МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу

Кафедра інженерії програмного забезпечення

Пояснювальна записка до дипломного проєкту

магістра
(освітній ступінь)

на тему «Аналіз ефективності використання методів обробки великих масивів даних у рекомендаційних системах»

ХАІ.603.6-96ПЗ1.121.169133.21В

Виконав: студент 6 курсу групи № 6-96пз1
Спеціальність 121 – Інженерія програмного забезпечення
(код та найменування)
Освітня програма Хмарні обчислення та Інтернет речей
(найменування)
Красюк Т.А.
(прізвище й ініціали студента)
Керівник Туркін І.Б.
(прізвище та ініціали)
Рецензент Коваленко А.А.
(прізвище та ініціали)

Харків – 2021

Міністерство світи і науки України
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Факультет програмної інженерії та бізнесу
(повне найменування)
Кафедра інженерії програмного забезпечення
(повне найменування)
Рівень вищої освіти другий (магістерський)

Спеціальність 121 – інженерія програмного забезпечення
(код та найменування)
Освітня програма хмарні обчислення та Інтернет речей
(найменування)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ _____І. Б. Туркін_____
(підпис) (ініціали та прізвище)
“_____” _____ 2020 року

З А В Д А Н Н Я
НА ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ) СТУДЕНТУ

Красюк Тетяна Анатоліївна
(прізвище, ім’я, по батькові)

1. Тема дипломного проекту *Аналіз ефективності використання методів обробки великих масивів даних у рекомендаційних системах*
керівник дипломного проекту *Туркін Ігор Борисович, д.т.н., професор*
(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від “__” ___2020 року № ___
2. Термін подання студентом роботи _____
3. Вихідні дані до роботи: *методи колаборативної фільтрації*
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Провести огляд і аналіз існуючих рекомендаційних систем в області електронної комерції;.
Провести критичний огляд й аналіз методів аналізу великих масивів даних користувачів соціальних мереж;.
Розробити метод надання рекомендацій, щодо придбання товарів в області електронної комерції, користувачеві на основі аналізу його профілю в соціальних мережах.
Розробити прототип програмного забезпечення для пошуку користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

5. Перелік графічного матеріалу
РПЗ – стор. 80, рисунків – 11 шт., таблиць – 3 шт., презентація – 15 слайдів.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Туркін І.Б., зав. каф. 603		
2	Туркін І.Б., зав. каф. 603		
3	Туркін І.Б., зав. каф. 603		

8. Нормоконтроль _____ __В.А. Постернакова__ «____» _____2020 р.
(підпис) (ініціали та прізвище)

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проекту	Строк виконання етапів проекту	Примітка
1	Отримання і затвердження теми диплому	03.09.2019	
2	Аналіз предметної області	04.09.2019	
3	Постановка задачі	25.11.2019	
4	Проведення теоретичних досліджень	27.11.2019	
5	Розробка прототипу ПЗ	09.09.2020	
6	Підготовка пояснювальної записки	28.10.2020	
7	Оформлення пояснювальної записки до дипломного проекту	19.11.2020	
8	Передзахист дипломного проекту	11.02.2021	
9	Захист дипломного проекту	24.02.2021	

Студент _____ ___Красюк Т.А. ____
(підпис) (прізвище та ініціали)

Керівник роботи _____ ___Туркін І.Б.
(підпис) (прізвище та ініціали)

РЕФЕРАТ

Пояснювальна записка до дипломного проєкту містить 80 стор., 11 рис., 26 джерел.

Об'єкт дослідження – рекомендаційні системи в області електронної комерції.

Предмет дослідження – методи аналізу великих масивів даних для надання рекомендацій користувачеві в області електронної комерції.

Метою дослідження є підвищення ефективності електронних продаж в мережі Інтернет шляхом надання рекомендацій користувачам стосовно пропозицій товарів в області електронної комерції за рахунок розробки програмного забезпечення з пошуку користувачів соціальних мереж за схожими інтересами.

Наукова новизна. Удосконалено метод надання рекомендацій в області електронної комерції користувачеві, якій на відміну від існуючих використовує колаборативну фільтрацію при групуванні користувачів соціальних мереж щодо їх вподобань, що дає змогу підвищити ефективність продаж товарів в області електронної комерції.

Практична значимість отриманих результатів. В результаті проведених досліджень створено прототип програмного забезпечення для пошуку користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

РЕКОМЕНДАЦІЙНІ СИСТЕМИ, ОБРОБКА ДАНИХ, ВЕЛИКІ МАСИВИ ДАНИХ, КОЛАБОРАТИВНА ФІЛЬТРАЦІЯ, ЕЛЕКТРОНА КОМЕРЦІЯ, ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

ABSTRACT

Explanatory note to the master's thesis 80 pp., 11 fig., 26 sources.

The object of research - recommendation systems in the field of e-commerce.

The subject of research - methods of analysis of large data sets to provide recommendations to the user in the field of e-commerce.

The aim of the study is to increase the effectiveness of e-sales on the Internet by providing recommendations to users on product offerings in the field of e-commerce by developing software to search for users of social networks by similar interests.

Scientific novelty. The method of providing recommendations in the field of e-commerce to the user has been improved, which, unlike the existing ones, uses collaborative filtering when grouping users of social networks according to their preferences, which allows to increase the efficiency of e-commerce sales.

The practical significance of the obtained results. As a result of the research, a prototype of software was created to search for users of social networks with similar interests and provide them with recommendations in the field of e-commerce.

RECOMMENDATION SYSTEMS, DATA PROCESSING, LARGE DATA ARRES, COLLABORATIVE FILTRATION, ELECTRONIC COMMERCE, SOFTWARE

ЗМІСТ

ВСТУП

З приходом нових технологій у сучасний світ, інструментів і засобів комунікацій між людьми стає все дедалі більше і більше. Багато людей спілкується у соціальних мережах. Кількість великих даних росте с кожним днем, вони зокрема, виробляються людьми, потік інформації зростає з кожним роком у геометричній прогресії. Співвідношення коефіцієнта корисності при цьому зменшується. Отже, вся інформація генерується та може бути використана для певних цілей тільки після ретельної обробки.

Термін «Big Data» означає великі колекції або великі потоки даних, які не можуть бути оброблені традиційними комп'ютерними техніками. Цей термін означає не саме поняття «великі дані», а предмет дослідження, який включає в себе різні інструменти, техніки і платформи.

Великі дані включають в себе інформацію, що генерується різними системами і додатками. Деякі зі сфер, які потрапляють під визначення «Big Data»: чорний ящик: інформаційна складова частина вертольота, літака, морського чи космічного корабля.

Дані подібного роду включають в себе запис голосів екіпажу (мікрофони та навушники), інформацію про характеристики об'єкта управління; соціальні медіа: включають дані, великі дані поширюються через соціальні мережі. Фондові біржі в свою чергу зберігають інформації про операції купівлі та продажу між компаніями партнерами. Енергосистеми також в подібному роді містять дані та інформацію про вузли та навантаження енергетичних систем. Транспортні системи в свою чергу моделюють характеристики відстані, за допомогою обробки великих даних.

Як наслідок, термін «Big Data» включає в себе великий обсяг, високу швидкість обробки та дуже широке розмаїття даних та можуть бути поділені на три типи, структурні дані які використовують реляційні БД.

Напівструктуровані дані, тобто XML-файли з якими можна маніпулювати у інтернеті, неструктуровані дані чи файли формату Word, PDF, Text або медіа-журнали.

Великі дані дійсно мають вирішальне значення для нашого повсякденного життя і стають однією з найважливіших технологій в сучасному світі. Найпоширенішими і відомими є лише кілька переваг. Наприклад, використання інформації, що зберігається в соціальних мережах, маркетингові агентства вивчають зворотний зв'язок між рекламою яку вони надають користувачам та як користувачі реагують на цю рекламу.

У свою чергу, використання інформації у соціальних медіа або соціальних системах, має дуже багато переваг та сприйняття продукту споживачами, залежить від того, як компанії надають кінцевому користувачу, пропозицій, щодо купівлі товару.

Відносно такої сфери, як медицина, застосовність великих даних про попередньої історії хвороби пацієнтів сприяє забезпеченню кращого та більш швидшого обслуговування, а згодом і відновлення пацієнтів.

Для використання можливостей операції з великими даними потрібно інфраструктура, яка може управляти і обробляти величезні обсяги структурованих і неструктурованих даних в реальному часі і може захистити конфіденційність і безпеку даних. Існують різні технології на ринку від різних постачальників, включаючи такі компанії, як Google, IBM, Microsoft, SAP та ін.

Великі Дані набули широкого поширення в багатьох галузях бізнесу. Їх використовують в охороні здоров'я, телекомунікації, транспортних мережах, логістиці, в фінансових компаніях, а також в державному управлінні.

За допомогою технологій Big Data підприємства можуть аналізувати величезні масиви даних і виявляти корисні закономірності, що дають їм конкурентні переваги. Для роботи з великими даними потрібні дуже потужні комп’ютери та структурування цих даних щодо того, де і як людина захоче їх використовувати і які завдання він ставить перед собою.

Об'єкт дослідження – рекомендаційні системи в області електронної комерції.

Предмет дослідження – методи аналізу великих масивів даних для надання рекомендацій користувачеві в області електронної комерції.

Метою дослідження є підвищення ефективності електронних продаж в мережі Інтернет шляхом надання рекомендацій користувачам стосовно пропозицій товарів в області електронної комерції за рахунок розробки програмного забезпечення з пошуку користувачів соціальних мереж за схожими інтересами.

Для досягнення поставленої мети необхідно вирішити ряд завдань: провести огляд і аналіз існуючих рекомендаційних систем в області електронної комерції; провести критичний огляд й аналіз методів аналізу великих масивів даних користувачів соціальних мереж; розробити метод надання рекомендацій, щодо придбання товарів в області електронної комерції, користувачеві на основі аналізу його профілю в соціальних мережах; розробити прототип програмного забезпечення для пошуку користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

Методи досліджень. У роботі було використано методи аналізу, синтезу, системного аналізу, порівняння та логічного узагальнення результатів, методи роботи з великими масивами даних.

Наукова новизна. Удосконалено метод надання рекомендацій в області електронної комерції користувачеві, якій на відміну від існуючих використовує колаборативну фільтрацію при групуванні користувачів соціальних мереж щодо їх вподобань, що дає змогу підвищити ефективність продаж товарів в області електронної комерції.

Практична значимість отриманих результатів. В результаті проведених досліджень створено прототип програмного забезпечення для пошуку

користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

1 АНАЛІЗ ПРОБЛЕМНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ

В останні роки кількість інформації росте з кожним днем та генерується бізнесом, урядом або науковою базою в якії працюють люди певної кваліфікації, надзвичайно зростає це явище, відоме як потік даних. У бізнесі транзакційні бази даних такі як Walmart, за оцінками експертів, містять більше 2,5 петабайта даних, що складаються з даних про поведінку та переваги клієнтів інтернет сервісів. Активності мережі та пристроїв і даних про тенденції розвитку на ринку. У різних цілях доцільно використовувати методи роботи з великими даними, наприклад у 24 року з’явився матеріал з Афганістану та Іраку в 2009 році [1] де наводится повний аналіз великих даних. У науці на Великому адронному коллайдері в ЦЕРНі в 2010 році було отримано 13 петабайт даних, це дуже багато за настільки короткий час [2].

Зокрема ми можемо зауважити, що дані датчиків, соціальних мереж, мобільних пристроїв та місцеположення користувачів соціальних мереж ростуть з дуже велико. швидкістю. Паралельно з цим можна зазначити рост даних, також з кожним роком дані стають все більш взаємопов'язаними між собою.

Наприклад, Facebook майже повністю підключений: 99,81 відсотка користувачів соціальної мережі належать до однієй пов’язаної мережі [3]. Цей дивовижний опит зростання і різноманітність глибоко вплинули на те, як люди обробляють і інтерпретують нові знання за допомогою великих даних. Оскільки велика частина цих даних створюється і зберігається в Інтернеті, однією з відкритих завдань є визначення того, як повинна у майбутньому розвиватися технологія інтернет обчислень, щоб ми мали змогу отримувати доступ або збирати, чи аналізувати і обробляти великі дані в цілому.

Прийнято вважати, що великі дані є першокласними компонентами в інтернеті. Спільне взаємодія між даними і обчислення даних у мережі

інтеренет, це насамперед інфраструктура має життєво важливе значення для забезпечення аналітики великих даних з малою похибкою та високою пропускною спроможністю. В свою чергу в соціальних мережах та аналітиці, ми можемо охоплювати багато комп'ютерних парадигми, в тому числі хмарні обчислення та хмарні сервіси.

В даний період часу більшість соціальних мереж об'єднують людей або групи людей, які демонструють схожі інтереси або ведуть себе заздалегідь передбачувано. У найближчому майбутньому ми можемо очікувати, що соціальні мережі з'єднають різні об'єкти системи, такі як програмні компоненти, веб-служби, ресурси даних і робочі процеси. Що ще більш важливо, взаємодія між людьми та нелюдськими ресурсами значно підвищила продуктивність даних якими у майбутньому можуть оперувати вчені. Аналітика великих даних з часом накопичує все більше і більше знань та підходів у взаємодіях людських ресурсів з цифровими ресурсами, це дозволяє насамперед виявляти закономірності, та надавати більш кращі принципи вирішення проблем пов'язаних для використання прямих пропозицій електронних продаж.

Наприклад, в недавніх подіях, пов'язаних з терактом на бостонському марафону в 2013 році, соціальні мережі, а саме профілі учасників марафону допомогли людям відшукати одне одного, це демонструє як великі дані допомагають людям, які пов’язані спільними інтересами відшукати один одного. Загальні високопродуктивні обчислювальні методи були об'єднані для кластеризації і аналізу великих наборів даних і справжніх фотографій і відеокадрів, що в кінцевому підсумку призвело до виявлення злочинців. Цей приклад ілюструє, як хмарні технології обробки можуть задовольнити обчислювальні потреби, в той час як аналітика посилюється завдяки спеціальному досвіду учасників соціальних мереж. Дуже швидкий темп зростання і в свою чергу різноманітність даних, дозволяє пов'язувати дані між собою та продовжує надавати глибоке вплив на те, як люди згодом зможуть

використовувати ці дані. Ми маємо змогу визначити цю взаємодію, як коло добропорядних людей, які використовують великі дані, лише з користю.

Фізичні або юридичні особи, в свою чергу мають змогу проводити особистий аналіз великих даних на онлайн підключенях, використовуючи хмарні засоби, які в свою чергу потрібні для цього і зроблені або підключенні комп'ютери до сеті та аналітика великих даних з цих підключених комп'ютерів може генерувати розумні пристрії, які потім поширюється та вертаються назад до користувачів у вигляді даних які вони і хотіли побачити.

Веб-сайти соціальних мереж, такі як Twitter, Facebook, LinkedIn, YouTube та Вікіпедія, не лише підключили велику кількість користувачів, але й захопили багато інформації, пов'язаної з їх щоденною взаємодією. Соціальна мережа має свої початки в роботі з соціальними проблемами які насамперед висвітлюють люди у своїх веб-сторінках на різноманітних інтернет ресурсах, але останнім часом, комп'ютерних науковців у дослідженні інформації або соціальних мереж, що підтримуються інтернетом можуть буду по різному інтерпретуватись залежачи від кінцевого користувача [4]. Таким чином, ми можемо відокремити основні дослідницькі проблеми в цих сферах.

Соціальні мережі, як не дивно починаючи ще з далеких 1920-х років, соціологи досліджували міжособистісні стосунки між людьми, оскільки вони стосуються великої мережевої соціальної групи або суспільних груп взаємопов'язаних людей між собою деякими інтересами. Ці дослідження дають змогу чітко сформулювати міцні стосунки між користувачами, але неявно визначили, як довіра відіграє взаємозв'язки цих відносин. Але якщо ми будемо на пряму відштовхуватись від цих систем або мереж, ми маємо змогу як соціологи застосувати деякі методи. Підходи до моделювання заздалегідь включає в себе збір даних, блокове моделювання системи, орієнтованих на мережу та на широку вибірки даних. Вимірювання цих засобів включає в себе різні міри централізації для групування користувачів за їх інтересами, оцінку соціальної мережі, де насамперед знаходяться користувачі або аналіз

листування для дворежимних мереж, також статистичну оцінку моделі за допомогою якої відбувається взаємодія користувачів між собою у соціальних мережах.

Зважаючи на те, що структура мережі нерегулярна, складна і динамічно розвивається завжди навіть у дійсний час, основним напрямком складної теорії мережі є розробка принципових, математичних підходів, що оцінюють мережі мільйонів вузлів. Важливим механізмом виведення поведінки цих мереж є аналіз довжин шляху та групування пов'язаних структур шляху. Складні мережі ми можемо уявити бути як найбільш фундаментальні форми, наприклад у вигляді графіків або мереж малого виміру простору, але доцільно буде вважати більш складні топографічні представлення, насамперед як зважені графи послідовності. Одним з найпопулярніших підходів до управління цими мережами, є змога ділитися інформацією за допомогою комп'ютерів, це насамперед є розподіл орієнтованих графів, які визначають мінімальну кількість ребер між двома наборами вершин у даному графі.

Ієрархічна кластеризація - це дуже ефективний метод насамперед для соціальних мереж, у яких можуть бути відсутні знання про кількість користувачів та кількість груп до яких можуть відноситись ці користувачі. Цей підхід в свою чергу намагається розділити вузли графів на кластери, де з'єднання всередині кластера є більш тісними, ніж з'єднання у різних кінцях цього кластеру або з вузлами призначеними для декого іншого кластеру. Якщо брати інщі підходи води намагаються шукати найбільшу відстань між вузлами, поки кластери не сформуються природним шляхом, як це і повинно бути.

1.1 Аналіз аналогів

На даний момент існують безліч веб-сервісів, що надають змогу функціонувати різним алгоритмам аналізуючи методи big data для обробки

даних вподобань користувачів різних соціальних мереж для того, щоб мати змогу робити прямі пропозиції для цільовой аудиторії.

Наприклад Hadoop є проектом верхнього рівня організації Apache Software Foundation, тому основним дистрибутивом і центральним репозиторієм для всіх напрацювань вважається саме Apache Hadoop. Однак цей же дистрибутив є основною причиною більшості незручностей з якими може зіткнуться користувач по замовчуванням установка цього додатку на кластер вимагає попереднього налаштування машин, ручного регулювання пакетів, редагування безлічі файлів конфігурації і купи інших різноманітних дій для повної взаємодій цього додатку з іншими веб-сервісами. При цьому документація найчастіше неповна або просто застаріла. Спочатку Hadoop був, в першу чергу, інструментом для зберігання даних і запуску MapReduce-задач.

На сьогоднішній день Hadoop вже став великим стеком різних технологій, пов'язаних з обробкою великих даних. технологія Hadoop, яка представляє собою програмний фреймворк, що дозволяє зберігати і обробляти дані за допомогою комп'ютерних кластерів, використовуючи парадигму MapReduce.

З приходом нової технологій у сучасний світ почали по іншому взаємодіяти соціальні мережі з користувачами. Ця парадигма працює з великою кількістю даних, її використовує компанія Amazon. Сервісом використовується збір персональних рекомендацій і динамічного ціноутворення. З огляду на тисячі чинників, електронний ритейлер змінює ціни кожні дві хвилини.

Крім цього, завдяки використанню нового фреймворку та Big Data почали використовуватись для детектування шахрайських дій. За повідомленнями самої компанії, вже через шість місяців після впровадження нових технологій кількість махінацій з пластиковими картами скоротилося на 50 відсотків, це дуже хороший показник для роботи за такою великою кількістю людей.

Рішення для роботи з великими даними використовується і профільними компаніями. Наприклад, постачальник товарів для офісу Staples, обслуговуючий майже всю територію США, впровадив нові технології, тому що потрібно було автоматизувати обробку і аналіз понад 10 мільйонів транзакцій в 1100 роздрібних магазинах по всій країні щодня.

Зображений цикл взаємного посилення, де проходить обробка великих даних (див.рис.1.1). Взаємопов'язані люди генерують потік даних, аналізований взаємопов'язаними комп'ютерами, а знання, які добуваються в результаті цього аналізу, передаються знову людям, тим самим ми можемо бачите, те як проходить взаємообмін великих даних серед людьми які знаходяться на великій відстані друг від друга.

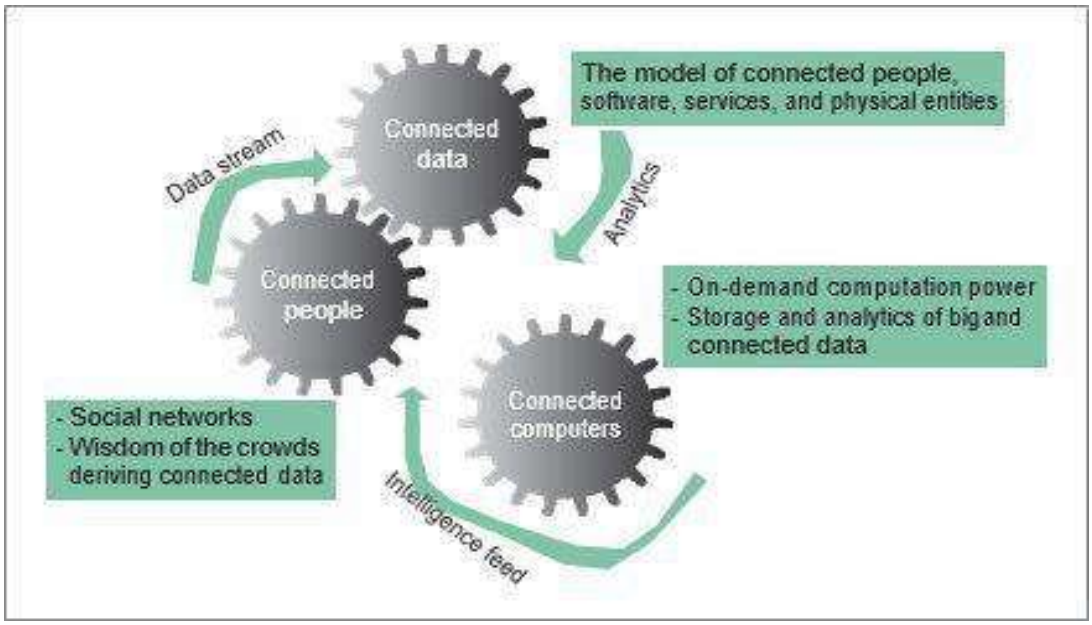

Рисунок 1.1 – Цикл взаємного посилання

В результаті була створена програма прогнозування денного і тижневого попиту, а витрати на маркетинг і рекламу знижені на 25 відсотків, так як вони стали націленими.

Продовольча компанія яка займається харчуванням в Великобританії. Ця компанія Waitrose спеціалізується на винах і бакалії. За рахунок аналізу великих даних з урахуванням сезонного попиту і наближенням свят компанії вдалося оптимізувати запаси і зменшити на 40 відсотків перезаказа товарів у постачальників, пов'язані з неправильною оцінкою попиту.

Наприклад, оператори фіскальних даних, приймають дані від касових апаратів рітейлерів. Аналізуючи ці показники, система може надати власнику бізнесу практично будь-яку інформацію, а також необхідні рекомендації [5]. Так можна зрозуміти, чи ефективна промоакція, а також можливі подальші дії, причому в режимі реального часу.

Крім аналізу даних, керівництво мережі може отримати інформацію про роботу кожної торгової точки, найбільш добре продаються товари в даний момент часу або збоїв в роботі.

1.2 Методи аналізу даних для пошуку цільової аудиторії

Існує безліч різноманітних методів аналізу масивів даних для пошуку цільової аудиторії серед користувачів соціальних мереж, в основі яких лежить безліч різноманітних інструментів, які так чи інакше зв’язані між собою приблизно однаковим підходом для аналізу великих даних: кластеризація, регресія, багатовимірний аналіз (OLAP), класифікація та пошук закономірностей.

Інкрементна обробка і приблизний результат. Обсяг і швидкість пред'являють суперечливі вимоги до систем великих даних. Великий обсяг даних вводиться в таку систему з високою швидкістю, в той час як аналіз і інтерпретація повинні виконуватися з однаковою швидкістю. У традиційній бізнес-аналітиці [6], дані транзакцій спочатку обробляються в системі онлайн-

обробки транзакцій (OLTP), а потім проходять через процес вилучення, перетворення, завантаження (ETL) в пакетному режимі. Зрештою, дані завантажуються в сховище даних онлайнової аналітичної обробки **(**OLAP), де вони аналізуються для забезпечення стратегічного розуміння. Цей підхід OLTPETL-OLAP [7] обмінює своєчасність на точність, з огляду на те що, між моментом коли стають доступними дані, і формуванням розуміння відбувається значна затримка серед обміном даними в часі. У деяких додатках з великими даними, таких як виявлення фінансового шахрайства та просування на ринку, тривалі затримки неприпустимі. Нещодавно з'явилася парадигма, звана потоковими обчисленнями, дозволяє здійснювати безперервні запити по потоковим даними, таким як канали соціальних мереж і записи даних викликів. Потокові обчислення відкривають шлях до аналітики в реальному часі, але залишаються деякі проблеми. Одним з них є взаємодія між створенням моделі пакетного режиму і вимірюванням потоків в реальному часі. З одного боку, накопичені історичні дані в сховище даних можуть допомогти інформаційним фахівцям побудувати статистичну модель для управління обробкою потоку - наприклад, вирішити які функції слід відстежувати і допомогти встановити поріг реагування. З іншого боку, знову надійшли дані з потокової системи повинні бути використані для настройки моделі у відповідності з останніми тенденціями. Механізм інкрементальної обробки даних і налаштування моделі має життєво важливе значення для цієї взаємодії. Що стосується проблем зі швидкістю передачі даних, інша перспектива полягає в наданні приблизних, своєчасних результатів для запитів або пріоритизації різних запитів шляхом виділення різної кількості ресурсів. Таким чином, можливі різні рівні узгодженості даних, в яких запити може бути точним, але повільним або кращим, але швидким.

Дані технології застосовуються для отримання та сприйняття людиною результатів, ефективних в умовах безперервного приросту, розподілу інформації по численним вузлам обчислювальної мережі.

Існує безліч різноманітних методик аналізу масивів даних, в основі яких лежить інструментарій, запозичений з статистики та інформатіки.

A/B testing. Методика, в якій контрольна вибірка по черзі порівнюється з іншими. Таким чином удасться виявити оптимальну комбінацію показників для досягнення, наприклад, найкращою відповідної реакції споживачів на маркетингову пропозицію. Великі дані дозволяють провести величезну кількість ітерації і таким чином отримати статистично достовірний результат.

Association rule learning. Набір методик для виявлення взаємозв'язків, тобто асоціативних правил, між змінними величинами в великих масивах даних. Використовується в data mining.

Classification. Набір методик, які дозволяють передбачити поведінку споживачів в певному сегменті ринку (прийняття рішень про покупку, відтік, обсяг споживання та ін.). Використовується в data mining.

Cluster analysis. Статистичний метод класифікації об'єктів по групах за рахунок виявлення наперед невідомих загальних ознак. Використовується в data mining.

Crowdsourcing. Методика збору даних з великої кількостю джерел.

Data fusion and data integration. Набір методик, який дозволяє аналізувати коментарі користувачів соціальних мереж та зіставляти з результатами продажів в режимі реального часу.

Data mining. Набір методик, який дозволяє визначити найбільш сприйнятливі для продукту, що просувається або послуги категорії споживачів, виявити особливості найбільш успішних працівників, передбачити поведінкову модель споживачів.

Ensemble learning. У цьому методі задіюється безліч предикативних моделей за рахунок чого підвищується якість зроблених прогнозів.

Genetic algorithms. У цій методиці можливі рішення представляють у вигляді «хромосом», які можуть комбінуватися та мутувати. Як і в процесі природної еволюції, виживає найбільш пристосована особина.

Machine learning. Напрямок в інформатиці (історично за ним закріпилася назва «штучний інтелект»), яке має на меті створення алгоритмів самонавчання на основі аналізу емпіричних даних.

Natural language processing (NLP). Набір запозичених з інформатики та лінгвістики методик розпізнавання природної мови людини.

Network analysis. Набір методик аналізу зв'язків між вузлами в мережах. Стосовно до соціальних мереж дозволяє аналізувати взаємозв'язку між окремими користувачами, компаніями, спільнотами і т.п.

Optimization. Набір чисельних методів для редизайну складних систем та процесів для поліпшення одного або декількох показників. Допомагає в прийнятті стратегічних рішень, наприклад, складу виведеної на ринок продуктової лінійки, проведенні інвестиційного аналізу та інше.

Pattern recognition. Набір методик з елементами самонавчання для передбачення поведінкової моделі споживачів. Regression. Набір статистичних методів для виявлення закономірності між зміної залежной змінної та однією або декількома незалежними. Часто застосовується для прогнозування та пророкувань. Використовується в data mining.

Simulation. Моделювання поведінки складних систем часто використовується для прогнозування, передбачення і опрацювання різних сценаріїв при плануванні.

Time series analysis. Набір запозичених з статистики та цифрової обробки сигналів методів аналізу, що повторюються з плином часу послідовностей даних. Одні з очевидних застосувань - Відстеження ринку цінних бумаг. В даний час безліч компаній стежать за розвитком технологій Big Data. Аналітична компанія IDC представила в 2017 р звіт «Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East», в якому передбачалося, що обсяги інформації будуть подвоюватися кожні 2 роки протягом наступних 8 років.

У найближчі 7 років кількість даних в світі досягне 40 ЗБ (1 ЗБ = 1021 байт), а це означає, що на кожного жителя Землі припадатиме по 5200 ГБ даних (див. рис. 1.2).

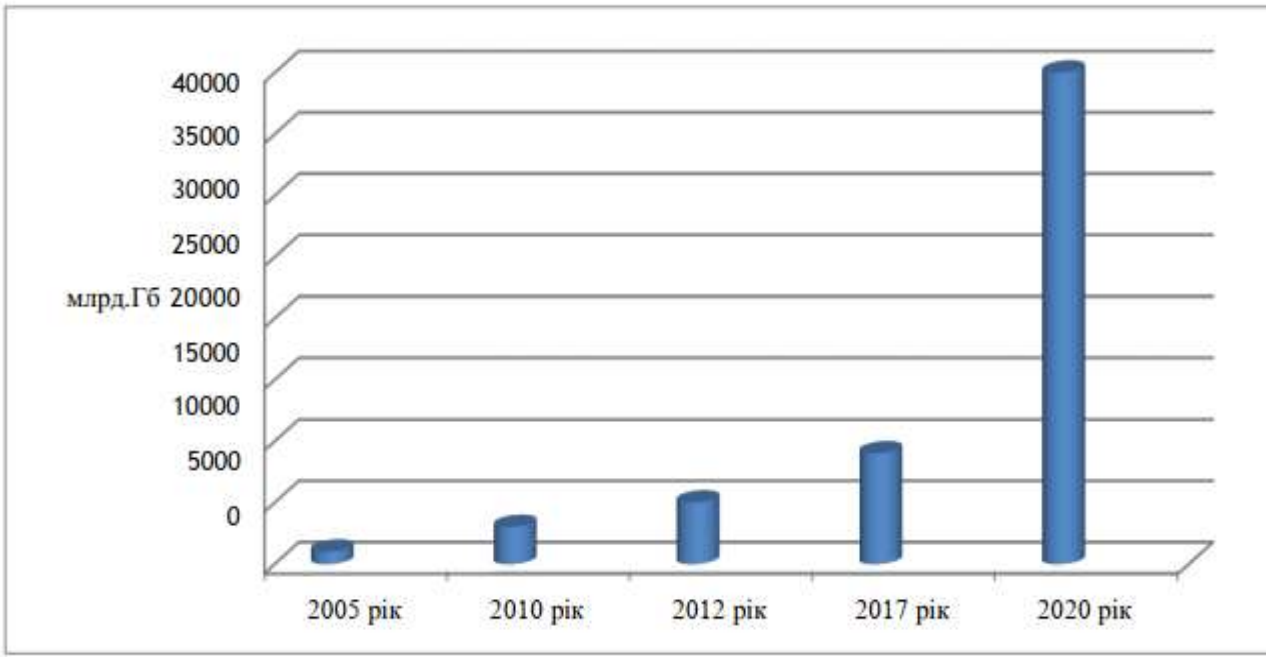

Рисунок 1.2 – Загальний обсяг цифрових даних в світі

Традиційні методи аналізу інформації не можуть наздогнати величезними обсягами постійно зростаючих і оновлюваних даних, що в підсумку і відкриває дорогу технологіям які дуже швидко розвиваються з використанням Big Data.

Можна виділити наступні особливості технологій Big Data:

– робота з інформацією величезного обсягу і різноманітного складу;

– інформація вельми часто оновлюється і знаходиться в різних джерелах;

– якісно відрізняється методи які відкриваються аналітика для виявлення практичних знань, які безпосередньо монетизуються в прибуток;

– наочне відображення звітів і можливості сценарного аналізу («що, якщо буде ...»);

– мета застосування технологій Big Data - збільшення ефективності роботи, створення нових продуктів та підвищення конкурентоспроможності.

Технології Big Data можуть бути корисні для вирішення наступних завдань:

- прогнозування ринкової ситуації;
- маркетинг і оптимізація продажів;
- вдосконалення товарів і послуг;
- прийняття більш обгрунтованих управлінських рішень на основі аналізу Big Data;
- оптимізація швидкого знаходження переваг користувача;
- підвищення ефективного впровадження нових товарів;
- ефективна логістика;
- моніторинг стану основних вподобань користувачів.

Методика та інструменти роботи зі структурованими даними вже давно створені.

Це реляційна модель даних і системи управління базами даних. Але в сучасних умовах соціальних мереж потрібно обробляти великі обсяги неструктурованих даних різних типів (рис. 1.3), а для цієї роботи колишні методи не зовсім підходять.

Simulation. Моделювання поведінки складних систем часто використовується для прогнозування, передбачення і опрацювання різних сценаріїв при плануванні.

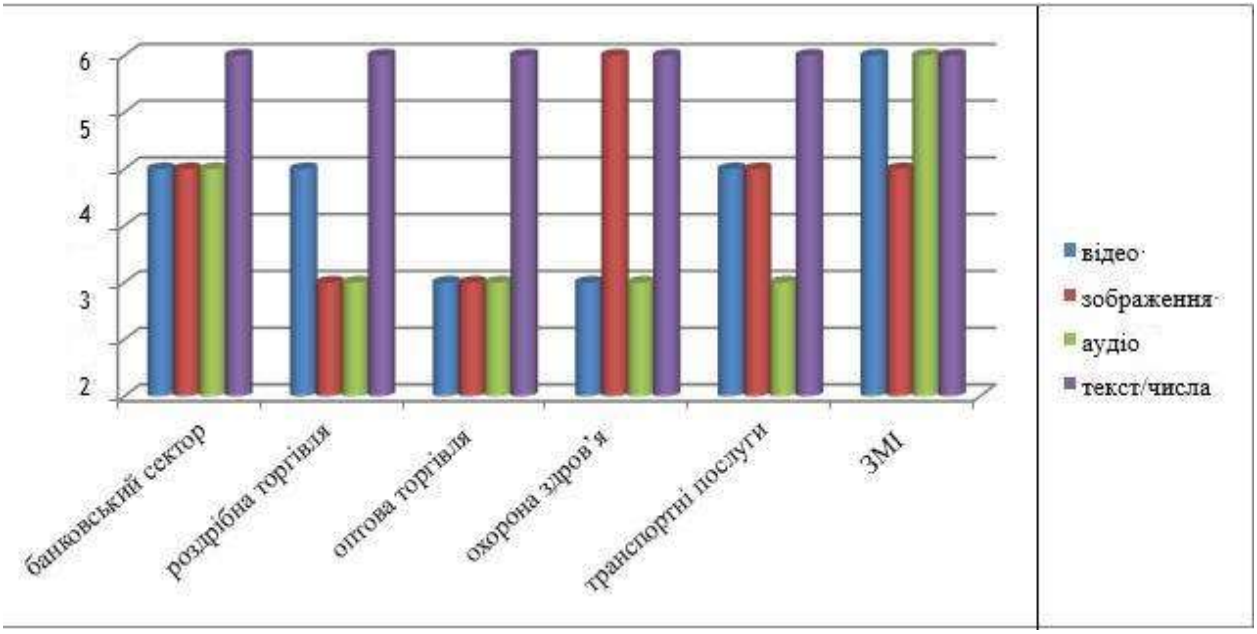

Рисунок 1.3 – Домінуючи типи інформації для різних сфер діяльності

1.3 Постановка мети й завдань дослідження

Метою дослідження є підвищення ефективності електронних продаж в мережі Інтернет шляхом надання рекомендацій користувачам стосовно пропозицій товарів в області електронної комерції за рахунок розробки програмного забезпечення з пошуку користувачів соціальних мереж за схожими інтересами.

Для досягнення поставленої мети необхідно вирішити ряд завдань:

– провести огляд і аналіз існуючих рекомендаційних систем в області електронної комерції;

– провести критичний огляд й аналіз методів аналізу великих масивів даних користувачів соціальних мереж;

– розробити метод надання рекомендацій, щодо придбання товарів в області електронної комерції, користувачеві на основі аналізу його профілю в соціальних мережах;

– розробити прототип програмного забезпечення для пошуку користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

1.4 Висновки по розділу 1

В першому розділі показано актуальність теми дослідження, проведено огляд і аналіз існуючих рекомендаційних систем в області електронної комерції. Проведено огляд методів аналізу масивів даних для пошуку цільової аудиторії серед користувачів соціальних мереж. Зроблено постановку мети й завдань дослідження.

2 АНАЛІЗ ВИКОРИСТАННЯ МЕТОДІВ ОБРОБКИ ВЕЛИКИХ МАСИВІВ ДАНИХ У РЕКОМЕНДАЦІЙНИХ СИСТЕМАХ

2.1 Виявлення підходів рекомендації

Заданий вхідний набір векторів користувачів, вектори враховують діяльність користувача в межах однієї мережі згідно їх інтересів, що розробляється та діяльність користувача в соціальних мережах, в даному випадку лише в соціальній мережі Facebook. Заданий набір даних певного типу: автомобілі, пральні машини, комп'ютерні ігри, смартфони. Кожен з цих видів товарів має певний рейтинг, а також кожен користувач може висловити свою думку про нього у вигляді оцінки та коментаря (в даному випадку важлива лише оцінка, так як для перетворення коментарів в числові значення потрібно розробляти додаткові модифікації).

Необхідно на основі поведінки користувача в межах системи та на основі даних отриманих поза системою підготувати цільові пропозиції, які будуть задовільними для нього. Основна ціль даного сервісу зі сторони бізнесу: збільшити аудиторію, також збільшити час перебування користувачів на сайті, покращити надання рекламних пропозицій, дані зміни покликані в свою чергу збільшити прибуток бізнесу. Ціль зі сторони користувача: знизити час на пошук необхідного матеріалу, покращити досвід користування сайтом.

Завдяки тому, що є дані з соціальної мережі можна уже на початковому етапі надавати користувачеві певні рекомендації, щодо товару якиї він у майбутньому захоче придбати, а також коли навіть буде здійснена достатня кількість дій в системі, інформація з соціальної мережі повинна допомогти точніше надавати рекомендації. Адже буде використовуватися більш повний профіль користувача.

Власна статистика мережі Twitter показує, що на сьогоднішній день активними є 284 мільйони користувачів щороку, і щодня надсилається близько 500 мільйонів твітів [8]. Twitter - це соціальна мережа мікро-блогів, за допомогою якої люди можуть спілкуватися та ділитися своїми думками, використовуючи повідомлення, також відомі як твіти. Twitter є платформою для спілкування не тільки звичайних користувачів, але і для знаменитостей та спілок людей. Таким чином, Twitter є також і потужним маркетинговим інструментом, який використовують організації для створення іміджу брендів та отримання відгуків від клієнтів.

Отже, багато організацій інвестують ресурси в аналіз Twitter з наступних причин:

– рекламувати свої продукти та послуги для залучення більшої кількості клієнтів;

– як дешеве джерело реклами та просування;

– отримати зворотній зв’язок від своїх користувачів;

– щоб зрозуміти свою позицію на ринку.

Аналіз даних твітів та групування їх у важливі категорії складне завдання, оскільки твіти:

– мають обмежену кількість слів через обмежену кількість символів, дозволених Twitter;

– мова їх написання зазвичай неформальна.

2.2 Огляд методів обробки великих масивів даних

Основною складністю при кластеризації користувачів є вибір методів, за якими буде порівнюватись їхня схожість.

Провівши аналіз існуючих способів кластеризації користувачів було вирішено використовувати схожість користувачів між собою, як оцінку відстані між ними у кластері. Для спрощення роботи та збільшення точності обробки, вирішено брати для аналізу агреговану колекцію всіх твітів користувача.

Індекс схожості визначає відстань між двома текстовими складовими користувачів. Кластерні алгоритми використовують деяку функцію вимірювання подібності як критерій віднесення елементу до певного кластеру. Ці функції допомагають визначити, наскільки схожі або несхожі два користувача.

Перш за все, документи повинні бути перетворені у векторну форму, оскільки алгоритми кластеризації та методи вимірювання дистанції не можуть інтерпретувати документи в їх оригінальній формі. Векторна модель (VSM) - широко використовуваний метод представлення тексту документа. Вага елементу (токену) вектору визначає свій внесок у семантику документа.

У будь-якому окремому твіті можуть міститись терміни або слова, які не є важливими для розгляду при кластеризації, також можна зазначити, що слід видалити інформацію, яка буду не потрібна у наступних маніпуляціях, щоб зберегти тільки ті ознаки, які будуть необхідні в дальнішому. Твіттер містить у собі зазвичай тільки короткі повідомлення,також доцільно було б зазначити, що користувачі у свою чергу, як правило, використовують не літературну мову.

Тому, попередня обробка повідомлень перед аналізом може стати складним завданням. Проте, вона є важливим кроком, оскільки це може вплинути на результати процесу аналізу схожості.

Токенізація – це процес розбиття строки або документу на невеликі логічні частини, які називаються токенами.

Наприклад повідомлення такого типу: «@AliBunkall : When Obama took office, 180000 US troops were in global conflict zones», буде перетворено на такий масив токенів:

[«@MuhamedAli», «:», «Then», «Putin», «take», «place», «126548», «UK», «grils», «was», «on», «junior», «confuze», «space»],

Тобто першим етапом твіти будуть розбиті на масиви токенів.

Пунктуація також може створювати зайвий шум, при тому що не несе у собі ніякого смислового навантаження, яке б можна було використати, отже на цьому етапі масив перетворюється у такий:

[«@MuhamedAli», «Then», «Putin», «take», «place», «126548», «UK», «grils», «was», «on», «junior», «confuze», «space»].

Видалення стоп слова також можуть створювати дуже великий шум, адже вони є дуже часто повторюваними, не маючи у собі ніякої тематики, тобто вони загальні.

Прикладом таких слів для англійської мови буде:

«i», «me», «my», «myself», «we», «our», «he», «you», «your», «yours», «yourself», «yourselves», «him», «herself», «its», «them», «themselves», «what», «who», «whom», «this», «that», «these», «those», «am», «is», «was», «were», «be», «being», «have», «has», «had», «having», «do», «did», «doing», «a», «an», «and», «but»,«how», «any», «both», «she», «few»,«how», «any», «both», «she», «few», «if», «or», «because», «as», «until», «while», «of», «at», «by», «with», «about», «against»,«how», «any», «both», «she», «few»,«how», «any», «both», «she», «few», «between», «into», «through», «during», «before», «after», «above», «below», «to», «up», «down», «in», «out», «on», «off», «over», «again», «further», «then», «itself», «here»,«how», «any», «both», «she», «few»,«how», «any», «both», «she», «few», «his», «there», «when», «where», «why», «how», «any», «both», «she», «few», «more»,«how», «any», «both», «she», «few»,«how», «any», «both», «she», «few», «most», «other», «some», «such», «no», «nor», «only», «own», «it», «so», «than»,

«too»,«how», «any», «both», «she», «few»,«how», «any», «both», «she», «few», «very», «s», «t», «can», «will», «just», «don», «hers», «should», «now» [9].

Отже після цього етапу масив токенів буде мати такий вигляд:

[«@AliBunkall», «Obama», «took», «office», «180000», «US», «troops», «global», «conflict», «zones»].

Після виконання попередніх кроків, все одно можуть залишатись деякі неважливі слова. Наприклад, слова з довжиною менше трьох символів. У деяких випадках короткі слова також можуть бути актуальними. Однак у цьому сценарії функції короткої довжини не є релевантними для процесу знаходження тематики і їх слід видалити. Таким чином, для видалення слів довжиною менше трьох символів потрібна подальша фільтрація. Іноді алфавітно-цифрові слова, такі як "abc123", "180000", також зустрічаються в документах чи твітах, які не є важливими для класифікації тексту. Ці слова також відкидаються.

Отже на цьому кроці масив токенів матиме такий вигляд:

[«@AliBunkall», «Obama», «took», «office», «troops», «global», «conflict», «zones»].

У Twitter посилання на зовнішні ресурси або ж на інших користувачів є дуже часто вживаними. Але для знаходження тематики вони будуть лише заважати. Зазвичай ім’я користувача починається з «@» символу а посилання починаються з «http».

Отже, провівши останній етап підготовки твітів отримується масив токенів, кожен із яких може мати якусь вагу у всьому документі та представляти певну тематику:

[«Obama», «took», «office», «troops», «global», «conflict», «zones»].

Методи кластеризації розподіляють дані за різними групами користувачів або кластерізують користувачів шляхом мінімізації цільової функції. Вони дуже довго відомі, як методи кластеризації на основі центроїдів. Для того, щоб розділити набір даних, об’єкти порівнюються з центрами кластерів, таким чином, щоб цільова функція була мінімальною або максимальною.

Метод k-means є одним з найпопулярніших алгоритмів кластеризації. Алгоритм складається з наступних кроків.

Поки кластерні центри не стануть стійкими (не змінюватимуться).

Крок 1. Обирається кількість кластерів на які буде розбито вхідні дані.

Крок 2. Випадковим чином обираються центри майбутніх кластерів.

Крок 3. Кожен елемент приписується до кластеру, відстань до центру якого найменша.

Крок 4. Розраховується новий центр кластеру як елемент, ознаки якого є середньо-арифметичними серед всіх елементів кластеру.

Кінець поки.

Інший популярний варіант алгоритму k-means - це бісективний алгоритм K-means. Він отримує результати наступним чином.

Поки необхідна кількість кластерів не досягнута.

Крок 1. Обирається кластер для розбиття.

Крок 2. Запускається загальна версія алгоритму k-means, щоб розбити кластер на два підкластери. Цей крок називається бісектуванням.

Крок 3. Повторювати крок 2 доки не буде досягнена максимальна подібність елементів у кластері.

Кінец поки.

У результаті [10] зроблено висновок, що бісективний алгоритм розподілу k-means перевершує традиційний k-means з точки зору точності та ефективності.

Також вказано, що алгоритм k-means та бісективний алгоритм k- means працює краще, ніж агломерна ієрархічна кластеризація. Висвітлено один великий недолік агломераційного ієрархічного кластеризації. Стверджується, що помилки можуть відбутися на попередніх етапах, і ці помилки дуже погано впливають на весь подальший процес.

Для вирішення проблеми шуму було розроблено алгоритм k-medoids. Він схожий на алгоритм k-means, але він не приймає середнє значення для всіх елементів у кластері, щоб знайти центроїд. Алгоритм k-mediod вибирає один з елементів у вигляді точки для порівняння [11]. Це вважається обчислювальним дорогим і також неефективним для великих наборів даних.

Час обчислення стандартного алгоритму k-means збільшується з збільшенням кількості елементів у великих наборах даних. Таким чином, у роботі запропоновано варіант Mini-Batch k-means, який виконує менше обчислень для великих наборів даних, ніж традиційний алгоритм k-means. Він використовує міні-партії для зменшення часу, але використовує ту ж цільову функцію, що й алгоритм k-means. Mini-Batch - це невеликі випадкові вибірки з вхідних даних, обраних під час кожної ітерації, що значно скорочує час обчислення.

Основні кроки алгоритму.

Поки існують елементи що не відносяться до кластеру.

Крок 1. Береться частина вхідних даних, яка разом формує Mini-Batch.

Крок 2. Mini-Batch призначається найближчому центру кластеру, далі знаходиться новий центр кластеру.

Кінець Поки.

На рисунку 2.1 показана швидкість обробки з k = 3 і k = 10 у порівнянні традиційного алгоритму k-means з алгоритмом Mini-Batch k-means. Результати показують, що алгоритм Mini-Batch k-means виконується швидше і дає кращі результати навіть на великих наборах даних (див.рис. 2.1). І навпаки,

традиційний алгоритм k-means виконується повільніше на великих наборах даних.

Основні кроки алгоритму.

Поки існують елементи що не відносяться до кластеру.

Крок 1. Береться частина вхідних даних, яка разом формує Mini-Batch.

Крок 2. Mini-Batch призначається найближчому центру кластеру, далі знаходиться новий центр кластеру.

Кінець поки.

На рисунку 2.1 показана швидкість обробки з k = 3 і k = 10 у порівнянні традиційного алгоритму k-means з алгоритмом Mini-Batch k-means. Результати показують, що алгоритм Mini-Batch k-means виконується швидше і дає кращі результати навіть на великих наборах даних. І навпаки, традиційний алгоритм k-means виконується повільніше на великих наборах даних.

Час обчислення стандартного алгоритму k-means збільшується з збільшенням кількості елементів у великих наборах даних. Таким чином, у роботі запропоновано варіант Mini-Batch k-means, який виконує менше обчислень для великих наборів даних, ніж традиційний алгоритм k-means. Він використовує міні-партії для зменшення часу, але використовує ту ж цільову функцію, що й алгоритм k-means. Mini-Batch - це невеликі випадкові вибірки з вхідних даних, обраних під час кожної ітерації, що значно скорочує час обчислення.

Для обрахування схожості між документами зручно їх звести до векторної моделі, а далі шукати індекс схожості порівнюючи ці вектори. Нижче наведено способи порівняння векторних моделей.

Евклідова відстань рахується як сума квадрату різниці між координатами двух об’єктів [12]. Тому, Евклідова відстань *d* між двома *n*-розмірними векторами *Xi* та *Xj* рахується як:

$$d = \sqrt{\sum_{LK?}^{J} (\chi_{"L} - \chi_{*L})^{@}},$$

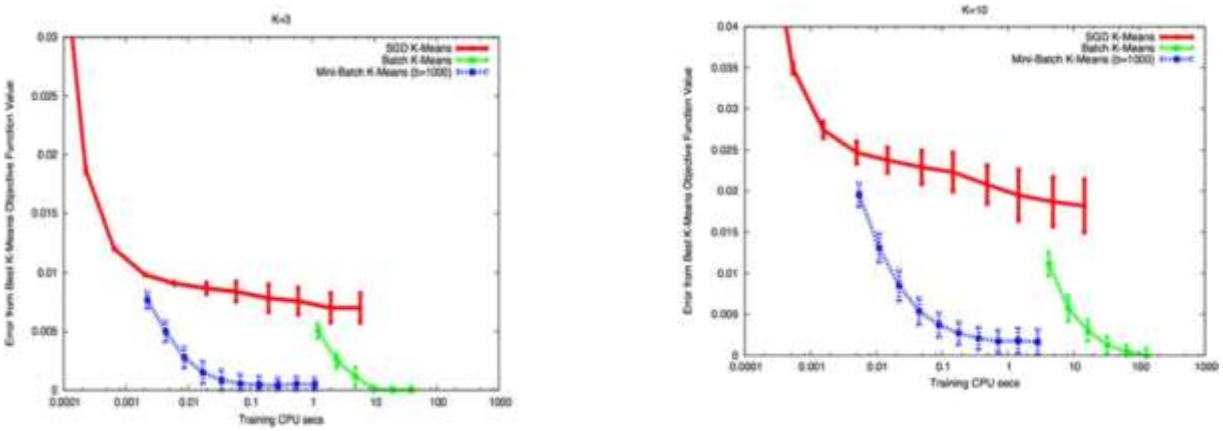

Рисунок 2.1 – Порівняння швидкодії алгоритмів

де $\chi_{"L}$ – *k*-тий елемент ектору X_i;

χ_{*L} – *k*-тий елемент вектору X_j.

Косинус подібності вимірюється як косинус кута між двома векторними моделями документів [13]:

$$\cos\theta = \frac{\chi_{"} \cdot \chi_{*}}{|\chi_{"}| \cdot |\chi_{*}|}.$$

Коефіцієнт Жаккарда вимірює відстань як перетин, поділений на об'єднання об'єктів. Він обчислює суму ваги загальних термів і порівнює його з сумою ваг термів, що зустрічаються в будь-якому з двох документів, але не є загальними [14].

$$jaccard(x, y) = |\chi_{"} \cap \chi_{*}| / |\chi_{"} \cup \chi_{*}|.$$

2.3 Аналіз даних, отриманих з соціальних мереж

Соціальні мережі з великою кількістю користувачів, такі як Facebook, Twitter, чи Youtube, створюють нові шляхи для спілкування людей, і створення

віртуальних спільнот. Результати в соціології та психології показують, що людські істоти схильні створювати зв'язки та ділитися інформацією з іншими людьми.

Можна побачити, як інформація з соціальних мереж може бути застосована для алгоритмів РС для підвищення точності рекомендації. Припускаємо, що користувачі є в соціальній мережі загального типу, такій як Facebook, або предметно-орієнтованій мережі, як Last.fm [17] для отримання рекомендацій щодо кіно, чи Epinions для широкого діапазону рекомендацій.

Зобразимо соціальну мережу як орієнтований граф G = (U, F), де U є набір користувачів з | U | = u0, а F - зв'язки дружби. Ця ж інформація представлена у вигляді як і наскільки користувач u довіряє або знає користувача v в соціальній мережі. Це можна зрозуміти з матриці S, розмірності u0 × u0, де показані коефіцієнти довіри між усіма користувачами з U. Кожен користувач u має набір Fu безпосередніх сусідів, що їм довіряє, і в той же час, u довіряють Fu - користувачів. Прямі соціальні відносини користувача u з користувачем v (наприклад користувач u знає користувача v) представлені позитивним значенням Su,v ∈ (0, 1]. Відсутні або приховані соціальні зв'язки - Su,v = sm, де, як правило, sm = 0. Соціальну вагу Su,v можна інтерпретувати як міру, що показує, наскільки користувач u довіряє користувачу v. Це може бути видно по явному зворотньому зв'язку користувача u та користувача v (наприклад, шляхом голосування), або з неявного зворотного зв'язку (наприклад, від ступеня взаємодії/зв'язку). Як правило, довіра між користувачами u та v - Su,v, невід'ємна. В особливих випадках, вона може приймати негативні значення, наприклад, якщо два користувача мають суперечливі смаки (див.рис. 2.2).

На цьому рисунку ілюстрована соціальна мережа з шістьома користувачами, кожен з яких має набір деякий друзів. Також можна зазначити, що кожна зв’язок між користувачами ми можемо назвати дружбою, також дружбу мі можемо відмітити як, позитивним значенням довіри. Це також можна побачити та відобразити у вигляді матриці S (див.табл 2.1).

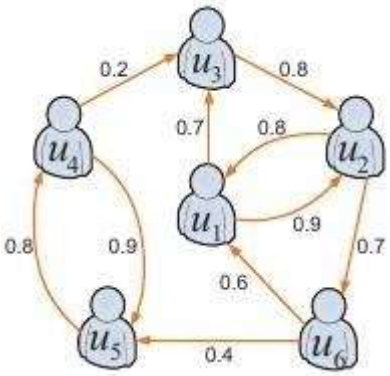

Рисунок 2.2 - Соціальний граф

Таблиця 2.1 – Довіра в соціальних мережах

	U1	U2	U3	U4	U5	U6
U1		0.9	0.7			
U2	0.8					0.7
U3		0.8				
U4			0.2		0.9	
U5				0.8		
U6	0.6				0.4	

Отже, існуючі алгоритми для РС пропонують рішення, що дозволяють використання даних з соціальних мереж:

– створення РС для соціальних мереж на основі Байесового виведення (Bayesian-inference). Це забезпечує хороший компроміс між захистом конфіденційності користувача і точності рекомендації;

– використання концепції «категорій друзів», для покращення точності рекомендацій;

Вивчення онлайн соціального голосування.

2.4 Методи колаборативної фільтрації

Колаборативна фільтрація (CF) є найпопулярнішим підходом до побудови рекомендаційних систем і успішно застосовується в багатьох додатках. Для рекомендації за допомогою CF використовують думки кількох користувачів, щоб передбачити інтерес іншого користувача[18]. В роботі запропонований імовірнісний підхід на основі CF до рекомендації оцінок для профіля нового користувача. За допомогою схеми активного навчання, профіль нового користувача може отримати достатньо даних для якісної рекомендації з мінімумом необхідного зусилля. В роботі розглядаються ключові рішення в оцінці спільних систем фільтрації рекомендувача. В роботі розглянуті ключові моменти щодо оцінки спільних інтересів користувачів враховуючи системи фільтрації користувачів за їх інтересами. В роботі розглядаються моделі розкладу матриць в методах з урахуванням CF. За допомогою прикладу Netflix-конкурсу, тут показано, що методи розкладання матриці стали одним з основних видів та методологій в аспекті рекомендації, за допомогою колаборативної фільтрації. Тут зазначено, що займалися рекомендаціями в соціальних мережах та розподіленою рекомендаційною системою.

З швидким зростанням онлайнових соціальних мереж, довірчий підхід CF став визначальним. В цільові основі CF підходів, оцінки користувачів, із соціальних мереж порівняні з оцінками «еталонних» користувачів, які також враховуються для рекомендацій.

Колаборативні методи аналізу сусідів вважаються основним та є найбільш ефективним з огляду на те, що при виявленні дуже локалізованих відносин між користувачами, зазначемо, що в той же час як CF на основі прихованих факторів - як правило, ефективні при використанні всієї рейтингової інформації користувача, як показано в роботі.

Позначимо число користувачів, як $u0$, а кількость елементів, як $i0$. Рейтинг Ru,i, який показує, на що користувач найчастіше звертає увагу при виборі товару, користувача позначемо, як u, де високі значення означають більш сильну перевагу. Розрізняючи передбачені оцінки із відомого числа, використовуючи позначення $\hat{R}u,i$, для прогнозованого значення Ru,i. Для методу, основаного на найближчих сусідах користувача:

$$\hat{R}_{u,i} = \bar{R}_u + \frac{\sum_{v \in N_u} sim(u,v)(R_{v,i} - \bar{R}_v)}{\sum_{v \in N_u} sim(u,v)},$$

де $\bar{R}_u$ є середня оцінка користувача u;

N_u - вибрані сусіди користувача u.

Використовується коефіцієнт кореляції Пірсона для вимірювання подібності двох користувачів. Коефіцієнт кореляції Пірсона між користувачем u і v визначається як:

$$sim(u,v) = \frac{\sum_{i \in I_C} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_C} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I_C} (R_{v,i} - \bar{R}_v)^2}},$$

де I_C набір загальних елементів у системі, оцінених u і v,

R_{ui} і R_{vi} їх рейтинги для виявлення товару який найбільш підійде користувачеві.

Назвемо цей метод KNN (K-Nearest Neighbor, К-найближчих сусідів).

Матричні моделі як користувачів у деякій системі, також і елементів (речі, подій) для визначення спільного простору розмірності j_0, будемо створювати таким чином, що взаємодії елемента (речі, події) користувача моделюються як внутрішні користувачі в просторі.

Відповідно, кожен елемент i пов'язаний з вектором $P_i \in \mathrm{R}^1 \times j^0$ і кожен користувач u пов'язаний з вектором $\mathrm{Q_u} \in \mathrm{R}^1 \times j^0{}_{ui}$. Скаляр $Q\,P^T$ описує взаємодію

між користувачем *u* та елементом *i* - загальний інтерес користувача в елементі, визначений по його характеристиках/

Передбачений рейтинг моделюється як:

$$R\hat{}_{u,i} = r_m + Q\,P^T,$$

де (u,i) - частини для яких $R_{u,i}$ є відомими (навчальний набір). Для запобігання перенавчання користувачів. Ми використовуємо норму Фробеніуса, яка позначається як $\|.\|F$, щоб упорядкувати вивчені деталі і користувальницькі приховані фактори. Треба звернути увагу, що зазвичай ряд особливостей $\min(u_0, i_0)$. В наших експериментах припускається j_0 = 10. Позначимо цю модель як SVD (Singular Value Decomposition).

Зокрема рекомендаційні системи на основі Байєсовського виведення і SVD – це два різні види систем. Байєсівське виведення [20] являє собою розподільчу систему, яка має змогу бути легко реалізована відштовхуючись від існуючих соціальних мереж. Користувачам не потрібно турбуватися про їхню оціночну конфіденційність, тому що вони діляться історією оцінок тільки з прямими друзями. У той час як в централізованих РС, користувацька конфеденційність є великою проблемою.

Тим часом, РС на основі Байєсового виведення використовує тільки локалізовану інформацію з огляду на те якою інформацією користувався користувач у своїх запросів пошуку у системі, також вона надає користувачеві дуже правдоподібний захист приватної інформації. Метод SVD є більш точним у прогнозуванні, що саме потрібно кінцевому користувачу, відштовхуючись від його потреб та рекомендує йому перелік товарів з огляду на його вподобання. Зокрема, в комерційних РС, користувачі дуже часто використовують функції оновлювання своїх профілів у соціальних мережах. РС що буде використовувати байєсові змінні, може дуже швидко інтегрувати нові дані в свії профіл або оновлювати свії профіль, якщо це буде доцільним для нього у

даному випадку. Але у той самий час, як SVD потребує перерахунку оцінок всіх характеристик користувачів та їх даних, якщо потрібно інтегрувати новостворені профілі.

Одна з головних ідея алгоритму колаборативной фільтрації, залежить від того, які само пропозиції було покладено за основу, але так само враховуючи нові елементи для обраного користувача з якойсь соціальної групи, зважаючи на основи попередніх запитів користувача або його однодумців. Інших вподобань користувачів або деяких інших користувачв з інших соціальних груп. На цей час, або краще сказати на сьогодняшний день дослідники розробили цілий ряд алгоритмів, які можна поділити на дві основні категорії.

Перший метод це метод який заснований на аналізі наявних оцінок, – анамнестичні методи (Memory-based). Ці алгоритми грунтуються на статистичних методах, щоб знайти групу користувачів близьких до цільового користувачеві. Цей підхід ще називають метод найближчих сусідів: використання попередніх оцінок, зроблених клієнтом, і аналіз оцінок інших користувачів, які мають подібні переваги. Тоді рекомендації (прогноз) для цільового користувача формуються на підставі обчислення якоїсь міри схожості по всьому накопиченим даними. Коли в наявності є дані великої кількості оцінок користувачів, вони можуть бути використані для визначення вподобань подібних користувачів (наприклад, користувачі, які преглянули електронні товари з одного сайту); користувачі з подібними смаками мають багато спільних покупок у своїх історіях у соціальних мережах, проте деякі електронні товари можуть відрізнятися – саме їх і можна рекомендувати користувачам, які мають такі ж смаки. Тоді рекомендації (прогноз) для цільового користувача формуються на підставі обчислення якоїсь міри схожості по всьому накопиченим даними [21].

Другий метод це метод який заснований на аналізі моделі даних, - модельні методи (Model-based). У цьому випадку спочатку за сукупністю оцінок формується описова модель переваг користувачів, товарів і

взаємозв'язку між ними, а потім формуються рекомендації на підставі отриманої моделі (див.рис. 2.2). Процес формування рекомендацій розбитий на два етапи: ресурсномістке навчання моделі в відкладеному режимі і досить просте обчислення рекомендацій на основі існуючої моделі в реальному часі. Ці алгоритми можуть бути засновані на імовірнісному підході, кластерному аналізі [22], аналізі прихованих чинників.

Методи, засновані на об'єднанні попередніх алгоритмів, - гібридні методи. Ці підходи в свою чергу можуть бути розбиті далі на групи методів. Так, методи на основі сусідства (близькості) поділяються на аналіз:

– подібності користувачів (User-based);
– подібності елементів (Item-based).

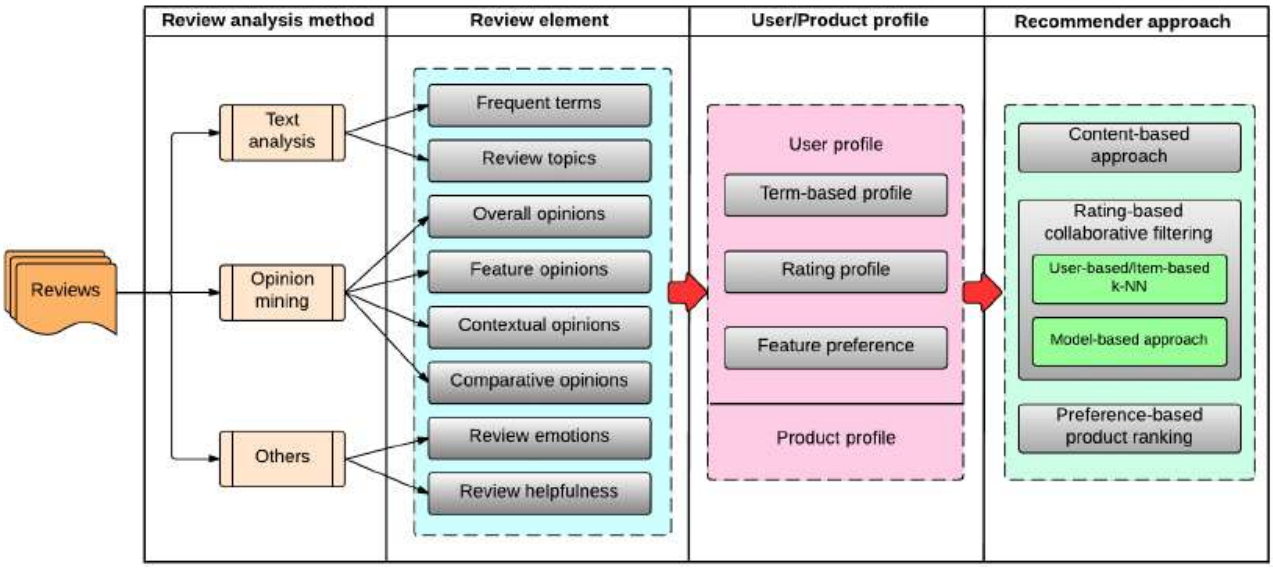

Рисунок 2.2 - Класифікація методів і алгоритмів коллаборативної фільтрації

Метою обох напрямків є виділення схожих об'єктів в групи на основі матриці оцінок. У першому випадку визначається схожість користувачів: знайти інших користувачів, чиї минулі оцінки поведінки схожі на ті, що і у поточного користувача, і використовувати їх оцінки інших елементів для прогнозування переваги поточного користувача. Другий підхід, на основі

подібності елементів, був запропанованний набагато раніше [23], і ця версія використовується в Amazon.com в даний час. У цьому випадку замість того щоб використовувати подобу між поведінкою користувальницьких оцінок для прогнозування переваги, використовується подібність між оцінками моделей елементів. Якщо два елементи, як правило, мають однакові оцінок, то вони схожі, і користувачі повинні мати аналогічні переваги для подібних елементів.

Для визначення подібності між користувачами або елементами можна використовувати такі підходи:

- відстань Евкліда, Хемминга;
- кореляція Пірсона;
- ранговая кореляція Спірмена;
- по коефіцієнт Жаккар;
- косуносное подобу.

Коллаборативна фільтрація на основі подібності користувачів (User- based) має високу точність. Однак, недоліком є ресурсомісткість (вимога до пам'яті) і складність (кількість обчислень, необхідну для отримання рекомендацій). До того ж обчислення ступеня близькості може проводитися тільки в реальному часі, так як дані про поточну транзакції стають доступними тільки в момент вироблення рекомендацій. Тому даний метод може застосовуватися тільки до відносно невеликих баз даних.

В алгоритмі на основі подібності елементів (Item-based) ступінь близькості аналізованого елемента до всіх інших може бути обчислена в відкладеному режимі за розкладом, так як вектора рейтингів всіх елементів доступні до моменту формування рекомендації. Таким чином цей алгоритм виявляється більш ефективним з точки зору часу формування рекомендацій завдяки можливості проведення відкладеної предобработки даних.

Для описаних вище методів є необхідність в зберіганні всієї матриці даних, тобто переваг користувачів про елементи. У зв'язку з цим виникають труднощі при прогнозі переваг для нових користувачів або при появі нових

елементів, тому що для них ще немає оцінок. Також обмежується можливість методів при обробці великих обсягів даних. У багатьох випадках зберігання всієї матриці переваг надлишково: як правило, користувачі і елементи діляться на групи з аналогічними профілями переваг. Наприклад, багато науково-фантастичні фільми будуть подобатися в аналогічній ступеня тим же набором користувачів. Тому виникає задача в зниженні розмірності матриці оцінок. Такі завдання вирішують методи другої групи (див.рис. 2.1).

В цьому випадку можливий варіант об'єднання користувачів (елементів) в кластери (профілі) за допомогою деякого індексу подібності. Елементи і оцінки, дані користувачами з одного кластера, використовуються моделі краще масштабуються, тому що звіряють профіль користувача з відносно невеликою кількістю сегментів, а не з цілої користувальницької базою. Складний і ємний кластерний підрахунок ведеться з оффлайн режимі. Це завдання може бути виконана на основі різних математичних підходів [24]. Саме тут розглядається використання методу на основі подібності елементів з нормалізацією даних.

Інформаційна область для систем колаборативної фільтрації складається з багатьох користувачів, які висловили переваги для тих чи інших товарів, які вони переглянули за деякий період часу. Перевага (оцінка) часто представляється у вигляді триплетів (користувач, товар, оцінка). Ці оцінки можуть приймати дуже різні форми, в залежності від системи в якій вони знаходяться. Деякі системи використовують речову або цілочислену оціночну шкалу, таку як від нуля до п’яти зірок, інші використовують бінарні від нуля до одного або потрійні заходи. Безліч всіх триплетів оцінок формує розріджену матрицю, яка називається матрицею оцінок. Пари (користувач, предмет), в яких користувачі не віддали перевагу предмету, є невідомими значеннями цієї матриці (див.табл. 2.2).

При використанні системи колаборативної фільтрації необхідно вирішити два завдання. По-перше спрогнозувати оцінку або перевагу, яку користувач віддасть обраному товару. Метою прогнозу є заповнення в матриці оцінок

відсутніх значень. По-друге надання рекомендацій, тобто формування ранжированного списку N елементів для даного користувача.

Таблиця 2.2 – Приклад матриці оцінок

	Елемент 1	Елемент 2	Елемент 3
Користувач 1	3	?	2
Користувач 2	?	4	3
Користувач 3	5	4	?

Визначимо математичні позначення для прив'язки різних елементів моделі рекомендаційних систем. Генеральна сукупність складається з набору користувачів U та набору елементів I. Перелік використовуваних змінних для формул:

- I_u – множина елементів, оцінених користувачем u;
- U_i множина користувачів, які оцінили елемент I;
- $r_{u,i}$ – оцінка користувача u для елемента i;
- r_i вектор усіх оцінок елемента I;
- d_u та d_i значення оцінок користувача u та елемента i відповідно;
- R– Рекомендаційний прогноз.

Перший крок: для кожного елемента j обчислюється міра близькості до елементу i. Для цього можна використовувати один із зазначених вище підходів, наприклад, коефіцієнт Пірсона:

$$s_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j))}{\sqrt{\sum_{u \in U}(r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U}(r_{u,j} - \bar{r}_j)^2}}$$

Другий крок: вибираємо множину елементів S, найбільш близьких до об'єкта i. У роботі [12] Савар визначив, що достатні результати виходять при k = 30 елементів безлічі S. Але ці дані залежні від розглядаємої задачі і розрідженості матриці.

Третій крок: передбачення рейтингу (оцінки) об'єкта на основі рейтингів близьких до нього об'єктів:

$$\hat{r}_{u,i} = \frac{\sum_{j \in S} s_{i,j} \cdot r_{u,j}}{\sum_{j \in S} | s_{i,j} |}$$

Даний алгоритм відображає теоретичну базу методу, але на практиці ряд факторів вимагає переосмислення розрахунків.

Як правило, переважна більшість оцінок невідомо, і розрідженість матриці оцінок досить висока. З іншого боку, дані, що вже є в матриці досить суб'єктивні. деякі користувачі – оптимісти, і їх оцінки завжди високі (середнє 4 з 5), інші користувачі – циніки, їх оцінки завжди занижені (середнє 2,5 з 5). Крім цього, завжди є елементи, які подобаються усім.

З метою боротьби з підгонкою розріджених даних з оцінками, проводиться регуляризація моделей таким чином, щоб скоротити ймовірність появи випадкових зв'язків між оцінками, які не відображають дійсність. Регуляризація контролюється константами, які позначаються як λ1, λ2, ..., λn. Точні значення цих констант визначаються перехресною перевіркою. По мірі їх зростання, регуляризація стає все важче.

Для того щоб оптимізувати продуктивність видачі рекомендацій, важливо нормалізувати оцінки до обчислення матриці подібності. Це може бути досягнуто шляхом обчислення базового прогнозу, в якому інкапсулюють відхилення користувача і елемента. Пари користувач-елемент (u, i) для яких

оцінки r_u, і відомі складають множину K. Базовий прогноз для невідомої оцінки ru, і позначається $b_{u,i}$ та визначається формулою:

$$b_{u,i} = \mu + b_u + b_i.$$

Загальна середня оцінка це сукупність параметрыв b_u і b_i - параметри, які показують спостерігається відхилення користувача u і елемента і відповідно від середнього значення.

Так як усі параметри взаємопов'язані, то розраховувати їх необхідно разом, вирішивши завдання найменших квадратів [23]. На основі представлених математичних вкладок були проведені дослідження з використання описаного методу. Для оцінки ефективності використовувалася середньоквадратичне відхилення.

Колаборативна фільтрація (CF) є найпопулярнішим підходом для побудови рекомендаційних систем і успішно застосовується в багатьох додатках. Для рекомендації за допомогою CF використовують думки кількох користувачів, щоб передбачити інтерес іншого користувача [24]. В роботі запропонований імовірнісний підхід на основі CF до рекомендації оцінок для профіля нового користувача. За допомогою схеми активного навчання, профіль нового користувача може отримати достатньо даних для якісної рекомендації з мінімумом необхідного зусилля. В роботі розглядаються ключові рішення в оцінці спільних систем фільтрації рекомендувача.

В роботі розглядаються моделі розкладу матриць в методах CF. За допомогою прикладу Netflix-конкурсу, автор показує, що методи розкладу матриці стали домінуючою методологією в рекомендаціях за допомогою колаборативної фільтрації. Автори робіт не займалися рекомендаціями в соціальних мережах і розподіленою рекомендаційною системою.

2.5 Рекомендаційні системи, що ґрунтуються на категоризації користувачів

Традиційні підходи колаборативної фільтрації передбачають інтереси користувачів за історією оцінок, які поставив користувач. Популярні соціальні мережі надають додаткову інформацію для покращення рекомендацій, в основі яких лежать рейтингові оцінки користувача. Деякі РС використовують інформацію з соціальних мереж для підвищення точності рекомендації. Загальна концепція – що вподобання користувача аналогічні вподобанням його друзів у соціальних мережах. Тим часом, життя користувачів у мережі та поза її межами, може суттєво відрізнятися. Багато соціальних мереж, на сьогодні мають функцію категоризації друзів. Найпершою мережею, що ввела таку фунцію була Google+, в якої це реалізовано в вигляді «кіл» користувачів. Користувач може розподіляти своїх друзів по різним колам, таким, як сім'я, одногрупники, колеги, друзі по рибалці, тощо. Facebook також запровадила схожу функцію, за допомогою якої користувач може сортувати своїх друзів за групами, і ділитися контентом з певним списком друзів. У Twitter, користувачі можуть організувати людей, на яких вони підписані (followees) у «списки». Коли користувач, переглядає певний список, він побачить твіти тільки людей, які є в цьому списку.

РС також мають вигравати від категорізації користувачів. Користувач довіряє різним людям з різних причин. Наприклад, в контексті рекомендацій в різних категоріях, користувач може довіряти іншому користувачу у всьому, що стосується автомобілів, але не довіряти в тому, що стосується наприклад смартфонів. Очевидно, що в такому разі, при рекомендації першому користувачеві в категорії автомобілів необхідно враховувати оцінки, які поставив другий користувач в цій категорії, і не враховувати оцінки цього користувача при рекомендаціїї, що стосується фільмів. В ідеальному випадку,

до рекомендації, відомо, які «кола» необхідно враховувати під час рекомендаціїї, а які ні. Нажаль, в більшості випадків (як і з доступними датасетами, так і у випадку реального використання), категорії не завжди чітко розділені між собою. Тому, якщо використовувати тільки одну категорію для передбачення, то можна пропустити важливу інформацію, що міститься в іншій категорії, хоча на перший погляд, інша категорія не повинна мати відношення до даної рекомендації. Тим не менше, навіть змішані категорії можуть допомогти при рекомендації, тому що таким категорія як «сім'я» чи «близькі друзі», людина схильна довіряти в багатьох випадках (див.рис. 2.3).

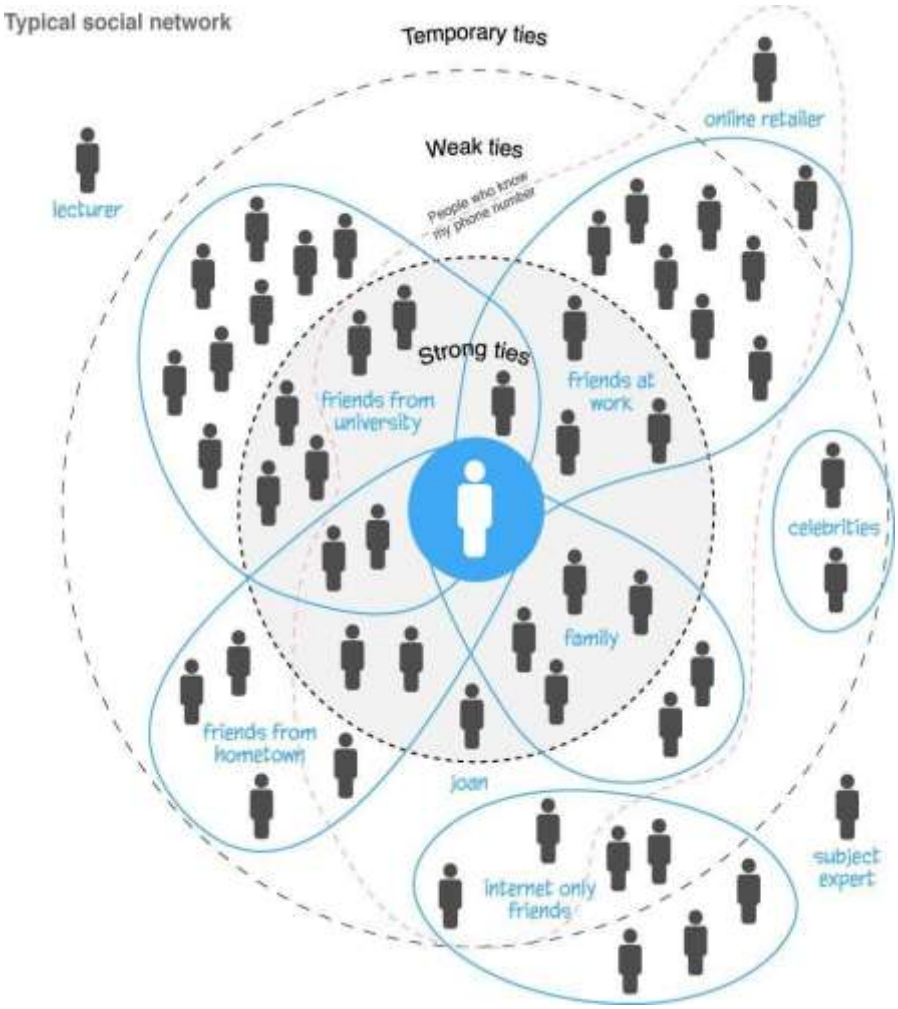

Рисунок 2.3 – Розподіл друзів за категоріями в соціальній мережі

Також доцільно було б зазначити, що ця моделі, в основу якій покладено розкладання низькорангових матриць, можна вважати одними з найкращих для колаборативної фільтрації користувачів, так як у них найнижчий RMSE. Нижче наведено стислий огляд таких моделей даних. Використання даних з соціальних мереж було доцільно використовувати з огляду на те, що для підвищення точності рекомендування користувачам прямих пропозиції продаж. Також дві різні моделі для інтеграції у джерела даних були запропоновані, такі як Social Recommendation (SoRec), Social Trust Ensemble (STE), Recommender Systems with Social Regularization, Adaptive social similarities for recommender systems. Для досягнення особливо низького значення RMSE була запропонована модель SocialMF та була використана як базова модель.

Моделі рекомендацій на основі категорізації користувачів. Перша розглянута модель - розширена модель SocialMF[26] з включенням категорій користувачів.

Гіпотеза заключається в тому, що користувач довіряє різним друзям в оцінці предметів, що можуть бути розділені на різні категорії, і що користувач може довіряти кожному другу стосовно тільки однієї категорії, а не багатьох. Наприклад коло друзів, яким довіряють стосовно автомобілів, може бути зовсім інше, ніж те, кому довіряють в виборі товару.

Умова1: однакова довіра

Найпростіший варіант визначення рівня довіри $S_{u,v} > 0$ в категорії друзів користувача u в категорії C: кожен користувач v в “колі” користувача u отримує однарове значення довіри, тобто $S_{u,v(c)*} = \text{const if } v \in C_u(c)$

Умова2: експертна оцінка

Найбільший рівень довіри надається користувачу, у якого найбільше досвіду в даній категорії. Як рівень його досвіду використовується кількість його оцінок в даній категорії. Суть полягає в тому, що більш досвідчений користувач виставляє більше оцінок в даній категорії.

2.6 Висновки по розділу 2

В другому розділі проведено критичний огляд й аналіз методів аналізу великих масивів даних користувачів соціальних мереж. Розроблено метод надання рекомендацій, щодо придбання товарів в області електронної комерції, користувачеві на основі аналізу його профілю в соціальних мережах.

3 ПРОЕКТУВАННЯ ТА РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

3.1 Побудова алгоритму

Завдання даної роботи передбачають створення методу та системи для проведення класифікації користувачів соціальних мереж згідно їх уподобань . У розд. 2.5 описано метод колаборативної фільтрації та математична модель класифікації користувачів. В даному розділі буде описана технічна реалізація системи на основі сформованих теоретичних пропозицій. Буде описана архітектура технологічного рішення, описана функціональність основних блоків, зазначено, які теоретичні рішення лежать в основі кожного елемента. У загальному випадку, розробляєма система вирішує наступні завдання:

- сбір даних користувачів з їх профілів;
- за основу взята задача класифікації користувачів по п’яти класам;
- результати класифікації, містять рішення з огляду на користувачів.

3.2 Архітектура моделі системи

Виходячи з поставлених перед технічною системою завдань, логічно уявити систему як сукупність двох блоків.

Перший блок це блок збору інформації про користувачів та видача результатів згідно їх вподобань. Як було описано раніше про колаборативну фільтрацію система основана на збору інформації являє собою сукупність розподілених агентів, вбудованих в кожну цільову веб-сторінки. В одного боку це вирішує проблеми різних форматів даних, з іншого боку дозволяє проводити

первинну обробку даних уже на клієнтському пристрої, знижуючи обчислювальну навантаження на блок аналізу.

Другий блок це блок аналізу даних які взяті з профілів користувачів Facebook. Також, в межах даного блоку проводитися зберігання інформації, одержуваної від кожного блоку збору даних для подальшого аналізу (рис 3.1).

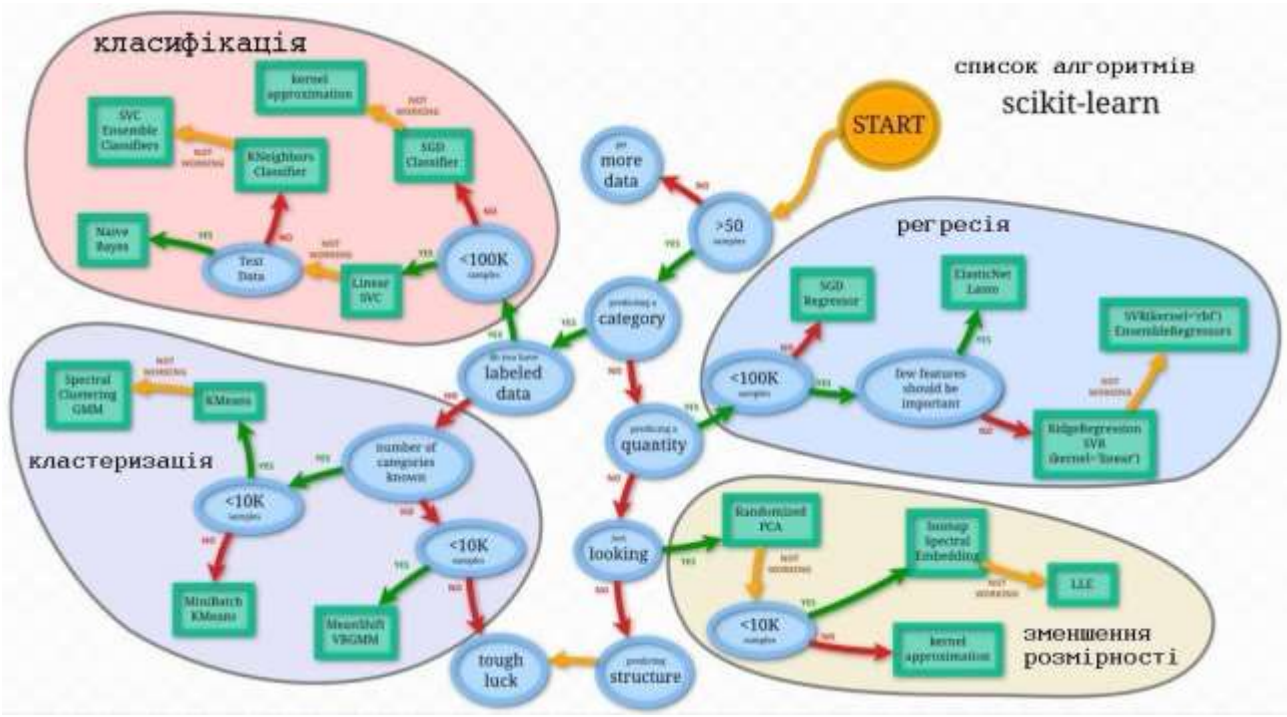

Рисунок 3.1 – Класифікація реалізованих методів бібліотеки scikit-learn

Отже, відштовхуючись від поставлених задач, при реалізації програмного забезпечення, було вирішити, що доцільно розгорнути середовище керування проектом (Redmine) репозиторій (Bitbucket Git) для керування версіями, Hadoop, на кластері, що складався з 4х нод для керування даними і використовувати Python як мову для написання програмної частини.

Блок збору даних архітектурно являє собою розподілену систему клієнтів (агентів), вбудованих в кожну цільову веб-сторінку. Під вбудовую розуміється факт того, що будь-який агент може перехоплювати запити користувачів до

даної сторінки, отримувати необхідні для аналізу дані, такі як: IPадреса, з якого відбувається запит на сторінку, ідентифікатор користувача, тип запиту, дані запиту і ін. Агенти повинні виконувати такі дії:

- IP адрес користувача, який відправив запит;
- ідентифікатор користувача в рамках системи;
- адреса запитаної веб-сторінки;
- типи запитів GET та POST;
- коди відповідей на запроси 200 чи 400.

На діаграмі компонентів зображено складові частини системи та способи їх взаємодії між собою (рис. 3.2).

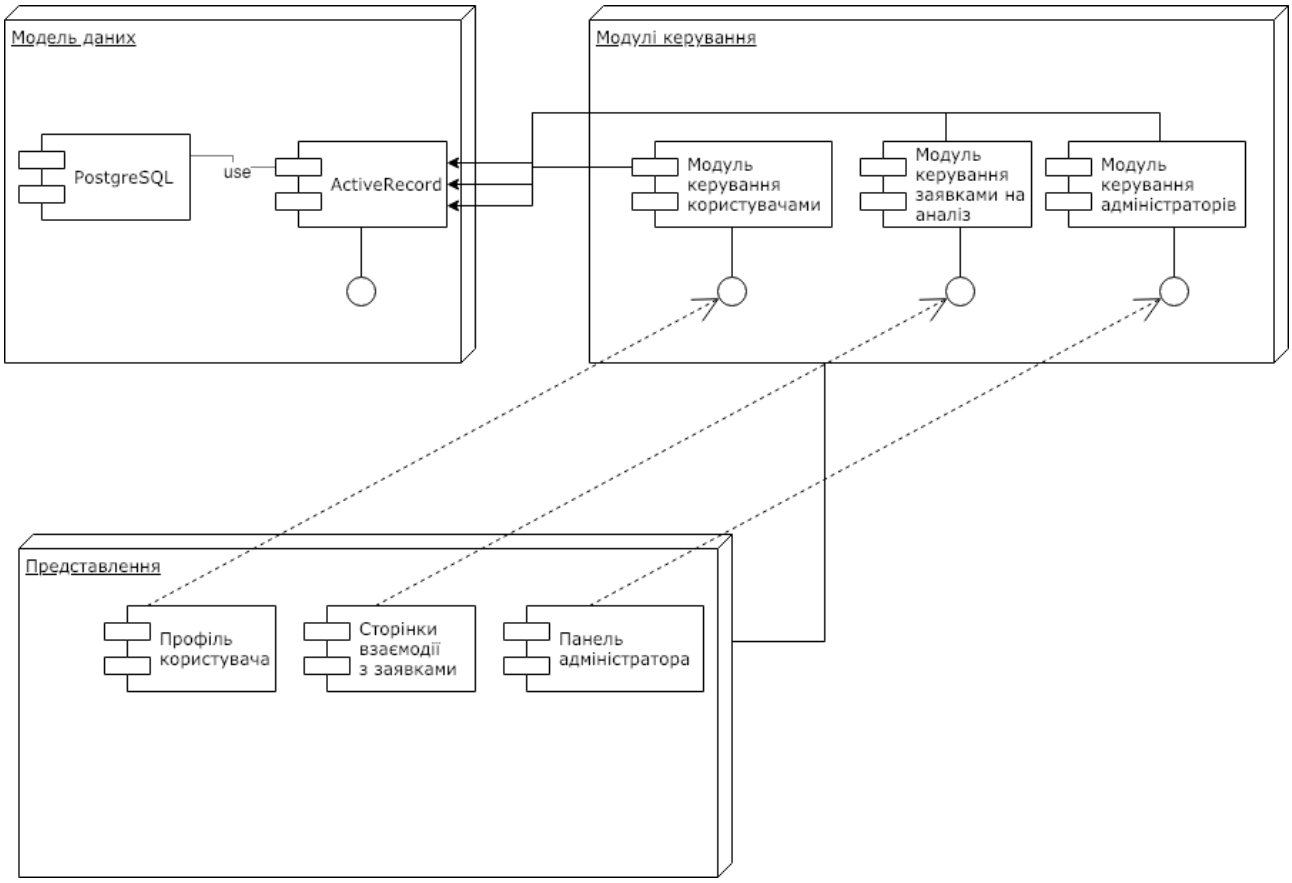

Рисунок 3.2 – Діаграма компонентів системи

3.3 Реалізація методу групування користувачів та надання їм рекомендацій

Для реалізації алгоритму беремо дані користувачів з їх акаунтів з facebook, виконуючи коллаборативної фільтрації для обробки описаних вище наборів даних. За основу бралася модель socialmf. Для кожного профілю користувача, який входить в якусь категорію, отримується окремий профіль і окремий профіль для кожної речі, яка відноситься в категорію. Це модифікована модель socialmf, в якій використовується інша функція для врахування соціальних категорій друзів.

Наведемо програмну реалізацію алгоритму, щоб брати інформацію з акантів користувачів:

```
Class foldcrCrossvValidation:
        def visitsPerSeason(self):
                handle = self.dbHandle()
  yValues = []
  xValues = []
  visitedDates = {}
                seasons = {}
                for record in handle.find():
  dateVal = datetime.strptime(record['date'],
'%d/%m/%y')
  timeVal = int(record['time'].split(':')[0])
                        month = dateVal.month
                        day = dateVal.day
                        if timeVal>= 11 and dateVal.weekday() in [3, 4, 5] and
timeVal<= 23:
if (3, 21) < (month, day) < (6, 20): visitedAP[record['ap_id']].setdefault('SPRING', 0)
visitedAP[record['ap_id']]['SPRING'] += elif (6, 21) < (month, day) < (9, 22):
visitedAP[record['ap_id']].setdefault('SUMMER', 0) visitedAP[record['ap_id']]['SUMMER'] += elif
```

```
(9, 23) < (month, day) < (12, 21): visitedAP[record['ap_id']].setdefault('FALL', 0)
visitedAP[record['ap_id']]['FALL'] += *
                elif ((12, 21) < (month, day) < (12, 31)) or ((1, 1)
< (month, day) < (3, 20)):
                visitedAP[record['ap_id']].setdefault('WINTER', 0)
                visitedAP[record['ap_id']]['WINTER'] += elif timeVal >= 11 and timeVal<= 22:
                                        if (3, 21) < (month, day) < (6, 20):
                visitedAP[record['ap_id']].setdefault('SPRING', 0)
                visitedAP[record['ap_id']]['SPRING'] += 1 with
open('data/VisitsPerSeasonDist.json', 'w') as outfile: json.dump(visitedAP, outfile, indent=4,
sort_keys=True)
                print('New File data/VisitsPerSeasonDist.json saved')
```

Для датасету EPINIONS використовувалися такі вагові коефіцієнти: останні “лайки” користувача, його вподобання (з профілю), вік, час коли користувач поставив лайк на якийсь товар унікальний та ідентифікатор.

Для датасету KAGGLE використовувалися вагові коефіцієнти: місцезнаходження користувача, його “лайки”, вподобання вік, час коли користувач поставив лайк на якийсь товар та унікальний ідентифікатор.

Для експерименту був використаний 10-разовий клас валідації (10-fold cross validation). Кожен раз 80% даних використовувалось для тренування моделі, а 20% що залишилися, використовувалися як тестові дані. Для оцінки точності рекомендацій використовувалась середньоквадратична похибка (root mean square error) і медіана абсолютної похибки (mean absolute rrror, так як це найбільш популярні метрики для оцінювання якості статистичних моделей. Схема роботи алгоритму (див. рис.3.3).

Зокрема, розглянуті методи класифікації існують в вигляді реалізації на мові Python, до того ж вихідні коди відкриті і доступні для модифікації – це дозволяє реалізувати новий запропонований метод колаборативної фільтрації користувачів за їх інтересами. Модель класифікації працює безпосередньо на

зібраних даних. Такий механізм дозволяє поліпшувати якість системи, ґрунтуючись на нових зібраних даних.

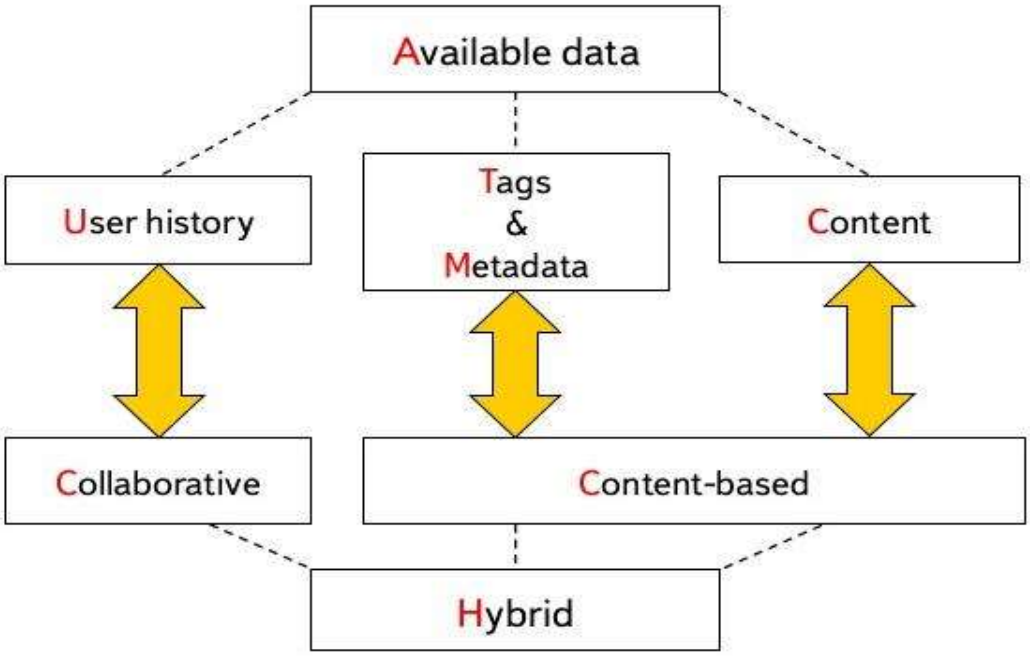

Рисунок 3.3 – Схема роботи методу колаборативної фільтрації користувачів за їх інтересами

Як інструмент зберігання даних, як уже було сказано раніше, підхід інтелектуального аналізу даних рекомендує індексовані бази даних. Використання таких технологій дозволяє істотно зменшити швидкість доступу до даних, що критично при аналізу великих обсягів інформації. В якості технологічного рішення для реалізації системи використовується графова база даних apache hbase. У цьому сховищі дані зберігаються у вигляді розрідженій матриці, рядки і стовпці якої використовуються як ключі. типовим застосуванням цього виду СУБД є завдання, пов'язані з великими даними, з зниженими вимогами до узгодженості даних. Для реалізації моделей і алгоритмів класифікації найкраще підходить мову програмування Python –

високорівнева мова програмування загального призначення. Python підтримує широкий спектр статистичних та чисельних методів і має гарну розширюваністю за допомогою пакетів.

3.4 Тестування та аналіз роботи розробленого методу

Є тільки два основні варіанти рекомендаційних систем: орієнтовані на аналіз контенту і на колоборативну фільтрацію.

Основна ідея підходу на основі аналізу контенту - використання властивості елемента для передбачення інтересу користувача до нього. Наприклад, для книги, можна використовувати ім'я автора, жанр, ключові слова і мітки. Ці властивості потім можна буде співставляти з вподобаннями цільових користувачів різних соціальних груп.

В роботі з РС за основу взято було колаборативну фільтрацію і тому це дуже багато уваги приділялось в літературі і було зосереджено на використанні даних явного оцінювання, у меншій мірі рейтингів (наприклад, за шкалою від однієї до п'яти зірок), які присвоювалися товарам чи групі товарів (наприклад, комп’ютери, смартфони, одяг). Завдання в тому, щоб передбачити значення інших елементів, рейтинг яких поставив користувач і точність рекомендації, як правило, вимірюється в термінах кореня середньоквадратичної похибки (Root Mean Square Error, RMSE), та середньої абсолютної похибки (Mean Absolute Error, MAE). Згідно з основною концепцією колаборативної фільтрації, були запропоновані різні види алгоритмів "найближчих сусідів" (k nearest neighbours), які можна буде використовуватиі в моделях користувач-користувач чи продукт-продукт окремо чи комбіновано. Одним з найкращих методів себе показав, метод розкладу матриці. Цей підхід добре себе зарекомендував у комп'ютерному зорі та аналізі тексту.

Саме базове розкладання матриці, що використовується – сингулярне розкладання, також були розроблені численні більш складні підходи. Основна ідея полягає у відображенні користувачів і речей в низькорозмірних просторах, і в них визначення подібності. Використання неявних даних отримало значно менше уваги в літературі. Відомі публікації в цій області, спрямовані на рекомендації телепередач користувачам на основі їх попередньої історії перегляду, наприклад, як багато часу вони витратили на кожний вид телевізійних програм.

Для проведення експерименту по перевірці гіпотези використання соціальних мереж в РС, висунутої в розділі 2, необхідні тренувальні та тестові дані. Для цієї цілі було вибрано відкриті набори даних, що є в вільному доступі. Перший з цих наборів — анонімізований набір даних користувачів з соціальних мереж, набір різних подій, що могли зацікавити користувача, і відношення користувачів до подій (у вигляді зацікавлений/незацікавлений). В даному наборі також є тестовий набір користувачів, для яких треба передбачити їх зацікавленість подіями.

Другий набір даних — датасет Epinions. Це рейтинги, що зібрані з оцінок фільмів, автомобілів та інших речей, користувачами сайту Epinions. Детальніше обидва набори даних описані нижче.

3.4.1 Набір даних Kaggle Event Recommendation Engine Challenge

Дані взяті для тестування з відкритого датасету, який був представлений на Event Recommendation Engine Challenge на порталі Kaggle.

Датасет складається з шести файлів: train.csv, test.csv, users.csv, user_friends.csv, events.csv і event_attendees.csv

train.csv має шість стовпчиків: user, event, invited, timestamp, interested, і not_interested.

test.csv містить ті ж стовпчики, як train.csv, крім interested, і not_interested.Кожен рядок файла відповідає певній події, перелік цих події:

– event – це id, що визначає подію;user – id, що визначає користувача;

– invited – бінарна змінна, чи був запрошений користувач на подію, чи ні;

– timestamp – це час в форматі ISO-8601 UTC, який показує приблизний час;

– interested – змінна, що показує чи користувач зацікавився даним товаром;

– not_interested – змінна, що показує, що користувачу не сподобався товар;

– user_id – це ідентифікатор користувача в системі;

– locale – рядок, що представляє мову користувача;

– birthyear – число, що представляє рік народження користувача

– gender – стать користувача;

– location – змінна, в якій зберігається місцезнаходження користувача;

– like – змінна, де зберігається товар користувача, що сподобався.

Epinions – це сайт загального споживчого огляду, створений в 1999 році. т, де збираються думки споживачів, на цьому сайті користувачі мають змогу рецензувати і оцінювати деякі речі, в тому числі фільми, автомобілі, книги, програмне забезпечення тощо. Набір даних складається з 356284 користувачів, 632548. Відгуки рейтингів та 1235897 різних предметів, для яких складено рейтинги. Рейтинги визначені за п'ятизіркового шкалою, де одна і дві зірки представляють негативні рейтинги, три зірки представляють рейтинги амбівалентності і чотири, п'ять зірок представляють позитивні рейтинг. Користувачі також висловлюють довіру до набору користувачів, чиї огляди та рейтинги виявилися цінним. Загальна кількість оцінок довіри 265849. Користувач отримує рекомендацію, тільки якщо вона пов'язана з іншими

користувачами через довірчі відносини. Після фільтрації ізольованих користувачів, набір даних має 52648 користувачів і 2658974рейтингів.

Кожен користувач має такі поля, що його характеризують: user_id, location, gender, birthyear, likes, що ідентичні за заначенням полям датасету Kaggle.

3.4.2 Профіль користувача

Ці набори даних, що описані вище, були вибрані, так як вони практично ідентичні з даними, які можна отримати про користувача через API різних соціальних мереж. Далі розглянуто API Facebook, щоб показати, що доступні через нього дані практично такі самі, як і в дата сетах з Kaggle або Epinions. Тобто, можна вважати що вибрані дата сети даних моделюють справжніх користувачів, які мали б змогу у майбутньому користуватися такою системою.

За допомогою API Facebook, про користувача можна зібрати різні дані: ім'я, вік, стать, мову якої користувач спілкується у Facebook, електронну пошту, віросповідання, день народження, його локацію, де користувач знаходиться наразі, його освіту та вподобання, які заповнив користувач у своєму профілі.

До вподобань користувача можна віднести такі категорії, як вподобані товари, фільми, автомобілі, смартфони тощо. Ці поля можуть бути як заповнені, так і можуть остатися порожніми. Саме тому при моделюванні профілю користувача доцільно структуризувати дані з його профілю, бо одні поля можуть бути як заповненими, так і можуть остатися порожніми, тобто відсутня цілісність даних.

Також, за допомогою API Facebook, можна дізнатися яким товарам користувач ставив “like”. Ці товари можуть бути дуже різноманітними, це також дає змогу визначати вподобання користувача. При завантаженні веб-сторінки перш за все потрібно перевірити, чи не увійшов чи вже людина на неї, використовуючи вхід через Facebook. Виклик FB.getLoginStatus відправляє

виклик до Facebook для отримання статусу входу. Потім Facebook відправляє отримані дані вашої функції зворотного виклику (див рис. 3.4).

Таким чином, в ході дослідження гіпотеза перевіряється на 2х різних наборах даних, що дозволить більш вірніше дослідити, як дані з профілю користувача можуть впливати на точність рекомендацій для даних з рейтінговим оцінюванням або без нього.

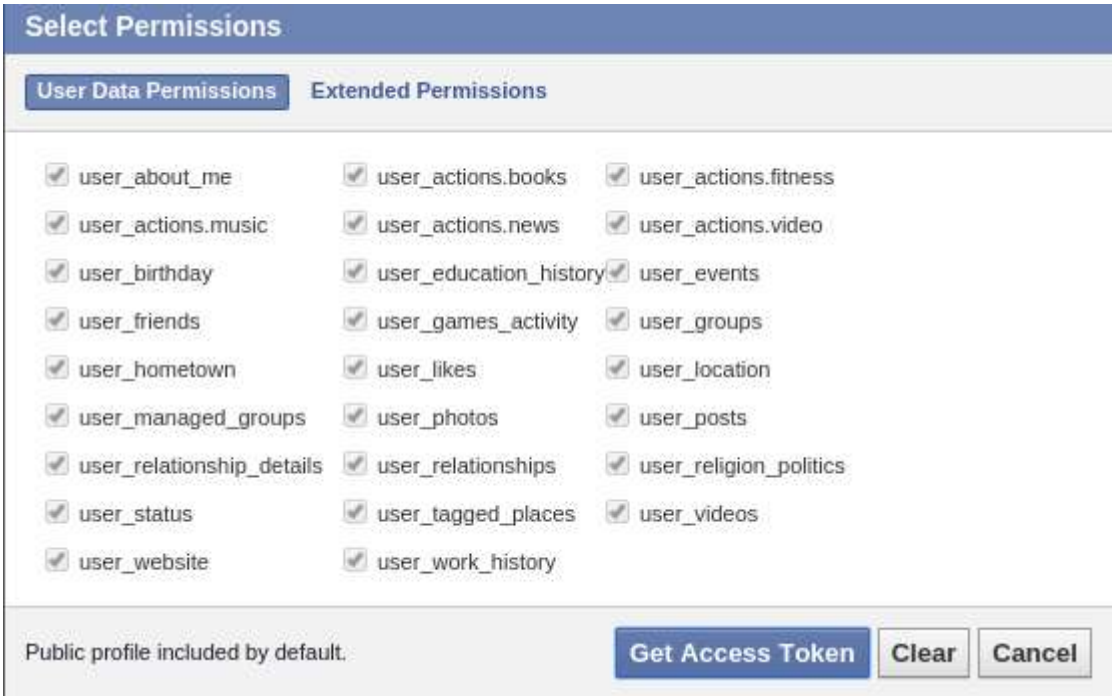

Рисунок 3.4 – Дані, які можна отримати за допомогою Facebook API

3.4.3 Результати виконання методу

Наведемо програмну реалізацію пошуку користувачів за інтересами та пропозицій їм продаж

```
def __init__(self, db_path, db_name,
collection_name):
    self.db_path = db_path
```

```
                    self.db_name = db_name
                    self.collection_name = collection_name
                self.my_client = Nonedef dbHandle(self):
                        if self.my_client is None:
                        self.my_client =
MongoClient(self.db_path)
                    my_db = self.my_client[self.db_name] collection =
                    my_db[self.collection_name] return collection
                def visitsPerDay(self):
                    handle = self.dbHandle()
                    yValues = []
                    xValues = []
                    visitedDates = {}
                    idTerm = ""
                        for record in handle.find():
                        dateVal =
            datetime.strptime(record['date'], '%d/%m/%y')
                        timeVal =
int(record['time'].split(':')[0])
                    if timeVal >= 11 and dateVal.weekday() in [3, 4, 5]
```

На рисунку нижче (див. рис 3.5) наведенні результати для дтасету Kaggle. В таблиці, де наводяться дані які вже тестовані (див. табл. 3.1).

З приведених результатів ми можемо побачити, що запропонований метод використання даних з профілю користувача та категоризації друзів очікувано зменшивши похибку рекомендації (RMSE). Також з результатів видно, що рекомендації більш точні на датасеті Kaggle, що і потрібно було довести, враховуючи велике значення вагового коефіцієнта місця знаходження користувача.

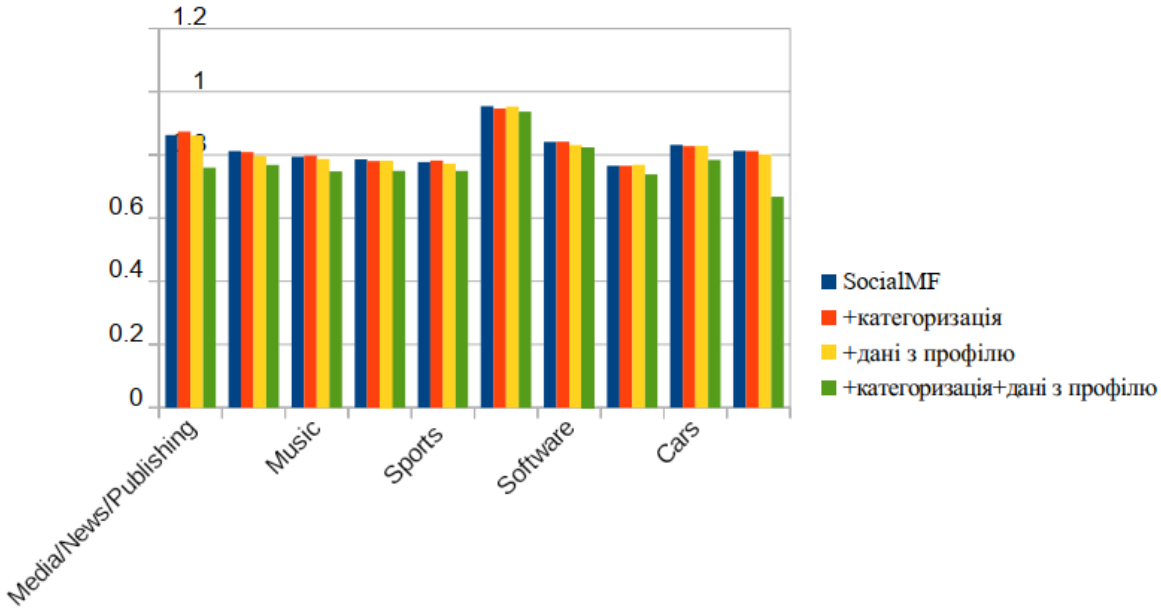

Рисунок 3.5 – Результати роботи методу на датасеті Kaggle

Таблиця 3.1 – Результати тестування на датасеті Epinions

Category	SocialMF	+категоризація	+дані з профілю	+категоризація+дані з профілю
Media/News/P	0.87	1	1.01	0.96
Films	1.034	1.07	1.06	1.045
Music	0.876	0.625	0.973	0.953
Books	0.568	0.568	0.767	0.735
Sports	2	0.485	0.988	0.964
Electronics	1.034	1.026	1.061	1.056
Pets	0.876	0.759	0.763	0.75
Cars	1.087	1.45	1.073	1.03
Games	1.056	1.068	1.042	1.036

3.5 Висновки по розділу 3

В третьому розділі проведено проектування та розробку прототипу програмного забезпечення для пошуку користувачів соціальних мереж зі схожими інтересами та надання їм рекомендацій в області електронної комерції.

Проведено тестування розробленого прототипу програмного забезпечення. Зроблено аналіз запропонованого методу використання даних з профілю користувача та категоризації друзів очікувано зменшивши похибку рекомендації. Також з результатів видно, що рекомендації більш точні на датасеті Kaggle, що і потрібно було довести, враховуючи велике значення вагового коефіцієнта місця знаходження користувача.

ВИСНОВКИ

В ході виконання дипломного проєкту була проаналізована предметна галузь в області методів роботи з великими масивами даних в соціальних мережах. Проаналізовано існуючі рекомендаційні системи та механізми їх роботи в області електронної комерції. Можна сказати що існуючі алгоритми та програми обробки Big Data усі базуються на різних платформах та дозволяють контролювати етапи введення даних, збирати статистику про користувачів соціальних мереж та підбирати для них доцільні пропозиції.

Проведено аналіз отриманих результатів, а саме порівняння продуктивності. Наведено моделі та методи, які використовуються у роботі. Зроблено порівняльний аналіз методів кластеризації та колаборативної фільтрації. Наведено опис методів вимірювання схожості документів між собою. Найкращим методом для роботи з великими даними для того щоб робити прямі пропозиції користувачам соціальних мереж зважаючи на їх вподобання є удосконалений метод колаборативної фільтрації, який дає змогу фільтрувати користувачів за групами по інтересам та надавати пропозиції згідно їх вподобань. Результатом роботи є данні переваг та недоліків використання прямих пропозицій електронних продаж. Розроблено програму, що знаходить користувачів за схожими інтересами у соціальних мережах та на основі цього використовуються прямі пропозиції продаж.

ПЕРЕЛІК ПОСИЛАНЬ

1. Karras T. A Style-Based Generator Architecture for Generative Adversarial Networks / T. Karras, S. Laine, T. Aila // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 4401-4410.

2. Gatys L. A. A Neural Algorithm for Big Data / L. A. Gatys, A. S. Ecker, M. Bethge [Electronic resource] // arXiv e-prints, arXiv:1508.06576v2 – 2015. – P. 1.

3. Image-to-Image Translation with Conditional Adversarial Networks / P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2017. – P. 1125-1134.

4. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks / J.-Y. Zhu, T. Park T., P. Isola, A. A. Efros // IEEE International Conference on Computer Vision (ICCV) – 2017. – P. 2223-2232.

5. Perarnau G. Fantastic GANs and where to find them [Electronic resource] / G. Perarnau // Gium Perarnau Blog – 2017. – Regime of access: http://guimperarnau.com/blog/2017/03/Fantastic-GANs-and-where-to-find-them/

6. Brock A. Large Scale GAN Training for High Fidelity Natural Image Synthesis / A. Brock, J. Donahue, K. Simonyan [Electronic resource] // arXiv e-prints, arXiv:1809.11096v2 – 2019. – P. 8.

7. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks / T. Xu, P. Zhang, Q. Huang, et al. // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2018. – P. 1316-1324.

8. Shafkat I. Intuitively Understanding Variational Autoencoders [Electronic resource] / I. Shafkat // Towards Data Science – 2018. – Regime of access: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf/

9. Semantic Image Synthesis with Spatially-Adaptive Normalization / T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 2337-2346.

10. Ouaknine A. Review of Deep Learning Algorithms for Image Semantic Segmentation [Electronic resource] / A. Ouaknine // Medium – 2018. – Regime of access: https://medium.com/@arthur_ouaknine/review-of-deep-learning-algorithms-for-image-semantic-segmentation-509a600f7b57/

11. Shaham T. R., Dekel T., Michaeli T. SinGAN: Learning a Generative Model from a Single Natural Image // IEEE International Conference on Computer Vision (ICCV) – 2019. – P. 4570-4580.

12. Automatic Keyword Extraction from Individual Documents / S. Rose, D. Engel, N. Cramer, W. Cowley // M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications – John Wiley & Sons, Hoboken, 2010. – P. 1-20.

13. Ondenyi E. Extractive Text Summarization Techniques With sumy [Electronic resource] / E. Ondenyi // Medium. – 2017. – Regime of access: https://link.medium.com/0SKJXLxGTNaskar A. Extract Custom Keywords using NLTK POS tagger in python sumy [Electronic resource] / A. Naskar // ThinkInfi. – 2018. – Regime of access: https://www.thinkinfi.com/2018/10/extract-custom-entity-using-nltk-pos.html

14. WordNet [Electronic resource] // Princeton University – Regime of access: https://wordnet.princeton.edu/

15. Pose Guided Person Image Generation / L. Ma, X. Jia, Q. Sun, et al. // Advances in Neural Information Processing Systems. – 2017. – P. 405-415.

16. COCO [Electronic resource] – Regime of access: http://cocodataset.org/

17. ImageNet [Electronic resource] – Regime of access: https://image-net.org/

18. PyTorch [Electronic resource] – Regime of access: https://pytorch.org/

19. Multi-Rake [Electronic resource] // The Python Package Index – Regime of access: https://pypi.org/project/multi-rake/

21. Mertens S. A complete anytime algorithm for balanced number partitioning // CoRR. 1999. Vol. cs.DS/9903011.

22. Zulawinski B. W., Iii W. F. P., Goodman E. D. The grouping genetic algorithm (gga) applied to the bin balancing problem.

23. Public Data Sets. URL: http://aws.amazon.com/publicdatasets/.

24. Tf-idf weighting. URL: http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html.

25. GanttProject. URL: http://www.ganttproject.biz/

26. Фаулер Рефакторинг. Улучшение существующего кода. 2010. Pp. 352 – 353.