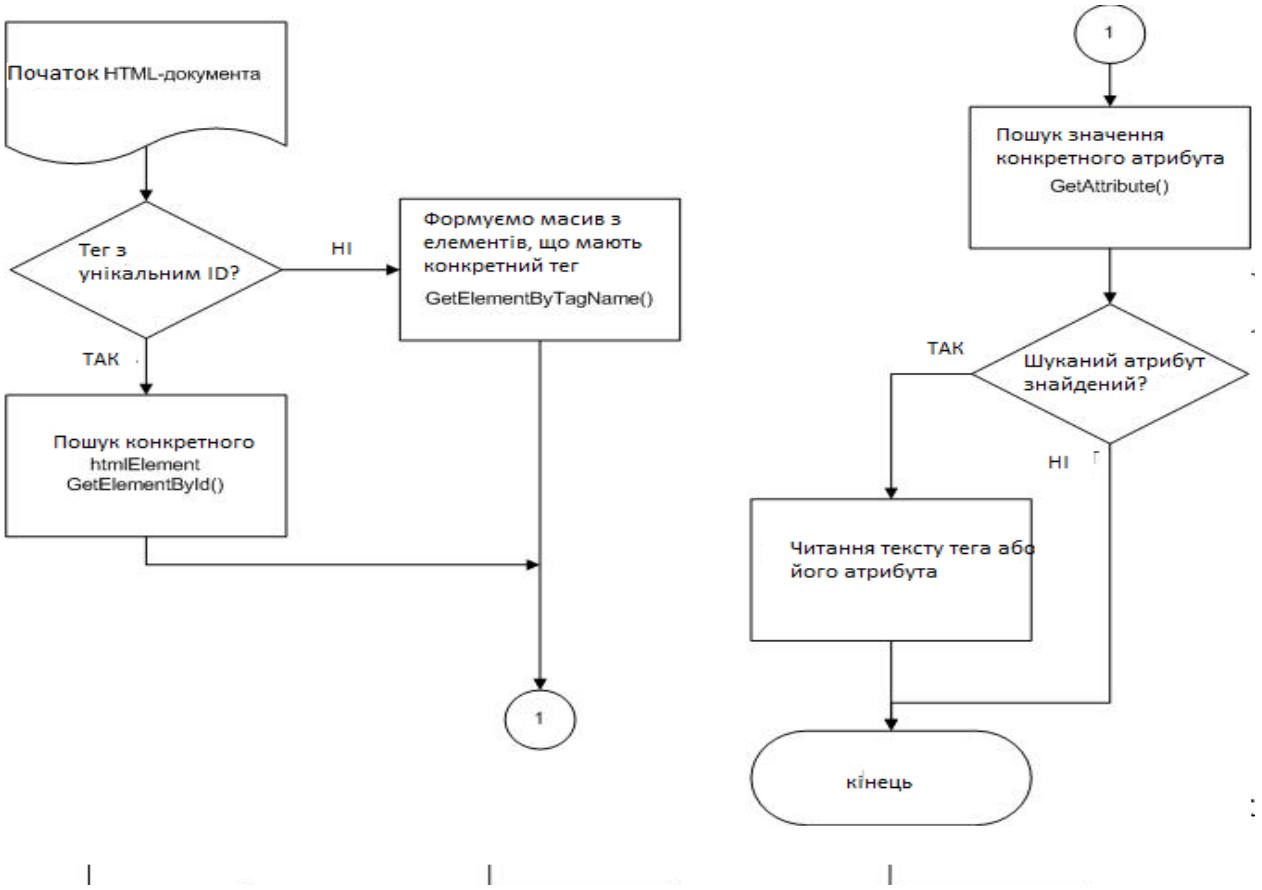

Рисунок 4.4 – Схема алгоритму доступу до Dom-елементів

Для структуризації змісту IP за тегами досить доповнити раніше розроблену структуру БД для ІК (рисунок 4.1) ще трьома сутностями. Для початку необхідно виділити окрему сутність для всіх URL досліджуваних IP, створивши таблицю Pages. Увесь довідник HTML-тегів [17] Dom-моделі

розміщено в словниковій таблиці `HTML_element`. Результати читання тегів розташовано в асоціативній сутності `HTML_value`. На рис. 4.5 показана структура із трьох таблиць для зберігання даних про теги та їх значення.

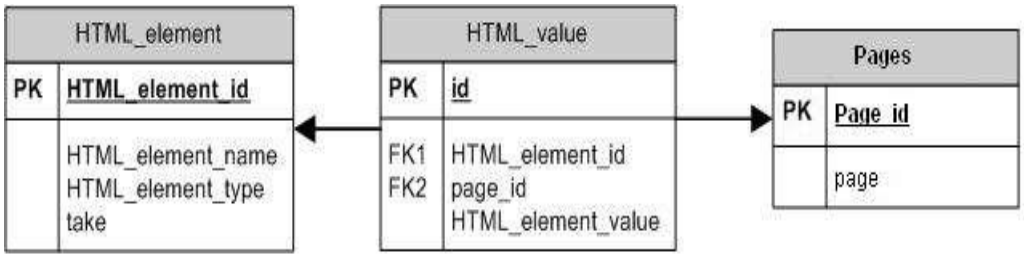

Рисунок 4.5 – Структура БД для зберігання даних про теги та їх значення

За списком URL можна запустити спеціально розробленого програмного робота HTMLdocdom (вихідний код програми в дод. Б) для зчитування вмісту Dom-моделі.

Структуризація об'єктів дослідження – ІК та ІР – є першим етапом рішення завдання персоналізації пошуку. Зі структурованими даними можна проводити дослідження, підтверджувати або спростовувати ефективність різних методів кластеризації й класифікації.

4.3 Опис програмних модулів internet_res_search і ie_analyzer

З урахуванням запропонованої вище структури корпоративної системи персонализації пошуку з використанням принципів об’єктно-орієнтованого програмування було розроблено два програмні модулі internet_res_search і ie_analyzer, що мають графічний користувацький інтерфейс. У якості операційних систем, для яких було розроблено програмне забезпечення СПП, обрані операційні системи ряду MS Windows, оскільки на сьогоднішній день вони мають найбільше поширення. У якості інструментальної системи обрана MS Visual

Studio, що забезпечує більші можливості для розробки й налагодження об'єктно-орієнтованих додатків під MS Windows, а також має багатий набір візуальних компонентів для побудови інтуїтивно зрозумілого й зручного графічного інтерфейсу користувача. Перевагою середовища розробки MS Visual Studio є також можливість програмування на таких мовах високого рівня, як Visual C# і Visual C++.

Графічний інтерфейс користувача програмного модуля internet_res_search – у верхньому лівому кутку вивантажується Url-адреса поточної завантаженої сторінки пошукової системи. Відразу під Url-адресою розташовується сам браузер (компонент Webbrowser). У правому верхньому кутку розташовується спеціальне поле для пошукових слів, а під ним кнопка «Пошук». Інтерфейс передбачає два методи автоматичної навігації: «метод імітації» і «метод URL». Кожен із зазначених методів виконує автоматичну пошукову навігацію залежно від зазначеного рівня глибини входження. Кожен з методів має свої особливості. Метод імітації – працює на рівні Dom-моделі сторінки, коли проводиться пошук конкретних елементів сторінки, а потім здійснюється імітація натискання кнопки миші. Наприклад, для імітації натискання кнопки «Пошук» на головній сторінці Google, необхідно спочатку знайти елемент <input class="b-form-button__input" type="submit" value="" tabindex="2" hidefocus="true"/>, а потім "click": HTMLelement.Invokemember("click"). Метод URL –підходить для навігації пошукової системою. У пошуковому рядку може бути інкапсульована множина керуючих команд, що дозволяють управляти пошуковою віддачею вилучених серверів. Наприклад, для переходу до 5-ої сторінки пошукових результатів потрібно вказати номер сторінки. Після натискання кнопки «Пошук» відбувається пошукова навігація і як тільки сторінка завантажується, здійснюється сканування HTML-коду й відбір гіперпосилань у тому ж порядку, у якому їх вивантажує сама пошукова система. Список гіперпосилань вивантажується в спеціальне поле, що розташоване в правій частині вікна графічного інтерфейсу користувача, з можливістю його збереження в текстовий файл для подальшої обробки.

Вихідним файлом є файл із розширенням txt. Кожний рядок цього файлу має наступний формат: <url>\n, де url – URL гіперпосилання, видане пошуковою системою.

Інтерфейс програми ie_analyzer складається з обмеженого числа візуальних компонентів (рис. 4.6).

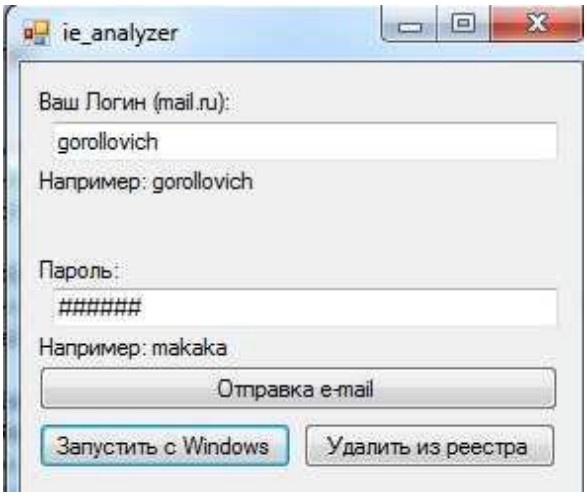

Рисунок 4.6 – Графічний інтерфейс програми ie_analyzer

При бажанні користувач може ввести свій особистий логін і пароль. Якщо в корпоративній мережі підприємства є поштовий сервер, то можна буде прив'язати до кожного користувача його особисту поштову адресу, звідки буде відбуватися розсилання Log-файлу зі списком відвідуваних IP на сервер кластерного аналізу.

Кнопки «Запуск із Windows» і «Вилучити з реєстру», призначені для включення й відключення автоматичного запуску програми при завантаженні Windows, тому що є можливість установки програми з постійним запуском при завантаженні ОС і тому немає ніякої необхідності перезапускати її заново. Кнопка «Відправлення e-mail» для відправлення Log-файлу ie_analyzer одразу на сервер кластерного аналізу або на корпоративний поштовий сервер, а потім на сервер кластерізації.

Програмний модуль дозволяє відслідковувати відвідуваність Інтернет-ресурсів ІК. Дані про всі IP, що були відвідані, будуть відправлені для подальшої обробки й кластеризації в СПП. Досить буде встановити програму ie_analyzer

один раз і додати її до реєстру Windows натисканням кнопки «Запуск із Windows» для постійного моніторингу активності ІК.

Вихідним файлом є файл із розширенням txt. Кожен рядок цього файлу має наступний формат: <a>\T<b>\T<c>\T<d>\T<e>, де a – назва браузера (наприклад, InternetExplorer); b – унікальний ідентифікатор об'єкта (handle) вікна браузера; c – URL браузера (адреса ІнтернеттРесурсу); d – час відвідування Інтернет-ресурсу; e – стан URL (активна Url-сторінка або закрита Url-сторінка).

Основним джерелом вихідної інформації про пошукову діяльність ІК є Log-файли, сформовані програмою `ie_analyzer`. Встановивши цю програму на ПК, ІК автоматично стає об'єктом кластерного аналізу СПП. Уся пошукова історія й історія заходів автоматично стає доступною для кластерного аналізу. У програмному коді `ie_analyzer` є можливість установки тимчасового лічильника (Timer) для автоматизації відправлення Log-файлів за пливом певних проміжків часу на сервер. Другим джерелом інформації для аналізу з боку ІК може служити програма `internet_res_search`, яка дозволяє враховувати глибину пошуку й формувати пошукову історію. Програма `internet_res_search` може бути встановлена на боці сервера кластерного аналізу для порівняння результатів пошуку до кластеризації, і після її застосування.

В розробленій програмі `ie_analyzer` відбувається моніторинг запущених додатків в оболонці ОС Windows Shellwindows. З моменту запуску програми здійснюється постійне спостереження за процесами (завданнями) всередині оболонки Shellwindows, тобто проводяться постійні спостереження за процесами, які можна знайти в диспетчерові завдань Windows. Як тільки запускається додаток iexplorer, одержуємо його handle і починаємо стежити за його активністю – будь-які події фіксуються в Log-файлі. Алгоритм роботи програмного модуля представлено на рис. 4.7.

Розроблена програма internet_res_search виконує два основні завдання: імітацію виконання пошуку й збір гіперпосилань зі сторінок пошукової системи. Як уже було сказано, ця програма може бути встановлена, як на боці ІК, так і на

боці сервера кластерного аналізу. Дана програма відслідковує стан вбудованого браузера й виробить список гіперпосилань, який формує пошукова система. Програма реалізує два методи – метод імітації й метод URL. Алгоритм роботи програмного модуля представлено на рисунку 4.8 і рисунку 4.9.

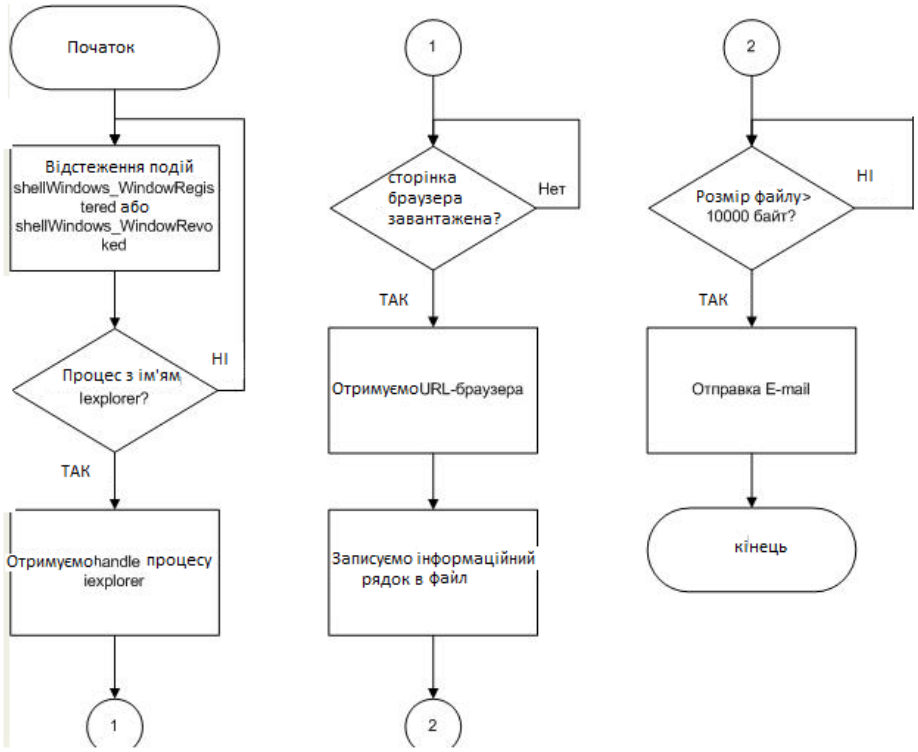

Рисунок 4.7 – Схема алгоритму роботи програмного модуля ie_analyzer

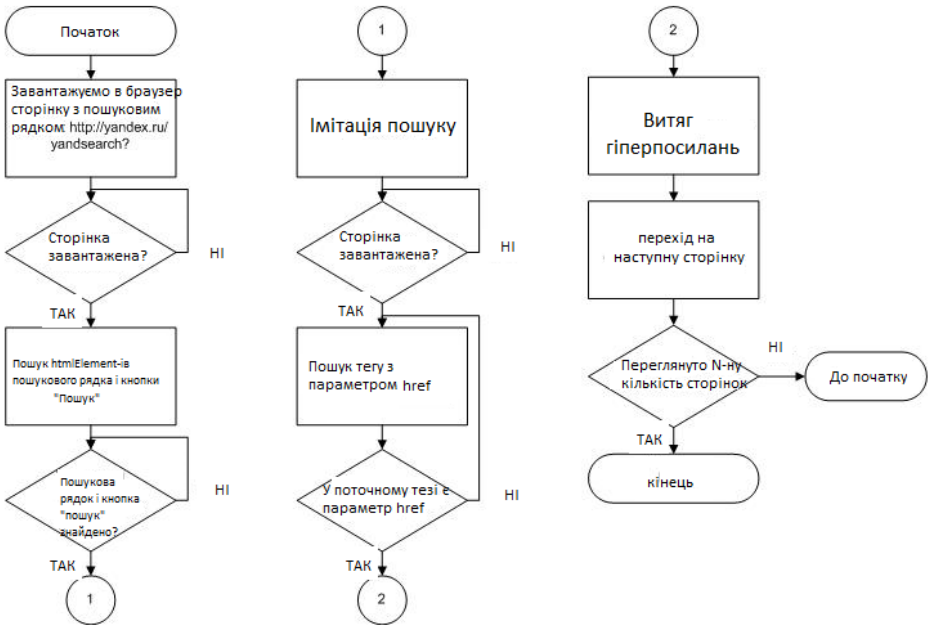

Рисунок 4.8 – Схема алгоритму роботи програмного модуля internet_res_search у режимі імітації виконання пошуку

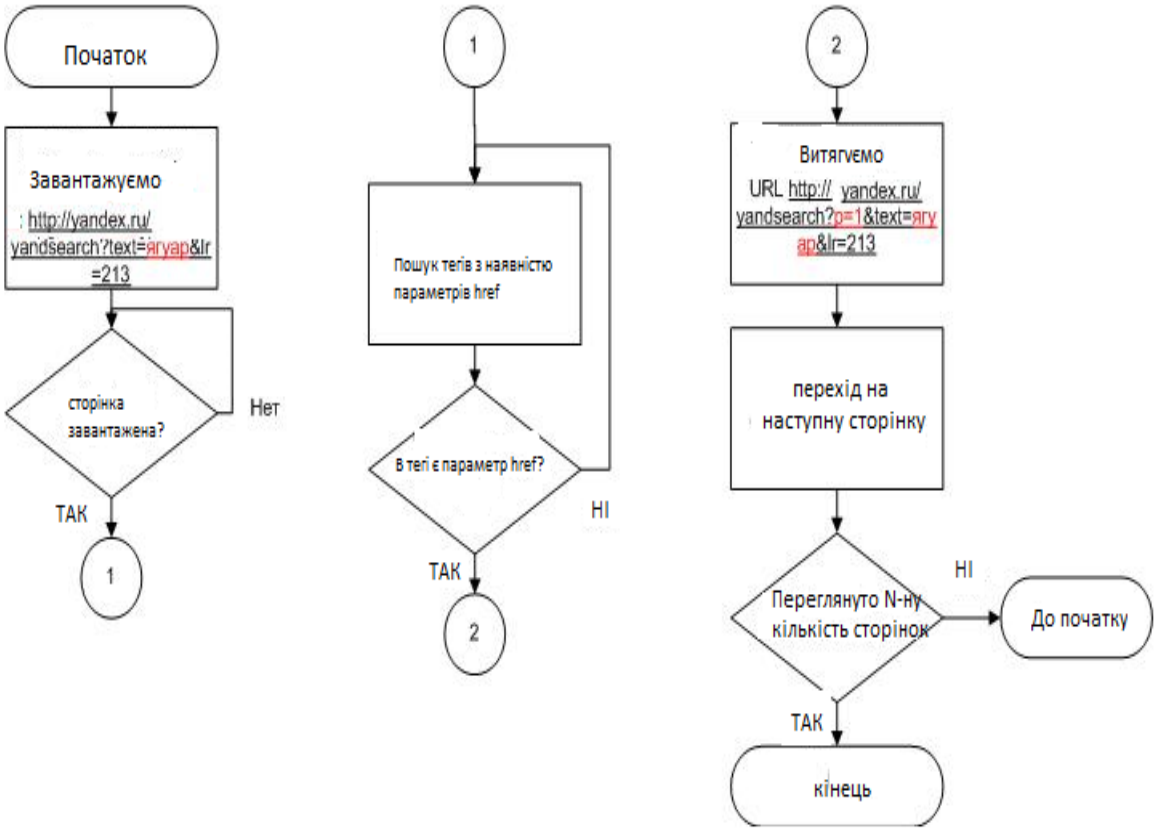

Рисунок 4.9 – Схема алгоритму роботи програмного модуля internet_res_search у режимі застосування URL

Алгоритм починає виконуватися після входу на сторінку браузера. Очікується момент повного завантаження сторінки, проводиться пошук HTML-еелментів пошукового рядка й кнопки «Знайти». Потім імітується запис «Що шукаємо» у пошуковому рядку й натискання кнопки «Знайти». Після очікування повного завантаження сторінки витягають усі гіперпосилання. Число ітерацій залежить від зазначеного рівня «Глибина заходу».

У цілому обидва методи (рисунок 4.8 і рисунок 4.9) досягають того самого результату, але через те, що на сайтах великих пошукових систем постійно можуть виникати зміни в коді сторінки й у її Dom-моделі, необхідно підтримувати працездатність модуля у двох режимах.

4.4 Підсистема кластерного аналізу й класифікації

Як уже було сказано, відповідно до концепції корпоративної системи персоналізації пошуку всі програми кластерного аналізу повинні виконуватися на сервері, працюючи із СУБД MS SQL Server. Для стабільної роботи системи (включаючи процеси тестування, дослідження й інтерпретації результатів) необхідно виділити більше 10 ГБ пам'яті на жорсткому диску, винятково під БД Internetdb. Ця вимога пов'язана з обсягом оброблюваних даних, що реально зберігаються.

Алгоритм кластерного аналізу й класифікації об'єктів реалізований у середовищі MS SQL Server. Об'єкти дослідження – ІК та ІР – представлені за допомогою характеристичних векторів і розміщені в таблицях БД. Для реалізації кластерного аналізу застосовувався метод *k*-середніх. Процес повністю автоматизований і не вимагає ніякого стороннього втручання. Експерт може лише варіювати вхідні параметри системи, якщо кластеризація не приводить до апріорі очікуваних результатів.

Структура підсистеми кластерного аналізу, який складається з головного модуля (запуску й керування) і чотирьох розрахункових процедур, показаних на рис. 4.10.

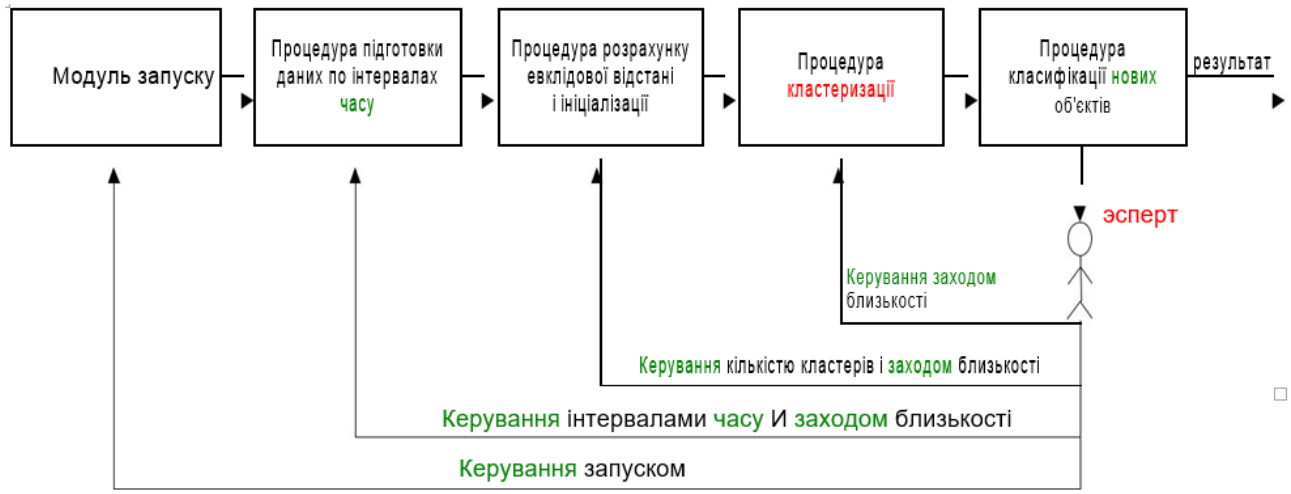

Рисунок 4.10 – Структура підсистеми кластерного аналізу

Експерт – аналітик, що одержує результати кластерного аналізу, має можливість контролювати такі параметри, як число кластерів у кластерній структурі, розмір тимчасових інтервалів спостереження за активністю ІК і захід близькості об'єктів кластеризації. Модуль запуску й керування реалізований як Sql-скрипт, що контролює хід виконання всього алгоритму підсистемою кластерного аналізу. У цьому модулі керування є всі необхідні змінні для керування кластерним аналізом:

Основне завдання процедури підготовки даних по інтервалах часу – перетворення даних про пошукову активність ІК і змісту ІР у зручну для обробки форму. У цій процедурі спочатку виконується розбивка даних залежно від зазначеного інтервалу часу за допомогою змінної @step1_hour_step, потім формується глобальний словник термінів з інтервальною розбивкою для динамічного застосування. Наприкінці процедури для всіх об'єктів дослідження формуються характеристичні вектори залежно від стану відповідного словника в конкретному інтервалі часу. Алгоритм роботи процедури показано на рисунку 4.11.

Процедура ініціалізації об'єктів і їх первісного розподілу по кластерах починає кластерний аналіз. Застосовуються спеціальні підходи, зручні для реалізації в середовищі MS SQL Server, що забезпечують економію пам'яті на жорсткому диску. Наприклад, застосовується негативне кодування для розрізнення Інтернет-об'єктів – ідентифікатори ІК стають негативними, а в ІР вони залишаються позитивними.

Схема алгоритму роботи процедури представлена на рис. 4.12.

Алгоритм процедури складається з 8-ми кроків. З метою економії дискового простору й виключення необхідності додавання нових стовпців, було здійснене кодування ідентифікаторів ІК негативними значеннями (крок 2). На кроці 3 експерт-аналітик вибирає метрику: евклідову відстань, її квадрат, мангеттенську відстань і ін. Для розрахунку заходу близькості об'єктів дослідження використовується обрана метрика й характеристичні вектора об'єктів, які були раніше завантажені в таблицю векторів БД.

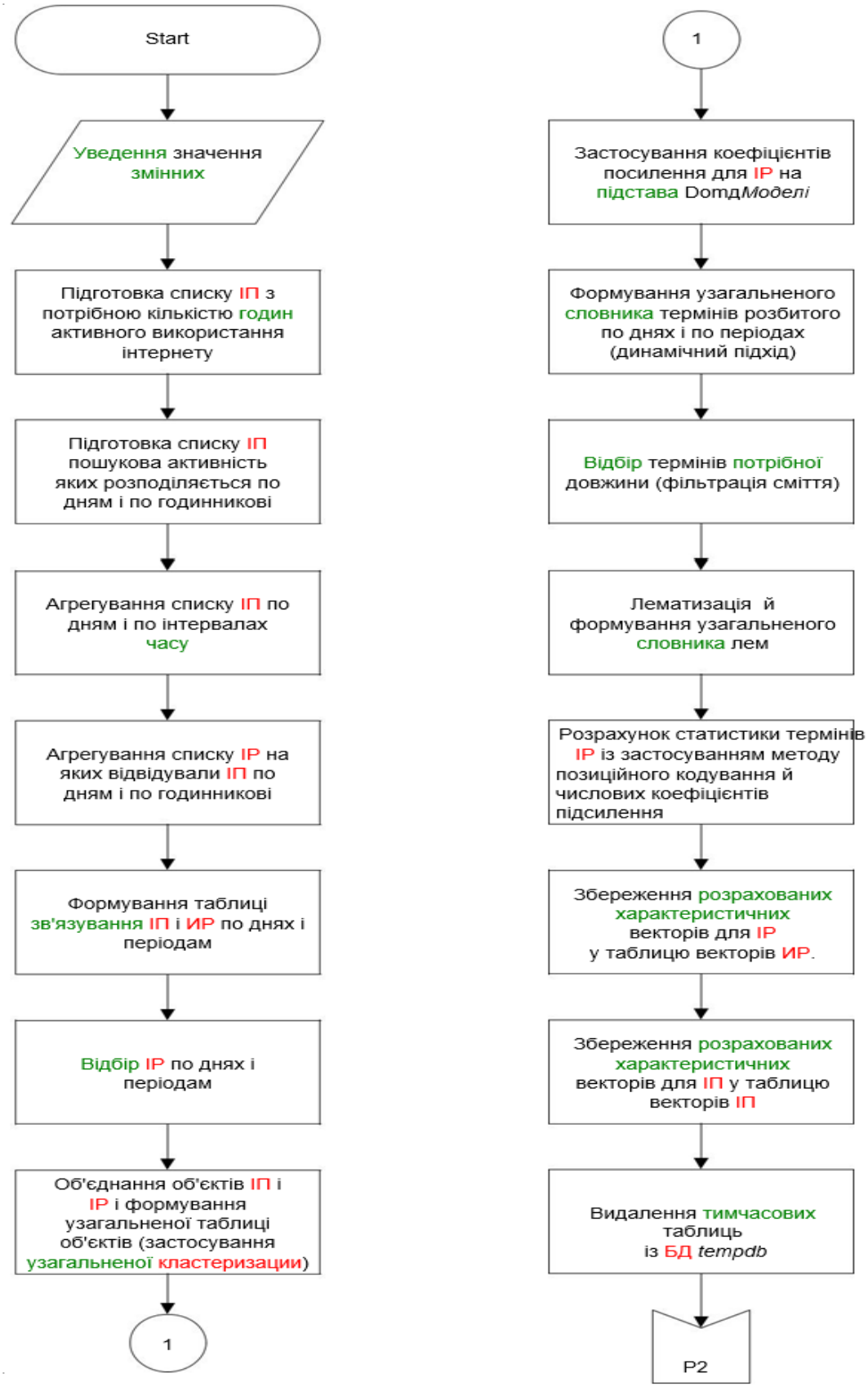

Рисунок 4.11 – Схема алгоритму підготовки даних для кластерного аналізу

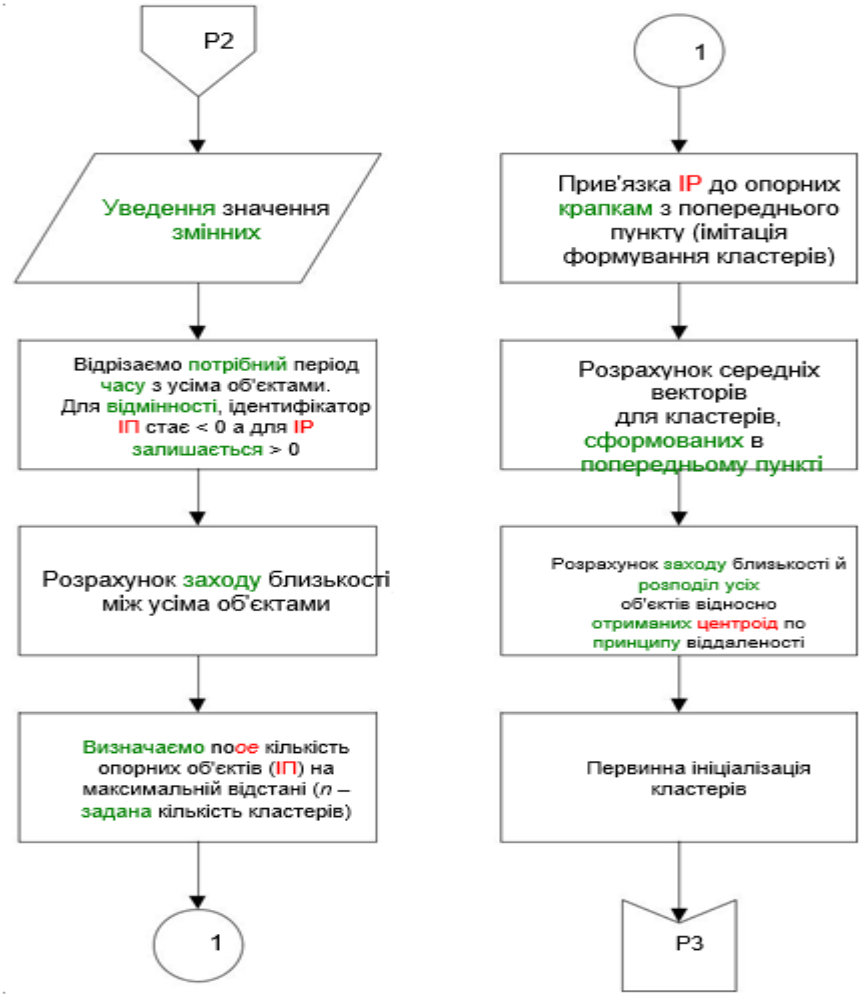

Рисунок 4.12 – Схема алгоритму ініціалізації об'єктів і їх первісного розподілу по кластерах

Із застосуванням таблиць БД завдання розрахунку середнього вектора зводиться до розрахунку середнього числового значення для кожної з координат об'єктів за допомогою функції AVG() і групування по кластерах. На кроці 4 згідно зазначених вхідних даних значенню необхідного числа кластерів n відбувається добір ІК, що перебувають на максимальній відстані одне від одного (вибір опорних точок). Як тільки опорні точки визначені, можна почати формування перших покластерів за допомогою прив'язки ІР відповідно до опорних точок (крок 5). Коли кластери сформовані, для кожного кластера за стандартною схемою розраховуються спочатку середній вектор (крок 6), а потім відстані від усіх об'єктів кластерів до середніх (крок 7). Результатом виконання

цієї процедури є первинна ініціалізація об'єктів – первинний розподіл їх по кластерах. Варто відзначити важливість первинної ініціалізації об'єктів, її вплив на подальший процес кластерного аналізу, якість якого прямо залежить від неї.

Третя процедура призначена винятково для виконання кластеризації – від моменту одержання первинного розподілу об'єктів по кластерах до досягнення стабільної кластерної структури з нерухливими центрами. Стабільна кластерна структура досягається за допомогою ітераційного циклу, який завершується, як тільки об'єкти перестають переміщатися між кластерами (рис. 4.13).

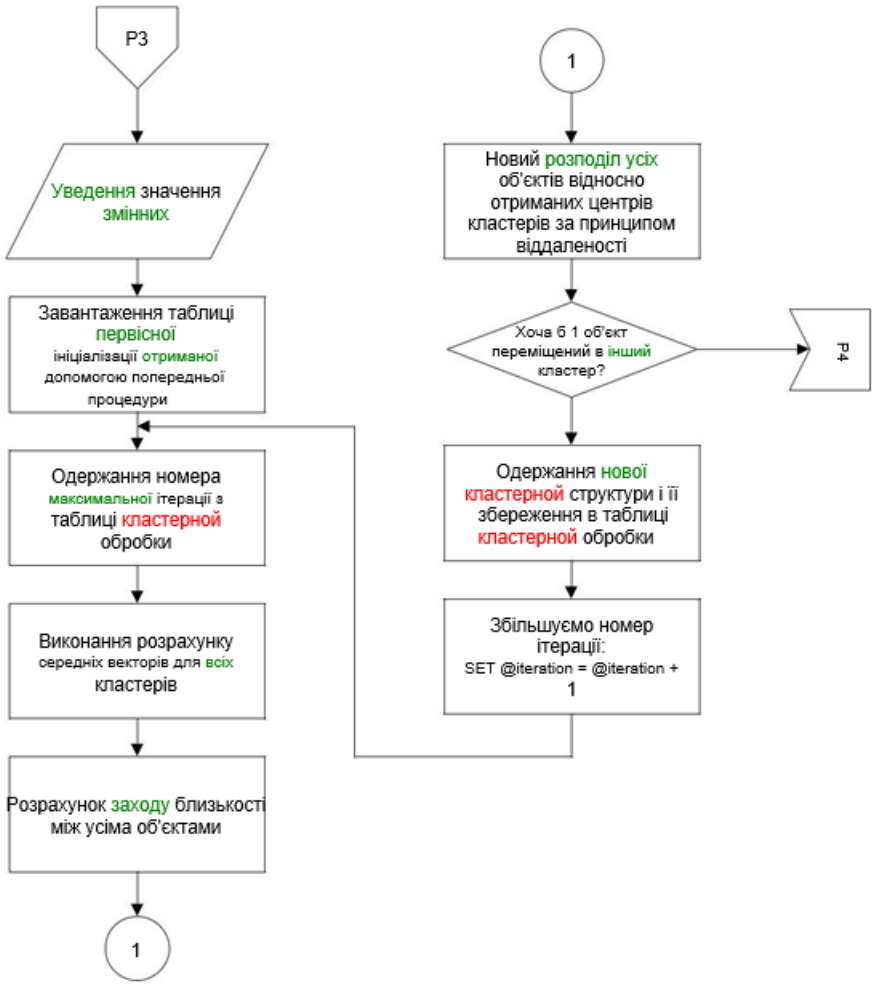

Рисунок 4.13 – Схема алгоритму кластеризації Інтернет-об'єктів

Виконується ітераційний цикл кластеризації. При кожній ітерації обробляється останній стан кластерної структури за значенням стовпця `iteration` (крок 3). Потім за стандартною схемою відбувається розрахунок

середнього вектора (крок 4) і заходу близькості між усіма об'єктами (крок 5). Результат 4 кроку може призвести до перерозподілу об'єктів у кластерній структурі. Як тільки об'єкти перестають мігрувати між кластерами, цикл завершуєтьсяі за поточним станом кластерної структури видається результат.

Четверта процедура призначена для прогнозування або розрахунку показника влучення в цільову групу (рис. 4.14). У результаті виконання попередньої процедури, отримуємо стабільну кластерну структуру, до складу якої входять, як ІК, так і ІР. Оцінити якість сформованих кластерів можна шляхом впровадження в неї нових об'єктів. Кластеризуємо ІР,за якими спостерігаємо в наступному тимчасовому вікні, і потім перевіряємо чи відвідували їх ІК, що перебувають у тому самому кластері.

Рисунок 4.14 – Схема алгоритму класифікації нових об'єктів

Алгоритм класифікації нових об'єктів у більшості своїх кроків повторює кроки алгоритму ініціалізації об'єктів і їх первісного розподілу по кластерах. Це свого роду одна чергова ітерація процесу кластерного аналізу. По завершенню цієї процедури отримуємо відсоток влучення в цільову групу, тобто відношення числа ІК, які насправді відвідали хоча б один з нових кластеризованих об'єктів до числа ІК, що беруть участь у кластеризації.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Основні обчислювальні завдання, пов'язані з персоналізацією пошуку розподіляються між сервером кластерного аналізу й сервером БД. На сервері кластерного аналізу відбувається формування й обробка Log-файлів і зберігається останній стан кластерної структури, який вираховується на сервері БД. Між сервером кластерного аналізу й сервером БД існує постійна синхронізація й відновлення.

Через обмеженість обсягу розділу, немає можливості розглянути всі експерименти й розрахунки можливих показників оцінки ступені персоналізації пошуку, проведених в процесі її підготовки. Проте, розглянемо кілька ключових показників у рамках завдання узагальненої кластеризації із застосуванням тимчасового вікна для ІК, числових коефіцієнтів підсилення Dom-моделі й тритактної кластеризації зі зворотного зв'язку для IP. Експерименти проводилися з використанням розроблених програмних модулів internet_res_search, ie_analyzer і HTMLdocdom на персональному комп'ютері із центральним процесором AMD FX-6100 Six-Core Processor, що мають тактову чистоту 3.30 ГГц, і оперативною пам'яттю ємністю 16 ГБ. Усі етапи кластеризації проводилися роздільно. Формування кластерів і розрахунки проводилися в середовищі MS SQL Server.

Експеримент по оцінці влучення в цільову групу – тксперимент проводився з даними про активність ІК, отриманими в період з 21 по 27 травня 2019 року. З усіх виконаних входжень більш 50% належить входженням на сайти соціальних мереж. У зв'язку з тим, що на них вирішується завдання персоналізації пошуку, коло розглянутих IP було звужене, і в експерименті оброблялися пошукові запити ІК до сайтів Google.com, gmail.com і fasebook.com, а також входження на сайти новин news.ukr.net, news.google.com. З одного боку, сайти новин можуть відбивати певні особливості й інтереси ІК, що виконують входження на їх сторінки, з іншого боку, вони містять велике число динамічних елементів, що

міняють текстовий зміст ІР, а це дозволяє застосовувати числові коефіцієнти підсилення й тритактну кластеризацію зі зворотним зв'язком.

Об'єкти дослідження – ІК та ІР – поєднувалися, і потім формувався узагальнений глобальний словник урізаних термінів і узагальнений характеристичний вектор. Експеримент проводився цілодобово (з 0 до 23 годин 59 хвилин) з інтервалом часу (вікном спостереження) 4 години. Слід зазначити, що при необхідності величину вікна спостереження можна міняти (змінюючи значення змінної @step1_hour_step) з метою підвищення показників персоналізації. У якості показників персоналізації можна використовувати коефіцієнт потрапляння в цільову групу, точність потрапляння, повноту вибірки, випадання й ін. У глобальному словнику термінів накопичувалися терміни, довжина яких однакова або перевищує 4 символи. Читання змісту ІР і фільтрація динамічних об'єктів виконувалася за допомогою розробленої програми HTMLdocdom, а вже на сервері БД застосовувалися числові коефіцієнти підсилення й узагальнена кластеризація об'єктів.

Підготовка узагальнених словників термінів і характеристичних векторів – lля виконання експерименту необхідно розбити весь масив отриманих за тиждень даних про ІК та ІР на підмасиви, віднесені до 4-годинних вікон спостереження, починаючи з нуля годин. Для кожного підмасива необхідно сформувати:

- динамічну таблицю, що містить глобальні словники термінів і лем;
- динамічну таблицю, що містить характеристичні вектори ІК;
- динамічну таблицю, що містить характеристичні вектори ІР;

Побудовані таблиці використовуються для виконання кластерного аналізу й розрахунку відсотка влучення в цільову групу.

У процесі формування зазначених таблиць можна спостерігати за графіками (рисунок 5.1) кількості об'єктів для кожного періоду спостереження.

Спостерігається чітка періодичність усіх представлених графіків – графіки досягають своїх локальних максимумів у проміжку часу з 12 до 16 годин. Локальні мінімуми з'являються також періодично в проміжку між 0 і 4 годинами.

Це пояснюється добовим циклом поведінки ІК – час активного життя змінюється часом відпочинку й сну.

Рисунок 5.1 – Графіки залежності кількості об'єктів від періоду спостереження

Експеримент 1: метод k-середніх, число кластерів 2 при Δt = 4 години. Для першого експерименту основними вхідними параметрами є число кластерів k = 2, тобто об'єкти будуть розподілені по двом кластерам, і Δt = 4 години (рис. 5.2).

Рисунок 5.2 – Графіки залежності відсотків потрапляння в цільову групу й кластеризації від періоду спостереження при k = 2 і Δt = 4 години

При підвищенні відсотка кластеризації збільшується й відсоток потрапляння в цільову групу. Не дивлячись на наявність пікових значень (порядку 60%), є й дуже низькі показники. Середнє значення відсотка потрапляння в цільову групу дорівнює 38,449%. Дисперсія дорівнює 114,862.

Після виконання цієї серії експериментів, побудови графіків (рис. 5.3), розрахунку середніх значень коефіцієнта потрапляння в цільову групу і його дисперсії, можна зробити висновок про те, що оптимальним рішенням для застосування кластерного аналізу є умови k = 3 або k = 4, тому що при цих умовах отримано добре значення коефіцієнта влучення в цільову групу з мінімальною дисперсією.

Результати експериментів сильно прив'язані до значення інтервалу часу Δt = 4 години. При $\Delta t < 4$ поведінка ІК стає більш локалізованою. Однак зменшується розмір як самих кластерів, так і характеристичних векторів і узагальненого глобального словника термінів, внаслідок чого одержувані значення коефіцієнтів повинні зрости. Потрібно відзначити ще одну особливість, що стосується періодів з низькими значеннями коефіцієнтів потрапляння в цільову групу – це динамічні процеси всередині кластерних структур: дифузія, поглинання кластерів або переміщення об'єктів з одного кластера в інший.

Випадання F вважається гарним показником для аналізу продуктивності пошукової системи (рис. 5.3).

Рисунок 5.3 – Гістограма випадання

Отримані показники для пошукової системи в першу чергу пов'язані з тим, що дослідження проводилося винятково на перших 50 гіперпосиланнях (немає можливості кластеризувати усі 3000000 IP). Не дивлячись на це, у результаті проведених експериментів, можна вважати доведеним перевагу застосування кластерного аналізу для персоналізації пошуку.

Слід зазначити, що значення зазначених показників будуть збільшуватися усе більше й більше, прагнучи до 1, при зростанні активності ІК.

ВИСНОВКИ

В атестаційній роботі запропонований комплексний підхід до кластеризації ІК і ІР, яка використовується для їхньої класифікації в рамках заходів для персоналізації Інтернет-пошуку.

У роботі наведений огляд існуючих некластерних методів класифікації об'єктів, досліджена можливість їх застосування до Інтернет-об'єктів – ІК і ІР. Некластерні методи можуть бути застосовані на перших етапах класифікації об'єктів до початку кластерного аналізу з метою попереднього розподілу об'єктів. Був зроблений висновок про те, що застосування алгоритмів кластерного аналізу є найкращим варіантом для рішення поставленого завдання. Кластерні методи, враховуючи їх різноманітність, здатні впоратися з величезними обсягами даних і працювати з векторами великої розмірності.

У ході роботи отримані наступні основні результати. 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.

З метою виявлення й фільтрації динамічних компонентів DOM-моделі запропонована схема кластеризації ІР зі зворотним зв'язком. Реалізація схеми дозволяє перетворювати динамічні ІР у статичні ІР і застосовувати до останніх стандартні алгоритми кластерного аналізу.

Розроблена математична модель представлення й процедури формування характеристичних векторів ІК і ІР, 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 ІР.

Пропонований підхід може бути використаний на рівні корпоративної мережі й охоплює практично всі етапи рішення цільового завдання:

– первинний збір інформації, у межах заданого ковзного тимчасового вікна, про пошукову активність ІК і відвідуваних ними ІР;

– багаторазове сканування Dom-моделі ІР, застосування числових коефіцієнтів підсилення й Dom-фільтрації;

– структуризацію об'єктів за допомогою спеціалізованої БД;

– формування глобальних словників термінів і лем;

– формування характеристичних векторів ІК та ІР, а також характеристичних векторів узагальнених Інтернет-об'єктів;

– розрахунок знаходження близькості між досліджуваними об'єктами;

– первісну ініціалізацію об'єктів;

– кластеризацію узагальнених об'єктів на основі алгоритму kсередніх.

Використання результатів кластеризації для персоналізації пошуку – результати аналізу пошукової активності ІК у поточному інтервалі часу можуть бути застосовані для прогнозу його інформаційних потреб у наступних інтервалах часу.

Розроблений набір програмних модулів для спостереження за активністю ІК і одержання текстового змісту ІР з обліком їх Dom-моделей. Розроблені спеціальні збережені процедури, що виконують усі необхідні розрахунки – від формування словників термінів до кінцевого розподілу об'єктів по кластерах. Зазначені модулі й збережені процедури утворюють єдину програмну систему, яка, бувши встановленою на сервери локальної мережі, дозволить, наприклад, організувати на підприємстві корпоротивну систему персоналізації пошуку.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Алгоритм Дейкстры // Електронний ресурс Викиконспекты. URL: http://neerc.ifmo.ru/wiki/index.php?title= D0%90%D0%BB%D0%B3%D0%BE%D1 %80%D0%B8%D1%82%D0%BC_%. (дата звернення 25.05.19 р.).

2. Алгоритм кластеризации *k*-means // Электронный ресурс // URL: http://robocraft.ru/blog/computervision/1061.html (дата звернення 25.05.19 р.)

3. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика. – 1989.

4. Основы моделирования и первичная обработка данных. М.: Финансы и статистика. – 1983.

5. Афонин А.А., Крейнес М.Г. Кластеризация текстовых коллекций: помощь при содержательном поиске и аналитический инструмент // Сборник научных статей «Интернет-порталы: содержание и технологии». Выпуск 4 / ФГУ «Информика». – М.: Просвещение. – 2007. – С. 510-537.

6. Барсегян А. А. Методы и модели анализа данных: OLAP и Data mining. / А. А. Барсегян, М. С. Куприяенов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-. – 2014.

7. Басакер Р., Саатн Т. Конечные графы и сети. Перевод с английского. – М: Наука. – 2003.

8. Библиотека работы с DOM HTML-документов для C# // Электронный ресурс // URL: http://htmlagilitypack.codeplex.com/ (дата звернення 25.05.19 р.)

9. Википедия. Яндекс // Електронний ресурс URL: http://ru.wikipedia.org/wiki/%DF%ED%E4%E5%EA%F1 (дата звернення 25.05.19 р.)

10. Воронцов К. В., Колосков А. О. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный интеллект. – 2006. № 2. – С. 30-33.

11. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования // Електронний ресурс URL: http://www.ccas.ru/voron/download/Clustering.pdf (дата звернення 30.05.19 р.)

12. Гайдышев И. Анализ и обработка данных: специальный справочник. – СПб.: Питер – 2001. – 752 с.

13. Гилл А. Введение в теорию конечных автоматов: теоретические основы технической кибернетики.– М.: Наука. – 1966.

14. Гулин В.В. Исследование и разработка методов и программных средств классификации текстовых документов // Електронний ресурс URL: http://www.mpei.ru/LANG/RUS/Publish/InfoAcadCncl/2013/GulinVV.pdf (дата звернення 25.05.19 р.)

15. Дунаев Е. В. Автоматическая рубрикация web-страниц в интернет-каталоге с иерархической структурой / Е. В. Дунаев, А. А. Шелестов // Интернет-математика 2005. Автоматическая обработка веб-данных. – М. 2005. – С. 382-398 .

16. Дюран Б. Кластерный анализ – М.: Статистика. – 1977. – 128 с.

17. Bien J., Tibshirani R. Hierarchical Clustering With Prototypes via Minimax Linkage // Journal of the American Statistical Association. 2011 // Електронний ресурс // URL: http://faculty.bscb.cornell.edu/~bien/papers/jasa2011minimax.pdf (дата звернення 25.05.19 р.)

18. Chakarbarti S. Mining the web: discovering knowledge from hypertext data. San Francisco: Morgan Kaufmann Publishers. – 2003.

19. Easily parse HTML Documents in C# // Електронний ресурс // URL: http://olussier.net/2010/03/30/easily-parse-html-documents-in-csharp/ (дата звернення 25.05.19 р.)

20. Kogan J. Introduction to Clustering Large and High-Dimensional data. – N. Y. : Cambridge University Press. – 2006.

21. Robertson S. Understanding Inverse Document Frequency: on theoretical arguments for IDF // Journal of Documentation, 2004, №5. – P. 503-520

22. Sculley D. Web-Scale K-Means Clustering // Конференция WWW 2010 //

23. Shamir R., Sharan R., Tsur D. Cluster graph modication problems Електронний ресурс // URL: http://www.ra.ethz.ch/cdstore/www2010/www/p1177.pdf (дата звернення 25.05.19 р.)

24. Engineering. 2014. № 3 // Електронний ресурс // URL: http://www.ijarcce.com/upload/2014/march/IJARCCE5H%20%20a%20pranjali%20%20MAIN%20CONTENT%20EXTRACTION.pdf (дата звернення 05.06.19 р.)

25. Guandong X., Yanchun Z., Lin L. Web Mining and Social Networking techniques and applications. – N. Y.: Springer. – 2018.

26. Gupta S., Kaiser G., Grimm P., Chiang M., Starren J. Automating Content Extraction of HTML Documents. Dordrecht: Kluwer Academic Publishers. 2018. // Електронний ресурс // URL: https://york.cs.columbia.edu/crunch/WWWJ.pdf (дата звернення 25.05.19 р.)