Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Центр післядипломної освіти_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження методів фільтрації даних в соціальних мережах

(тема)

Виконав: студент 2 курсу, групи ІПЗмзд-18-1__
спеціальності _121- Інженерія програмного забезпечення_____
(код і повна назва спеціальності)
освітньо-наукової програми _Інженерія програмного_забезпечення_ _____
(повна назва освітньої програми)
_____ Забіяка Я.І._____
(прізвище, ініціали)
Керівник _____проф. Шостак І.В._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти_____

Кафедра програмної інженерії_____

Рівень вищої освіти _____другий (магістерський)_____

Спеціальність _121 – Інженерія програмного забезпечення_____
(код і повна назва)

Освітньо-наукова програма Інженерія програмного забезпечення _____
(повна назва)

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

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові _Забіяці Яні Ігоровні_____
(прізвище, ім'я, по батькові)

1. Тема роботи _Дослідження методів фільтрації даних в соціальних мережах

затверджена наказом по університету від _«__»_____2020 р № __-__

2. Термін подання студентом роботи до екзаменаційної комісії «21» травня 2020 р.

3. Вихідні дані до роботи *Алгоритми керування, методи взаємодії багатомірних систем, методи обробки великих даних та пояснювальна записка. Використовувати ОС Windows, середовище об'єктно-орієнтованого проектування.*

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

5. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Шостак І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	25 березня 2020 р.	
2.	Огляд існуючих методів	31 березня 2020 р.	
3.	Методи кооперації штучних агентів	15 квітня 2020 р.	
4.	Підготовка пояснювальної записки	20 квітня 2020 р.	
5.	Спецчастина	28 квітня 2020 р.	
6.	Підготовка презентації та доповіді	03 травня 2020 р.	
7.	Попередній захист	15 травня 2020 р.	
8.	Нормоконтроль, рецензування	17 травня 2020 р.	
9.	Занесення диплома в електронний архів	18 травня 2020 р.	
10.	Допуск до захисту у зав. кафедри	20 травня 2020 р.	

Дата видачі завдання _ «____»_ _____2020 р.

Студент _____
(підпис)

Керівник роботи _____ __проф. Шостак І.В._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 101 с., 8 табл., 41 рисунок, 3 дод., 24 джерела.

АГРЕГАТ, АГРЕГАЦІЯ ДАНИХ, СТРУКТУРА ДАНИХ, МАСШТАБУВАННЯ, ДОСТУПНІСТЬ, ТРАНЗАКЦІЇ, КЛАСТЕР, СОЦІАЛЬНІ МЕРЕЖИ, ФРАГМЕНТАЦІЯ.

Об'єкт дослідження – використання методів агрегації даних для організації даних для просування бізнес-стратегій за допомогою соціальних мереж.

Мета роботи – аналіз методів агрегації даних для створення необхідних пакетів даних користувача соціальних мереж.

Методи дослідження – методи, що використовуються для агрегації даних в соціальних мережах.

У результаті роботи було проведено: об'єктний аналіз поставленої задачі, дослідження методів для організації сховища даних, аналіз методів агрегації даних у соціальних мережах, реалізація методів агрегації даних в соціальних медіа.

AGGREGATE, DATA AGGREGATION, DATA STRUCTURE, SCALING, ACCESSIBILITY, TRANSACTIONS, CLUSTER, SOCIAL NETWORKS, FRAGMENTATION, RECEPTION..

The purpose of the work – analysis of data aggregation methods to create the necessary user data packages of social networks and software implementation of non-relational database such as "key-value"

The object of research – non-relational data models, distribution methods, methods of ensuring consistency and methods of data aggregation for.

The analysis of non-relational data models is carried out, the practical part of the work is the creation of a non-relational database of the type "key-value" for user data packages of social networks.

ЗМІСТ

ВСТУП

В наш час існує багато соціальних мережевих додатків, доступних в Інтернеті – Facebook, LinkedIn, Google+, My Space і т.д. Акаунти у соціальних мережах мають мільйони користувачів, і кожен середній користувач має профіль у більш ніж одній з цих мереж. Деякі дані з профілю користувача соціальної мережі є конфіденційними, а деякі – відкритими. Більшість даних сайтів пропонують можливість для налаштування параметрів конфіденційності. Таким чином, існує величезна кількість загальнодоступних даних, які доступні з цими постачальниками соціальних мереж. Такі дані можуть бути об'єднані і використані для створення профілю користувача, а також визначення способу комунікації з ним.

Сукупні дані профілю користувача соціальної мережі можуть використовуватися для різних завдань:

– ефективне встановлення зв'язку з людьми;

– ефективність продажів через соціальну мережу, залежно від уподобань та інтересів користувача соціальної мережі;

– формування маркетингової стратегії просування продуктів і послуг, яка націлена на людей з урахуванням їхніх інтересів;

– формування перевірки потенційних кандидатів в режимі он-лайн для роботодавців.

Агрегація даних на основі веб-платформи включає агрегування загальнодоступних даних про людину з веб-сайтів соціальних мереж, таких як Facebook, LinkedIn і т.д. Платформа може взаємодіяти з постачальниками соціальних мереж, щоб отримати дані в режимі реального часу на основі імені особи і змісту профілю. Інформація про контакти користувача мережі для створення віртуальної соціальної мережі, яка об'єднує користувачів за певними критеріями. Крім того, враховується інформація про місцезнаходження користувача.

Інтерес представляють наступні функціональні можливості:

– пошук – на основі імені, прізвища, організації тощо;

– інформація про профіль – побудова профілю з агрегованими даними користувача, якого шукають;

– Social Network Connections – список контактів або взаємодій користувача в мережі;

– область наукових інтересів – публічно доступні дані можна пов'язати з ключовими словами або знаками з профілю, які можуть бути співвіднесені з продуктами і послугами на сайтах електронної комерції;

– індивідуальний маркетинг – на основі інформації, що переглядається користувачем на сайтах електронної комерції визначається розташування людини, вартість товарів і послуг, що переглядаються, перелік потенційних продуктів і послуг, залежно від інтересів користувача.

Метою роботи є аналіз методів агрегації даних для створення необхідних пакетів даних користувача соціальних мереж.

Аналіз методів агрегації даних, як концепція може бути розширена до формування змісту профілю користувача соціальної мережі. Відомості про профіль можуть бути інтегровані практично з усіма постачальниками соціальних мереж, таких як Facebook, LinkedIn, Google+, Twitter, MySpace тощо. Більше інтеграції дають кращі результати.

Також це може бути інтегровано з пошуковою системою, яка має можливість кластеризації результатів за допомогою ключових слів на основі параметрів і інтересів користувачів соціальної мережі тощо. Дані, витягнуті з профайлів користувачів соціальної мережі можуть бути використані в пошукових системах, а також для побудови більш значущих профілів користувачів інших соціальних мереж з урахуванням характеру людини, кола знайомих, його інтересів і переваг, що сприятиме індивідуальному маркетингу – менеджер з продажу набагато впевненіше може знайти своїх потенційних клієнтів, використовуючи силу соціальних мереж.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз предметної галузі

Соціальні медіа є комп'ютерно-опосередковані технології, які дозволяють створювати і обмінюватися інформацією, ідеями, професійними інтересами та інші формами вираження через віртуальні спільноти і мережі.

Соціальні медіа можуть допомогти поліпшити почуття фізичних осіб в реальних і / або Інтернет-спільнотах, є ефективним інструментом для корпорацій, підприємців, некомерційних організацій, в тому числі правозахисних організацій і політичних партій і урядів в області маркетингу та ведення бізнесу.

Термін соціальні медіа, як правило, використовується для опису сайтів соціальних мереж, таких як: Facebook, Twitter, LinkedIn, Pinterest, Snapchat тощо.

Соціальні мережі постійно розвиваються. Є різниця між онлайн соціальними мережами і діловими соціальними мережами.

Соціальні мережі включають в себе такі сайти, як Friendster, MySpace, Facebook і т.д. Головна первинна мета цих сайтів це те, що багато користувачів Інтернету використовують їх для он-лайн спілкування зі світом. Звичайно Friendster стара і невизнана всіма людьми соціальна мережа, яка була створена для організації взаємодії людей за інтересами. Багато людей розуміють термін соціальні мережі, але виникає плутанина, якщо соціальна мережа об’єднує людей, за власними інтересами та інтересами бізнесу [1].

Бізнес соціальні мережі об’єднують однодумців з бізнесу, привертають увагу інших до різних подій та активностей, що пропонуються бізнесом, збирають необхідну аудиторію користувачів, та формують сегмент ринку для бізнесу. Прикладами таких мереж є LinkedIn, що дозволяє спілкуватися з діловими контактами, з якими ви або працюєте зараз або працювали в минулому, формування бізнес-мережі з розширенням кількості бізнес-контактів та бізнес-версія Facebook.

Це основний орієнтир при підході бізнес і он-лайнових соціальних мереж. Існує різниця між ними, але вони пов'язані один з одним. Онлайн взаємодія буде рости в результаті ділової взаємодії користувача. Мережі показують, хто їхні користувачі, контактні дані, комунікацій та відносин користувачів. Більшість спеціалістів пов'язують і бізнес, і он-лайнові соціальні мережі разом, але є різниця між ними.

Сучасні соціальні мережі не допомагають користувачам організовувати пакети інформації. Вони вимагають різних облікових даних для входу в різні соціальні мережі, це не є зручним оскільки користувач має запам’ятати кілька логінів, паролів, а також запам’ятовувати перелік друзів які використовують ту чи іншу соціальну мережу.

Агрегатори соціальних мереж – це відносно новий вид додатків, за допомогою яких намагаються консолідувати всі свої різні профілі соціальних мереж в одне ціле, але з незначним успіхом.

1.2 Бізнес потенціал соціальних медіа

Традиційні соціальні медіа пропонують безліч можливостей для компаній в широкому діапазоні секторів бізнесу, економічного сектора, мобільних соціальних медіа-для маркетингових досліджень, комунікацій, акцій продажу (знижки, розвиток, відносини програм лояльності) [2].

Маркетингові дослідження. Мобільні додатки для соціальних медіа пропонують дані про автономні рухи споживачів на рівні деталізації. Будь-яка фірма може знати: точний час, коли клієнт увійшов в один з багатьох торгових пунктів, а також про зауваження, що були висловлені під час візиту.

Зв'язок. Мобільне соціальне середовище передачі даних приймає дві форми: компанія-споживач (в якій компанія може встановити з'єднання зі споживачем на

основі його розташування і надавати відгуки про місця поблизу) і призначеного для користувача контенту.

Стимулювання продажів і знижки. Незважаючи на те, що клієнти повинні були використовувати надруковані купони в минулому, мобільні соціальні медіа дозволяють компаніям адаптувати рекламні акції для конкретних користувачів в певні моменти часу.

Програми розвитку і лояльності відносин. З метою підвищення довгострокових відносин з клієнтами, компанії можуть побудувати програми лояльності, які дозволяють регулярним клієнтам заробити знижки або пільги.

Електронна комерція. Соціальні медіа – це сайти для реалізації маркетингових чистих стратегій, створення платформ, які є взаємовигідними для користувачів, бізнесу, а самі мережі в популярності і доступності електронної комерції або он-лайн покупок [3].

Компанії електронної комерції можуть посилатися на соціальні медіа, як споживчі генеровані засоби масової інформації (Consumer-Generated Media – CGM). Узагальненням є змішання технологій і соціальної взаємодії для спільного створення цінності.

Люди отримують цінну інформацію, освіту, новини та інші дані з електронних і друкованих ЗМІ. Соціальні медіа відрізняються від промислових або традиційних ЗМІ, таких як газети, журнали, телебачення, кіно, оскільки вони є порівняно недорогими і доступними. Вони дозволяють будь-якому (навіть приватній особі) публікувати або отримувати доступ до інформації. Промислові ЗМІ зазвичай вимагають значних ресурсів для публікації інформації, як і в більшості випадків статті проходять через багато переглядів перед публікацією. Цей процес збільшує вартість і в результаті ринкову ціну.

Спочатку соціальні медіа використовувалися тільки приватними особами, але тепер вони використовуються підприємствами, благодійними організаціями, а також в політиці.

Однією з характерних рис – це здатність досягати малих або великих аудиторій [4]. Деякі з властивостей, які допомагають описати відмінності між соціальною і промисловою ЗМІ є:

– якість – у традиційному видавництві діапазон якості істотно вуже, ніж в соціальних медіа, однак, контент в соціальних медіа має високу дисперсію – від самих якісних до низької якості, інколи образливого змісту [5];

– досяжність – промислові та соціальні медіа-технології забезпечують можливість масштабування і здатні досягти глобальної аудиторії. Промислові ЗМІ, однак, як правило, використовують централізовану структуру для організації, виробництва і поширення інформації, в той час як соціальні медіа є за своєю природою більш децентралізованими, менш ієрархічними;

– частота – кількість разів реклама відображається на платформах соціальних медіа;

– доступність – соціальні інструменти ЗМІ, як правило, доступні для громадськості за невелику плату або безкоштовно;

– зручність – промислове виробництво ЗМІ, як правило, вимагає спеціальних навичок і підготовки. З іншого боку, велика частина виробництва соціальних медіа вимагає лише скромних існуючих навичок з якими будь-яка людина з доступом може працювати зі ЗМІ в соціальних медіа;

– безпосередність – часовий інтервал між повідомленнями, що створені промисловими ЗМІ можуть бути довгими (кілька днів, тижнів або навіть місяців) в порівнянні з соціальними медіа;

– сталість – промислові ЗМІ, після створення не можуть бути змінені (один раз в журналі друкуються і розповсюджується статті, зміни не можуть бути зроблені в тій же статті), тоді як соціальні медіа можуть бути змінені практично миттєво за допомогою коментарів або редагування.

Соціальні медіа роблять сильний вплив на ділову активність та ефективності бізнесу. Є чотири канали, за допомогою яких ресурси соціальні медіа трансформуються в можливості продуктивності бізнесу [6]:

– соціальний капітал: представляє ступінь, в якій соціальні медіа впливають на відносини фірм з суспільством і підвищують корпоративні можливості соціального впливу;

– виявлені переваги: представляє ступінь, в якій соціальні медіа виставляють «симпатії» клієнтів і збільшують фінансові можливості;

– соціальний маркетинг: представляє, в якому ступені соціального маркетингу ресурси використовуються для збільшення фінансових можливостей;

– соціальні корпоративні мережі – відносяться до неформальних зв'язків і зв'язків корпоративного персоналу через соціальні мережі. Соціальні корпоративні мережі можуть збільшити експлуатаційні можливості продуктивності.

Є інструменти, які залучають експертів, клієнтів, постачальників і співробітників в розробці продуктів і послуг з використанням соціальних медіа. Компанії можуть використовувати ці інструменти, щоб поліпшити їх бізнес – потенціал і підвищити ефективність: управління взаємовідносинами з клієнтами, інновація, навчання, управління знаннями [7].

1.3 Інструменти персоналізації пошуку в ІПС

Персоналізація пошуку становить величезний інтерес у якості засобу для зменшення неоднозначності пошуку та сприяє підвищенню релевантності результатів запиту конкретного користувача, тим самим забезпечуючи ефективний пошук і доступ до необхідної користувача інформації [21, 22]. Одним із ключових факторів для точного персоналізованого доступу до інформації є наявність користувацьких контекстних даних. Кожний користувач має конкретну мету при пошуку інформації.

Існує три види пошукових систем, які можуть забезпечити персоналізовану інформацію:

– системи зворотного зв'язку по релевантності запиту;
– системи демографічної інформації;
– системи, що базуються на аналізі попередніх запитів користувачів.

В першої, користувачі повинні зареєструватися та указати особисту інформацію, таку як ім'я, адреса електронної пошти, ID і ін. Крім того, користувачі повинні забезпечити зворотний зв'язок для визначення релевантної інформації.

У другому типі систем, користувач повинен указати свої інтереси та свої переваги на опозиційній шкалі. Цей підхід вимагає більших витрат за часом і користувачі віддають перевагу більш простим методам.

В системах третього типу немає необхідності явно вказувати користувацькі дані, тому що система здатна зберегти та проаналізувати журнал обігів у браузері та cookies, враховуючи зміни кожного користувача і його переваги у фоновому режимі.

Різні пошукові системи використовують різні методи для витягу результатів відповіді на запит.

Під персоналізацією пошуку розуміється надання користувачеві індивідуальних, персоналізованих результатів залежно від його інформаційних потреб, пріоритетів, інтересів, географічного положення, соціального стану, віку та інших особливостей.

Як правило, методи, використовувані для персоналізації пошуку, ділять на дві групи.

Методи персоналізації пошуку, засновані на явному зворотному зв'язку з користувачами. У цьому випадку персоналізація пошуку виконується на основі аналізу даних користувацького профілю, зазначених при реєстрації, переданих користувачем даних в ІПС і зібраної статистики (пошукової історії користувача, соціальної та корпоративної інформації про користувача, його переваг і ін.) і даних про документи, що зберігаються в індексі пошукової машини. Для аналізу зазначених даних застосовуються статистичні методи, кластерний аналіз, механізми коллаборативної фільтрації і ін.

Методи персоналізації пошуку, засновані на неявному зворотному зв'язку з користувачами. Дані методи орієнтовані на аналіз контекстної інформації та поведінки користувача у фоновому режимі при роботі з ІПС (характеристик кличів і переглядів, запит текстових характеристик, журналу користувацького браузера, cookies і ін.). Для обліку зазначеної інформації застосовуються як зазначені вище методи, так і евристичні методи, поведінковий таргетинг, різні моделі ранжирування та семантичні методи.

Існує зростаюча тенденція до використання інструментів моніторингу соціальних медіа, які дозволяють маркетологам і компаніям шукати, відстежувати і аналізувати бесіди он-лайн про їх торгову марку, продукти, про теми , що пов'язані з їхнім бізнесом і що для них представляє інтерес [8]-[9]. Це може бути корисно в зв'язку з громадськістю управління і рекламної кампанії спостереження, що дозволяє користувачеві вимірювати віддачу від інвестицій, аудит конкурента, а також спільну участь в громадськості. Відстеження в соціальному медіа також дозволяє компаніям швидко реагувати на он-лайн повідомлення, які критикують їх продукт, послугу. Тим самим допомагати користувачеві вирішити проблеми і зменшити негативні наслідки з он-лайн скарг про продукт, послугу, сервісне обслуговування.

«Сотова структура» – це функціональні блоки за допомогою яких визначають аудиторію в соціальних медіа для ведення бізнесу [10]. Ці будівельні блоки допомагають пояснити потреби залучення соціальної аудиторії ЗМІ. Наприклад, користувачі LinkedIn піклуються головним чином про власну особу, репутації і відносин, в той час як YouTube виконує функції обміну, розмови, формує групи і репутацію [11]. Багато компаній будують свої власні соціальні «контейнери», в яких намагаються пов'язати сім функціональних блоків навколо своїх брендів. Це окремі люди з приватних спільнот, які займаються навколо конкретного бренду, професії або хобі, а не контейнери соціальних медіа, таких як Google+, Facebook і Twitter. PR – відділи стикаються з серйозними проблемами в боротьбі з вірусними негативними настроями, що спрямовані на організації або

окремих осіб платформ соціальних медіа (це отримало назву «sentimentitis»), яка може бути реакцією на оголошення або події [12].

Елементи «Сотової структури» включають в себе такі блоки [10]:

– ідентичність – представляє собою ступінь, в якій користувачі розкривають свою особистість в умовах соціальних медіа. Це може включати в себе розкриття інформації, такі як ім'я, вік, стать, професія, місце, а також інформація, яка зображує користувачів певним чином;

– розмови – це ступінь спілкування користувача з іншими користувачами в соціальних медіа. Багато сайтів соціальних медіа призначені перш за все для полегшення розмов між окремими особами і групами. Ці розмови трапляються з різних причин. Люди публікують блоги, роблять он-лайн коментарі і відправляють повідомлення іншим користувачам, щоб зустріти нових однодумців, щоб побудувати їх почуття власної гідності, або бути попереду нових ідей або трендів. Проте, інші бачать соціальні медіа як спосіб зробити їх послання почуття і позитивно впливає на гуманітарні цілі, екологічні проблеми, економічні питання або політичні дебати;

– спільне використання – цей блок представляє собою ступінь, в якій користувачі обмінюються, поширюють і отримують контент, починаючи від короткого тексту поста або цифрової фотографії. Термін «соціальний» має на увазі, що обмін інформацією між людьми має вирішальне значення. У багатьох випадках, однак, соціальність про об'єкти, які опосередковують ці людино – причинні зв'язки;

– наявність – цей блок представляє собою ступінь доступності користувача як в реальному так і віртуальному просторі. Деякі сайти соціальних медіа мають значки, що вказують, коли користувачі знаходяться в мережі, або відмічають їх розташування у реальному просторі;

– відносини – цей блок представляє собою ступінь зв’язку користувачів;

– репутація – представляє собою ступінь, в якій користувачі можуть ідентифікувати положення інших людей, в тому числі себе, в соціальних медіа. Репутація може мати різні значення на платформах соціальних медіа. У більшості

випадків, репутація є питання довіри. Використовуються користувачем агрегатні інструменти, для генерування інформації для визначення достовірності;

– групи – представляє собою ступінь, в якій можуть утворювати спільноти користувачів і суб-спільнот людей за будь-якими критеріями (вік, гендерність, інтереси, професіональні якості тощо). Чим більше стає «соціальна» мережа, тим більша кількість різних груп.

1.4 Платформи агрегації даних у соціальних мережах

Агрегація даних з соціальних мереж є процес збору контенту з цих соціальних мереж, сервісів, таких як Instagram, Tumblr, Flickr, LinkedIn, Twitch, YouTube, і т.д. в єдине представлення. Це часто здійснюється за допомогою агрегатора соціальної мережі (наприклад, Taggbox, Meta, Іній, Hootsuite і FriendFeed), який поєднує інформацію в одному місці, або допомагає користувачеві об'єднати кілька профілів соціальних мереж в одному профілі. Різні агрегатні послуги надають інструменти або віджети, щоб дозволити користувачам консолідувати повідомлення, відстежувати друзів, об'єднувати закладки, здійснювати пошук по декількох сайтах соціальних мереж, читання RSS – канали для кількох соціальних мереж, мати можливість бачити, коли їх ім'я згадується на різних сайтах, надавати доступ до різних профілів з однім інтерфейсом, забезпечити «життє-потокі», тощо. [19].

Є й інші споріднені види використання соціальних медіа – агрегаторів. Деякі агрегатори (такі як Taggbox Juicer.io і Ubernet) призначені, щоб допомогти компаніям (блогерам) поліпшити взаємодію з їх брендом (и) шляхом створення агрегованих соціальних потоків, які можуть бути вбудовані в існуючий веб – сайт і налаштовувати візуальний вигляд сайту. Це дозволяє потенційним клієнтам взаємодіяти з усіма соціальними медіа, що підтримується брендом, не вимагаючи від них переходу від сайту до сайту. Це має перевагу в утримані клієнтів на сайті

бренду протягом більш тривалого періоду часу (метрикою є збільшення «час на сайті»).

Агрегатори для соціальної мережі це новий вид додатків, які намагаються консолідувати інформацію з різних профілей соціальних мереж в одне ціле.

Панель інструментів браузера Minggl, який працює з Firefox, IE і Flock і допомагає організувати управління вашої соціальної веб. Ідея Minggl полягає в «підключенні» профіля соціальних мереж для Minggl, а потім контролювати за ними з одного місця. Недоліком є підключення профілю через запрошення Minggl, а потім потреба в необхідності часу для тестування отриманої інформації.

Інша програма iStalkr заснована на концепції групи lifestreaming дозволяє стежити за своєю власною сторінкою і діяльністю своїх друзів в соціальних мережах залежно від часу, і впливати на нього безпосередньо з інтерфейсу iStalkr в. Недоліки роботи з iStalkr – це не своєчасне оновлення, що може бути пов'язано з обмеженнями різних API, які використовуються або проблем самого додатку.

Correlate.us дає хороший огляд тільки вашої діяльності в соціальних мережах. Додаток простий і це може бути хорошою основою для більш великого проекту.

Замість того, щоб агрегувати інформацію в соціальних мережах, Explode.us (рисунок 1.1) дозволяє здійснювати пошук у всіх соціальних мережах за однією формою. Для кожного знайденого користувача можна побачити теги, друзів, а також останні розміщені коментарі цього користувача. Explode.us підтримує, серед іншого, LiveJournal, Flickr, Twitter, Jaiku і 43Things.

Spokeo це соціальна мережа трекер, який дозволяє відстежувати, що ваші друзі роблять на різних соціальних мережах від інтерфейсу Spokeo. Це найпростіший з агрегаторів, пропонуючи свого роду «RSS Reader для соціальних мереж».

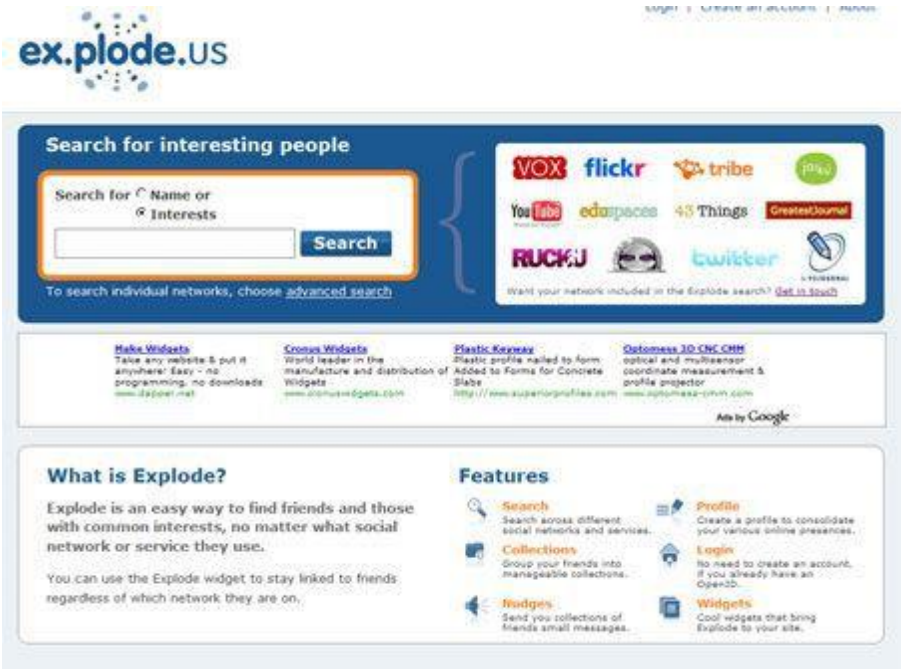


Рисунок 1.1 – Вікно додатку Explode.us

Profilefly створює профіль з особистої інформації, вирізок з Інтернету і особистої соціальної мережі життя потоку людини. Profilefly (рисунок 3.2) працює як віджет або як додаток Facebook, і він підтримує величезну кількість соціальних мереж, в тому числі MySpace, Digg, Hi5, Facebook, Last.FM, Second Life і багато інших. Фактичний профіль трохи м'який, з життяпотоку – який повинен бути центром такого додатка – пропонуючи дуже обмежені можливості; наприклад, тимчасові мітки і будь-якого роду інтерактивностей не вистачає.

PeopleAggregator прагне стати центром соціальної ідентичності. Він працює з допомогою програмного додатка робочого столу, який в даний час працює тільки на Linux. І хоча це залежить від OpenID і відкритих стандартів в цілому, рішення про початок служби, яка спрямована для підключення користувачів соціальних мереж і їх профілі, на платформі Linux, схоже на самогубство. Не вирішує проблему і офіційний FAQ, який містить багато спам – посилань.

Рисунок 1.2 – Вікно додатку Profilefly

SocialURL допоможе організувати свою ідентичність і повернутися в контакт з усіма своїми друзями та однокласниками. Це єдиний профіль з підтримкою фото і відео галерей, а також центральний портал з посиланнями, що вказує на всі ваші інші профілі соціальної мережі. Він містить додаткові функції, таких як електронна пошта, нагадування, закладки, тощо.

Tabber (рисунок 1.3.) особиста сторінка профілю, яка відображає деяку інформацію про користуівача разом з останньою активністю на Digg, del.icio.us, власний блог, Twitter або RSS-канал. Це дуже схоже за своєю концепцією на ProfileFly, і так само, йому не вистачає будь-якої можливості взаємодіяти з вашим життє-потоком.

Naymz ще один персональний сайт профілю, який йде на крок далі, ніж Tabber або ProfileFly, даючи можливість активно контролювати певні місця для згадки вашого імені. Naymz також активно просуває профіль, намагаючись зробити його більш помітним на Google.

Рисунок 1.3 – Вікно додатку Tabber

На відміну від більшості інших послуг, описаних тут, 8hands є настільним додатком. Це дозволяє отримати доступ до профілів в соціальних мережах (в даний час підтримується в Facebook, MySpace, Flickr, YouTube, Twitter і багато іншого). Ідея полягає в тому, щоб мати загальне уявлення про те, що відбувається в соціальних мережах, і відправляти миттєві повідомлення іншим користувачам. Програмне забезпечення 8hands в даний час має має деяку нестабільність в роботі.

Сайт називається Second Brain (рисунок3.4.) і використовує переваги API, що надаються подібними Flickr, Blogger, YouTube, і одинадцять інших веб – служб. З кожним, ви можете надати ім'я користувача і пароль, а Second Brain почне відстежувати зміст, який Ви розміщуєте там. Це зміст може бути конфіденційними чи спільний з іншими.

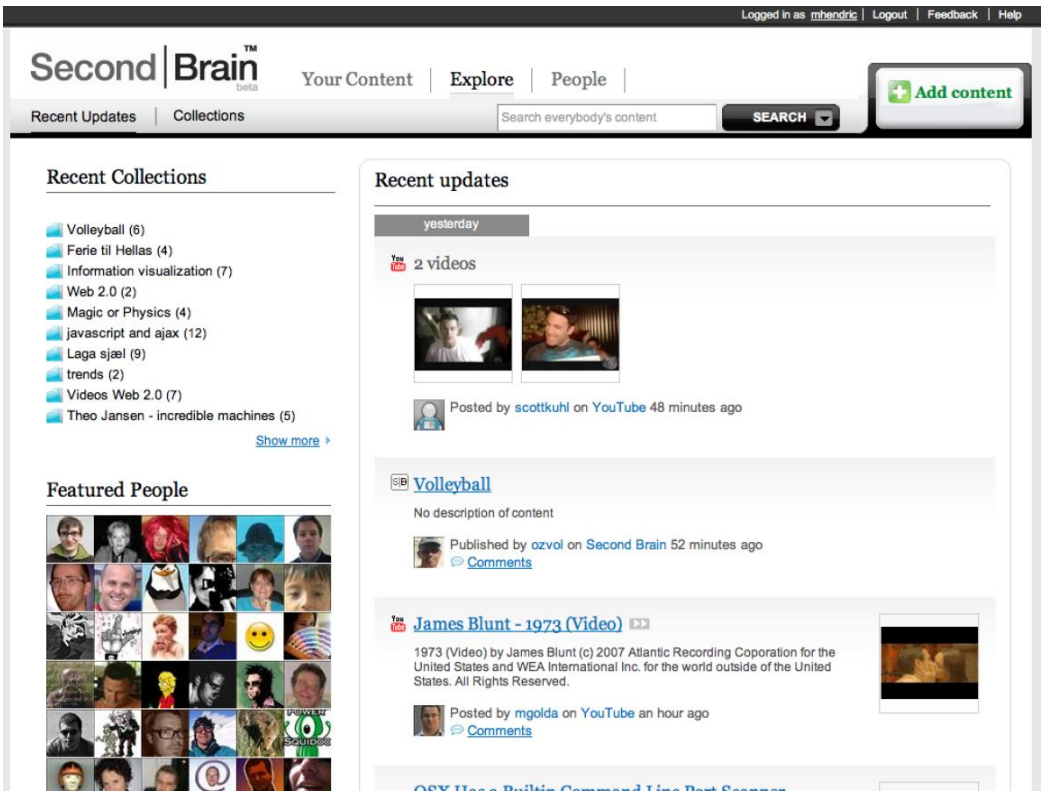


Рисунок 1.4 – Вікно додатку Second Brain

Послуга по суті є соціальною мережею для обміну User-generated content (UGC) з друзями, яка таким чином робить його більш просунутою версією функціональних можливостей спільного використання веб – сайту і інших соціальних мережах.

Громадський зміст контенту отримує Second Brain, і цей контент з'являється в останніх оновленнях на сторінках ваших друзів і на домашній сторінці. Весь контент можна розділити на контент, що дозволяє угрупованню аналогічного змісту знайденого на різних веб-сервісах (Flickr фото і відео на YouTube про технології, наприклад, можуть бути згруповані разом Second Brain). Можна коментувати вміст, організовувати дискусії.

Second Brain приймає зовсім інший підхід до агрегації, ніж інші програми в цьому списку. Організація власних даних – це включення даних з профілів соціальних мереж, таких як фотографії Flickr або відео на YouTube, – в колекції . Колекція може містити безліч посилань, фотографії, або інші біти і шматки даних кинуті на приладовій панелі; можна створювати власні колекції або досліджувати

те, що зібрали інші. Хоча концепція здається сильним, я знайшов себе, намагаючись знайти що – то робити з усім цим.

Сервіс від Second Brain є привабливим для людей, які завантажують контент на кілька напрямків по всьому Інтернету.

Більшість агрегаторів соціальних мереж працюють не стабільно. Деякі не праціють взагалі.

Наприклад: Taggbox, Meta, Іній, Hootsuite мають веб-сайти і можливість зареєструватися та користуватися всіма сервісами.

1.5 Огляд соціальних медіа

Соціальних дані засобів масової інформації, безсумнівно, є найбільш динамічною базою людської поведінки, що надає в результаті нові можливості для розуміння окремих осіб, груп і суспільства. Багато вчених і фахівців галузі все частіше знаходять нові способи автоматичного збору, об'єднання і аналізу цього величезного обсягу даних.

Аналіз соціальних медіа є актуальним для трьох галузей бізнес, біонауки і соціальні науки.

Роздрібні компанії використовують соціальні медіа, для поширення інформації про торгову марку, поліпшення якості продукції / послуг клієнта, рекламу / маркетингові стратегії, аналіз мережевої структури, поширення новин та навіть виявлення шахрайства. В області фінансів, соціальні засоби масової інформації використовуються для вимірювання настроїв учасників ринку, поширення новин та отримання даних для просування фінансових послуг.

В біонауці, соціальні медіа використовується для збору даних про великі когорти для ініціатив поведінкових змін і моніторингу впливу, таких як боротьба з курінням і ожирінням або моніторингу захворювань. Прикладом може служити біологи з Університету штату Пенсільванія, які розробили інноваційні системи та

методи для відстеження поширення інфекційних захворювань, за допомогою веб-сайтів новин, блогів та соціальних медіа. Обчислювальні додатки в області соціальних наук включають в себе: моніторинг громадських відповідей на оголошення, виступи і події особливо політичних коментарів і ініціатив; здатність проникнення в суть поведінки спільноти; соціальні медіа опитування (важко контакту) груп; раннє виявлення виникаючих подій, як в Twitter. Наприклад, Лерман і ін. (2008) використовують комп'ютерну лінгвістику, щоб автоматично прогнозувати вплив новин на суспільне сприйняття політичних кандидатів. Yessenov і Misailovic (2009) використовували фільм огляд коментарів для вивчення впливу різних підходів в добуванні особливості тексту на точність чотирьох методів машинного навчання, дерева рішень, максимальної ентропії і K-середньої кластеризації.

Проведено огляд і аналіз соціальних медіа:

– соціальні медіа даних, за типами даних (наприклад, соціальні мережі засобів масової інформації, вікі, блоги, RSS-канали та новини, і т.д.) і форматів (наприклад, XML і JSON). Це включає в себе набори даних і канали все більш важливі дані в режимі реального часу, таких як фінансові дані, дані по операціях клієнтів, операторів зв'язку і просторових даних;

– соціальні медіа програмні послуги доступу даних та інструменти для пошуку і вишкрібання (текстових) даних із соціальних мереж засобів масової інформації,

RSS-канали, новини і т.д. вони можуть бути з користю поділені на:

– джерела даних, послуги і інструменти – дані доступні за допомогою інструментів, які захищають вихідні дані або надають просту аналітику. Приклади включають в себе: тенденції Google, SocialMention, SocialPointer і соціал пошук, які забезпечують потік інформації, що об'єднує різні соціальні медіа-канали;

– дані каналів за допомогою API-інтерфейсів, десь набори даних і канали доступні через програмовані HTTP на основі API-інтерфейсів і повернення зазначених даних з використанням XML або JSON і т.д. Приклади включають в Wikipedia, Twitter і Facebook.

– очищення та зберігання текстових інструментів – інструменти для очищення і зберігання текстових даних. Google Detailed і DataWrangler є прикладами для очищення даних.

– інструменти аналізу тексту – індивідуальні або бібліотеки інструментів для аналізу даних соціальних медіа, які були зняті і почищені. Це в основному для обробки природної мови, аналізу та класифікації інструментів, які описані нижче:

– прості інструменти, які можуть перетворювати введений текст даних в таблиці, карти, діаграми (line, pie, scatter, bar, etc.), графік або навіть руху (анімація за термінами), такі як Tables, Zoho Reports, Tableau Public або IBM’s Many Eyes;

– інструменти аналізу – більш просунуті аналітичні інструменти для аналізу соціальних даних, виявлення зв'язків і створення мереж, таких як Gephi (з відкритим вихідним кодом) або Excel плагін NodeXL;

– соціальні медіа-платформи – середовища, які забезпечують всебічні дані соціальних медіа та бібліотеки інструментів для аналітики.

Приклади включають в себе: Thomson Reuters Machine Readable News, Radian 6 and Lexalytics:

– соціальні мережі медіа-платформи платформи, які забезпечують видобуток і аналітику даних на Twitter, Facebook і широкий спектр інших джерел соціальної мережі засобів масової інформації;

– новинні платформи – платформи, таких як Thomson Reuters надання комерційних архівів новин / каналів та пов'язана з ними аналітику.

Існує дві основні перешкоди на шляху використання соціальних медіа для наукових досліджень, по-перше доступ до комплексних наборів даних і друге інструментів, які дозволяють проводити аналіз «глибоких» даних без необхідності мати можливість програмувати на мові, таких як Java.

Більшість соціальних медіа-ресурсів є комерційними і компанії, природно, намагаються монетизувати свої дані. Як вже було помічено раніше, дослідники мають доступ до відкритого вихідного коду соціальних медіа, набору даних і засобів для проведення експериментів. В іншому випадку, дослідження

соціальних медіа може стати виключною прерогативою великих компаній, урядових установ і привілейованих академічних дослідників, що керують особистими даними, з яких вони виробляють документи, які не можуть бути повторені або критиковані.

1.6 Постановка задач дослідження

На основі аналізу предметної галузі основною задачею є аналіз методів агрегації даних з використанням підходу Інтернет аналітичної обробки даних, що представляє простий тип агрегації даних для створення необхідних пакетів даних користувача соціальних мереж. Цей он-лайновий механізм пакетів даних в подальшому використовуватиме аналітик, маркетолог, менеджер для звітності, обробки інформації тощо.

Це дозволить аналітикам і менеджерам електронного бізнесу отримувати уявлення про користувача завдяки отриманої інформації та за допомогою

Для досягнення зазначеної мети необхідно розв'язати наступні завдання:

– вивчення основних характеристик і особливостей ІПС, а також різних методів пошуку в ІПС;

– аналіз проблем побудови ІПС Інтернету та можливих шляхів їх рішення;

– розробка методів інтелектуалізації та персоналізації ІПС;

– розробка архітектури інтелектуальної ІПС на основі прецедентів;

– програмна реалізація прототипу ІПС на основі прецедентів.

Поставлені завдання вирішуються з використанням методів дискретної математики, математичної логіки, штучного інтелекту, теорії програмування та теорії інформаційного пошуку.

2 ДОСЛІДЖЕННЯ МЕТОДІВ ДЛЯ ОРГАНІЗАЦІЇ СХОВИЩА ДАНИХ

2.1 Сховище даних

Агрегація даних – це будь-який процес об'єднання даних, в якому зібрана інформація виражається в короткій формі, для таких цілей, як статистичний аналіз. Поширена мета агрегації полягає в тому, щоб отримати більше інформації про конкретні групи, що засновані на конкретних змінних, таких як вік, професія, прибуток, тощо. Інформацію про такі групи можна використати для персоналізації веб-сайту, щоб вибрати вміст і рекламу ймовірно, щоб звернутися до людини, що належить до однієї або декількох груп, для яких були зібрані дані. Наприклад, сайт, який продає музичні компакт-диски може рекламувати певні компакт-диски на основі віку користувача і сукупності даних для їх вікової групи. Інтернет аналітична обробка даних (OLAP) – це простий тип агрегації даних, в якому маркетолог використовує он-лайновий механізм звітності для обробки інформації [13].

Об'єднання даних може бути встановлено користувачем на основі: послуги агрегації персональних даних, що надають користувачеві єдину точку для збору їх особистої інформації, отриманої від інших веб-сайтів. Клієнт використовує персональний ідентифікаційний номер (PIN), щоб дати їм доступ до різних акаунтів (наприклад, для фінансових установ, авіакомпаній, книжкових і музичних клубів, декількох соціальних мереж, тощо). Такий тип агрегування даних, іноді називають «читання з екрану» (SS – screen scraping).

Сховище даних (Data Warehouse англ.) – предметно-орієнтована інформаційна база даних, спеціально розроблена і призначена для підготовки звітів і бізнес-аналізу з метою підтримки прийняття рішень в організації. Будується на базі систем управління базами даних і систем підтримки прийняття рішень. Дані, що надходять в сховище даних, як правило, доступні тільки для читання [14].

Дані з OLTP-системи копіюються в сховище даних таким чином, щоб при побудові звітів і OLAP-аналізі не використовувалися ресурси транзакційної системи і не порушувалася її стабільність. Є два варіанти оновлення даних в сховищі:

– повне оновлення даних в сховищі. Спочатку старі дані видаляються, потім відбувається завантаження нових даних. Процес відбувається з певною періодичністю, при цьому актуальність даних може трохи відставати від OLTP-системи;

– інкрементне оновлення – оновлюються тільки ті дані, які змінилися в OLTP-системі.

Сховище даних будується шляхом інтеграції даних з декількох різнорідних джерел, які підтримують аналітичну звітність, структуровані і / або спеціальні запити і прийняття рішень. Зберігання даних включає в себе очищення даних, інтеграцію і консолідацію [14].

Ключові особливості сховища даних:

– орієнтоване – сховище даних предметно – орієнтоване, оскільки воно надає інформацію навколо предмета, а не поточної діяльності організації. Ці предмети можуть бути продукт, клієнти, постачальники, продаж, дохід і т.д. Сховище даних не фокусується на поточних операціях, а основна увага приділяється моделюванню й аналізу даних для прийняття рішень;

– інтегроване – сховище даних будується шляхом інтеграції даних з різнорідних джерел, таких як реляційні бази даних, плоскі файли і т.д. Така інтеграція підвищує ефективний аналіз даних;

– залежить від часу – дані, зібрані в сховище даних ідентифікується з певним періодом часу. Дані в сховищі даних містять інформацію з історичної точки зору;

– незалежне – означає, що попередні дані не видаляються, коли нові дані додаються до нього. Сховище даних зберігається окремо від робочої бази даних і, отже, часті зміни в оперативній базі даних не відображаються в сховище даних.

Сховище даних не вимагає обробки транзакцій, відновлення і управління паралелізмом, так як фізично зберігається окремо від операційної бази даних.

Сховище даних допомагає бізнес-керівникам організовувати, аналізувати і використовувати їх дані для прийняття рішень. Сховище даних служить єдиною частиною плану-виконання-оцінки «замкнутого циклу» системи зворотного зв'язку для управління підприємством. Сховища даних широко використовуються в наступних областях:

– фінансові послуги;
– банківські послуги;
– споживчі товари;
– роздрібні сектора;
– контрольоване виробництво.

Існують три типи сховища даних:

– обробка інформації – сховище даних дозволяє обробляти дані, що зберігаються в ньому. Ці дані можуть бути оброблені за допомогою пошуку інформації статистичного аналізу, звітності з використанням крос – таблиць, таблиць, діаграм і графіків;

– аналітична обробка – сховище даних підтримує аналітичну обробку інформації, що зберігається в ньому. Дані можуть бути проаналізовані за допомогою основних операцій OLAP;

– інтелектуальний аналіз даних підтримує виявлення знань, знаходячи приховані закономірності та асоціації, побудова аналітичних моделей, що виконують класифікацію та прогнозування. Ці результати робіт можуть бути представлені за допомогою засобів візуалізації.

Порівняльний аналіз сховища даних з оперативною БД наведений у таблиці 2.1

Таблиця 2.1 – Порівняльний аналіз сховища даних з оперативною БД

№	Сховище даних (OLAP)	Оперативна БД (OLTP)

1	Включає в себе історичну обробку інформації.	Включає в себе обробку день у день.
2	OLAP Системи використовуються керівниками менеджерами і аналітиками.	OLTP системи використовують клерки або фахівців баз даних.
3	Використовується для аналізу бізнесу.	Використовується для запуску бізнесу
4	Зосереджено на інформації із зовні.	Зосереджена на даних.
5	Засновано на схемі зірки, Snowflake Schema і Fact Сузір'я схеми.	Заснована на Entity Relationship Model.
6	Зосереджено на інформації із зовні.	Додаток орієнтований.
7	Містить історичні дані.	Містить поточні дані
8	Забезпечує підсумовані і консолідовані дані.	Забезпечує примітивні і дуже докладні дані.
9	Забезпечує підсумовані і багатовимірне представлення даних.	Забезпечує детальне і плоске реляційне уявлення даних.
10	Кількість користувачів сотні.	Кількість користувачів тисячі.
11	Кількість записів, доступ до яких знаходиться в мільйонах.	Кількість записів, доступ до яких знаходиться в десятки разів.
12	Розмір БД від 100 ГБ до 100 ТБ.	Розмір БД від 100 МБ до 100 ГБ.

Є підтримка прийняття рішень технології, які допомагають використовувати наявні дані в сховище даних. Ці технології допомагають керівникам використовувати цю інформацію швидко і ефективно. Вони можуть збирати дані, аналізувати їх і приймати рішення, засновані на представленої інформації в будь-який з наступних областей:

– налаштування стратегії виробництва. Стратегії продукту можуть бути добре налаштовані шляхом перестановки продуктів і управління портфелем продуктів шляхом порівняння продажів щоквартально або щорічно.;

– аналіз клієнтів. Аналіз клієнтів здійснюється шляхом аналізу купівельних перевагах клієнта, час покупки, бюджетних циклів і т.д.;

– операції аналізу. Інформація дозволяє аналізувати бізнес – операції, допомагає в управлінні взаємовідносинами з клієнтами.

Компоненти, що входять до типового сховища, представлені на рисунок 2.1 [15].

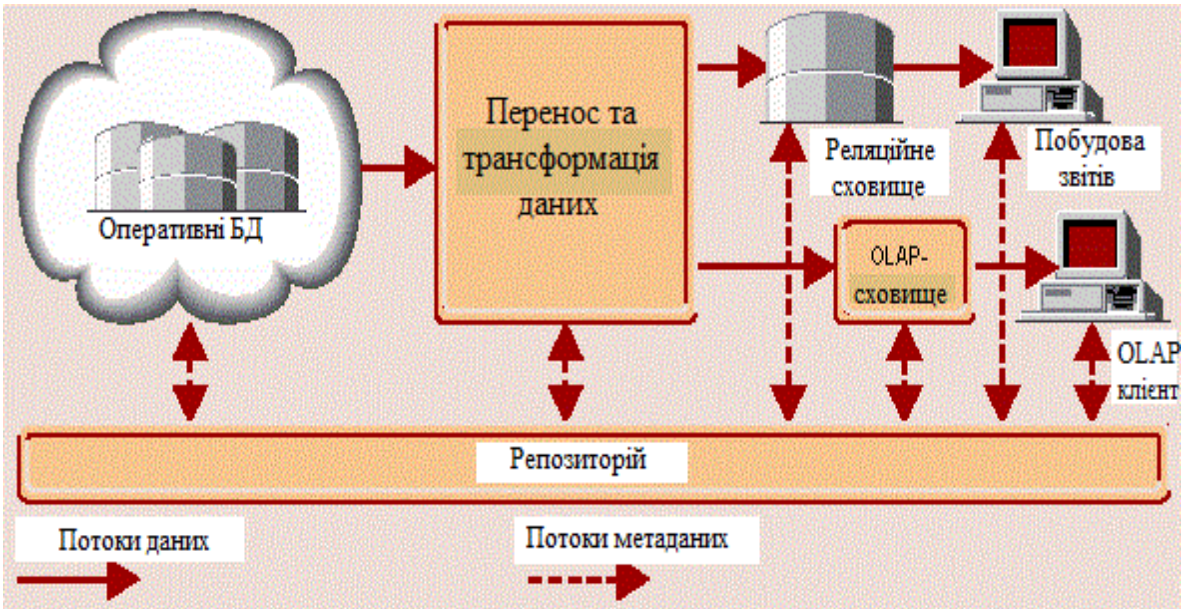


Рисунок 2.1 – Структура сховища даних [15]

OLAP не представляє собою необхідний атрибут сховища даних, він все частіше і частіше застосовується для аналізу накопичених в цьому сховищі відомостей.

2.2 Дизайн сховищ даних

Загальна метамодель сховища даних (Common Warehouse Metamodel – CWM) (рисунок2.2)– це стандарт, що описує обмін метаданими при використанні технологій сховища даних, Business Intelligence, Knowledge Management.

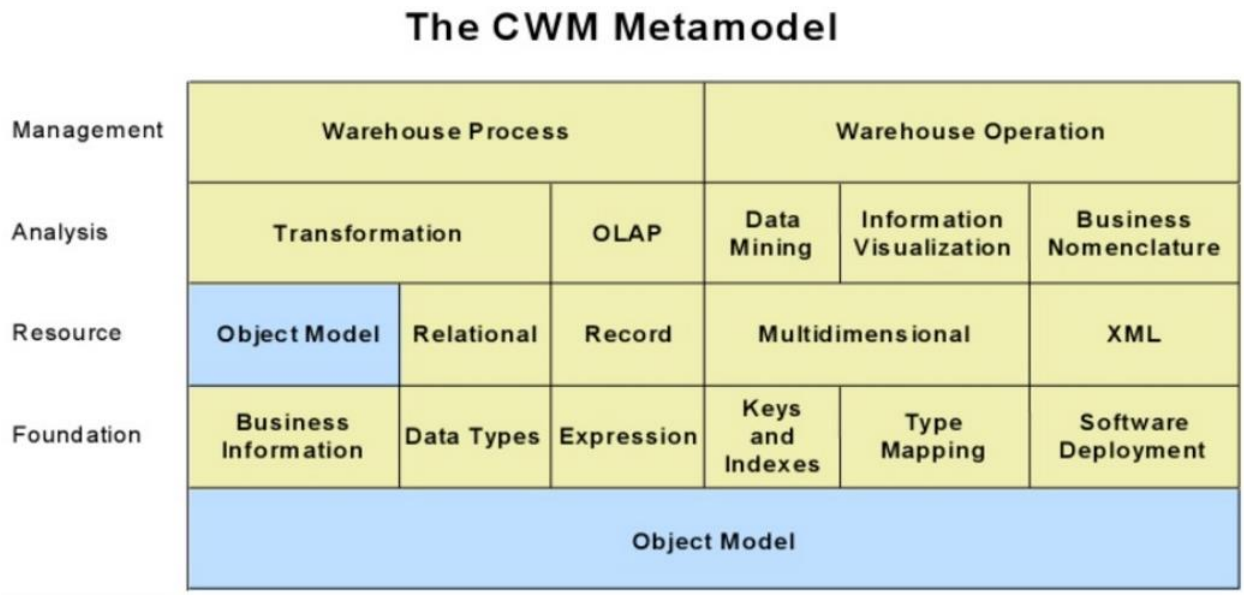


Рисунок 2.2 – Загальна метамодель сховища даних (Common Warehouse Metamodel – CWM)

Існують два архітектурних напрямки – нормалізовані сховища даних і сховища з вимірами [14].

У нормалізованих сховищах, дані знаходяться в предметно-орієнтованих таблицях третьої нормальної форми. Нормалізовані сховища характеризуються як прості в створенні та управлінні, недоліки нормалізованих сховищ – велика кількість таблиць як наслідок нормалізації, через що для отримання будь-якої інформації потрібно робити вибірку з багатьох таблиць одночасно, що призводить до погіршення продуктивності системи. Для вирішення цієї проблеми використовуються денормалізовані таблиці – вітрини даних, на основі яких вже виводяться звітні форми. При величезних обсягах даних можуть використовувати кілька рівнів «вітрин» / «сховищ».

Сховища з вимірами використовують схему «зірка» або схему «сніжинка». При цьому в центрі «зірки» знаходяться дані (fact table), а вимірювання утворюють промені зірки. Різні таблиці фактів спільно використовують таблиці вимірювань, що значно полегшує операції об'єднання даних з декількох

предметних таблиць фактів (приклад – факти продажу та поставок товару). Таблиці даних і відповідні вимірювання утворюють архітектуру «шина». Вимірювання часто створюються в третій нормальній формі, в тому числі, для протоколювання зміни в вимірах. Основною перевагою сховищ з вимірами є простота і зрозумілість для розробників і користувачів, також, завдяки більш ефективному зберігання даних і формалізованим вимірам, полегшується і прискорюється доступ до даних, особливо при складних аналізах. Основним недоліком є більш складні процедури підготовки і завантаження даних, а також управління і зміна вимірювань даних.

Схема «зірка» – спеціальна організація реляційних таблиць, зручна для зберігання багатовимірних показників. Лежить в основі реляційного OLAP. Модель даних складається з двох типів таблиць: однієї таблиці фактів (fact table) – центр «зірки» – і декількох таблиць вимірів (dimension table) по числу вимірювань в моделі даних – промені «зірки» (рисунок 2.3).

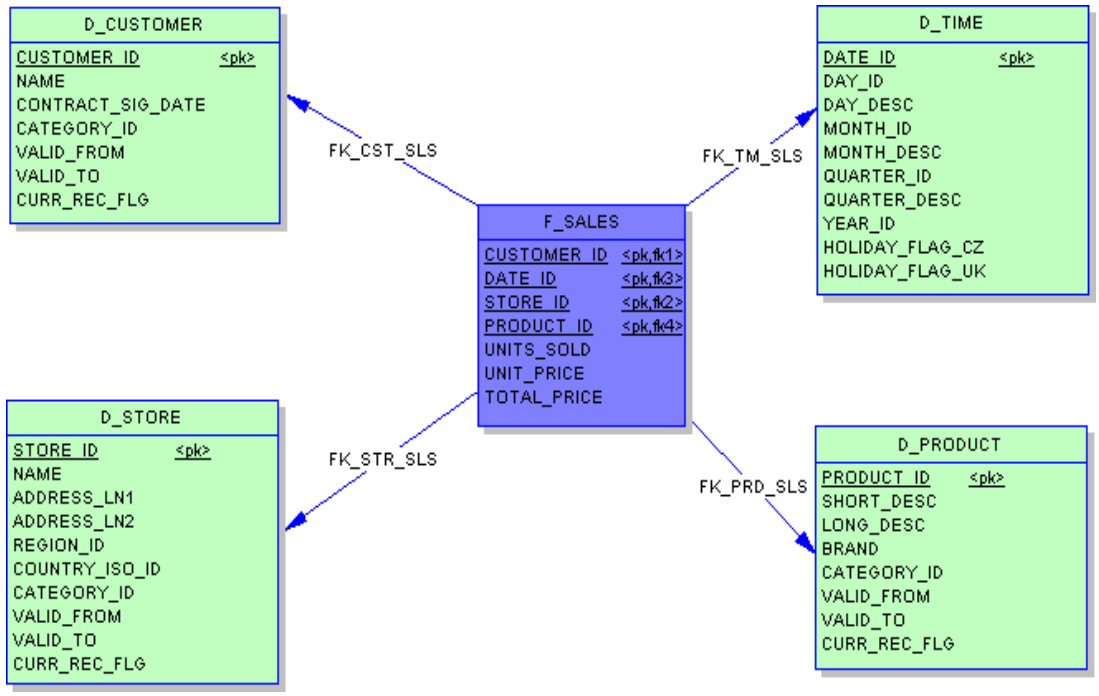


Рисунок 2.3 – Модель даних, схема «Зірка»

Таблиця фактів зазвичай містить одну або кілька колонок типу DECIMAL, що дають числову характеристику якогось аспекту предметної області (наприклад, обсяг продажів для торговельної компанії або сума платежів для банку), і кілька цілочисельних колонок-ключів для доступу до таблиць вимірів.

Таблиці вимірювань розшифровують ключі, на які посилається таблиця фактів; наприклад, «Таблиця продуктів» виміру «товари» бази даних торгової компанії може містити відомості про назву товару, його виробника, тип товару. За рахунок використання спеціальної структури таблиці вимірювань реалізується ієрархія вимірів, в тому числі розгалужені.

Зазвичай дані в таблицях-вимірах денормалізовані: ціною кілька неефективного використання дискового простору вдається зменшити число що беруть участь в операції з'єднання таблиць, що зазвичай призводить до сильного зменшення часу виконання запиту. Іноді, проте, потрібно провести нормалізацію таблиць-вимірювань; така схема носить назву «сніжинка» (snowflake schema).

Логічне та фізичне представлення схеми «Зірка» зображено на рисунок 2.4 [15].

SQL-запит до схеми «зірка» зазвичай містить в собі:

– одне або декілька з'єднань таблиці фактів з таблицями вимірювань;

– кілька фільтрів (SQL-оператор WHERE), що застосовуються до таблиці фактів або таблиць вимірів;

– угруповання та його узагальнення за необхідними елементами ієрархії вимірів (елементи виміру).

Наприклад:

```
SELECT
  d_product.brand,
  d_store.country_iso_id,
  SUM (f_sales.units_sold) AS summa
 FROM
  f_sales, d_customer, d_time, d_store, d_product
 WHERE
  f_sales.customer_id = d_customer.customer_id  AND
  f_sales.date_id = d_time.date_id  AND
  f_sales.store_id = d_store.store_id AND
  f_sales.product_id = d_product.product_id AND
  d_time.year_id = 1997  AND
  d_product.category_id = "tv"
 GROUP BY
  d_product.brand,  d_store.country_iso_id
```

Розвитком схеми «Зірка» стала схема «Сніжинка» (Snowflake scheme). ЇЇ відрізняє від першої схеми наявність підпорядкованих таблиць при описі розмірностей для реалізації декількох рівнів ієрархії.

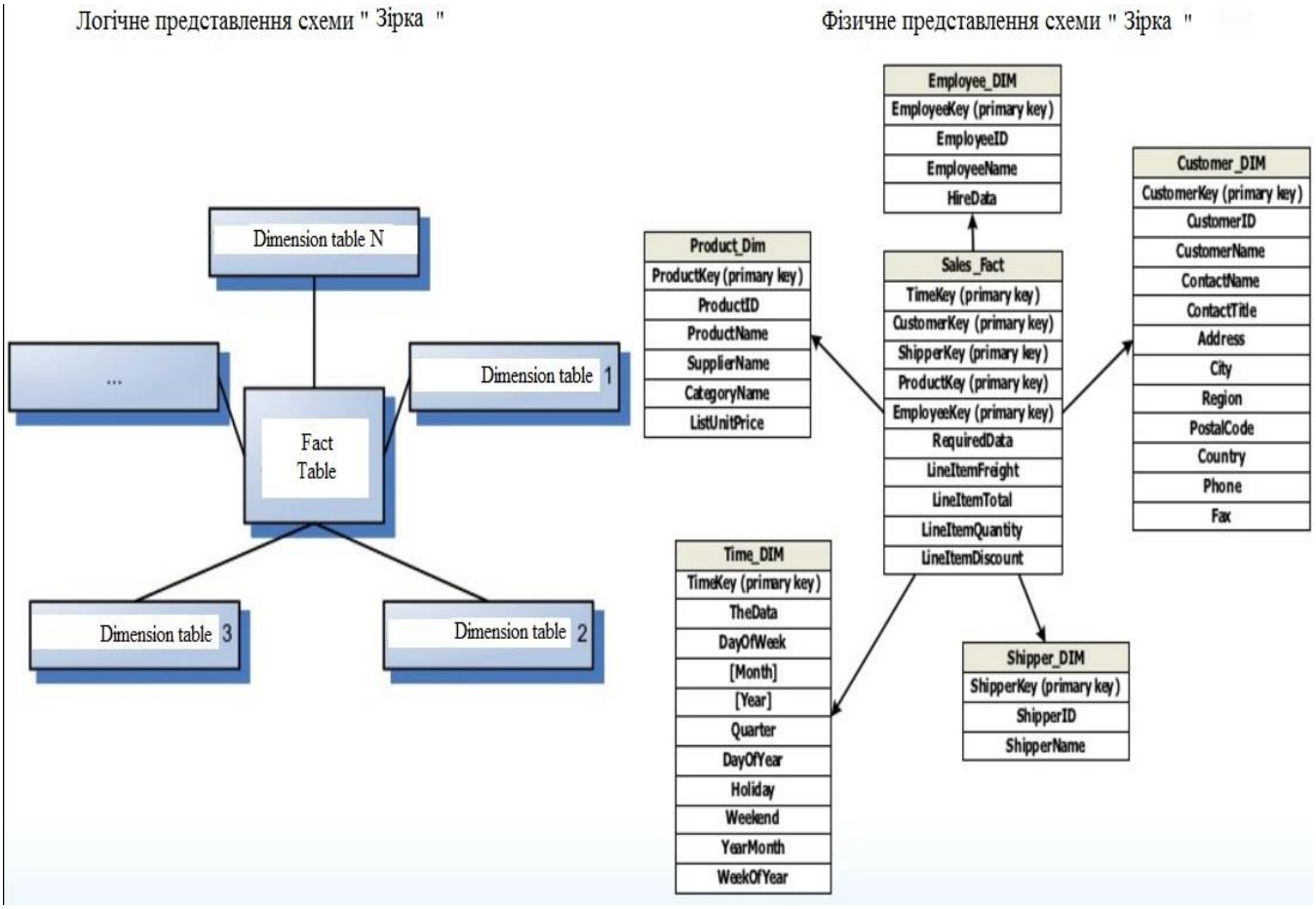


Рисунок 2.4 – Логічне та фізичне представлення схеми «Зірка»

Схема «Сніжинка» отримала свою назву за свою форму, у вигляді якої відображається логічна схема таблиць в багатовимірної базі даних. Так само як і в схемі зірки, схема сніжинки представлена централізованої таблицею фактів, з'єднаної з таблицями вимірювань. Відмінністю є те, що тут таблиці вимірювань нормалізовані з рядом інших пов'язаних вимірювальних таблиць, – в той час як в схемі зірки таблиці вимірювань повністю денормалізовані, з кожним виміром представленим у вигляді єдиної таблиці, без з'єднань на пов'язані таблиці в схемі сніжинки. Чим більше ступінь нормалізації таблиць вимірів, тим складніше

виглядає структура схеми сніжинки. Створюваний «ефект сніжинки» зачіпає тільки таблиці вимірювань, і не застосуємо до таблиць фактів.

Схема «Сніжинка», також як і схема зірки, найбільш часто зустрічається в таких сховищах даних, для яких швидкість отримання даних важливіша, ніж ефективність їх маніпуляції. Отже, таблиці повинні бути нормалізовані в малому ступені, і часто розробляються з застосуванням не вище третього рівня нормалізації. Логічне та фізичне представлення схеми «Сніжинка» зображено на рисунку 2.5 [15].

Рішення в сторону використання схеми зірки або ж схеми сніжинки, обумовлюється відносної потужністю платформи БД, і інструментарію для реалізації запитів. Схема зірки підходить оточенню, в якому інструментарій реалізації запитів надає користувачам широкий доступ до структури таблиць, а також в середовищах, де більшість запитів прості за своєю природою. Схема сніжинки більш підходить для випадків застосування більш складного інструментарію для реалізації запитів, який більшою мірою ізолює користувачів від детальної структури таблиць, а також для середовища з безліччю запитів складної структури.

Поєднанням схем «Зірка» та «Сніжинка» є схема «Сузір’я» див. рисунок 2.6 [15].

Багатовимірні схеми можуть бути визначені за допомогою інтелектуального аналізу даних Query Language (DMQL).

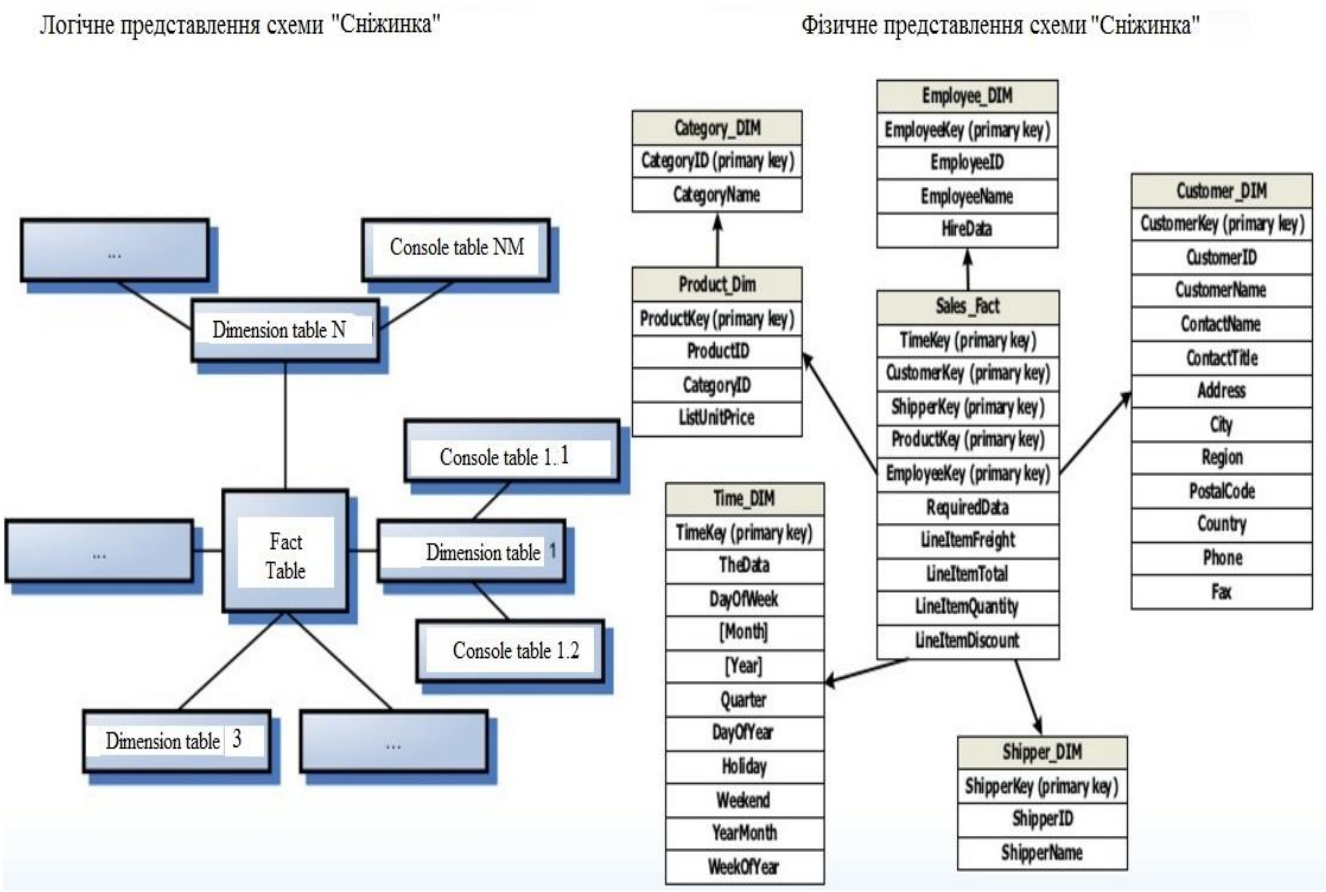


Рисунок 2.5 – Логічне та фізичне представлення схеми «Сніжинка»

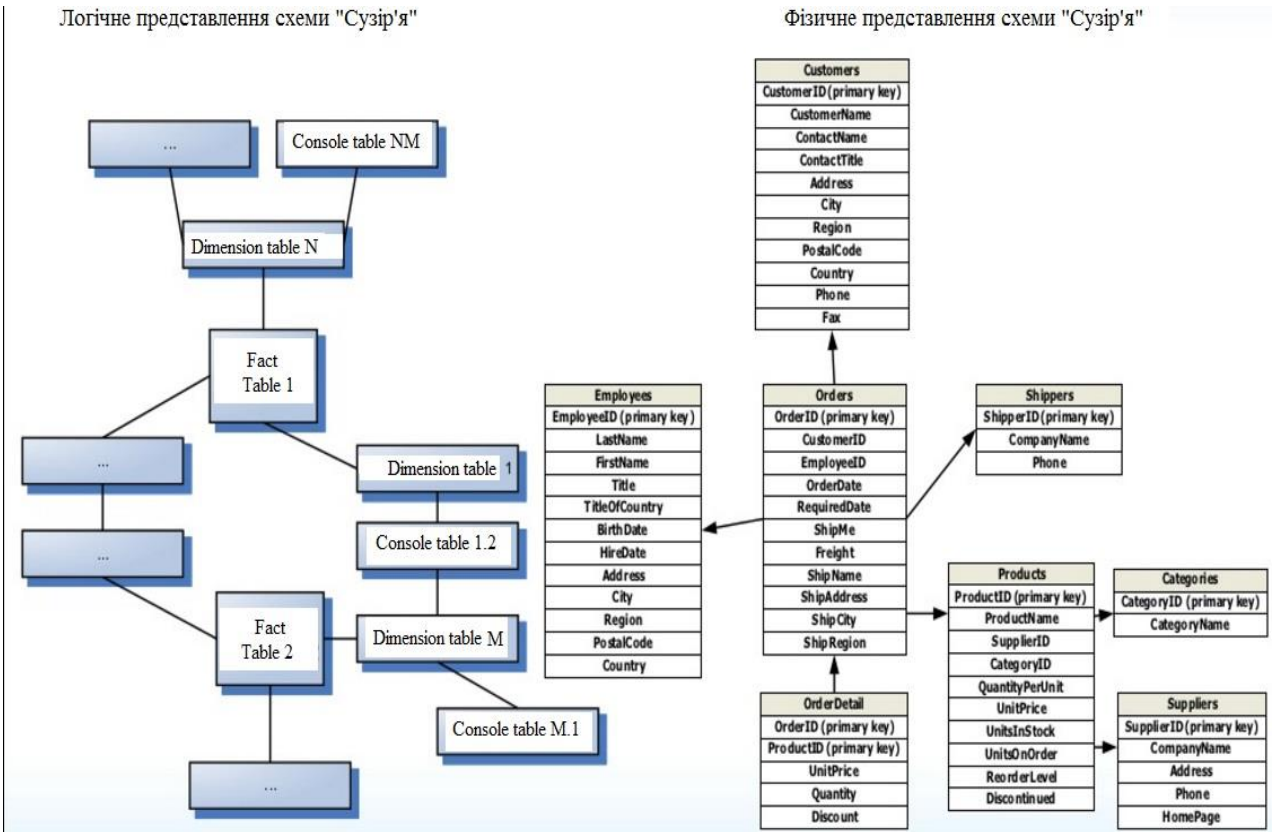


Рисунок 2.6 – Логічне та фізичне представлення схеми «Сузір’я»

2.3 Аналітична обробка в реальному часі – OLAP

Причина використання OLAP для обробки запитів – швидкість. Реляційні БД зберігають сутності в окремих таблицях, які зазвичай добре нормалізовані. Ця структура зручна для операційних БД (системи OLTP), але складні багатотабличні запити в ній виконуються відносно повільно.

OLAP-структура, створена з робочих даних, називається OLAP куб. Куб створюється із з'єднання таблиць із застосуванням схеми зірки або схеми сніжинки. У центрі схеми зірки знаходиться таблиця фактів, яка містить ключові факти, за якими робляться запити. Множинні таблиці з вимірами приєднані до таблиці фактів. Ці таблиці показують, як можуть аналізуватися агреговані реляційні дані. Кількість можливих агрегувань визначається кількістю способів, якими початкові дані можуть бути ієрархічно відображені.

OLAP-куб містить базові дані та інформацію про вимірювання (агрегати). Куб потенційно містить всю інформацію, яка може знадобитися для відповідей на будь-які запити. При величезній кількості агрегатів часто повний розрахунок відбувається тільки для деяких вимірювань, для інших же проводиться «на вимогу».

Існують три типи OLAP [16]:

– багатовимірний OLAP (OLAP багатовимірні – MOLAP). MOLAP – класична форма OLAP, так що її часто називають просто OLAP. Вона використовує підсумовує БД, спеціальний варіант процесора просторових БД і створює необхідну просторову схему даних зі збереженням як базових даних, так і агрегатів;

– реляційний OLAP (OLAP Реляційні – ROLAP). ROLAP працює безпосередньо з реляційних сховищем, факти і таблиці з вимірами зберігаються в реляційних таблицях, і для зберігання агрегатів створюються додаткові реляційні таблиці;

– гібридний OLAP (OLAP Hybrid – HOLAP). HOLAP використовує реляційні таблиці для зберігання базових даних і багатовимірні таблиці для агрегатів.

Особливим випадком ROLAP є «ROLAP Реального часу» (в режимі реального часу ROLAP – R-ROLAP). На відміну від ROLAP в R-ROLAP для зберігання агрегатів не створюються додаткові реляційні таблиці, а агрегати розраховуються в момент запиту. При цьому багатовимірний запит до OLAP-системі автоматично перетвориться в SQL-запит до реляційних даних.

Кожен тип зберігання має певні переваги, хоча є розбіжності в їх оцінці у різних виробників. MOLAP найкраще підходить для невеликих наборів даних, він швидко розраховує агрегати і повертає відповіді, але при цьому генеруються величезні обсяги даних. ROLAP оцінюється як більш масштабується рішення, яке використовує до того ж найменше можливе простір. При цьому швидкість обробки значно знижується. HOLAP знаходиться посеред цих двох підходів, він досить добре масштабується і швидко обробляється. Архітектура R-ROLAP дозволяє виробляти багатовимірний аналіз OLTP-даних в режимі реального часу.

Складність в застосуванні OLAP полягає в створенні запитів, виборі базових даних і розробці схеми, в результаті чого більшість сучасних продуктів OLAP поставляються разом з величезною кількістю попередньо налаштованих запитів. Інша проблема – в базових даних. Вони повинні бути повними та не суперечливими.

Розмірності (dimensions) куба асоціюються з фактами (також званими вимірами «measures»). Згідно реляційної термінології у фактів мають місце відносини «багато до одного» з розмірностями. Наприклад, компанія Acme Computer Supplies має базу даних для продавців. Її розмірності, як правило, – Customers (клієнти), Products (Продукти). Time Element (період: month – місяць, quarter – квартал і т.д.). Сума продажів деякого продукту (Cat5e cables) деякого клієнту (Oracle Corp.) за деякий період часу (Aug 2008) – це один факт (міра). Розмірності зберігаються в окремих таблицях, також зберігаються і факти – в даному випадку сума продажів. Так що, використовуючи реляційну термінологію,

таблиця фактів (з даною сумою продажів) – це породжена (child) таблиця для таблиць розмірності.

Приклад OLAP-кубу представлений на рисунок 2.7

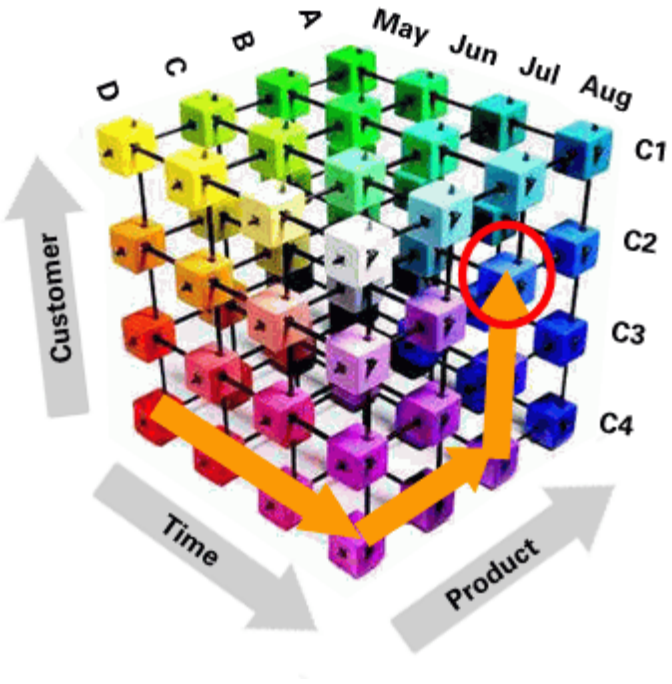


Рисунок 2.7 – Приклад OLAP-кубу

Доступ до вимірювань заходи в реляційної схемою здійснювався б через індекси, створені на шпальтах клієнт, продукт або час таблиці факт. При використанні OLAP-підходу доступ до специфікованих клітин (cell) – вимірам здійснюється завдяки перетинанню цього куба: в цьому прикладі завдяки переходу до шару, який містить заданий період – Aug 08; потім продукт – Cat5e; і, нарешті, клієнт – Oracle.

Oracle знає як дістатися до цих шарів, завдяки обчислень мети в масиві (куба), не в таблиці. Наприклад, припустимо, що розмірності організовані, як показано нижче:

```
Dimension Time := {'May','Jun','Jul','Aug'}
Dimension Customer := {'Microsoft','IBM','Oracle','HP'}
Dimension Product := {'Fiber','Cat6e','Cat5e','Serial'}
```

Щоб виявити факт для Oracle + Aug + Cat5e, OLAP-движок виконує навігацію приблизно такого типу:

Aug08 – це четвертий елемент масиву Time, так що переходимо по вимірюванню часу даного куба до четвертої клітці;

Cat5e – це третій елемент масиву Product, так що переходимо до цього третього елементу;

Oracle – це третій елемент масиву Customer, тому переходимо до цього елементу.

Ось так можна дістатися до потрібного виміру. Це зроблено без індексів, так як значення розмірностей служать як покажчики масиву. Аналогічно, якщо необхідно обчислити загальні суми продажів по всіх клієнтах в Aug08, можна зробити те ж саме, що зроблено вище, за винятком того, що на кроці Крок 3 необхідно просумувати всі факти елементів масиву без переходу до специфікованої клітці.

Контраст між таким підходом і реляційним доступом в чисто реляційної формі даних, що зберігаються в класичній схема зірка, показаний на рисунок2.8.

У підході із застосуванням реляційної бази даних необхідно з'єднати (join) цю «fact» таблицю з усіма розмірностями. Кожен раз, коли потрібні дані, потрібно їх вибрати (select) з цієї fact таблиці, можливо через індекси, і з'єднати їх з кожної розмірністю одна за одною, знову-таки через індекси. Хоча технічно цілком це можливо, такий підхід все-таки нереальний з великими базами даних. В якості альтернативи щодо створення матеріалізованих уявлень – MV (Materialized Views) для всіх цих вибірок, користувач міг би використовувати будь-які комбінації елементів в розмірностях:

Продажі Cat5e в Aug всім клієнтам.

Продажі в Oracle продукту Serial Cable в Aug, як відсоток продажів в IBM того ж самого продукту і за той же період.

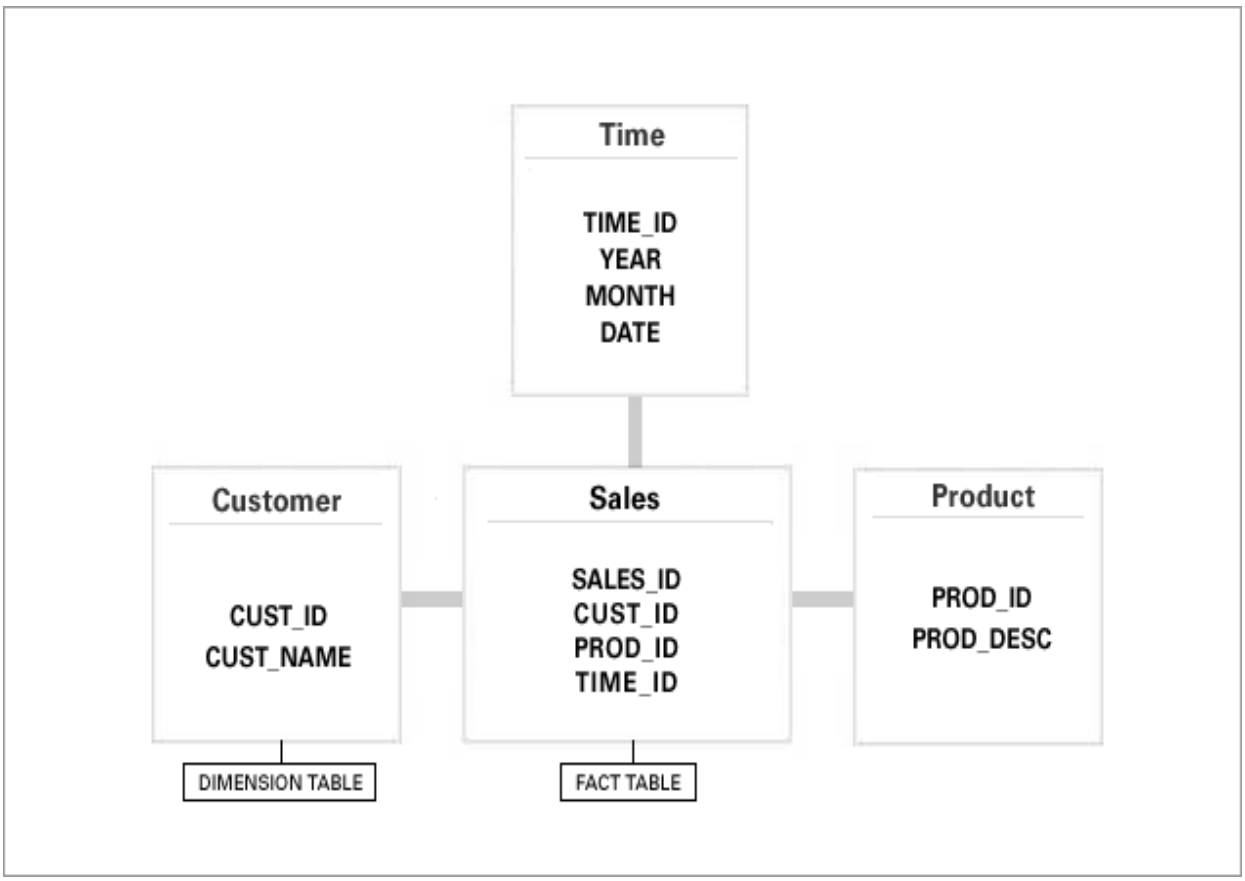


Рисунок 2.8 – Різниця між OLAP- підходом і реляційним доступом в реляційної формі даних, що зберігаються в класичній схемі «Зірка» [17]

Продажі продукту Fiber Cables в HP, як відсоток продажів продукту Serial Cables в Microsoft і т.д. Теоретично можна створити MV, одне для кожної комбінації (4 x 4 x 4 = 64 MV). Крім простору, має місце потреба в часі і ресурсах бази даних для освіження матеріалізованих уявлень, коли їх дані зміняться, і в усе це залучені тисячі елементів. Отже, число створюваних і керованих M.B. стає величезним.

За контрастом з цим, куб – це єдиний сегмент, який однаково легко справляється з будь-якими типами запитів. Хоча обидва вони (MV і куб) використовуються при проектуванні сховищ даних для більш швидкої обробки сумарних, а не для (детальних) даних OLTP, проте між ними є фундаментальна різниця: в той час як MV зберігають попередньо обчислені (pre-computed) результати, щоб уникнути з'єднань (joins) та інших операцій агрегування, куби зберігають сирі / вихідні дані і обчислюють більшість сум на льоту, так як деякі

суми вже створені. Куб вирішує, які агрегування корисні, і створює тільки їх. У всіх інших випадках суми обчислюються на льоту. Так як шляхи доступу через масиви (кубів) засновані на обчисленнях, то вибірка даних в кубах відбувається значно швидше, ніж в реляційних таблицях типу MV.

OLAP-об'єкти, такі як куби, зберігаються в спеціальних областях бази даних, які називаються аналітичними просторами – Analytic Workspace (AW). База даних може включати одне або більше AW. AW зберігаються в таблицях, з префіксом AW $.

Хоча куби даних в Oracle Database не нові, в версіях, що передують Oracle Database 11g, доступ до них виконувався трохи інакше. (Адже СУБД Oracle початково і перш за все реляційна СУБД.) Подання куба в цих версіях – це нерідна (non-native) концепція, особливо, коли це зачіпає MV.

MV володіють деякими цікавими можливостями: автоматичне переписування запиту (automatic query rewrite), коли користувальницький запит переписується автоматично, інкрементні (incremental refreshes), коли оновлюються тільки частини MV і так далі. З іншого боку, MV – це уявлення реляційної природи, нерідне для OLAP-кубів.

3 АЛГОРИТМИ АГРЕГАЦІЇ ДАНИХ У СОЦІАЛЬНИХ МЕРЕЖАХ

3.1 Алгоритми дослідження соціальних медіа

Вимоги до методології дослідження можуть бути згруповані в: дані, аналітику і можливості [20]. Розглянуто більш детально кожний з них.

Дані. Дослідникам для проведення дослідження необхідний онлайновий доступ до історичних і даних в реальному часі в соціальних мережах, особливо даних з головних джерел:

– соціальна мережа медіа – доступ до комплексних наборів історичних даних, а також доступ в реальному часі до джерел, можливо, з 15-ти хвилиною затримкою часу, як і з Thomson Reuters і фінансових даних Bloomberg.

– новини даних доступ до історичних даних і в режимі реального часу набори даних новин;

– громадські дані доступ до скрапенгу (scraped) і архіву важливих суспільних даних, що доступні через RSS-канали, блоги або відкриті урядові бази даних;

– програмовані інтерфейси – дослідники також повинні мати простий доступ до інтерфейсу прикладного програмування (API), щоб очистити і зберігати інші доступні джерела дані, які не можуть бути автоматично зібрані.

Аналітика. В даний час дані соціальні медіа, як правило, або доступні за допомогою простих загальних процедур або вимагають від дослідника програмувати їх аналітику мовами: MATLAB, Java або Python. Як вже говорилося вище, дослідники вимагають:

– аналітику dashboards – ніпрограмні інтерфейси необхідні для надання того, що називають «глибокий» доступ до «сирих» даних;

– цілісні дані аналізу – інструменти, необхідні для об'єднання (і ведення аналітики по горизонталі) кількох соціальних медіа та інших наборів даних;

– візуалізацію даних (Data Visualization) як інструменти візуалізації, за допомогою яких необхідна інформація може бути візуалізована в деякій

схематичною формі з метою передачі інформації чітко і ефективно за допомогою графічних засобів

Можливості. І, нарешті, величезний обсяг даних соціальних медіа, що генерується, наводить аргументи на користь національних і міжнародних установ, які будуть створені для підтримки соціальних досліджень ЗМІ (наприклад. Wharton Research Services Data https://wrds-web.wharton.upenn.edu):

– при зберіганні даних обсяг даних соціальних медіа знаходиться за межами більшості окремих установ. Зберігання потрібно як для основних джерел даних (наприклад, Twitter), так і для джерел, зібраних за індивідуальними проектами і архівуватимумться для подальшого використання іншими дослідниками, в тому числі у освітніми установами;

– обчислювальні можливості – дистанційно доступні обчислювальні засоби необхідні для: захисту доступу до збережених даних; хостингу аналітики і інструментів візуалізації; забезпечення обчислювальних ресурсів.

Проблеми. Більшість нинішніх ресурсів соціальних медіа є для отримання доступу. Аналітичні інструменти, що надаються постачальниками часто прив'язані до одного набору даних, є обмеження в аналітичній здатності, і вартість даних роблять їх дорогими у використанні. Більше число потужних комерційних платформ, таких як ті, що поставляються SAS і Thomson Reuters не мають «освітньої ліцензії» і унеможливлюють процес досліджень освітніми установами.

3.2 Технології для аналізу даних користувача з соціальних мереж

Аналітичне агентство Gartner в 2012 році опублікувало звіт під назвою «Цикл ажіотажу для країн, що розвиваються технологій» [21]. Згідно звіту, технології «Соціальна аналітика» і «Великі дані» в даний час перебувають на т.зв. «Піке завищених очікувань». Зокрема, дослідженнями соціальних даних активно

займаються університети -Меллон Карнегі, Стенфорд, Оксфорд, INRIA, а також компанії Facebook, Google, Yahoo!, LinkedIn і багато інших. Компанії-власники сервісів онлайнових соціальних мереж (Facebook, Twitter) активно інвестують в розробку вдосконалених інфраструктурних (Cassandra, Presto, FlockDB) і алгоритмічних (нові алгоритми пошуку і рекомендації користувачів, товарів і послуг) рішень для обробки великих масивів призначених для користувача даних. Виникають і успішно розвиваються комерційні компанії, що надають послуги з доступу до сховищ соціальних даних (GNIP), збору соціальних даних за заданими сценаріями (80legs), соціальної аналітики (DataSift), а також розширенню існуючих платформ за допомогою соціальних даних (FlipTop).

Таким чином, фахівці з дослідницьких центрів і компаній по всьому світу використовують дані соціальних мереж для моделювання соціальних, економічних, політичних та інших процесів від персонального до державного рівня з метою розробки механізмів впливу на ці процеси, а також створення інноваційних аналітичних і бізнес-додатків і сервісів.

Разом з тим, при роботі з соціальними даними потрібно брати до уваги такі фактори, як нестабільність якості призначеного для користувача контенту (спам і неправдиві акаунти), проблеми із забезпеченням приватності особистих даних користувачів при зберіганні і обробці, а також часті поновлення користувальницької моделі і функціоналу. Все це вимагає постійного вдосконалення алгоритмів розв'язання різних аналітичних і бізнес-задач.

Обробка соціальних даних вимагає також розробки відповідних алгоритмічних і інфраструктурних рішень, що дозволяють враховувати їх розмірність. Наприклад, база даних соціальної мережі Facebook на сьогоднішній день містить більше 1 мільярда користувальницьких акаунтів і понад 100 мільярдів зв'язків між ними. Кожен день користувачі додають більше 200 мільйонів фотографій і залишають більше 2 мільярдів коментарів до різних об'єктів мережі. На сьогоднішній день більшість існуючих алгоритмів, що дозволяють ефективно вирішувати актуальні завдання, не здатні обробляти дані подібної розмірності за прийнятний час. У зв'язку з цим, виникає потреба в нових

рішеннях, що дозволяють здійснювати розподілену обробку і зберігання даних без істотної втрати якості результатів.

Стек технологій для аналізу даних користувача з соціальних мереж містить:

– фреймворк для збору реальних призначених для користувача даних шляхом звернення до веб-інтерфейсів соціальних сервісів;

– інструмент для генерації випадкових соціальних графів із заданими структурними властивостями;

– методи обробки текстових даних користувачів соціальних мереж: визначення демографічних атрибутів шляхом лінгвістичного аналізу профілів і текстів повідомлень, а також пошук описів подій в повідомленнях;

– методи обробки мережевих даних (соціальних зв'язків між користувачами), а саме метод ідентифікації користувачів різних соціальних мереж та метод пошуку співтовариств користувачів;

– метод вимірювання інформаційного впливу і пошуку найбільш впливових користувачів.

3.3 Фреймворк для збору даних

Веб-інтерфейси соціальних мереж є джерелами даних реального часу і призначені для перегляду і взаємодії зі сторінками соціальної мережі в веб-браузері або для використання даних користувачів спеціалізованими додатками. оскільки сценарії використання інтерфейсів соціальних мереж не передбачають автоматичного збору даних безлічі користувачів з метою побудови соціального графа, то виникає ряд проблем:

– приватність даних – часто доступ до даних користувачів дозволений тільки для зареєстрованих і авторизованих учасників мережі, що вимагає підтримки емуляції користувальницької сесії за допомогою спеціальних облікових записів (акаунтів);

– слабка структурованість даних – у багатьох випадках програмні інтерфейси (API) соціальних мереж мають обмежений функціонал, що вимагає підтримки отримання з допомогою призначеного для користувача веб-інтерфейсу статичних копій HTML-сторінок, коректної обробки їх динамічної частини (включаючи виконання асинхронних запитів до сервера соціальної мережі), вилучення потрібних даних за допомогою алгоритму і / або шаблону і побудови їх структурованого уявлення, зручного для подальшої автоматичної обробки;

– обмеження доступу і блокування – з метою запобігання несанкціонованого автоматичного збору даних і обмеження навантаження на інфраструктуру сервісу соціальної мережі власники сервісів часто вводять явні чи приховані обмеження на допустима кількість запитів від одного користувача акаунта і / або IP-адреси в одиницю часу, що вимагає врахування кількості посилаються запитів, а також підтримки динамічної ротації використовуваних для збору даних користувача акаунтів і IP-адрес;

– розмірність даних обумовлює необхідність в паралельному методі збору даних, а також в методах отримання репрезентативної вибірки користувачів соціальної мережі (семплірованіє).

У зв'язку з постійною необхідністю отримання великих наборів даних з соціальних мереж, був розроблений фреймворк для збору даних з різних інтернет-сервісів.

Реалізовано кілька способів отримання репрезентативних вибірок користувачів соціальних мереж: семплірованіє методом обходу в ширину (пошук в ширину, breadth-first search BFS) [22], по Метрополісу-Гастінгс (Metropolis-Hastings Random Walk – MHRW) [23] і методом «лісової пожежі» (Forest Fire -FF) [24].

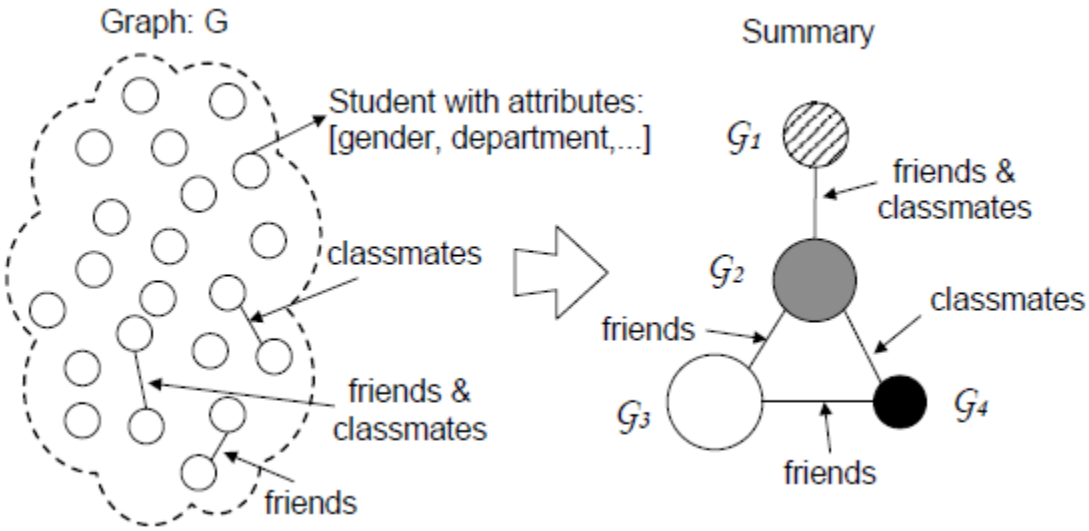
При реалізації механізму автоматичного вибору облікового запису соціальної мережі для кожного запиту, а також підтримка – проксі з'єднань. Це забезпечує стійкість до блокувань по IP-адресами і облікових записів. Крім того, фреймворк підтримує багатопоточний скачування.

Однією з ключових особливостей такого фреймворка є можливість швидко реалізувати нові сценарії скачування і методи семплінгу.

3.4 Математична модель агрегації для узагальнення графа

Графи це потужний інструмент для моделювання даних в різних галузях. Вузли в графах, як правило, представляють собою реальні об'єкти світу і ребра вказують на відносини між об'єктами. Приклади змодельованих даних у вигляді графів включають в себе соціальні мережі, біологічні мережі, тощо. Часто, вузли мають атрибути, пов'язані з ними.

Наприклад, на рисунку 3.1 (а), вузол, який представляє студента з атрибутами: стать і відділ. Крім того, граф може містити безліч різних типів відносин, таких як друзі і однокласники відносин, показаних на рисунку 3.1 (а).



(а) (б)

Рисунок 3.1 – Узагальнення графа *G* за допомогою агрегації (Graph Summarization)

В багатьох випадках графи дуже великі, з тисячами або навіть мільйони вузлів і ребер. В результаті майже неможливо зрозуміти інформацію, закодовану в великих графах за допомогою простого візуального огляду. Таким чином, ефективні методи узагальнення графа необхідні, щоб допомогти користувачам отримати і зрозуміти основну інформацію.

Більшість існуючих методів узагальнення графа використовують прості статистичні дані для опису характеристик графа.

Користувачам потрібен більш керованим і інтуїтивний метод для узагальнення графів. Метод узагальнення має дозволити користувачам вільно вибирати атрибути і відносини, що представляють інтерес, а потім використовувати ці функції, щоб виробляти невеликі і інформативні резюме. Крім того, користувачі повинні мати можливість контролювати дозвіл одержуваних зведень і «drill-down» (деталізація) або «roll-up» («згортання») інформації, так само як і методи агрегування OLAP в традиційних системах баз даних.

Пропонується модель узагальнення графа (МУГ) на основі угруповування вузлів на атрибутах і парних відносин, це дає короткий граф вхідного графа за допомогою угруповання вузлів на основі вибраних користувачем атрибутів вузлів і зв'язків.

На рисунку 3.1 схематично показана ідея роботи МУГ. На рисунок 3.1 а) представлений граф про студентів (з атрибутами: gender-стать, department – відділ і т.д.) і відносини (classmates – однокласників і friends – друзів) між ними. Слід зазначити, що лише деякі з ребер показані на рисунок 3.1 а). На основі обраних користувачем атрибутів: gender-стать і department-відділ та відносин серед classmates – однокласників і friends – друзів, операція МУГ виробляє короткий граф, показаний на рисунок 3.1 б). Це резюме містить чотири групи студентів і відносини між цими групами. Студенти в кожній групі мають однакову стать і знаходяться в тому ж відділі, і вони ставляться до студентів, що належать одному і тому ж набору відносин груп з друзями та однокласниками. Наприклад, на рисунок3.1 б), кожен студент в групі G_1 має принаймні одного однокласника в

групі G_2. Це компактне узагальнення розкриває основні характеристики про вузли та їх відносин у вихідному графі.

Формально позначено граф G, як (V, Υ), де V є множина вузлів, і$\Upsilon = \{E_1, E_2, ..., E_r\}$ множина типів ребер, з кожним $E_i \subseteq V \times V$, що становить множину ребер певного типу.

Вузли в графі є набір атрибутів, пов'язаних з ними, який позначається як $\Lambda = \{a_1, a_2, ..., a_t\}$. Кожен вузол має значення для кожного атрибута. Ці атрибути використовуються для опису особливостей об'єктів, які представляють вузли. Наприклад, на рисунку 3.1 а), вузол, який представляє Студент може мати атрибути, які представляють стать студента і відділ.

Різні типи ребер в графі відповідають різним типам відносин між вузлами, такими як друзі та однокласник. Два вузли можуть бути з'єднані різними типами ребер, наприклад, два студента можуть бути однокласниками і друзями одночасно.

Для простоти викладу буде позначити множину вершин графа G, як $V(G)$, набір атрибутів, як $\Lambda(G)$, фактичне значення атрибута a_i. для вузла v як $a_i.(v)$, безліч типів ребер таких як $\Upsilon(G)$, а множина ребер типу Ei як $Ei(G)$. Крім того, ми будемо позначати потужність множини S в якості $|S|$.

Операція узагальнення графа на основі угрупування вузлів на атрибутах і парних відносин виробляє короткий граф через однорідне угрупування вузлів вхідного графа, засноване на обраних користувачем атрибутах вузлів і зв'язках.

Для того, щоб почати формальне визначення операції УГОУВ, ми спочатку визначимо поняття вузла-угруповання.

Крок 1. Вузол – угруповання графа G $\Phi = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_k\}$ називається вузлом-угруповання G, тоді і тільки тоді, коли:

$$\forall \mathcal{G}_i \in \Phi, \mathcal{G}_i \subseteq V(G) \ and \ \mathcal{G}_i \neq \emptyset, \tag{3.1}$$

$$\bigcup_{\mathcal{G}_i \in \Phi} \mathcal{G}_i = V(G), \tag{3.2}$$

для $\forall \mathcal{G}_i, \mathcal{G}_j \in \Phi$ і $(i \neq j), \mathcal{G}_i \cap \mathcal{G}_j = \emptyset.$ (3.3)

Вузол-угрупування розділяє вузли в графі на непересічні підмножини. Кожна підмножина $\mathcal{G}_i$ називається групою. Коли немає ніякої двозначності, ми просто називаємо угруповання вузла – угрупованням. Для цього угруповання *Ф* з *G*, група, який належить вузол *v* і позначається як *Ф(v)*. Крім того, визначено розмір угруповання, як кількість груп, що вона містить.

Тепер визначається часткове відношення до порядку $\preccurlyeq$ на безлічі всіх груп графа.

Крок 2. Домінування зв'язків. Для графа *G*, угруповання *Ф* домінує угруповання *Ф'*, позначається *Ф'*$\preccurlyeq$ *Ф* тоді і тільки тоді, коли $\forall \mathcal{G}'_i \in \Phi', \exists \mathcal{G}_j \in \Phi$, де $\mathcal{G}'_i \subseteq \mathcal{G}_j$. Ставлення домінування $\preccurlyeq$ рефлексивно, анти-симетрично і транзитивне, отже, є відношенням часткового порядку. Далі визначено особливий вид угруповання на основі набору обраних користувачем атрибутів.

Крок 3. Атрибути сумісності угруповання.

Для набору атрибутів $A \subseteq \Lambda(G)$ угруповання *Ф* сумісні з атрибутами *A* чи просто *A*-сумісні, якщо він задовольняє наступним чином: $\forall u, v \in V$ якщо $\Phi(u) = \Phi(v)$ тоді $\forall a_i \in A$, $a_i(u) = a_i(v)$

Якщо угрупування *Ф* сумісне з *A*, позначити його як Φ_A. У кожній групі *A*-сумісного угруповання, кожен вузол має точно такі ж значення для набору атрибутів *A*.

Не може бути більше одного сумісного угрупування з *A*. Фактично тривіальне угруповання, в якій кожен вузол це група завжди сумісна з будь-яким набором атрибутів.

Далі, доведено, що серед усіх *A*-сумісних груп графа, існує глобальний максимум угруповання по відношенню до домінантності відносини $\preccurlyeq$.

Крок 4. Доведено, що у множині всіх *A*-сумісних груп графа *G*, позначається як S_A, $\exists \Phi_A \in S_A$, $\forall \Phi'_A \in S_A$, $\Phi'_A \preccurlyeq \Phi_A$

Припускається, що не існує глобального максимуму *A*-сумісного угруповання, але більше, ніж одного максимального угруповання. Тоді для будь-яких двох таких максимальних угруповань Φ_1 і Φ_2, буде побудовано нове *A*-сумісне угруповання Φ_3 таким чином, що $\Phi_1 \preccurlyeq \Phi_3$ і $\Phi_2 \preccurlyeq \Phi_3$, що суперечить

припущенню, що суперечить припущенню, що Φ_1 і Φ_2 є максимальним сумісним угрупованням.

Припускається, що $\Phi_1 = \{\mathcal{G}_1^1, \mathcal{G}_2^1, ..., \mathcal{G}_s^1\}$ і $\Phi_2 = \{\mathcal{G}_1^2, \mathcal{G}_2^2, ..., \mathcal{G}_t^2\}$.

Побудовано двочастковий граф $\Phi_1 \cup \Phi_2$, як показано на рисунок 3.2.

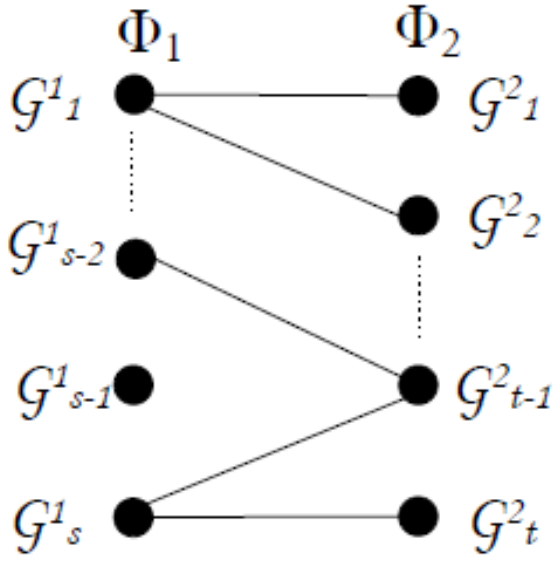


Рисунок 3.2 – Двочастковий граф Φ_3

Вузли в двочастковому графі є групи Φ_1 і Φ_2. І є ребро між $\mathcal{G}_i^1 \in \Phi_1$ і $\mathcal{G}_j^2 \in \Phi_2$ якщо $\mathcal{G}_i^1 \cap \mathcal{G}_j^2 \neq \emptyset$. Після побудови двочасткового графу, розкладається цей граф на компоненти зв'язності C_1, C_2, ..., C_m.

Для кожного підключеного компонента C_k, об'єднано групи всередині цього компонента $\cup(C_k)$. Тепер можна побудувати нове угруповання $\Phi_3 = \{\cup(C_1), \cup(C_2), ..., \cup(C_m)\}$. Легко бачити, що $\Phi_1 \preccurlyeq \Phi_3$ і $\Phi_2 \preccurlyeq \Phi_3$

Тепер доведемо, що Φ_3 сумісний з A. З визначення A-сумісних угруповань, якщо $\mathcal{G}_i^1 \cap \mathcal{G}_j^2 \neq \emptyset$ вузли в $\mathcal{G}_i^1 \cup \mathcal{G}_j^2$ всі мають однакові значення атрибутів.

Тепер, побудоване нове А-сумісне угруповання Φ_3 таким чином, що $\Phi_1 \preccurlyeq \Phi_3$ і $\Phi_2 \preccurlyeq \Phi_3$ Це суперечить попереднім припущенням, що Φ_1 і $\Phi_2$2 два різних максимальні A-сумісні угруповання. Таким чином, існує глобальний максимум A-сумісного угруповання.

Позначивши цей глобальний максимум A-сумісне угруповання Φ_A^{max}. Таким чином, Φ_A^{max} також А-сумісне угруповання з мінімальною кількістю елементів. Насправді, якщо розглядати кожен вузол в графі як запис даних, а потім Φ_A^{max}

дуже схожий в результаті з групової операцією для цих записів даних на атрибути, які є в системах реляційних баз даних.

А-сумісні угруповання облікового запису є тільки для атрибутів вузла. Проте, вузли не тільки атрибути, а й беруть участь у парних відносинах, представлених ребрами. Далі розглянемо відносини при угрупованню вузлів.

Для угруповання Φ позначено сусіда-групи вузла v в E_i як $NeighborGroups_{\Phi,E_i}(v) = \{\Phi(u)|(u,v) \in E_i\}$.

Тепер визначено угруповання, що сумісне з обома атрибутами вузла і відносин.

Крок 5. Атрибути і відносини сумісності угруповання.

Для набору атрибутів $A \subseteq \Lambda(G)$ і набір типів відносин $R \subseteq \Upsilon(G)$, угруповання Φ сумісний з атрибутами A і відносини типів R або просто (A, R) сумісний, якщо він задовольняє:

$$\Phi - A\text{-сумісний,} \tag{3.4}$$

$$\forall u, v \in V(G), \text{якщо } \Phi(u) = \Phi(v), \text{ потім } \forall E_i \in R, \tag{3.5}$$

$$NeighborGroups_{\Phi,E_i}(u) = NeighborGroups_{\Phi,E_i}(v) \tag{3.6}$$

Якщо угруповання Φ сумісне з A і R, також потрібно позначати його як $\Phi_{(A, R)}$. У кожній групі (A, R) групування сумісне, всі вузли є однорідними з точки зору як атрибути A і відносини в R. Іншими словами, кожен вузол всередині групи має точно такі ж значення для атрибутів A, і знаходиться поруч з вузлами в тому ж наборі груп для всіх відносин в R.

Як приклад, припущено, що угруповання сумісне з атрибутами «стать» і «відділ», і відносини «однокласник» і «друг». Потім, наприклад, кожен студент (вузол) в групі $\mathcal{G}_2$, має ту ж саму стать і відділ значень атрибутів, і є одним деяким студентом (ами) в $\mathcal{G}_3$, однокласник деякого студента (ів) в $\mathcal{G}_4$, і один до деякого студента (ів), а також однокласником до деякий студент (и) в $\mathcal{G}_1$.

З огляду на угруповання $\Phi_{(A, R)}$, можна зробити висновок, відносини між групами з відносин між вузлами в R. Для кожного типу ребра $E_i \in R$, визначимо

відповідну групу відносин як $E_i(G, \Phi_{(A,R)}) = \{(\mathcal{G}_i, \mathcal{G}_j) |\ \mathcal{G}_i, \mathcal{G}_i \in \Phi_{(A,R)}$ і $\exists u \in \mathcal{G}_i, v \in \mathcal{G}_j \text{ s.t. } (u, v) \in E_i\}$. Насправді, за визначенням з *(A, R)*, сумісними з угрупованнями, якщо є один вузол група примикає до деякого вузла (ів) в іншій групі, а потім кожен вузол в першій групі примикає до деякого вузла (ів) в другій.

Аналогічно атрибути сумісних груп, не може бути більше одного угруповання сумісного з заданими атрибутами і відносинами. Угруповання, в якому кожен вузол утворює групу завжди сумісні з будь-якими заданими атрибутами і відносинами.

Крок 6. Серед всіх *(A, R)* сумісними з угрупованнями існує глобальний максимум угруповання по відношенню до домінантності відносини $\preccurlyeq$.

У множині всіх *(A, R)* сумісних угруповань графа *G*, позначається як $S_{(A,R)}$, $\exists \Phi_{(A,R)} \in S_{(A,R)}$, $\forall \Phi'_{(A,R)} \in S_{(A,R)}$, $\Phi'_{(A,R)} \preccurlyeq \Phi_{(A,R)}$

Отже операція УГОУВ приймає в якості вхідних даних граф *G*, набір атрибутів $A \subseteq \Lambda(G)$, і набір ребер $R \subseteq \Upsilon(G)$і отримується сумарний узагальнюючий граф G_{total} , де $V(G_{total}) = \Phi^{max}_{(A,R)}$ і $\Upsilon(G_{total}) = \{E_i(G, \Phi^{max}_{(A,R)}) | E_i \in R\}$

Таким чином, зрозуміло, що операція УГОУВ виробляє узагальнений граф вхідного графа на основі вибраних користувачем атрибутів і зв'язків.

Вузли цього короткого графа відповідають групам в максимумі *(A, R)*, сумісні з угрупованнями.

А ребра цього короткого графа є групові відносини виведені з вузла відносин в *R*.

Отже представлена операція агрегації УГОУВ заснована на угрупування графа. Цей метод дозволяє користувачам вільно вибирати атрибути вузлів і відносини, які становлять інтерес, і виробляють угрупування на основі певних функцій.

3.5 Генерація випадкових соціальних графів

Незважаючи на наявність коштів для збору даних з соціальних мереж і великої кількості доступних наборів даних, актуальною є задача створення моделей випадкових соціальних графів і інструментів для генерації випадкових графів із заданим набором властивостей. для достовірного тестування методів аналізу соціальних даних вони повинні бути застосовані до безлічі наборів даних з різними властивостями. До розміру початкового графа, середнього ступеня вершини, коефіцієнта кластеризації та інших структурних властивостей. Збір необхідних для достовірного тестування реальних даних утруднений не тільки внаслідок тимчасових витрат на скачування і обробку великих масивів слабоструктурованої інформації, але і в силу складності управління процесом збору з метою отримання набору даних з конкретним набором властивостей.

Були розроблені модель і оригінальний метод для генерації випадкових графів, що володіють основними властивостями соціальних мереж (розподіл ступенів, діаметр, коефіцієнт кластеризації і т.д.) і заданою структурою спільнот користувачів. Для кожного користувача здійснюється генерація атрибутів профілю, соціальних зв'язків, спільнот і текстових повідомлень. Запропонований метод має розподілену реалізацію на основі фреймворку Apache Spark1 (http://spark.incubator.apache.org/), що дозволяє створювати випадкові графи великої розмірності для тестування продуктивності і точності методів аналізу соціальних даних.

При заповненні свого профілю в соціальній мережі користувачі найчастіше помилково або навмисно не заповнюють деякі поля або дають неправдиву інформацію про факти своєї біографії, інтереси та вподобання. Крім того, в тематичних мережах (Twitter, YouTube) призначений для користувача профіль часто обмежений набором базових атрибутів, недостатнім для вирішення багатьох задач, які передбачають персоналізацію результатів.

Таким чином, актуальні методи часткової ідентифікації авторів повідомлень за значеннями їх демографічних атрибутів. Зокрема, в інтернет-системах маркетингу і рекомендацій особливу важливість представляє визначення

демографічних атрибутів користувача для цільового просування товарів і послуг в групах користувачів з однаковими значеннями атрибутів. Крім інтернет-сервісів, такі демографічні характеристики знаходять застосування в різних дисциплінах: соціологія, психологія, кримінологія, економіка, управління персоналом та ін.

Демографічні атрибути можна умовно розділити на категоріальні (пол, національність, раса, сімейний стан, рівень освіти, професія, працевлаштованість, релігійні і політичні погляди) і чисельні (вік, рівень доходів). Умовність поділу пов'язана з тим, що значення числового атрибута можна відобразити в набір категорій і надалі розглядати цей атрибут як категоріальний. Зокрема, значення віку можна розділити на кілька вікових категорій, що часто застосовується на практиці.

Метод визначення демографічних атрибутів користувачів мережі за текстами їх повідомлень складається з наступних етапів:

- побудова вихідного набору даних;
- попередня обробка тексту;
- побудова простору ознак опису;
- відбір інформативних ознак;
- навчання;
- класифікація.

Всі етапи, за винятком першого, виконуються окремо для кожного атрибута.

На етапі побудови вихідного набору даних проводиться збір даних користувачів з мережі. Для кожного користувача спочатку запитується тільки його профіль у мережі. Наприклад Twitter. При наявності в ньому посилання на профіль того ж користувача в мережі Facebook (в якій набір призначених для користувача атрибутів істотно більше, ніж в Twitter) запитуються і зберігаються всі доступні повідомлення користувача з мережі Twitter. Після чого для поточного користувача запитується і зберігається його профіль у мережі Facebook, з якого здобуваються зазначені користувачем значення його атрибутів.

На етапі попередньої обробки тексту до текстів отриманого на попередньому етапі набору даних застосовується метод визначення мовної

приналежності тексту. Після цього дані користувачів розподіляються в різні набори даних в залежності від мови користувача.

Крім того, на цьому етапі здійснюється фільтрація повідомлень, авторство яких не належить користувачу (ретвіти). Оскільки цитування повідомлень інших користувачів є досить популярним способом поширення інформації в мережі Twitter, цей крок попередньої обробки особливо важливий для підвищення точності методу.

Таким чином, елементом набору даних для кожного атрибута і мови є набір символьних рядків, отриманих з текстів повідомлень і профілю одного користувача в Twitter, а також значення атрибута у даного користувача в Facebook.

На етапі побудови простору ознак опису з повідомлень користувачів витягуються лінгвістичні ознаки. З отриманих токенов будується набір ознак у вигляді N-грам розміром від 1 до 3 з урахуванням порядку токенов.

Кожен тип ознак представлений двома підтипами: з урахуванням і без урахування регістра символів.

Підсумковий вектор ознак для користувача є бінарним, тобто містить тільки інформацію про наявність чи відсутність ознаки в його текстових даних. Кількість примірників однієї ознаки ігнорується.

На етапі відбору інформативних ознак застосовується метод, заснований на розрахунку умовної взаємної інформації. Виробляється ітеративний відбір тих ознак, які містять найбільшу кількість інформації про значення атрибута і при цьому істотно відрізняються від ознак, обраних на попередніх ітераціях. Таким чином, кожна ознака результуючого набору високо інформативний і слабо залежить від інших ознак.

На етапі навчання проводиться побудова моделі класифікації з використанням он-лайнового пасивно-агресивного алгоритму.

На етапі класифікації в якості вхідних даних використовуються тексти повідомлень і поля профілю довільного користувача. Виконується алгоритм

класифікація для заданого мови і атрибута. Результатом є значення атрибута обраного користувача.

Повідомлення користувачів соціальних мереж складають істотну частку текстового контенту сучасного Інтернету. Крім того, соціальні мережі часто виступають в ролі неформальних ЗМІ, де будь-який користувач може опублікувати повинне повідомлення про події, що відбуваються (Інформаційних приводи).

Разом з тим, щоб автоматично завантажувати набори повідомлень про невідому заздалегідь подію є нетривіальним завданням в силу наступних чинників:

– великий обсяг вхідних даних (наприклад, Twitter користувачі публікують кілька тисяч повідомлень щосекунди);

– велика кількість нерелевантних/неінформативних повідомлень;

– користувачі можуть по-різному описувати одну і ту саме подію;

– різні події можуть збігатися за часом;

– складність поділу події і його підподій (наприклад, Олімпійські ігри і конкретний футбольний матч в рамках цієї події).

Для пошуку подій в контенті повідомлень користувачів Twitter була розроблена спеціалізована система, робота якої ґрунтується на послідовному виконанні наступних кроків [29,30]:

– побудова сигналів для кожного токена (послідовності символів) з використанням інформації про частоту його появи в корпусі в різні моменти часу;

– застосування вейвлетного аналізу до отриманих сигналів;

– видалення незначних токенов з використанням авто-кореляції сигналів;

– побудова матриці крос-кореляції сигналів токенов;

– пошук подій як наборів токенов шляхом кластеризації отриманої матриці;

– пошук повідомлень, що описують кожну подію, за допомогою методу - мульті документного реферування по документам, що містить токени з кожного набору.

Ця система має наступні переваги: не вимагає даних про користувачів і доступу до зовнішніх баз знань; не вимагає навчання; можливість інкрементальної обробки при надходженні нових повідомлень; можливість пошуку подій в різних часових масштабах – годину, день, тиждень тощо.

Як приклади знайдених подій можна привести такі набори токенів з повідомлень користувачів:

– «Вибори»: `#electionday, #electionday, congratulate, decide, elect, Florida, friends, marijuana, nice, people, report,Romney;`

– «Спорт»: `award, back, black, final, Friday, game,team, turn, watching, world`.

Потенційною сферою застосування є пошук і складання короткого реферату реакції користувачів на невідомі або заздалегідь певні оффлайн – і онлайн-події. Прикладами таких подій можуть служити черговий випуск телевізійного шоу, спортивні події, стихійні лиха, політичні події, запуск нового сервісу для користувачів соціальної мережі і т.д.

Однією з фундаментальних проблем при використанні соціальної інформації про користувача є її фрагментованість серед множини різних он-лайн соціальних мереж. Щороку з'являється безліч як універсальних, так і нішевих соціальних сервісів, і для активних користувачів Інтернет типово мати кілька профілів в різних соціальних мережах. Незважаючи на те, що існують спроби по забезпеченню єдиного способу взаємодії між різними соціальними платформами (наприклад, OpenSocial), вони не отримали широкого застосування, а нові соціальні сервіси продовжують з'являтися.

Ідентифікація користувача в різних соціальних мережах дає змогу отримати більш повну картину про соціальну поведінку даного користувача в мережі Інтернет. Виявлення акаунтів, що належать одній людині, в декількох соціальних мережах, дозволяє отримати більш повний соціальний граф, що може бути корисно в багатьох задачах, таких як інформаційний пошук, інтернет-реклама, рекомендаційні системи і т.д.

Оскільки пошук акаунтів користувача в різних мережах в загальному випадку вимагає наявності актуальних даних про всіх користувачів даних мереж, доцільно обмежити простір пошуку найближчими сусідами якого або користувача, акаунти якого в досліджуваних мережах відомі.

Таким чином, завдання ідентифікації користувачів в різних соціальних мережах в локальній перспективі має на увазі зіставлення акаунтів користувачів в рамках списків контактів деякого центрального користувача в різних соціальних мережах. Таке завдання часто виникає при роботі з контактами користувачів в соціальних мета-сервісах, які, зокрема, можуть служити для об'єднання новинних потоків в підтримуваних соціальних сервісах або надання єдиної системи обміну повідомленнями. Подібна задача виникає також при використанні функції автоматичного об'єднання контактів з різних джерел (телефонна книга, соціальні мережі, месенджери), поширеною в сучасних мобільних пристроях.

Був розроблений метод розв'язання задачі ідентифікації користувачів різних соціальних мереж, яка зводиться до пошуку різних варіантів віртуальних особистостей одного і того ж користувача в декількох соціальних мережах. На основі графічної ймовірнісної моделі умовного випадкового поля була розроблена оригінальна модель, заснована на схожості віртуальних особистостей користувачів по атрибутам їх профілів і зв'язків з іншими користувачами. Розроблений метод використовує соціальні зв'язки обох розглянутих соціальних мереж шляхом порівняння оригінальних списків контактів, природним чином комбінуючи їх з інформацією атрибутів профілів, завдяки чому позбавлений багатьох недоліків існуючих методів ідентифікації користувачів.

Метод пошуку співтовариств користувачів.

Природною властивістю людського суспільства є тенденція до об'єднанню в різні спільноти. Аналогічна картина спостерігається в соціальних мережах, де користувачі об'єднуються явно (використовуючи засоби мережі для створення груп і взаємодії всередині них) або неявно (встановлюючи зв'язку на основі загальної або схожою діяльності, ролі, соціального кола, інтересу або інших властивостей).

Пошук спільнот користувачів є важливим інструментом вивчення й аналізу соціальних мереж, що дозволяє досліджувати модульну організацію мережі і використовувати отриману інформацію для вирішення різних завдань [22]. Наприклад, знання про структуру спільнот незамінні для передбачення зв'язків і атрибутів користувачів, розрахунку близькості користувачів в соціальному графі, оптимізації потоків даних в соціальної мережі, деяких аналітичних додатків і т.д.

Інформація про спільноти (модульну структуру) соціальної мережі на глобальному рівні знаходить застосування в системах рекомендацій, фільтрації спаму і багатьох інших додатках. Автоматично певні спільноти найближчих контактів користувача в соціальній мережі можуть застосовуватися для оптимізації потоків вхідної та вихідної інформації (відправити повідомлення тільки спільноті «Колеги», прочитати новини тільки від спільноти «Близькі друзі»).

Метод пошуку неявних спільнот користувачів соціальних мереж на основі соціальних зв'язків між ними використовує алгоритм, що локально імітує людське спілкування між парами індивідуумів, і глобально моделює інфекційний процес. Основою алгоритму є процес обміну мітками спільнот між вершинами відповідно до динамічних правил взаємодії, в ході якого заохочується об'єднання спільнот найближчих контактів окремих користувачів в глобальні спільноти. Додатковим кроком алгоритму є визначення спільнот з недостатньою внутрішньої пов'язаністю і поділ їх на більш зв'язкові підспільноти. Даний метод має наступні особливості: придатність до орієнтованих і неорієнтованих графів; облік ваг на ребрах; пошук як пересічних, так і непересічних спільнот; пошук як локальних (серед найближчих контактів користувача), так і глобальних спільнот; низька обчислювальна складність; можливість розподіленої реалізації в рамках обчислювальної моделі Pregel.

Метою методу вимірювання інформаційного впливу між користувачами в соціальних мережах з орієнтованими зв'язками і переважанням текстового вмісту (на прикладі Twitter) (http://imdemo.at.ispras.ru/demo) є модель, що враховує такі індикатори інформаційного впливу, як близькість інтересів користувачів,

кількість оригінальних повідомлень і цитувань, опублікованих користувачем під впливом інших користувачів, близькість користувачів в соціальному графі, а також факт знаходження користувачів в одних і тих же співтовариствах. Цей метод має низьку обчислювальну складність і має розподілену реалізацію на основі фреймворку Apache Spark, що дозволяє обробляти графи соціальних мереж з населенням понад 1 мільярда користувачів. Може застосовуватися в системах соціальної рекомендації, а також для пошуку тематичних експертів і знаменитостей, які володіють значним інформаційним впливом на конкретного користувача або в масштабі всієї мережі.

4 ОПИС РОЗРОБЛЕНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

4.1 Інформаційні ресурси соціальних медіа

Інформаційні ресурси соціальних медіа в широкому сенсі підрозділяються на тих, хто надає:

– вільно доступні бази даних-сховищ, які можна вільно скачати, наприклад, Wikipedia (http://dumps.wikimedia.org) і дані Enron електронної пошти доступні через комплект http://www.cs.cmu.edu/*enron/ ;

– доступ до даних за допомогою інструментів-джерел, які забезпечують контрольований доступ до своїх даних в соціальних мережах за допомогою спеціальних інструментів, як для полегшення опитування, а також, щоб зупинити скачування всіх початкових даних користувачів зі сховища. Прикладом може служити Google.

Вони поділяються на:

– безкоштовні джерела-сховища, які вільно доступні, але інструменти захисту можуть обмежити доступ до «сирих» даних в сховищі, наприклад, діапазон інструментів, що надаються Google;

– реселлери комерційні джерела-даних, які стягуватимуть плату за доступ до своїх даних в соціальних медіа. Gnip і DataSift забезпечують комерційний доступ до даних Twitter на основі партнерства, і Thomson Reuters – доступ до даних новин;

– доступ до даних за допомогою використання API-соціальних сховищ даних ЗМІ, що забезпечують програмований HTTP-доступ до даних через API (наприклад, Twitter, Facebook і Wikipedia).

Основним відкритим вихідним кодом соціальних медіа є Вікіпедія, яка пропонує безкоштовні копії всього наявного контенту для зацікавлених користувачів. Ці бази даних можуть бути використані для дзеркального відображення запитів до бази даних і аналітики соціальних медіа. Ще один приклад вільно доступних даних для дослідження є дані Світового банку, тобто,

Всесвітній банк – банк даних (http://databank.worldbank.org/data/databases.aspx), який забезпечує понад 40 баз даних, таких як бази гендерної статистиці, харчування, охорони здоров'я та демографічної статистики, глобальні економічні перспективи, показники світового розвитку і глобального розвитку фінансів, тощо. Більшість баз даних можуть бути відфільтровані по країнам / регіонам / часу. Крім того, надаються інструменти, для формування звітів у вигляді таблиць, діаграм або форматів мап.

Але більшість надає комерційні послуги щодо надання доступу до даних в соціальних мережах за допомогою онлайн-інструментів для управління доступом до вихідних даних.

Google за допомогою таких інструментів, як Trends і InSights є хорошим прикладом цієї категорії. Стратегія Google полягає в наданні широкого спектру пакетів, таких як Google Analytics, а не з точки зору дослідників більш корисні програмовані HTTP на основі API. На рисунку 4.1 показано, як Google Trends відображає конкретний пошуковий запит, в даному випадку «nure». Використовуючи Google Trends ви можете порівняти, як часто згадувалися теми і в яких географічних регіонах теми були найбільше популярними для пошуку.

Існує все більше число комерційних послуг, засобів масової інформації, що надають платні щодо доступу до даних за допомогою простих аналітичних інструментів. (Більш комплексні платформи з великою аналітикою розглянуті в розд. 8.) Крім того, такі компанії, як Twitter обидва обмежують вільний доступ до своїх даних і ліцензування своїх даних для комерційних посередників даних, таких як Gnip і DataSift.

Gnip є найбільшим в світі постачальником соціальних даних, крім того був першим партнером Twitter, Tumblr, Foursquare, WordPress, Disqus, StockTwits та інших провідних соціальних платформ для організації доступу до їх соціальних даних. Gnip забезпечує соціальні дані для клієнтів в більш ніж 40 країнах, і клієнти Gnip в надають соціальну аналітику медіа більш ніж на 95% від списку Fortune 500. У режимі реального часу дані з Gnip можуть бути доставлені для

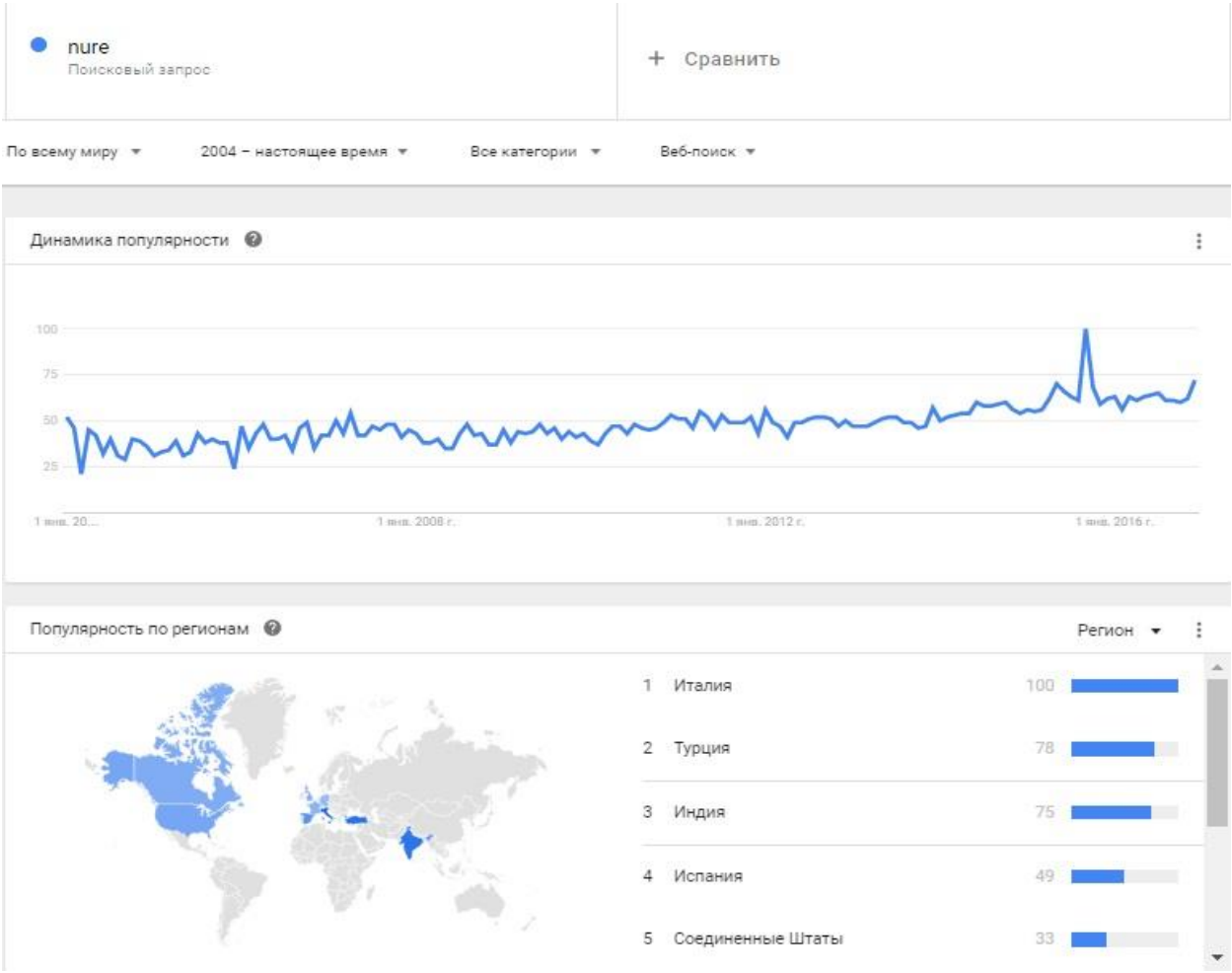


Рисунок 4.1 – Google Trends

кожного окремого виду діяльності або через PowerTrack, патентований інструмент фільтрації, який дозволяє користувачам створювати запити навколо тільки необхідних їм даних. Правила Powertrack можуть фільтрувати потоки даних на основі ключових слів, географічних меж, фраз, типу контенту або засобів масової інформації в своїй діяльності. Потім компанія пропонує додаткові дані, таких як Geo Profile, розширення URL і визначення мови для подальшого підвищення вартості наданих даних. Додатково до доступу до даних в режимі реального часу, компанія також пропонує історичні Powertrack і пошук API доступ до Twitter, які дають клієнтам можливість витягнути будь-який твіт після першого повідомлення.

Gnip забезпечує доступ до преміум («Повний доступ» джерела Gnip є видавці, які мають угоду з Gnip перепродавати свої дані) і каналів безкоштовних даних (джерела Gnip з налагодженим державним доступ до API, що забезпечують

доступ до нормованих і консолідованих вільних даних від їх API, але це платні послуги для колекторів даних) за допомогою своєї приладовій панелі (рисунок 4.2). По-перше, користувач бачить тільки канали в приладовій панелі, які були оплачені в рамках договору купівлі-продажу. Щоб вибрати канал, користувач натискає на видавця, а потім обирає конкретний канал від цього видавця, як показано на рисунок 4.2.

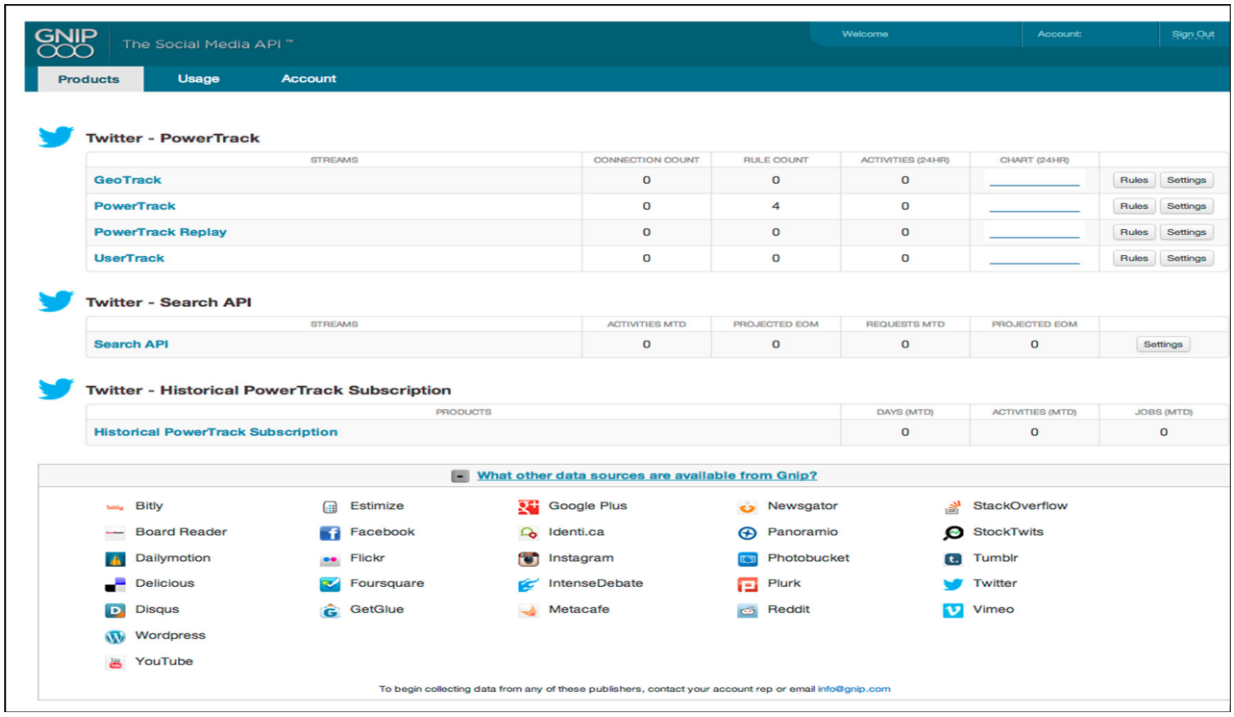


Рисунок 4.2 – Інформаційна панель Gnip, видавців та веб-каналів

Різні види каналів служать різним типам варіантів використання і відповідають різним типам запитів і API кінцевих точок на вихідному API видавця. Після вибору каналу, користувач допомагає Gnip налаштувати його з будь-якими необхідними параметрами, перш ніж вона починає збір даних. Це включає додавання щонайменше одного правила. У розділі «Get Data – Отримання даних» – «Advanced settings – Додаткові налаштування» можна налаштувати, як часто ваш канал запитує вихідний API для даних («query rate – швидкості запиту»). Вибір між рідним форматом даних видавця і формату активності потоків Gnip (XML для Enterprise Data Collector feeds).

4.2 Доступ до даних за допомогою API-інтерфейсів

Вікіпедія (і в цілому Wiki) надає сховище контенту з великими відкритими початковими кодами для користувача (Crowdsourced). Те, що не так широко відомо, що Вікіпедія надає http-інтерфейси, засновані на тому що дозволяють програмований доступ і пошук інформації (тобто, scraping), який повертає дані в різних форматах, включаючи XML. Насправді, API не є унікальним для Вікіпедії, але частина в MediaWiki з відкритим вихідним кодом набору інструментальних засобів може бути використаний з будь-яким MediaWiki на основі вікі-сторінки. API-інтерфейс працює http-вікі на основі запитів, погоджуючись з тим, що містять один або кілька вхідних аргументів і повертають рядки, часто в форматі XML, які можуть бути розібрані і використані клієнтом, що запитує. Інші формати, що підтримуються, включають JSon, WDDX, YAML або PHP серіалізації.

Запит http повинен містити:

– запит, вибір операції редагувати або видаляти;

– запит аутентифікації;

– будь-яку дію, що підтримується, нариклад, запит повертає XML-рядок зі списком перших 10 категорії Вікіпедії з префіксом «Голівуд».

Існує багато інструкцій щодо scraping у Вікіпедії за допомогою / середовища розробки PHP Apache і клієнта HTTP, який здатний передавати GET і PUT запити і відповіді обробки.

Як і в Wikipedia, популярні соціальні мережі, такі як Facebook, Twitter і Foursquare, зробили частину своїх даних доступних через API. Хоча багато медіа сайтів соціальних мереж надають API, але не всі сайти (наприклад, Bing, LinkedIn і Skype) забезпечують доступ до API scraping data – вишкрібання даних. (Scraping data – це метод, в якому комп'ютерна програма вилучає дані з людського сприйняття виводу, що надходить з іншої програми). У той час як все більше і більше соціальних мереж переходять до загальнодоступного контенту, багато провідних мережі обмежують вільний доступ, навіть для вчених. наприклад,

Foursquare з грудня 2013 року не дозволяє приватні реєстрації користувачів під iOS, і в даний час співпрацює з Gnip, щоб забезпечити безперервний потік анонімної реєстрації даних. Дані доступні в двох пакетах: повний Firehose рівень (пожарний шлангової рівень) доступу і відфільтрований – через службу Powertrack Gnip.

Налаштування облікового запису за замовчуванням у Twitter зберігає публічні твіти користувачів, хоча користувачі можуть захистити свій твіт і зробити їх видимими тільки для їх затверджених послідовників Twitter. Проте, менше 10% всіх рахунків Twitter є приватними. Публічні твіти (в тому числі відповіді і згадки) доступні в форматі JSON через API Twitter пошуку для пакетних запитів минулих даних і потокового API для недавніх даних в режимі реального часу.

Search API-запитів Twitter для останніх твітів, що містять певні ключові слова. Є частиною API v1.1 Twitter Rest (він намагається відповідати проектним принципам REST архітектурного стилю, який виступає за Representational State Transfer) і вимагає авторизований додаток (з використанням oAuth, відкритий стандарт для авторизації) перед витяганням будь-якого результату від API.

Streaming API-режимі реального часу потоку твітів, фільтрується ID користувача, ключові слова, географічне розташування або випадкової вибірки.

Можна отримати твіти, що містять певні ключові слова через API пошуку в Twitter (частина REST API v1.1) з наступними API виклику: https://api.twitter.com/1.1/search/tweets.json?q=APPLE і дані в реальному часу з використанням потокового API виклику:

https://stream.twitter.com/1/statuses/sample.json .

Streaming API твіти дозволяє даним бути доступним через фільтрацію (за ключовими словами, ідентифікатори користувачів або місце розташування) або шляхом відбору проб всіх оновлень від обраної кількості користувачів. Рівень доступу за замовчуванням «Spritzer» дозволяє отримати вибірку приблизно 1% від усіх публічних статусів, з можливістю отримання 10% всіх статусів через рівень доступу «Gardenhose» (більше підходить для інтелектуального аналізу даних і

дослідницьких програм). У соціальних медіа, потокові API-інтерфейси часто називають Firehose-канал синдикації, який публікує всі громадські заходи, як вони відбуваються в одному великому потоці. Twitter нещодавно оголосив про програму грантів даних Twitter, де дослідники можуть звернутися, щоб отримати доступ до громадських твітів в Twitter і історичних даних для того, щоб отримати уявлення з масивного набору даних (Twitter налічує більше 500 мільйонів твітів на день); науково-дослідні інститути і вчені не отримуватимуть Firehose-канал – шлангової рівень доступу; замість цього, вони будуть тільки отримувати набір даних, необхідних для їх дослідницького проекту. Дослідники можуть звернутися до нього за наступною адресою: https://engineering.twitter.com/research/data-grants . Результати твітів зберігаються в масиві JSON об'єктів, що містять поля, що наведені нижче:

```
// 1. Приклад Виводу в форматі JSON для
// Twitter API v1 REST

{
// Results page-specific nodes:
"completed_in":0.019, // Seconds taken to generate the results page
"max_id":270492897391034368, // Tweets maximum ID to be displayed up
to
"max_id_str":"270492897391034368", // String version of the max ID
"next_page":"?page=2&max_id=270492897391034368&q=financial%20times&r
pp=1&include_entities=1&r
esult_type=mixed", // Next results page parameters
"page":1, // Current results page
"query":"financial+times", // Search query
"refresh_url":"?since_id=270492897391034368&q=financial%20times&resu
lt_type=mixed&include_entities
=1", // Current results page parameters
// Results node consisting of a list of objects, i.e. Tweets:
"results":[
{
// Tweet-specific nodes:
"created_at":"Sun, 18 Nov 2012 16:51:58 +0000", // Timestamp Tweet
was created at
"entities":{"hashtags":[],"urls":[],"user_mentions":[]}, // Tweet
metadata node
"from_user":"zerohedge", // Tweet author username
"from_user_id":18856867, // Tweet author user ID
"from_user_id_str":"18856867", // String representation of the user
ID
"from_user_name":"zerohedge", // Tweet author username
"geo":null, // Geotags (optional)
"id":270207733444263936, // Tweet ID
```

```
"id_str":"270207733444263936", // String representation of the Tweet
ID
"iso_language_code":"en", // Tweet language (English)
"metadata":{"recent_retweets":6,"result_type":"popular"}, // Tweet
metadata
// Tweet author profile image URL (secure and non-secure HTTP):
"profile_image_url":"http:\/\/a0.twimg.com\/profile_images\/72647502
\/tyler_normal.jpg",
"profile_image_url_https":"https:\/\/si0.twimg.com\/profile_images\/
72647502\/tyler_normal.jpg",
// Tweet source (whether it was posted from Twitter Web or
// another interface):
"source":"&lt;a
href=&quot;http:\/\/www.tweetdeck.com&quot;&gt;TweetDeck&lt;\/a&gt;"
,
"text":"Investment Banks to Cut 40,000 More Jobs, Financial Times
Says", // Tweet content
// Recipient details (if any):
"to_user":null,
"to_user_id":0,
"to_user_id_str":"0",
"to_user_name":null
}
],
// Other results page-specific nodes:
"results_per_page":1, // Number of Tweets displayed per results page
"since_id":0, // Minimum Tweet ID
"since_id_str":"0" // String representation of the 'since_id' value
}
```

Масив JSON складається зі списку об'єктів, що відповідають фільтрам і рядкам пошуку, що поставляються, де кожен об'єкт є твітом і його структура чітко визначається полями об'єкта, наприклад, «created_at» і «from_user». Приклад 1 складається з виходу виклику GET API пошуку Твітера через http://search.twitter.com/search.json?q=financial%20times&rpp=1&include_entities=true&result_type=mixed де параметри визначають, що пошуковий запит «Financial Times», один результат на сторінці, кожен Tweet повинен мати вузол під назвою «об'єкти» (тобто метадані про твіт) і перелік типів результатів «mixed», тобто включають в себе як популярні, так результати і в реальному часі відклику.

Питання конфіденційності Facebook є більш складними, ніж Twitter, а це означає, що багато повідомлень про стан отримати важче, ніж твіти, що вимагає «відкритої авторизації» статусу від користувачів. Facebook в даний час зберігає всі дані у вигляді «objects» (це може бути людина, сторінка, зображення або подія.) і має ряд API-інтерфейсів, починаючи від Public Feed, згадування

ключових слів Insight API. Для того, щоб отримати доступ до властивостей об'єкта, його унікальний ідентифікатор повинен бути відомий, щоб зробити виклик API. Пошук API Facebook (частина Graph API Facebook,) можна отримати за запитом:

https://graph.facebook.com/search?q=QUERY&type=page .

Докладний формат API запиту показаний нижче.

```
// 2. Приклад Facebook Graph API Search Query Format

GET graph.facebook.com
/search?
q={your-query}&
[type={object-type}](#searchtypes)
```

Тут, «QUERY» може бути замінений будь-яким терміном пошуку, а «page» може бути замінена на «post», «user», «page», «event», «group», «place», «checkin», «location» or «placetopic». Результати цього пошуку будуть містити унікальний ідентифікатор для кожного об'єкта. При поверненні ідентифікатором для конкретного результату пошуку, можна використовувати https://graph.facebook.com/ID , щоб отримати більш детальну інформацію сторінок, таких як кількість «likes». Такого роду інформація представляє інтерес для компаній, коли мова йде про впізнаваність бренду і моніторинг конкуренції. В пошукових запитах Facebook Graph API потрібен маркер доступу, що має бути включений в запит. При пошуку сторінок і місць потрібен маркер доступу до додатка, в той час як пошук інших типів вимагає маркера доступу користувача. Замінивши «page» з «post» у вищезгаданому URL пошуку буде повертати все публічні статуси, що містять цей термін пошуку

```
// Facebook Graph API Search Results for
// q='Centrica' and type='page'


{
"id":"96184651725",
"name":"Centrica",
"picture":"http:\/\/profile.ak.fbcdn.net\/hprofile-ak-
snc4\/71177_96184651725_7616434_s.jpg",
"link":"http:\/\/www.facebook.com\/centricaplc",
"likes":427,
```

```
"category":"Energy\/utility",
"website":"http:\/\/www.centrica.com",
"username":"centricaplc",
"about":"We're Centrica, meeting our customers' energy needs
now...and in the future. As a leading
integrated energy company, we're investing more now than ever in
new sources of gas and power.
http:\/\/www.centrica.com",
"location":{
"street":"Millstream, Maidenhead Road",
"city":"Windsor",
"country":"United Kingdom",
"zip":"SL4 5GD ",
"latitude":51.485694848812,
"longitude":-0.63927860415725
},
"phone":"+44 (0)1753 494000",
"checkins":228,
"talking_about_count":5
}
```


Пакетні запити можуть бути відправлені відповідно до процедури, описаної в: https://developers.facebook.com/docs/reference/api/batch/. Інформація про отримання оновлення в реальному часі можна знайти в https://developers.facebook.com/docs/reference/api/realtime/.

Facebook також повертає дані в форматі JSON і тому вони можуть бути отримані і збережені з використанням тих же методів, що використовуються з даними з Твітера, хоча поля різні залежно від типу пошуку, як показано нижче.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Оцінка інтегрального показника ефективності інформаційного пошуку ІПС та ефективність інформаційного пошуку документів, забезпечувана ІПС, оцінюється по двом показникам:

– $k_{\text{п}}$ – коефіцієнт інформаційної повноти;

– $k_{\text{ш}}$ – коефіцієнт інформаційного шуму.

Коефіцієнти $k_{\text{п}}$ і $k_{\text{ш}}$ приймають значення в інтервалі від 0 до 1 (рисунок 5.1). У деяких джерелах ці коефіцієнти виражають у відсотках.

Введено наступні позначення: D – множина документів в інформаційному сховищі, $di \in D$ – i-й документ, $Dj \subseteq D$ – підмножина документів.

У даному контексті під документом будемо розуміти, як власне текстовий або гіпертекстовий документ, так і окремий запис у БД.

Нехай ІПС пред'явлений і-*й* запит. ІПС містить множину документів *Di*, релевантних цьому запиту. У результаті пошуку буде отримана множина *Di**.

Визначимо коефіцієнти повноти та шуму:

$$k_{\text{П}} = \frac{|D_i \cap D_i^*|}{|D_i|}, k_{\text{Ш}} = \frac{|D_i^* \setminus D_i|}{|D_i^*|}.$$

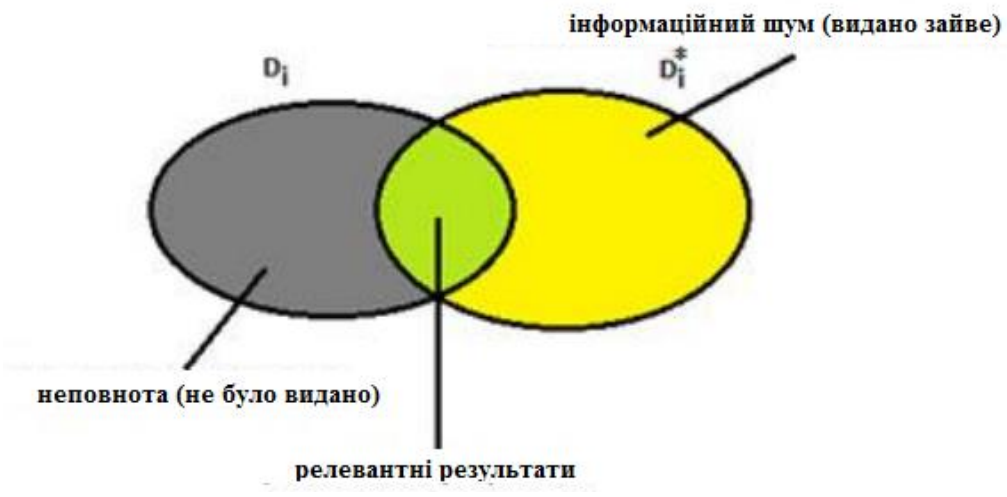


Рисунок 5.1 – Зміст коефіцієнтів повноти та шуму

Ефективність інформаційного пошуку E_1 виражається через коефіцієнти Ш та П, що дозволяє розглядати її як інтегральний показник ефективності інформаційного пошуку ІПС.

В літературі для функції $E1$(Ш, П) замість $k_{ш}$ прийнято використовувати зворотний йому показник – коефіцієнт точності Т.

$$k_{\text{Т}} = 1 - k_{\text{Ш}} = \frac{|D_i \cap D_i^*|}{|D_i^0|}.$$

Таким чином, запишемо дану функцію у вигляді:

$$E_1 = \frac{2k_{\text{Т}}k_{\text{П}}}{k_{\text{Т}}+k_{\text{П}}}.$$

Оцінка ефективності роботи ІПС на основі прецедентів виконувалася на наборі з різних запитів (15 запитів для формування БП CBR-агентів і 5 тестових запитів) до ІПС, у БД якій було проіндексовано пошуковим роботом 742 документа. Первісний список URL адрес для пошукового робота був сформований на основі списку TOP 100 результатів, виданих трьома пошуковими машинами (Google, Yandex, Bing) по запиту «Середовище розробки Visual Studio .NET і мова C#».

В таблиці 5.1 та на рисунках 5.2 – 5.3 наведено екранні форми результатів обчислення показників ефективності інформаційного пошуку для стандартної ІПС і ІПС на основі поповнюваної БП одного CBR-агента.

Таблиця 5.1 – Значення показників ефективності для стандартної ІПС і ІПС на основі поповнюваної БП одного CBR-агенту

	ИПС	1 прецедент у БП CBR-агента	2 прецедент у БП CBR-агента	3 прецедент у БП CBR-агента	4 прецедент у БП CBR-агента	5 прецедент у БП CBR-агента	6 прецедент у БП CBR-агента	7 прецедент у БП CBR-агента	8 прецедент у БП CBR-агента	9 прецедент у БП CBR-агента	10 прецедент у БП CBR-агента	11 прецедент у БП CBR-агента	12 прецедент у БП CBR-агента	13 прецедент у БП CBR-агента	14 прецедент уБП CBR-агента	15 прецедент у БП CBR-агента
$K_{п}$	0,68	0,37	0,39	0,40	0,43	0,47	0,50	0,52	0,54	0,56	0,59	0,60	0,61	0,63	064	0,66
$K_{ш}$	0,33	0,17	0,17	0,18	0,19	0,21	0,22	0,23	0,23	0,24	0,25	0,25	0,25	0,26	0,27	0,29
E_1	0,70	0,51	0,52	0,53	0,55	0,57	0,69	0,60	0,60	0,61	0,63	0,63	0,64	0,65	0,65	0,67

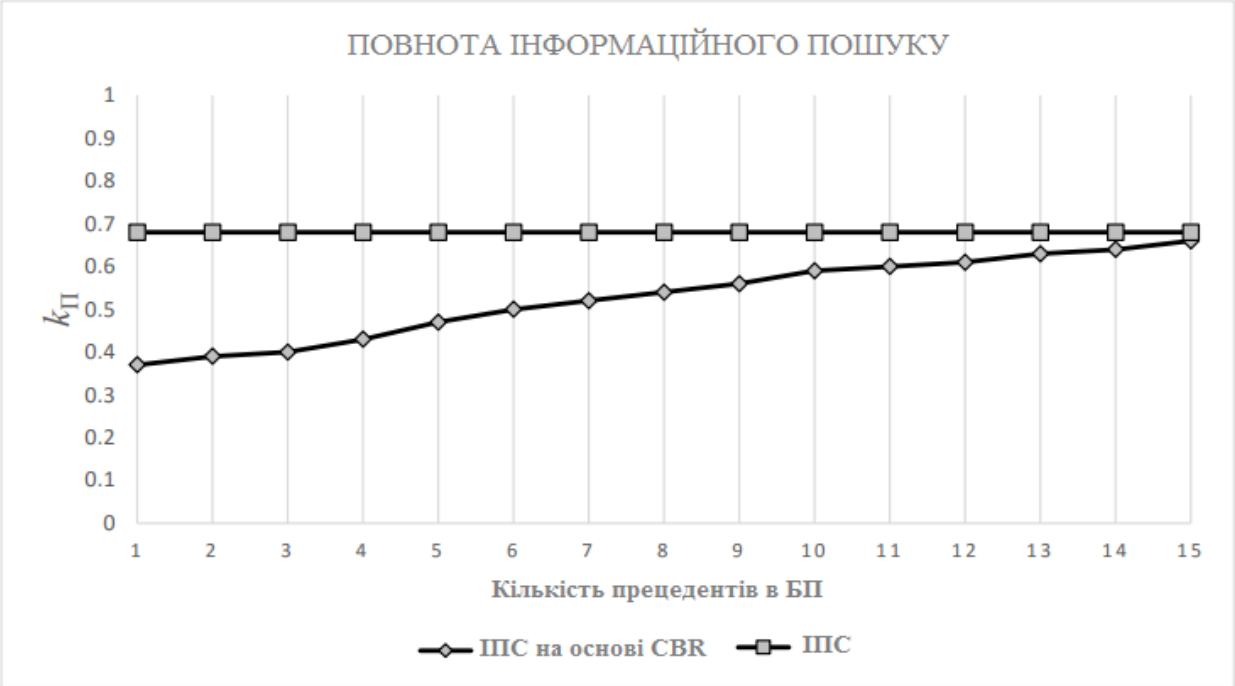


Рисунок 5.2 – Значення $k_П$ для ІПС «STRAY SEARCH» з використанням CBR-агентів і без використання CBR-агентів

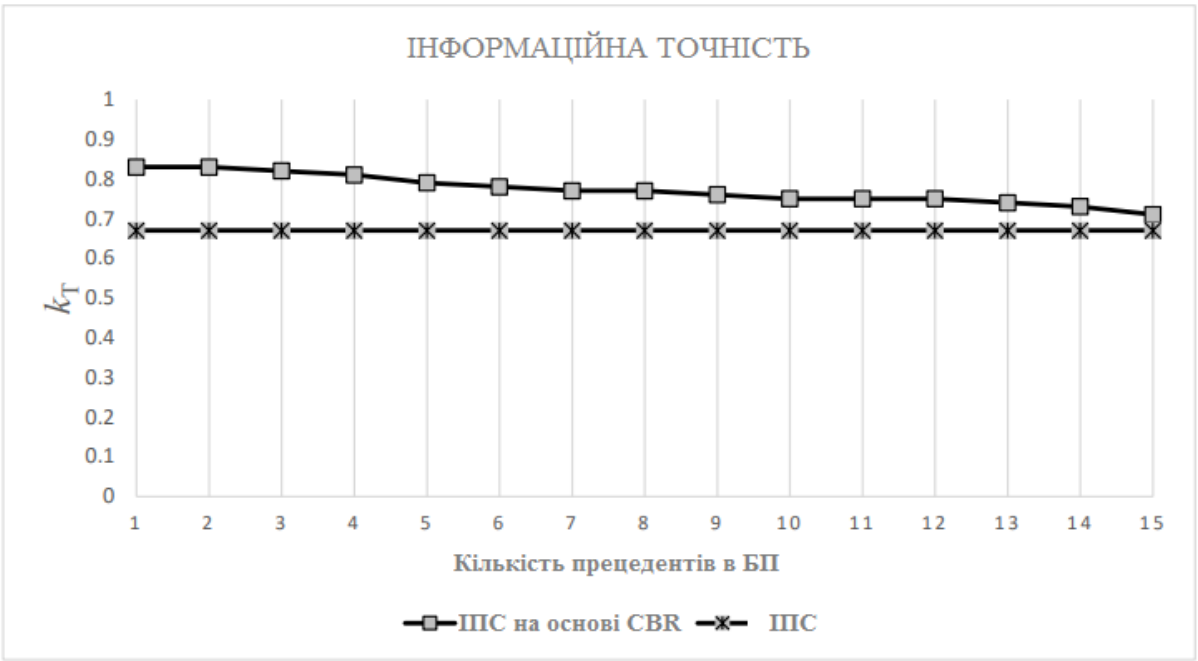


Рисунок 5.3 – Значення $k_Т$ для ІПС «STRAY SEARCH» з використанням CBR-агентів і без використання CBR-агентів

Із графіків видно, що пошук з використанням CBR-агента, який накопичує досвід (інформацію про користувацькі запити), дозволяє знизити інформаційний

шум, але інтегральні показники відрізняються не настільки значно, тому що стандартні механізми пошуку забезпечують більше значення коефіцієнта повноти.

Слід зазначити, що застосування CBR механізмів дозволяє знизити кількість звертань до сервера пошукової машини (індексу), що знижує навантаження на пошукову машину, тому що частина запитів може оброблятися безпосередньо на стороні клієнта, а також даний підхід дозволяє деякою мірою розв'язати питання з конфіденційністю даних.

Використання запропонованої архітектури для інформаційно-пошукової системи припускає, що результат відповіді на запит може бути отриманий у результаті взаємодії агентів.

ВИСНОВКИ

У ході виконання атестаційної роботи магістра був проведений аналіз існуючих методів зберігання даних. Розглянуті вимоги до основних методів зберігання, розподілу й забезпечення погодженості даних.

Основна увага була приділена аналізу зібраних даних з метою їх наочної інтерпретації, методам розподілу даних, зокрема – фрагментації й реплікації. Був зроблений аналіз методів реплікації й комбінації реплікації й фрагментації. Були визначені гідності й недоліки розглянутих методів.

В роботі було проведений аналіз методів агрегації даних для створення необхідних пакетів даних користувача соціальних мереж з використанням підходу Інтернет аналітичної обробки даних, як простого типу агрегації даних для створення необхідних пакетів даних користувача соціальних мереж.

Наведений детальний аналіз методів для організації сховища даних, платформ агрегації даних у соціальних мережах.

Детально представлена практична реалізація методів агрегації даних в соціальних медіа. Всі програмні приклади наведено для JSON (JavaScript Object Notation) – простого формату обміну даними, який зручний для читання і написання як людиною, так і комп'ютером. Він заснований на підмножині мови програмування JavaScript, визначеного в стандарті.

В роботі запропоновані:

– моделі пошуку співтовариств користувачів та інтеграції соціальних даних через сайти соціальних мереж. Ці моделі використовуватимуться для вирішення проблем відстежування соціальної активності користувачів соціальних мереж та перевантаження великою кількістю соціальних даних (поновлення друзів і інших видів діяльності);

– математична модель агрегації для узагальнення графа. Математично визначена операція агрегації узагальнення графа на основі угруповування вузлів на атрибутах і парних відносин (УГОУВ) виробляє короткий граф через

однорідне угруповання вузлів вхідного графа, заснований на обраних користувачем атрибутах вузлів і зв'язках, що засновано на угрупуванні графа.

Цей метод дозволяє користувачам вільно вибирати атрибути вузлів і відносини, які становлять інтерес, і виробляють угрупування на основі певних функцій.

Дослідженні і запропоновані моделі, підходи дозволять стейкхолдерам електронного бізнесу отримувати уявлення про користувача на основі отриманої інформації та за допомогою швидкого, послідовного, і інтерактивного доступу до інформації підвищити ефективність ведення бізнесу для визначеного сегменту споживачів – користувачів соціальних мереж, що сприятиме поширенню і розвитку електронного бізнесу за допомогою даних соціальних мереж.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Магдануров Г., Юнев В. ASP.NET MVC Framework; БХВ–Киев 2014. – 320 с.

2. Фримен А. ASP.NET MVC 4 с примерами на C# 5.0 для профессионалов; Вильямс – Москва, 2013. – 688 с.

3. Чедвик Д., Снайдер Т., Панда Х. Разработка реальных веб–приложений с помощью ASP.NET MVC; Вильямс – Москва, 2013. – 432 с.

4. Эспозито Д. Программирование на основе Microsoft ASP.NET MVC; БХВ–, 2012. – 464 с.

5. Freeman A. Pro ASP.NET MVC 4; Apress – , 2013. – 756 с.

6. Evjen B., Christian N., Duffy T. .NET 2.0 Wrox Box: Professional ASP.NET 2.0, Professional C# 2005, Professional .NET 2.0 Generics, and Professional .NET Framework 2.0; Wrox – 2015. – 709 с.

7. NoSQL Relational Database Management System URL: http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/nosql/.

8. Sadalage P. NoSQL distilled : A brief guide to the emerging world of polyglot persistence P. Sadalage, M. Fowler – Addison-Wesley, 2012. – 192 с.

9. NoSQL meetups URL: http://nosql.eventbrite.com

10. NoSQL debrief – Braindump URL: http://blog.oskarsson.nu/post/22996140866/nosql-debrief.

11. Gefen D, Govindarajulu C. Advanced Visual Basic.NET: Programming Web and Desktop Applications in ADO.NET and ASP.NET; – БХВ, 2017. – 595 с.

12. Emad I. ASP.NET MVC 1.0 Test Driven Development;– БХВ, 2017. – 312 с.

13. Коршунов А. Задачи и методы определения атрибутов пользователей социальных сетей // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2013.

14. Francois F. Fast Binary Feature Selection with Conditional Mutual Information // JMLR, 5:1531–1555, 2014

15. Koby C., Ofer D., Online Passive-Aggressive Algorithms // JMLR, 7(Mar):551–585, 2006.

16. Jianshu W., Bu-Sung L. Event Detection in Twitter // ICWSM 2011

17. Zhu X. Goldberg A. Improving Diversity in Ranking using Absorbing Random Walks // HLT-NAACL, 97-104, 2007

18. Bartunov S Joint Link-Attribute User Identity Resolution in Online Social Networks // Proceedings of The Sixth SIGKDD Workshop on Social Network Mining and Analysis (SNAKDD'12)

19. Buzun N., Korshunov A. Innovative Methods and Measures in Overlapping Community Detection // Proceedings of the International Workshop on Experimental Economics and Machine Learning (EEML 2019), Brussel, Belgium.

20. Malewicz G., Austern M., A system for largescale graph processing. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.

22. Salton G., Wong A., Yang C. S. A vector space model for automatic indexing.- Communications of the ACM, 18(11). – 2015- P.613–620.

23. Shostak I., Matyushenko I., Romanenkov Yu., Danova M., Kuznetsova Yu. Computer Support for Decision-Making on Defining the Strategy of Green IT Development at the State Level. In book: Green-IT Engineering: Social, Business and Industrial Applications, Vol. 171. Berlin, Heidelberg: Springer International Publishing, 533–559 (2018), https://doi.org/10.1007/978-3-030-00253-4

24. Shostak I., Kapitan R,. Volobuyeva L., and Danova M., Ontological Approach to the Construction of Multi-Agent Systems for the Maintenance Supporting Processes of Production Equipment. In Proc. : IEEE International Scientific and Practical Conference «Problems of Infocommunications. Science and Technology» (PICS&T-2018). Ukraine, Kharkiv, October 9-12, 2018. P. 209 – 214