ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

БД	база даних;
КД	композиційні додатки;
СБ	система збору;
СМ	соціальні медіа;
СУБД	система управління базами даних;
ВСК	віртуальне суспільство користувачів;
ІП	інформаційний процес;
AaaS	додаток як сервіс;
API	інтерфейс програмування програм;
SaaS	програмне забезпечення як послуга;
JSON	нотація об'єктів мови JavaScript;
XML	розширювана мова розмітки;
HTML	стандартизована мова розмітки документів;
CLAVIRE	віртуальне середовище для хмарних додатків;
SQL	мова програмування;
URL	уніфікований покажчик ресурсу;
JAVA	об'єктно-орієнтована мова програмування;
WF	формалізми потоків робіт.

ВСТУП

Актуальність дослідження. В даний час соціальні мережі - є динамічними джерелами різнорідної інформації, що відбиває різні процеси, які виникають в реальному суспільстві. Для їх кількісного аналізу приміняються спеціальні технології, орієнтовані на збір і обробку величзних обсягів неструктурованих даних, до яких відноситься технологія BigData. Особливості адаптації технологій BigData до соціальних мереж пов'язані з необхідністю обліку топологічної структури віртуального суспільства користувачів. В цілому це визначає динамічні властивості соціальних мереж, облік яких вимагає розвитку окремого класу математичного і програмного забезпечення BigData. Саме завдяки використанні BigData операторам мобільного зв’язку є можливість ефективно проводити аналіз потреб абонентів. Оцінюючі потребу того чи іншого регіону, формування гнучких тарифних планів і індивідуальних послуг, оперативного опрацювання заперечень, а також для припинення мобільного шахрайства при підміні номерів.

Крім того трійка лідерів завдяки інструментам Big Data просувають свої послуги. Аналіз споживаних послуг кожним абонентом дозволяє ефективно вибирати місце для нового офісу, вести моніторинг пересування абонентів, усувати шахрайство і боротися з відтоком клієнтів. У минулому році, як приклад, «ВФ Україна» уклав партнерську угоду з компанією Google з метою розширення розв'язуваних Big Data практичних завдань. Рішення Google допоможуть регулювати відтік клієнтів, сегментувати абонентів за потребою в ШПД, поліпшити якість обслуговування в Call-центрах і, як наслідок, пропонувати тарифні плани під потреби абонентів і реалізовувати більш точний маркетинг.

Таким чином компанія планує оптимізувати витрати і збільшити ефективність компанії.

Предмет дослідження магістерської роботи є методи та засоби підвищення ефективності систем збору і обробки великих обсягів даних.

Метою роботи є дослідження підвищення ефективності збору та обробки великих обсягів даних з соціальних мереж.

Для виконання поставленої мети були поставлені наступні завдання:

1. Аналіз особливостей реалізації та використання Big Data;

2. Аналіз процедур ефективного збору даних оператором зв`язку в соціальних мережах;

3. Практична реалізація та особливості функціонування програмної платформи збору даних.

Ступінь наукової розробки полягає в використанні хмарної моделі AaaS для реалізації адаптивної процедури збору і обробки даних із соціальних мереж , їх обробки і зберігання в рамках концепції BigData, ефективного планування збору даних, а також композитних додатків, що реалізують прикладні завдання дослідження соціальних мереж.

Практичне значення одержаних результатів полягає у реалізації програмної платформи для збору даних з соціальних мереж Facebook, Instagram, з урахуванням актуального стану їх топологічної структури.

1. ДОСЛІДЖЕННЯ ПЕРСПЕКТИВ ВПРОВАДЖЕННЯ ТА ВИКОРИСТАННЯ BIG DATA

1.1 Актуальність використання Big Data у різних сферах

У сучасному світі існує безліч джерел великих даних. Ними можуть виступати дані, що безперервно надходять з вимірювальних пристроїв, потоки повідомлень із соціальних мереж, метеорологічні дані, дані супутникового дистанційного спостереження, потоки інформації про місце знаходження абонентів мереж стільникового зв'язку, пристроїв аудіо- та відео реєстрації тощо. Масове поширення перерахованих вище технологій і принципово нових моделей використання різного роду пристроїв та інтернет-сервісів послужило відправною точкою для проникнення великих даних майже в усі сфери діяльності людини. Насамперед − у науково-дослідну діяльність, комерційний сектор і державне управління. Сучасний «цифровий всесвіт» (Digital Universe) містить як цифрові зображення і відео, завантажені з мобільних телефонів, наприклад, на You Tube, так і HD відео, передане по мережах провайдерів. Це як корпоративні дані, що генеруються бізнес-додатками, так і дані, які створює великий андронний коллайдер. Значну частку даних, вироблених до 2020 р., згенерують не люди, а машини в ході взаємодії між собою та іншими мережами даних. Сюди належать, наприклад, сенсори та інтелектуальні пристрої, які можуть обмінюватись потоками інформації з іншими девайсами.

Важко знайти галузь, для якої проблематика великих даних не була б актуальною. Вміння оперувати великими обсягами даних, аналізувати взаємозв'язки між ними і ухвалювати зважені рішення, з одного боку, надає потенціал для компаній з різних вертикалей для збільшення показників прибутковості і підвищення ефективності. З іншого боку, це чудова можливість для додаткового заробітку партнерів вендорів-інтеграторів і консультантів.

Попри на малий термін існування сектору Big Data, на сьогодні відомі оцінки ефективного використання цих технологій, засновані на реальних прикладах. Один з найвищих показників отримано в енергетиці – за оцінками аналітиків, аналітичні

технології Big Data здатні на 99 % підвищити точність розподілу потужностей генераторів, а охорона здоров'я США завдяки Big Data може заощадити до $ 300 млрд.

Інтернет став першою індустрією, що оперує Big Data. Завдяки цій технології вся інформація, поширена у глобальній мережі, та «події», що відбуваються офлайн (offline), залишають цифровий відбиток. Зберігається все – відомості про операції через електронні платіжні системи, дані про запити у пошуковику Google, лайки (likes) у соціальних мережах, розмови у чатах, дзвінки зі Skype та Viber, звичайна прогулянка зі смартфоном тощо.

Абсолютно все, що робить користувач онлайн (on-line), нікуди не зникає. Кожен лайк чи замовлення товару залишає цифровий слід у мережі. Навіть якщо користувач нічого не замовив, а лише переглянув інформацію, кілька наступних днів на різних сайтах, які він буде відвідувати, відображатимуться рекламні повідомлення з пропозицією придбати річ, якою він цікавився. Такий прийом називається контекстною рекламою. Ще більше інформації про себе, самі того не підозрюючи, користувачі залишають у соціальних мережах. Кожен поставлений лайк – вкрай цінна інформація, на основі якої можна створити точний психологічний портрет конкретної людини. Аналіз кількох десятків чи сотень таких лайків дає можливість отримати інформацію про прибутки користувача, його сімейний стан, релігійні переконання, емоційну стабільність, сексуальну орієнтацію, колір шкіри та політичні погляди. Це терабайти інформації, і не відомо, хто і як може її використати.

Уся зібрана інформація використовується для створення індивідуальних профілів потенційних покупців з метою адресної розсилки відповідної реклами, відслідковування можливих загроз державній безпеці, проведення виборчих кампаній та розробки новітніх технологій маніпулювання суспільною свідомістю. Big Data стали для людства значним досягненням та великою небезпекою.

Ефективні рішення в області роботи з великими даними для самих різних напрямків діяльності здійснюються завдяки багатьом комбінаціям програмного і апаратного забезпечення. Важливе значення Big Data - можливість застосовувати нові інструменти з тими, які вже використовуються в цій сфері.

Нескінченно великий інформаційний потік, який складає основу Big Data,

дозволяє нам отримувати кардинально нові знання, які були недоступні ще кілька років тому. Наприклад, вже зараз проекти, що базуються на Big Data допомагають:

Лікувати хвороби. Завдяки аналізу величезної кількості медичних записів та обробки медичних знімків можливо точніше і раніше ставити діагнози, краще розуміти природу різноманітних захворювань та винаходити нові ліки та методи лікування.

Боротися з голодом. Сільське господарство переживає справжню революцію Big Data, яка допомагає використовувати ресурси так, щоб максимально збільшити врожаї при мінімальному втручанні в екосистему. А також здешевити вирощені продути внаслідок оптимального використання обладнання та добрив.

Відкривати нові далекі планети. НАСА завдяки аналізу великої кількості даних отриманих з телескопів має змогу визначати хімічні склади атмосфер планет, що знаходяться на відстані багатьох світлових років, та робити припущення про їх придатність для життя.

Прогнозувати надзвичайні ситуації. Наприклад, завдяки даним з численних сенсорів та супутників науковці можуть передбачати де і коли можуть відбутися землетруси або природні катаклізми, чи змоделювати людську поведінку під час надзвичайних ситуацій, завдяки чому збільшаться шанси на виживання.

Запобігати злочинам. Внаслідок використання нових технологій та аналізу обширних даних з'явилася можливість автоматично попереджувати фінансові махінації з пластиковими картками та відмиванням грошей.

Оптимізувати прибутки. Торгові мережі можуть випускати нові продути з високим попитом та впроваджувати глобальні маркетингові компанії, які будуть орієнтовані під окремий сегмент покупців.

Покращити ефективність держави. Міністерство праці Німеччини використовує Big Data для аналізу заявок на отримання допомоги по безробіттю. При початковому аналізі виявилося, що 20 відсотків допомог виплачувалися незаслужено. Завдяки цьому уряд скоротив видатки на 10 мільярдів євро.

Це лише декілька прикладів. Насправді сфер, в яких використовується Big Data на сьогодні дуже велика кількість, яка з кожним днем буде збільшуватися допоки не охопить всі наявні.

1.2 Реалізація концепції Big Data в мобільному зв’язку

Телекомунікаційна галузь одна з небагатьох, яка в даний час освоїла і активно інвестує в рішення Big Data. Оператори зв'язку вже не тільки їх впровадили і навчилися використовувати, але і пожинають фінансові плоди. Згідно з дослідженням IDC, за підсумками 2018 року дохід світового ринку Великих Даних збільшиться більш ніж на 11%. І така тенденція збережеться в найближчі 3 роки. Саме телекомунікаційні компанії, за прогнозами аналітиків, стануть лідерами за середньорічним приростом доходів від Big Data. І саме в цій сфері безліч точок зростання і джерел інформації для Big Data.

Фахівці відзначають, що Big Data - це не сама мета. Це тільки інструмент, покликаний якісно модернізувати ту чи іншу галузь. Сама ідеологія Великих Даних реалізується через специфіковані під бізнес-потреби додатки. Зібрати дані недостатньо. Їх слід аналізувати для досягнення конкретної мети. Очевидно, що бізнесу важливі не нові інструменти, а дохід від них, тому проекти з впровадження Big Data оцінюються в тому числі з точки зору їх монетизації, яка часто не проявляється в коротко- і навіть середньостроковій перспективі. Тривалу окупність і ризик збитковості називають головною причиною, яка гальмує розгортання Big Data. Така поведінка може спровокувати два сценарії в телекомі: стагнацію галузі або поява нових «сміливих» гравців.

Згідно з дослідженнями IBM Institute for Business Value, понад половини компаній, в тому числі телекомунікаційних, застосовують Big Data для ефективного маркетингу. Вона дозволяє відстежувати споживчу поведінку, організовувати точкові маркетингові заходи і передбачати потреби клієнтів.

Використання технології Великих Даних продиктовано і тенденціями в телеком-сфері. Наприклад, аналітика Cisco повідомляє, що обсяг споживаного тільки в Україні мольної інтернет-трафіку до 2021 року зросте на 800% в порівнянні з 2018 роком, що створює колосальне навантаження на мережі. Крім того, в найближчому часі почне глобальне поширення концепція інтернету речей, яка вимагає загальної інтернет-

середовища і обміну даними між «розумними» об'єктами. Мобільні оператори прогнозують стрибок обсягу збережених і оброблюваних даних, тому націлені на передові технології.

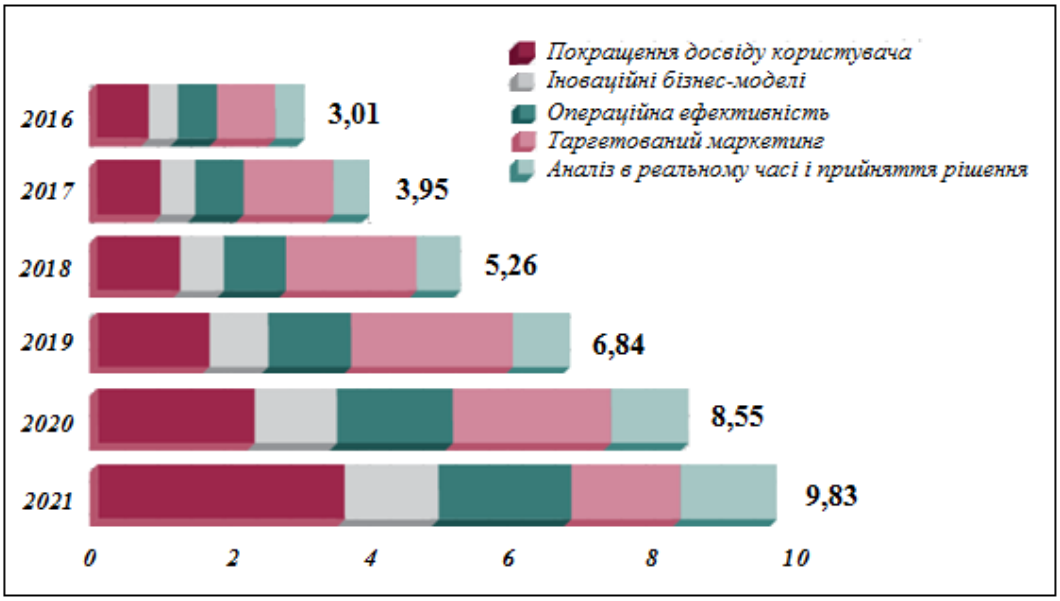

Рисунок 1.1. Прогноз світового ринку Big Data по категоріям використання

Області застосування концепції Великих Даних українськими операторами мобільного зв'язку

Трійка лідерів мобільного зв’язку в Україні не відстають від телеком-тенденцій, покращуючи тим самим рівень сервісу, що надається і підвищуючи ефективність процесів.

Компанія ТОВ «Лайфселл» використовує Big Data для створення ефективної інфраструктури мережі. Компанія оцінює потребу того чи іншого регіону в будівництві додаткових базових станцій, заміряючи обсяг попиту на послуги. В області маркетингу та роботи з клієнтом ТОВ «Лайфселл» використовує Великі Дані для формування гнучких тарифних планів і індивідуальних послуг, оперативного опрацювання заперечень, а також для припинення мобільного шахрайства при підміні номерів.

ПАТ «Vodafone Україна» реалізує Big Data і як B2B-рішення, знижуючи вартість прийняття рішення в галузях економічного та міського планування.

Наприклад, надає зібрані мережею знеособлені дані, які дозволяють робити висновок про структуру населення міста і взаємодії його жителів з міською інфраструктурою. Компанія співпрацює з РЖД, надаючи платформу для аналізу пасажироперевезень.

ПАТ «Vodafone Україна» також знайшов широке застосування можливостям Великих Даних. Для формування операторської мережі компанія аналізує динаміку руху абонентів всередині міста і, виявляючи тим самим точки масового скупчення абонентів, що дозволяє запобігти перевантаження за рахунок вдосконалення мережевої інфраструктури. Крім того, компанія веде пілотний проект в Києві з розвитку роздрібної мережі за допомогою рішень Big Data. Технологія заміряє пішохідний трафік в планованому місці відкриття офісу і на підставі отриманих даних приймається рішення про доцільність його відкриття, що вимагається кількості консультантів, площі приміщення, а також про види акцій, товарний асортимент і пріоритетних для клієнтів мобільних, фіксованих і банківських послуг. Компанія застосовує інструменти технології для бізнес-аналітики профілю клієнта, щоб сформувати для нього індивідуальну пропозицію, а також вважає перспективним напрямком фінансові сервіси. Фінансові програми дозволяють оцінювати платоспроможність клієнта для видачі кредиту. Вони проявили настільки високу ефективність, що оператор планує співпрацювати з банками, поставляючи їм сервіс з моніторингу підозрілих активностей з SIM-картою абонента, до номеру якого прив'язані банківські карти.

ПрАТ «Київстар» , в свою чергу, інструментами Big Data просуває свої послуги. Аналіз споживаних послуг кожним абонентом дозволяє ефективно вибирати місце для нового офісу, вести моніторинг пересування абонентів, усувати шахрайство і боротися з відтоком клієнтів. У минулому році оператор уклав партнерську угоду з компанією Google з метою розширення розв'язуваних Big Data практичних завдань. Рішення Google допоможуть регулювати відтік клієнтів, сегментувати абонентів за потребою в ШПД, поліпшити якість обслуговування в call-центрах і, як наслідок, пропонувати тарифні плани під потреби абонентів і реалізовувати більш точний маркетинг. Таким чином компанія планує оптимізувати витрати і збільшити ефективність компанії.

1.3 Приклади реалізації концепції Big Data в мобільному зв'язку

Якщо повернутись до методів Big Data, які направлені на отримання ефекту для бізнесу оператору зв'язку, то тут, зокрема, розглядаються чотири основні напрямки – перші три націлені на покращення внутрішньої роботи самої компанії, а останнє являє собою додатковим ринковим продуктом для зовнішніх клієнтів:

– високоточний маркетинг (precise marketing) – адресна пропозиція продуктів та послуг тим абонентам, які найбільш готові до їх придбання (нові тарифні плани, додаткові сервіси, платіжні термінали, індивідуальні роумінгові пропозиції та інше);

– управлінням якості послуг для абонентів (Customer Experience Management) для підвищення задоволеності з метою запобігання відтоку абонентів. Дана проблема актуальна для операторів, оскільки залучення абонентів дорожче, чим його утримання;

– оптимізація внутрішньої роботи оператора і планування розвитку (ROI-based Network Optimization and Planning) на основі обліку всіх обєктивних факторів і суджень абонентів для забезпечення максимальних гарантій повернення інвестицій в найкоротші терміни:

- тестування кращих варіантів пропозицій (A/B-тестування);
- аналітика і підстроювання контексту дослідження в реальному часі;

– монетизція інформаційних активів (Data Asset Monetization) – продаж в тій чи іншій формі (в тому числі у вигляді пайової участі в проектах) даних, які мають оператори, своїм парнтерам, щоб вони могли з їх допомогою вирішувати свої задачі. Зокрема, збір геолокаційних даних про абонентів довзоляє :

– адаптувати рекламу, виходячи з місця розташування людини чи її передбачуваного місця призначення;

– виявляти затори на основі кількості та швидкості переміщення абонентських терміналів уздовж шосе. Інформація про місце розташування і швидкість переміщення можуть слугувати також для визначення вартості нерухомості або розцінок для рекламних щитів;

– отримувати інформацію про те, в який районах міста кипить нічне життя або скільки протестуючих зібралось на демонстрації;

– прогнозувати поведінку поведінку людей, проаналізувавши їх переміщення і список дзвінків.

Приклад реалізації можливостей Big Data на основі відкритого ПО Hadoop приведений на рисунку 1.2.

Набір бібліотек та утиліт Hadoop включає в себе такі інструменти, як :

– CEM – Customer Experience Management – управління клієнтським досвідом;

– RA – Revenue Assurance – система «гарантованих доходів»;

– API – Application Programming Interface – додаток з відкритим кодом для вбудовування в інтерфейси других програм;

– Та інші.

Рисунок 1.2. Варіант реалізації Big Data в мережі оператору мобільного зв'язку

Опитування телеком-компаній також свідчать, що, збираючи великі обсяги даних, пов'язаних з активністю клієнтів, і величезна транзакційні логи елементів мережі, телеком-компанії далеко не в повній мірі використовують ту цінність, яка в цих даних міститься. Так, дослідження компанії Heavy Reading рисунок 1.3. показало, що проблеми неможливості інтегрованого аналізу всіх наявних у телеком-операторів джерел інформації і недостатня оперативність аналізу цих даних – основні бар'єри, які стоять перед компаніями при розвитку систем оперативної аналітики.

Рисунок 1.3. Відповіді респондентів з числа телеком-операторів: 3 – найбільш важливо, 0 – найменш важливе джерело

Дійсно, у телеком-компаніях рівень розвитку IT нижче, ніж в інтернет-компаніях, кількість IT-спеціалістів менше, а професіоналів в області великих даних може взагалі не бути в штаті. Тому телеком-компанії потребують не тільки в зовнішньому постачальнику технології Big Data, але і в консультанті, який би розбирався в їх бізнесі, їх IT-інфраструктурі, специфіку побудови систем Big Data для телеком-індустрії. Плануючи проект впровадження рішення Big Data, важливо продумати бізнес-схеми отримання прибутку від аналізу даних рисунку 1.4.

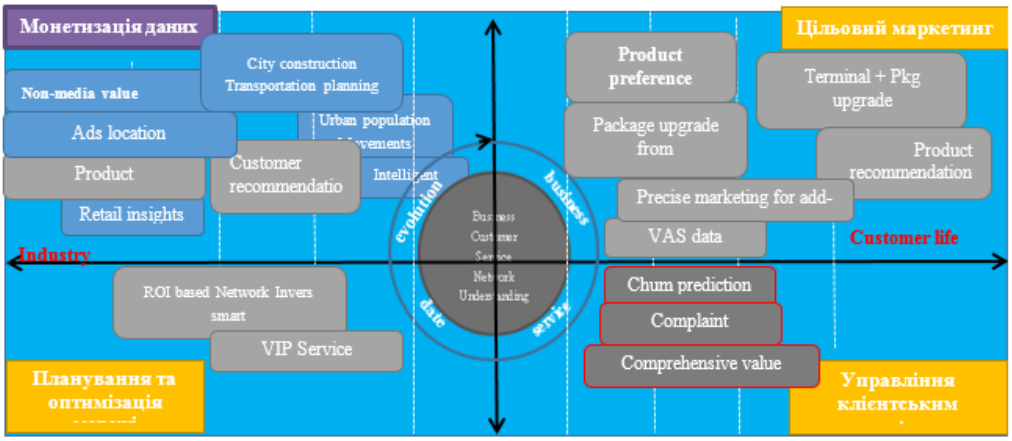

Рисунок 1.4. Можливі схеми отримання доходу від аналізу великих даних

Як випливає з рисуноку 1.4 впровадження проекту Big Data у телеком-бізнесі має цілий ряд схем повернення інвестицій як за рахунок оптимізації власних бізнес-

процесів, так і від надання інформації третім особам. Для кожної схеми є свої рішення на базі технологій Big Data, які може використовувати оператор Вищезазначені схеми на рис. 3 представлені в чотирьох сегментах: планування та оптимізація мережі, монетизація даних, цільовий маркетинг та управління клієнтським досвідом. Розглянемо ці дані більш докладно

Управління клієнтським досвідом. Одна з найважливіших завдань оператора – прогнозування відтоку абонентів (Churn prediction) і управління ним. Оператор, запровадив систему Big Data, має можливість суттєво підвищити лояльність абонентів. У Європі мобільні абоненти користуються послугою зв'язку за передоплатою, не пов'язані довгостроковим контрактом з оператором і можуть змінити його зі збереженням свого номера телефону. Аналіз поведінки абонентів (наприклад, схожості профілю раніше відключилися абонентів) дозволяє визначити групи ризику. Потенційних кандидатів на перехід до іншого оператора можна запропонувати вигідні для них тарифні плани або інші опції.

Ще одне завдання – прогнозування скарг (Compliant prediction). Відомо, що аналіз тональності мільйонів повідомлень в соціальних мережах (невдоволення сервісом, повідомлення про поломки тощо) дозволяє визначити негативно налаштованих абонентів, а в соціальних мережах – виявляти так званих лідерів думок, рішення яких про відхід може спровокувати на аналогічний крок багатьох інших користувачів. Можливість вплинути на думку лідерів думок дозволяє впливати на ставлення до сервісів компанії.

З іншого боку, скарги, свідоцтва про недоліки сервісу – це цінний інформаційний ресурс, важливе джерело інформації для їх виявлення і усунення.

Цільовий маркетинг. Product performance best match – оптимізація продуктів для клієнта. Завдання полягає в тому, щоб надавати клієнтам продукти і тарифи, які оптимально відповідають їх запитам і при цьому дозволяють оптимізувати дохід оператора. Наприклад, абонент, маючи телефон з можливостями 4G, підписаний на тариф 3G, тобто частина можливостей не використовує. Якщо оператор відслідковує подібних користувачів і пропонує їм новий тариф, вони починають більш повно задіяти можливості телефону, наприклад підключаються до відеосервісам, отримують

сучасні послуги, а оператор підвищує дохід за рахунок збільшення ARPU.

Приватними випадками розглянутого сервісу є пропозиції тарифів, оптимізованих на різні групи населення: Terminal +Pkg upgrade promotion (пропозиція апгрейда термінального пристрою і тарифу), Package upgrade from 3G to 4G (апгрейд від 3G до 4G), Product recommendation (рекомендації продукту).

Тут також присутній пункт Precise marketing for add-ons (цільовий маркетинг за пропозицією додаткових послуг) – мова йде про можливості пропозиції додаткової реклами на основі аналізу профілю абонентів. Наприклад, якщо клієнт стільникового оператора живе в зоні покриття ШПД і не є передплатником цієї послуги, має сенс прорекламувати її, націлився саме на ту категорію осіб, яка може бути зацікавлена в подібній рекламі.

Ще один пункт – VAS data service – додаткові сервіси передачі даних.

Планування та оптимізація мережі

У даному розділі в якості прикладу згадані дві схеми, які не вичерпують набір.

ROI based network Investment – інвестування в розвиток мережі з урахуванням ROI. Якщо ми говоримо про мережі мобільного оператора, то важливо оптимізувати цей процес: наприклад вирішити такі питання, як проектування ефективного підключення нових базових станцій, використовуючи дані геолокації і якості сервісу. Big-Data-рішення дозволяє визначити, як краще спланувати розміщення базових станцій.

VIP assurance services – гарантування доступності сервісу для VIP-клієнтів. У будь-якого оператора є клієнти, яким важливо забезпечити сервіс в першу чергу. У крупного оператора це досить велика категорія користувачів, і виникає необхідність вести статистику та управління не тільки по клієнтам взагалі, але з урахуванням значущості певних груп. Технологія Big Data дозволяє оптимізувати цю задачу.

Монетизація даних. У даному сегменті розглянемо наступні пункти. Non media value evaluation (оцінка джерел крім медіа) і Додає location recommendation (рекомендації по розміщенню реклами). Тут може йти мова про оптимізацію розміщення зовнішньої реклами на базі аналізу поведінки абонентів оператора, наприклад з урахуванням місця їх проживання, роботи або шляху прямування.

Близький за змістом пункт Urban population movement pattern (витяг інформації про паттернах пересування городян). Такі дані можуть використовуватися в різних індустріях, наприклад інформація про потоки громадян – для планування транспортних магістралей (Transportation planning) або міської забудови (City construction): тут монетизація може базуватися на продаж рекомендацій по оптимальному розташуванню торгових точок з урахуванням асортименту, відповідного потоків споживачів.

До вищезазначеної теми має відношення ще один пункт – Retail insights (рекомендації щодо вдосконалення купівельного маркетингу), де мається на увазі можливість вдосконалити асортимент товарів для більш повного задоволення попиту на основі прогнозування поведінки користувачів. До цієї ж групи можна віднести таку глобальну тему, як Intelligent city (інтелектуальний місто), яка зачіпає цілий ряд управлінських проблем великих міських центрів.

Product recommendation – рекомендації по удосконаленню продукту. Тут може йти мова про пропозицію додаткових продуктів (мобільні додатки, мелодії, відео, рінгтони і тощо).

Як уже зазначено, телеком-оператори знаходяться на початку шляху по впровадженню технології великих даних і потребують партнера, що має не лише технологічне рішення, але і досвід роботи у телеком-індустрії. Наприклад, компанія Huawei володіє компетенціями в області ІКТ, великих даних, а також знанням проблемних зон телеком-операторів.

Можна зробить висновок, що Big Data чекає велике майбутнє і той, хто навчиться збирати та аналізувати величезні масиви даних таким чином, щоб на їх основі передбачати виникнення тих чи інших подій, відкриє додаткове джерело монетизації великих даних, що вони мають. Це, в свою чергу, може дати унікальні конкурентні переваги, допомогти побудувати більш ефективну державу, надати нові можливості людям і зрештою зроблять життя людей більш зручним. Тим не менше слід відмітити, що розвиток і впровадження технологій Big Data має супроводжуватись вдосконаленням підходів в області інфобезпеки і конфіденційності даних абонентів.

1.4 Аналіз проблем використання Big Data

Попри величезні можливості цієї технології ми маємо розповісти про проблеми, які з'являються з розповсюдженням використання Big Data.

Впровадження Big Data ставить під загрозу приватність користувачів, що може негативно відобразитись на особистій безпеці. Паралельно необхідне дотримання правильного балансу між бізнес-процесами та дотримання законодавства по обмеженню і контролю доступу до особистих даних.

Крім цього, проблеми розробки надійних алгоритмів Big Data звязані з вимогою обробки як постійних так і даних, що різко збільшуються, пошуку і класифікації даних в умовах росту складності та числа їх окремих елементів. Так до 2003 року в усьому світі було згенеровано і збережено близько 5 ексабайт (1 ексабайт = 1018 байт) даних. А вже в 2011 році ті ж самі 5 ексабайт були створені менше ніж за 2 дня. Через декілька років за 10 хвилин, та темпи продовжують рости. Garter та IDC (International Data Corporating) прогнозують, що до 2020 року загальна кількість згенерованих даних в світі буде складати вже не ексабайти, а 35-40 зетабайт (1 зетабайт = 1021 байт).

Варто врахувати також, що при спробах впровадження багато світових компаній зіштовхнулись з традиційними проблемами рис.1.5. Нехватка бюджету та кваліфікованих кадрів, складності інтерграції системи та її монетизції, та інше.

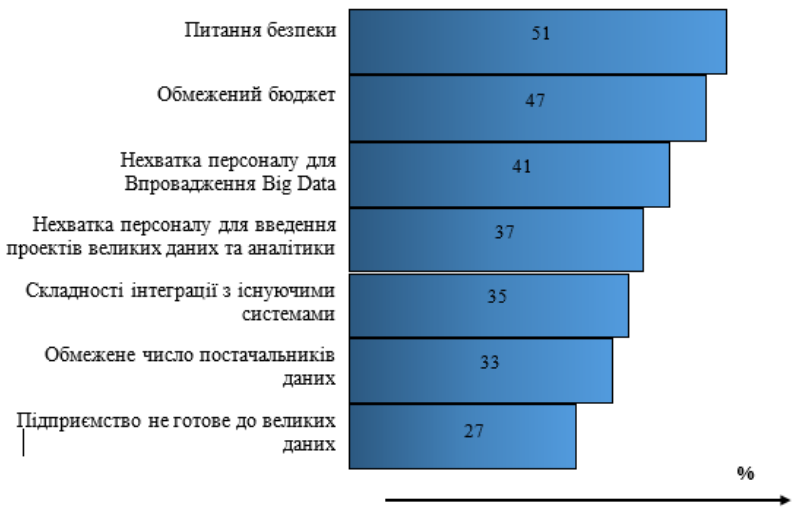

Рисунок 1.5. Основні проблеми при впровадженні проектів Big Data

Персональні дані та їхня недоторканість. Big Data збирає неймовірну кількість інформації, що стосується нашого приватного життя, яку ми б воліли зберігати в таємниці. Тому резонно постає питання балансу між тим, чим ми готові ділитися, та тим, наскільки комфортнішим може стати наше життя завдяки відкритості. Великі корпорації можуть маніпулювати цими даними роблячи нас певною мірою своїми заручниками. Хорошим прикладом цього є науково-фантастичний фільм "Сфера", в якому піднімаються подібні питання.

Безпека. Навіть якщо припустити, що Big Data буде використовуватися лише в шляхетних цілях, немає ніякої гарантії, що персональні дані зможуть бути надійно захищені від зловмисників та хакерів.

Дискримінація. Коли все відомо завдяки Big Data, окремі люди можуть обмежуватися наприклад в доступі до банківських кредитів через можливу недостатню надійність, або переплачувати за медстрахування через можливості певних захворювань пов'язаних зі схильністю до генетичних хвороб, або через несприятливі умови проживання в певних місцевостях. Або навіть державні структури та приватні компанії захочуть обмежити нас в доступі до певних сервісів та ресурсів.

Питання про зберігання великих обсягів інформації пов'язане з необхідністю організації певних умов, тобто зі створенням простору і можливостей.

Що стосується швидкості, то вона пов'язана не стільки з уповільненням і гальмуванням при використанні застарілих методів обробки, скільки з інтерактивністю: результат тим продуктивніше, чим швидше відбувається процес обробки інформації.

Проблема неструктурованості виходить з роздільності джерел, їх формату і якості. Для успішного об'єднання і обробки Big Data потрібна робота з їх підготовки, аналітичні інструменти або системи.

Великий вплив робить і межа «кількості» даних. Визначити обсяг досить складно, а виходячи з цього - проблематично прорахувати, які потрібні фінансові вкладення і необхідні технології. Проте, для певних величин, наприклад, терабайт, на сьогоднішній день успішно застосовуються нові методи обробки, які постійно

вдосконалюються.

Відсутність загальноприйнятих принципів роботи з Big Data - ще одна проблема, яка ускладнюється вищезгаданою неоднорідністю потоків. Для вирішення цієї проблеми створюються нові методи аналізу Big Data.

Складнощі також викликають підбір даних для аналізу і алгоритм дій. На сьогоднішній день немає чіткого розуміння того, які дані несуть цінну інформацію і вимагають аналітики Big Data, а які можна не брати до уваги. Також, на ринку недостатньо професіоналів, які впораються з глибинним аналізом, сформують звіт про рішення задачі і, відповідно, принесуть прибуток.

Моральна сторона питання: чи відрізняється збір даних без відома користувача від грубого вторгнення в приватне життя. Збір даних покращує якість життя: наприклад, безперервний збір даних в системах Google і Facebook допомагає компаніям покращувати свої сервіси в залежності від потреб споживачів.

Системи цих сервісів відзначають кожен клік користувача, його місце розташування і відвідувані ресурси, всі повідомлення і покупки. Це надає можливість демонстрації реклами, виходячи з поведінки користувача. Якщо користувач не надавав своєї згоди на збір даних, тоді таких зручностей не буде.

З цього випливає наступна проблема: наскільки безпечно зберігається інформація? Наприклад, відомості про потенційних покупців, історія їх покупок і переходів на різні сайти може допомогти вирішити багато бізнес-завдань, але чи є платформа, якою користуються покупці, безпечною. На сьогоднішній день жодне сховище даних - навіть сервери військових служб - не захищені від атак хакерів в достатній мірі.

1.5 Принципи збору даних в Інтернет

Технологи збору даних в соціальних мережах ґрунтуються на методах краулінга.

Пошуковим роботом, або краулер (від слова to crawl - просуватися), називаєтьс програмний компонент, який здійснює обхід веб-сторінок і занесення інформації про них в базу даних. Оскільки апріорі список всіх веб-сторінок не відомий, то пошук

посилань на нові сторінки здійснюється на вже завантажених. Таким чином, краулінга являє собою процес обходу веб-графа. Метою роботи краулер є збір даних про якомога більшій кількості «якісних» веб-сторінках. Залежно від розв'язуваної задачі термін «якісність» веб-сторінки може бути сформульований по-різному.

Спочатку краулери виникли як елементи пошукових систем, що збирають інформацію про веб-сторінках для подальшого індексування, що дозволяє здійснювати швидкий пошук у величезних обсягах даних. Однак для будь-якого аналізу даних з Інтернету необхідно в першу чергу отримати ці самі дані, для чого використовуються краулери. Залежно від специфіки розв'язуваної задачі конкретні реалізації краулерів можуть відрізнятися один від одного, але при цьому вони мають схожу алгоритмом роботи, наведеному на рисуноку 1.6.

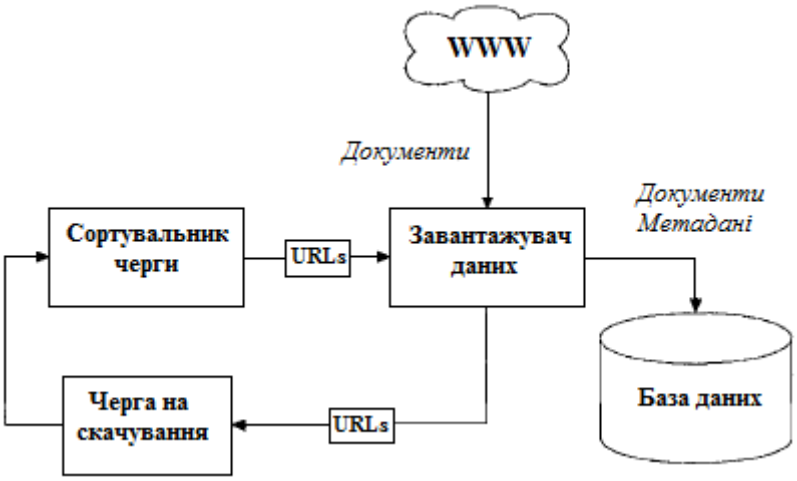

Рисунок 1.6. Узагальнений алгоритм роботи краулер

Алгоритм роботи краулер може бути описаний таким чином. па кожної ітерації роботи краулер дістає з черги один або безліч URL, для яких завантажуються дані з Інтернет. Зібрані дані зберігаються в базу даних, а так само з них витягуються прямі посилання на інші сторінки. отримані посилання додаються в чергу на завантаження, яка їх сортує виходячи з їх пріоритету. На цьому закінчується поточна ітерація роботи краулер і починається наступна. Крауле починає роботу з заданого набір початкових URL і закінчує роботу' при виконанні деякого критерію, наприклад при досягненні бази даних певного розміру.

Незважаючи на простий алгоритм роботи, до краулер пред'являється безліч вимог, які ускладнюють як його архітектуру і реалізацію, гак і часто вимагає використання складного математичного апарату.

Краулери можуть бути розділені на два основних вила: краулери реального часу, що дозволяють приступити до обробки інформації відразу після се збору, і batch-краулери, що дозволяють приступити до аналізу даних, тільки після збору досить великого обсягу даних. Краулери реального часу можуть бути використані при реалізації систем моніторингу за процесами, що відбуваються в мережі, і швидкого реагування на них. Вимога високої швидкості роботи з даними призводить до того, що краулери реального часу працюють переважно з оперативною пам'яттю комп'ютера і дані обробляються незалежно один від одного. Batch-краулери здійснюють однотипні маніпуляції відразу над цілим масивом даних, а не обробляють їх окремо. Цей тип архітектур дозволяє створювати ефективні системи для роботи з великим обсягом даних, але тим самим зменшується час доступу до нової інформації. Для роботи з великими обсягами даних даний тип краулерів активно використовує ефективні алгоритми роботи з жорсткими дисками, і, оскільки жорсткі диски є одним з найбільш повільних елементів сучасних комп'ютерів, при реалізації batch -краулерів використовують тільки операції читання/сортування/об'єднання збережених на жорстких дисках даних.

1.6 Аналіз вимоги які пред'являються до систем збору даних

До основних вимог, що пред'являються до краулера можна віднести наступні:

1) Ввічливість. Потрібно дотримуватися накладаються сервером правила, що регулюють частоту звернення і запитувані ресурси.

2) Розподіленість і Масштабованість. Дня обробки великих обсягів даних потрібно можливість функціонування краулер на декількох машинах. При цьому потрібно здатність збільшення робочого навантаження, шляхом додавання нових машин або збільшення пропускної здатності.

3) Можливість розширення. Потрібно «легкість» додавання або зміни

компонентів краулера. Наприклад, зміна алгоритму впливає на роботу системи, підтримка нового формату даних або мережевого протоколу передачі даних.

4) Стійкість. Система повинна стабільно обробляти різного роду мережеві помилки, що неминуче виникають при обміні даних з серверами. при зупинці роботи, краулер повинен підтримувати відновлення роботи з останньої точки.

5) Продуктивність. Робота з великими обсягами даних вимагає ефективного використання всіх ресурсів машини.

6) Якість. Зібрана краулер база даних документів повинна задовольняти вимогам клієнтів системи. Наприклад, при вирішенні задач моніторингу мережі, база даних повинна містити свіжі копії сторінок.

Конкретні вимоги до системи залежать від специфіки розв'язуваної задачі. Однак при роботі з сайтами, можна виділити дві головні вимоги до всіх систем збору даних: ввічливість і стійкість. Сайти накладають явні і неявні обмеження на число звернень до них. Більшість сайтів не рекомендують встановлювати більше ніж одне з'єднання з ними, при цьому між двома послідовними з'єднаннями повинен бути витриманий часовий інтервал, достатній для того щоб не створювати сайту 'проблем. Ці обмеження вимагають ефективного використання доступного числа звернень до сайтів, щоб якість одержуваної бази було якомога вище.

Збір даних з мільярдів сайтів вимагає реалізації розподіленої системи збору даних, при якій ресурси кожної окремої машини будуть використовуватися максимально ефективно і самі машини ефективно обмінюються даними один з другом. Однак якщо специфіка розв'язуваної краулер завдання вимагає збору даних невеликого обсягу з невеликого числа сайтів, то реалізація розподіленої системи не є обов'язковою.

1.7 Особливості збору даних з соціальних мереж

Збір даних з соціальних мереж має наступні особливості.

Збір даних з невеликої кількості сайтів. Збір даних з інтернету здійснюється з мільярдів сторінок, тоді як кількість соціальних медіа налічує всього сотні сайтів.

Кожен з них є унікальним з точки зору наданих даних і способів їх отримання. У підсумку збір даних здійснюється з невеликої кількості сайтів, що призводить до відсутності вимоги ефективної роботи з мережевим каналом.

У той час як краулер для збору даних з інтернету повинен легко підтримувати 100 одночасних з'єднань і здійснювати збір даних до 600 Мб в секунду, збір даних з соціальних медіа менш вимогливий до мережевого каналу. Для цього існує дві причини: невелика кількість сайтів соціальних медіа та обмеження на завантаження даних.

Динамічні сторінки. Сторінки сайтів соціальних медіа є динамічними, і змінюються в процесі перегляду в залежності від дій користувача. Складна динамічна структура HTML сторінок, що містять масу скриптів, що визначають наведену інформацію, ускладнює процес вилучення інформації. Витяг інформації з таких сторінок вимагає акуратного і уважного програмування, що значно ускладнює процес розробки програмного забезпечення. Верстка сторінок змінюється з часом, що призводить в необхідності переписувати алгоритми вилучення інформації.

Використання API. Соціальні медіа надають Інтерфейси Програмування додатків (API - Application Programming Interface), для отримання різної інформації про сутності мережі в зручному для обробки форматі.

Використовуються різні протоколи звернення до API, часто унікальні для кожного соціального медіа. Так само різниться формат даних, що надаються, серед найбільш популярних форматів можна виділити JSON і XML. Для використання API може знадобитися реєстрація додатки в системі і використання механізмів аутентифікації в системі. Однак, незважаючи на зазначені особливості API, їх використання значно спрощує збір і подальший аналіз даних з соціальних мереж.

Обмеження на число запитів. Соціальні медіа обмежують число звернень до системи, не дозволяючи збирати великий обсяг інформації. Наприклад, на

момент написання цього тексту, в соціальній мережі Facebook квота методів API вимагають аутентифікації дорівнює трьом запитам в секунду. А для мережі мікроблогінгу Twitter, квота на ряд методів дорівнює 180 запитам в 15 хвилин. оскільки популярні соціальні медіа містять мільйони користувачів, то отримання

даних досить великого обсягу для вирішення ряду завдань важко. Тому використання доступною квоти має бути максимально ефективним.

Великий обсяг даних. Соціальні медіа містять великий обсяг структурованих і неструктурованих даних, які повинні ефективно зберігатися і оброблятися. Прикладом структурованих даних можуть бути: ім'я та прізвище користувачів, список улюблених фільмів/музики/фільмів, освіту користувача, і т.д. До неструктурованих даних можуть бути віднесені текстові повідомлення користувачів, листування між собою. Обсяги даних чималі щоб зробити можливість використання реляційних баз даних для їх зберігання і обробки. Однак може бути органічно використаний підхід, заснований на концепції BigData обробки і зберігання даних, що полягає у використанні не реляційних (NoSQL сховищ даних.

Складний контекст користувача. Користувачі або об'єкти в соціальних мереж; описуються безліччю різних даних: особиста інформація про користувачів, його інтереси, різні зв'язки з іншими об'єктами СМ, фотографії, музика, геодані.

Повний набір даних описують конкретного користувача називається його контекстом, в які входять до нього конкретні дані - атрибутами. На відміну від збору даних з мережі інтернет, політики обходу розглядають вага сторінки незалежно один від одного, в соціальних медіа необхідно розглядати контексти користувачів. При цьому деякі політики обходу можуть вимагати одночасного аналізу декількох профілів користувачів. Так само повний контекст користувача є «Багаторівневим» об'єктом, оскільки входять до нього атрибути так можуть формувати ієрархічну структуру. Наприклад, користувач має атрибутом «Записи», які в свою чергу мають атрибутом «коментарі до запису».

Ключовим моментом є процес формування контексту профілю користувача. Оскільки різні атрибути користувача витягуються з різних веб-сторінок, то виникає задача об'єднання цих «розрізнених» атрибутів в єдиний контекст користувача. Це завдання ускладняється великим обсягом даних, проте легко вирішується в рамках реляційних баз даних.

Політики обходу різноманітніші. Політики обходу в соціальних медіа володіють великим розмаїттям, оскільки для своєї роботи оперують з контекстами користувачів,

що містять безліч різних типів даних. деякі політики обходу можуть вибирати наступних для відвідування користувачів на основі їхнього списку друзів, інші можуть оперувати інтересами користувача. При цьому частини політик обходу для своєї роботи можуть вимагатися дані, які для іншої політики обходу не є обов'язковими. У процесі розширення функціональних можливостей системи збору даних можуть з'являтися як нові типи атрибутів користувачів, так і нові політики обходу, що вимагають для своєї роботи аналізу цих даних. Щоб розширення системи було порівняно легке, необхідно вміти щодо «легко» інтегрувати ці атрибути в контекст користувача і політики обходу.

Жорсткі обмеження на число можливих запитів до системи і щодо невелике число соціальних медіа не вимагає ефективного і швидкого використання мережевого каналу. Однак для збору і обробки великого обсягу різнорідних даних потрібна реалізація розподіленої системи. Як мінімум потрібно використання

1.8 Класифікація задач збору і обробки даних в соціальних мережах

Можна виділити наступні завдання, для яких можуть бути використані краулери: збір даних з мережі для подальшого аналізу, моніторинг інформації та задача семпліювання представлені на рисунку 1.7, завдання позначені прямокутниками, а рішення - восьмикутниками. Завдання збору даних є найбільш загальною з позначених трьох завдань і полягає в тому, що необхідно перебрати певні сторінки і занести їх в базу даних. Можна виділити різновид цього завдання, а саме тематичний збір даних - полягає в тому, що необхідно зібрати дані релевантні певній темі. Це завдання може бути вирішена двома принципово різними способами: шляхом перебору всіх даних (краулінга всіх даних і фільтрація релевантних даних), або шляхом використання пошукових систем. В якості пошукових систем можуть бути використані системи як вбудовані в сайт

Соціальної мережі, так і зовнішніх по відношенню до нього. Кожен спосіб вирішення володіє своїми перевагами і недоліками. Незважаючи на велику швидкість отримання результатів при використанні пошукових систем, повнота результатів може

бути нижче, ніж при здійсненні краулінга. Так само деякі теми пошуку можуть бути погано формалізуються з точки зору мови запитів до пошукової системи, що фактично виключає їх використання. Таким чином оптимальність використання того чи іншого підходу так само залежить від конкретної розв'язуваної задачі.

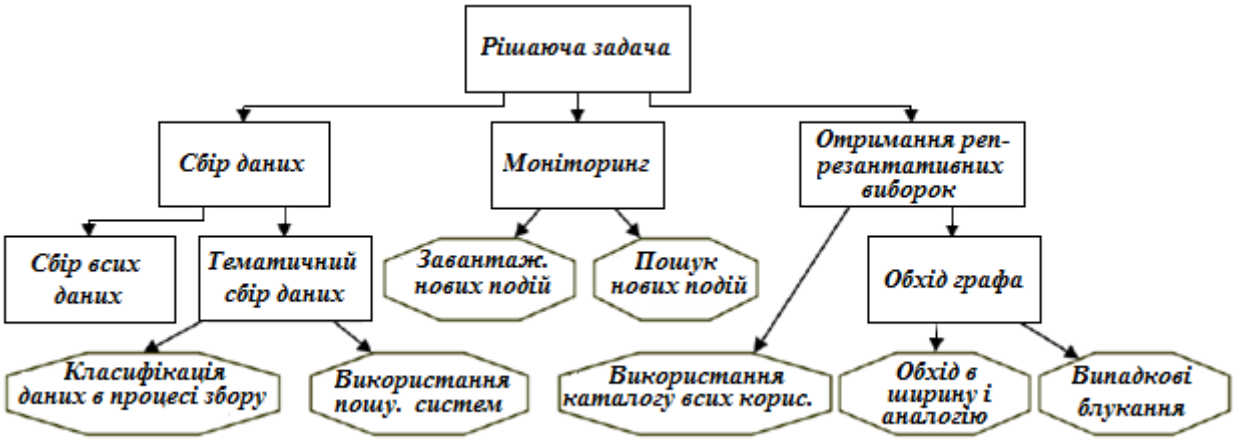

Рисунок 1.7. Таксономія завдань, що вирішуються при зборі даних з мережі і способів їх рішення

Завдання моніторингу соціального медіа полягає в тому, що необхідно максимально оперативно отримувати відомості про нові події, що сталися в сервісі.

Наприклад, може виникнути завдання моніторингу інтересу до певної теми в мережі, і виявлення різко збільшеного числа повідомлень на задану тематику. Залежно від наданих соціальних медіа API, дана задача так само може бути вирішена двома принципово різними методами: завантаження нових подій і пошук нових подій в системі. Деякі соціальні медіа, наприклад Twitter, дозволяють підписатися на потік всіх повідомлень, що виникають в сервісі. Аналізуючи цей потік в режимі реального часу, може бути здійснено моніторинг потрібну тему. Інший підхід полягає в тому, що нові повідомлення відшукуються в системі, для чого завантажується інформація про різні елементи системи і здійснюється пошук в цій інформації нових подій, для чого новий стан елемента порівнюється з його попереднім станом. Однак через квоти на завантаження даних, можуть бути оновлені відомості тільки про невелике число елементів системи, при цьому деякі з них можуть виявитися незмінною. Тому постає

завдання моделювання змін елементів в мережі і найбільш ефективного використання квот на завантаження. Різні математичні моделі можуть бути використані для передбачення появи нових подій.

Завдання семпліювання даних соціальної мережі полягає в необхідності отримання вибірки володіє необхідними властивостями. Наприклад, для соціальних мереж часто постає завдання отримання вибірки вузлів, для якої розподіл ступенів вузлів збігається з розподілом для все мережі. Так само постають завдання отримання вибірки, максимально покриває мережа, або отримання вибірки з деякого топологічного суспільства. Для вирішення описаних задач можуть бути використані алгоритми обходу графа, що дають володіють необхідними якостями вибірки. Відзначимо, що для мереж, в яких присутній каталог всіх елементів, наприклад соціальної мережі Facebook, завдання семпліювання можуть бути вирішені без використання алгоритмів обходу графів.

2. АНАЛІЗ ПРОЦЕДУР ЕФЕКТИВНОГО ЗБОРУ ДАНИХ ОПЕРАТОРОМ ЗВ`ЯЗКУ В СОЦІАЛЬНИХ МЕРЕЖАХ

2.1 Адаптивна процедура збору даних

Аналіз існуючих систем збору даних показав, що процедура збору даних і архітектура систем кожен раз адаптуються під нові завдання. Однак можуть бути вирішені завдання в соціальних медіа мають великою різноманітністю, і представляється досить складним адаптувати процедуру збору даних під кожну нову задачу. Зміни в процедурі збору даних в свою чергу будуть вимагати внесення низькорівневих змін в архітектуру системи. Внаслідок цього буде відсутній уніфікований підхід до збору даних з соціальних медіа.

Основні труднощі створення гнучких і універсальних процедур збору даних полягає в тому, що нові політики обходу вимагатимуть нових типів характеристик об'єктів соціальних медіа. Характеристики можуть бути досить різноманітними, наприклад, число знайдених посилань, що вказують на об'єкт, частота поновлення об'єкта, добовий шаблон поведінки об'єкта і «незвичайні» інтереси користувача, приналежність його до певної спільноти і т.д. дані характеристики визначаються на підставі ретроспективних даних, вже наявних в загальній базі даних, і змінюються від ітерації до ітерації роботи алгоритму.

У системах збору даних, які було розглянуто мною, процес обчислення характеристик вбудований в загальний цикл роботи краулер. В таких процедурах збору даних для обчислення нового типу характеристик буде необхідно вносити зміни безпосередньо в цикл роботи краулер. Це спричинить досить низькорівневі зміни в коді. В результаті дані системи будуть мати досить низькими розширюваністю і гнучкістю.

Тому для створення гнучких процедур збору даних необхідно розділит процес обчислення характеристик від самих політик обходу. обчислення всіх допоміжних даних, необхідних для роботи політики обходу, має бути відокремлено від неї. Сама політика обходу в цьому випадку реалізує тільки алгоритм створення списків URL, а все пораховані за даними характеристики отримує від інших компонент системи. Політику обходу можна розглядати як свого роду «Клієнта», а систему збору даних -

службою надає клієнту все необхідні дані. Для реалізації описаного підходу може бути ефективно використаний принцип AaaS.

Моделі побудови програмного забезпечення AaaS (Application as a Service) [9], використовуються в хмарних середовищах і надають клієнту певні функціональності, реалізовані у вигляді обчислювальних пакетів об'єднаних в композитні додатки (КД). Одним з поширених підходів для подання КД є формалізми потоків робіт (workflow) WF [9,10] який дозволяє описати зв'язки між окремими операціями у вигляді орієнтованого графа.

Композитний додаток створюється за рахунок зв'язування існуючих програмних пакетів в одне складне додаток, яке виповнюється на розподілених ресурсах.

Таким чином, обчислення різних характеристик об'єктів СМ, необхідних для роботи політик обходу може бути реалізовано за допомогою композитних додатків, що дозволить значно підвищити гнучкість системи.

Для роботи політик обходу необхідний аналіз великого обсягу різнотипних даних про об'єкти соціальних медіа. Для створення єдиної системи, здатної працювати з декількома соціальними медіа і вирішувати різні завдання, може бути використана концепція Big Data. Одним з принципів цієї концепції є обробка даних в тому місці, де вони зберігаються, що дозволяє значно знизити навантаження на мережеві канали при обміні даними між машинами. В рамках цієї концепції над даними зазвичай виробляються тільки найпростіші операції - послідовне читання всіх даних, додавання даних в «кінець» файлу і сортування.

Цих операцій досить для створення batch-краулерів, в яких послідовно обробляються досить великі обсяги даних.

Жорсткі обмеження, що накладаються СМ на завантаження даних, вимагають ефективного використання доступних квот. Ефективність збору даних може бути підвищена при одночасному вирішенні трьох завдань: фільтрації даних, ефективному розподіл квот між об'єктами СМ, а так само при використанні політики обходу топології СМ, яка дозволить отримати вибірку релевантну завданням збору даних.

Фільтрація даних релевантних важливість справ може бути реалізована за допомогою методів тематичного збору даних. Розподіл квот між об'єктами

вирішується тими ж методами, що і при моніторингу соціальних медіа. політики обходу топології СМ дозволяють отримати репрезентативні вибірки з графів або ж вибірки володіють необхідними топологічними властивостями.

Іншим важливим вимога підвищення ефективності збору є адаптація процедури збору під конкретні дані. В силу гетерогенності СМ локальні характеристики в різних ділянках можуть досить сильно відрізнятися один від одного, тому в процесі збору даних необхідно проводити адаптацію.

Підсумовуючи все викладені вище ідеї для збору даних з соціальних мережах пропонується наступна процедура, представлена на рис. 2.1. процедура є адаптивної, оскільки на кожній ітерації заново визначаються різні характеристики, що впливають на вибір об'єктів, для яких будуть завантажені дані.

Обсяг даних досить великий, тому використовується концепція Big Data для їх зберігання і обробки. Основними характеристиками, що впливають на вибір об'єктів, є оцінки активностей користувачів і топологічні характеристики. Однак внаслідок великої різноманітності можливих політик обходу система передбачає вбудовування композитних додатків, які здійснюють обчислення різних додаткових характеристик об'єктів. Для підвищення гнучкості і розширюваності ця функціональність вбудовується в систему за допомогою моделі хмарних обчислень AaaS. Управління краулер здійснюється на підставі рішення трьох завдань: фільтрації даних (тематичного збору даних), розподілу квот доступу (Моніторинг об'єктів) та використання найбільш ефективної для вирішуваних завдань політики обходу топології СМ (отримання репрезентативних вибірок). На підставі вирішення зазначених завдань відбувається визначення наступних об'єктів СМ, для яких на наступній ітерації будуть завантажені дані. У наступних розділах описуються методи вирішення трьох означених завдань.

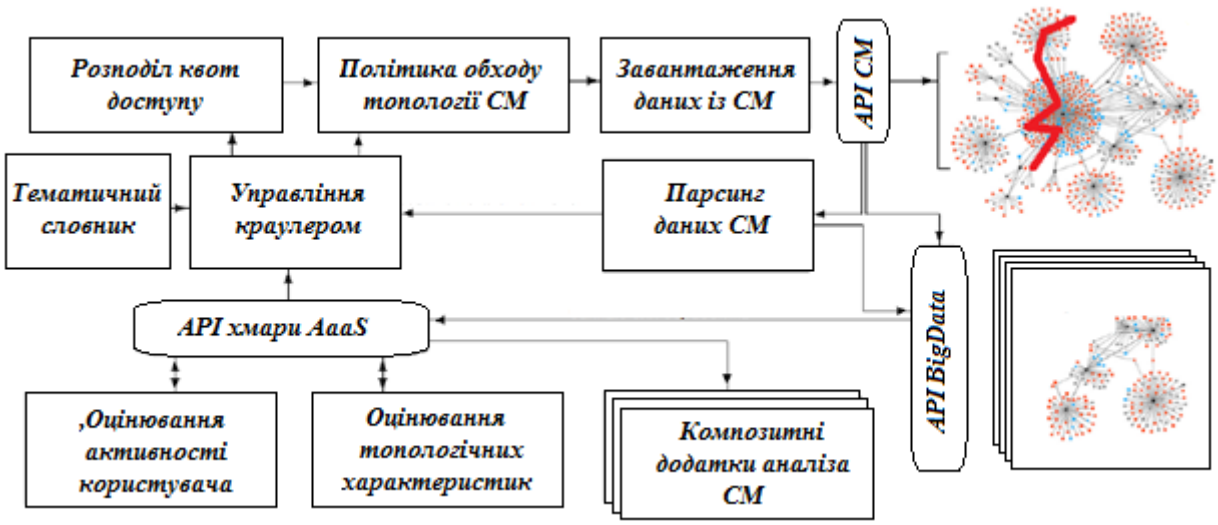

Рис. 2.1. Загальна схема адаптивної процедури ефективного збору даних в соціальних мережах

2.2 Тематичний збір даних в соціальних медіа

Тематичний збір даних ставить собі за мету збір даних про конкретну тематиці. Прикладами тематичних краулерів можуть бути системи з пошуку домашніх сторінок вчених, людей з певними інтересами, а так же текстів присвячених певної тематики. Контекстом об'єкта мережі називається вся доступна про нього інформація. При реалізації тематичного збору даних визначається релевантність контексту об'єкта мережі опису тематики. Фактично здійснюється збір даних про об'єкті (його контексту) і подальше визначення релевантності контексту опису теми. Для здійснення тематичного збору даних необхідно визначити наступні компоненти:

- контекст об'єкта соціального медіа (елемента мережі);
- опис тематики;
- визначення релевантності елемента мережі опису тематики.

Основна гіпотезу, що лежить в основі тематичного збору даних полягає в тому, що користувачі, зацікавлені в однаковій темі, утворюють неявні топологічні спільноти в соціальних медіа. Між цими користувачами багато зв'язків один з одним і маю зв'язків з іншими користувачами. З цієї гіпотези можна зробити висновок: при виявленні об'єкта в СМ, який зацікавлений в певної теми, ймовірність знайти серед

його друзів так само зацікавлених в даній темі, підвищується. Таким чином, алгоритм тематичного обходу мережі виглядає наступним чином.

1) Аналіз контексту елементом та мережі. Перевірка релевантності його опису тематики.

2) Якщо контекст елемента сіли релевантний заданій темі, то пріоритет його «друзів» підвищується.

3) Якщо контекст не віднесено до певної теми, то продовжити роботу.

Наведений в розділі 3 аналіз даних декількох соціальних медіа, для спільнот дійсно вдасться виділити «теми», яка згрупувала користувачів до спільноти. За зворотне не завжди може бути правильно: деякі теми не групують навколо себе користувачів в топологічні спільноти. Тому для таких тем описаний алгоритм обходу не підходить і фактично повинен бути здійснений повний перебір даних. Апріорі зробити припущення про ефективність запропонованого методу не є можливим.

Контекст елементів сети. Елементи мережі описуються своїми атрибутами, і ці дані формують контекст елемента. У разі роботи з соціальними мережами, де аналізуються елементами є користувачами мережі, атрибутами користувачів є: стать, вік, освіта, інтереси, групи в яких варто користувач, створені користувачем записи. При роботі з іншими типами СМ контекст об'єктів зазвичай не такий багатий.

В якості контексту об'єкта так само можуть виступати провідні на нього посилання і супроводжуваний їх текст [11]. Даний тип інформації є не менш важливим, ніж самі атрибути об'єкта, і довідкова інформація використовується в багатьох системах, наприклад в пошукових системах [12]. Посилання на об'єкт зазвичай супроводжує деякий тест - текст посилання (anchor - анкор, якір), параграф в якому вона зустрічається або навіть вся сторінка. Залежно від характеру теми, для якої здійснюється тематичний збір, опис посилання може бути ефективно використано для визначення релевантності об'єкта.

У описуваної системі використовувалися обидва підходи: аналізувалися як атрибути самого об'єкта, так і опис посилань ведучих на об'єкт. Наприклад, при пошуку користувачів, які пишуть тексти на тему наркокультури, використовувався гібридний підхід, при якому одночасно при аналізі тексту визначалася його

релевантність темі, а так само проводився пошук посилань, що вказують на інших користувачів. Користувачі, згадані в записах стосуються теми зв’язку, отримували трохи більший пріоритет, і тому збір даних про них відбувався трохи швидше.

Опис теми збору даних. Для опису тим використовується підхід, заснований на словниках ключових слів і фраз, так само можуть бути використані онтології. Експерт предметної області складає список ключових слів - найбільш відомі і однозначні терміни, які чітко визначають тематику тексту, а також різні багатозначні жаргонізми. Однак при складанні опису теми виникає ряд складнощів: морфемного різноманіття і семантична неоднозначність.

Для складання словника ключових слів, який би повно описував предметну область, необхідно привести всі можливі словоформи ключових слів, що найчастіше неможливо в силу великого морфемного різноманіття. Щоб зменшити труднощі в складанні словників були використані «патерни» слів, що дозволяють вказувати тільки одну словоформу для одного ключового слова. Були виділені наступні патерни слів, які показані на прикладі складання словника різних термінів наркокультури:

1) Точна (конкретна) форма слова - наприклад, патерн «конект».

2) Префікс слова - фіксується тільки початок слова, суфікси і закінчення можуть бути будь-якими. Наприклад, патерн: "зв’язку *", під який потрапляють «мобільний», «тариф».

3) Суфікс слова - фіксується закінчення слова, приставки можуть бути будь-якими. Наприклад, патерн "*розмовляти", під який потрапляють: «передача інформації», «тариф».

4) Задана середина слова - і приставки, і суфікси можуть варіювати. наприклад, патерн «*відновив зв’язок *».

5) стемінг слова [13] - за допомогою різних евристик, специфічних для кожного мови, від слова відкидаються закінчення і суфікси, і виходить словоформа, близька за своєю структурою до об'єднання приставки і кореня слова. Дані евристики не завжди мають високу точність, але володіють швидкодією, що в даному випадку є критичним параметром. існують різні набори евристик для кожної мови, нами було реалізовано сімейство евристик, званих Стеммер Портера.

Іншою проблемою при складанні описі теми є семантична неоднозначність використовуваних ключових слів. Ключові слова, які відносяться до різним семантичним темам, знижують точність збору даних - зібрана база буде містити дані відразу з кількох тем. усунути семантичну неоднозначність можна за допомогою використання онтологій. багато досліджень показали, що використання онтологій може підвищити точність класифікації документів [14].

Онтологія складається зі словника ключових слів, що описують предметну область, а так же логічних взаємини між ними. Описують предметні області онтології часто мають ієрархічну структуру, де найпростіші поняття об'єднуються в більш великі концепції. Дані взаємини між поняттями дозволяють автоматично сформулювати правила освіти «фраз», які можуть усунути семантичну неоднозначність і підвищити точність класифікації текстів.

Помстимося, що спосіб опису теми дуже сильно залежить від самої теми: для якихось понять може бути досить словника ключових слів, тоді як для семантично багатозначних тим має сенс використовувати онтології. побудова і пристрій онтології так само залежить від самої теми.

Описи тим зазвичай містять велику кількість ключових слів і фраз, при цьому необхідно в режимі реального часу визначати їх наявність в текстах. Ця інформація використовується для подальшого підрахунку релевантності тексту до опису теми. Тому до алгоритмів пошуку слів і фраз в тексті висуваються жорсткі вимоги на швидкість роботи.

Визначення релевантності опису теми. Для визначення релевантності контексту об'єкта опису теми можуть бути використані різні методи інформаційного пошуку або машинного навчання, а так само методи аналізу природних мов. Залежно від наявності навчальної колекції, методи можуть бути засновані як на «навчанні з учителем», і «навчанні без вчителя ».

У даній системі були використані методи ранжирування, засновані на навчанні без учителя, оскільки в більшості вирішуваних завдань була відсутня навчальна вибірка. Для визначення релевантності використовувалися зважування термінів і модель векторного простору [15]. Контекст користувача і опис теми користувача

представляються у вигляді векторів єдиного об'єднаного багатовимірного простору. Кожна розмірність цього простору відповідає одному з слів або термінів, які зустрічаються в описі теми і контексті користувача.

Була використана як булева модель зважування, при якій всі компоненти векторів описують контекст користувача і тему мають значення нуль або один. Так ж були використані і функції зважування, засновані на використанні статистичних даних про народження слова в усій колекції документів.

Пошук користувачів зацікавлених в темі «Роумінг». Даною системою було здійснено тематичний краулінга - був здійснений пошук користувачів які мають тексти, присвячені темі використання різних засобів зв’язку та програмного забезпечення які надають можливості здійснювати дзвінки за кордон. Експертом з Київського інформаційно-аналітичного центру була складена онтологія, яка містить різні аспекти. Даний контент містить 368 офіційних і «сленгових» виразів, які поділені на 12 груп, в залежності від семантичного значення.

Були виділені наступні групи: роумінг, Viber, Skype, Facebook, мобільний зв’язок, роумінг, сленгові слова приготування, загальні слова.

Дані групи так само формували більші концепції і поняття. На підставі семантичних груп слів були запропоновані наступні правила формування фраз: «Дзвінок» + «спосіб зв’язку», «використання технології» + «нкраїна», «Назва програми» + «якість зв’язку», «напрямок зв’язку» + «синонім». Разом з набором вручну заданих фраз, даний словник включай 8359 фраз, які складалися з двох або трьох слів.

Так само слова були поділені на три групи в залежності від ступеня їх приналежності до теми зв’язку. Була проведена дана класифікація, оскільки сленгові слова часто вживаються і в інших значеннях, що не відносяться до теми зв'язку. Кожній групі слів експертом був призначений певне політичне значення.

Побудованим на основі онтології фразам був призначений більш високу вагу, ніж сума ваг входять до неї слів. Це відповідає інтуїції, що зустрінута в тексті фраза є більш сильним сигналом приналежності тексту темі спілкування у роумінгу. Ваги груп слів і розбиття слів на групи кілька разів переглядалися експертом, щоб збільшити

точність методу класифікації. Для цього аналізувалися частоти появи слів і фраз в колекції, а так само проводився ручний аналіз текстів.

Релевантність текстів даного опису обчислювалася як сума ваг всіх знайдених в тексті слів і фраз. В якості документів відносяться до теми зв’язку було вибрано 20% самих релевантних документів. У число упізнаних документів потрапили документи, що містять або одну фразу, або три ключових слова. На рисунку 2.2 для кожної групи слів з онтології представлено кількість містять їх документів.

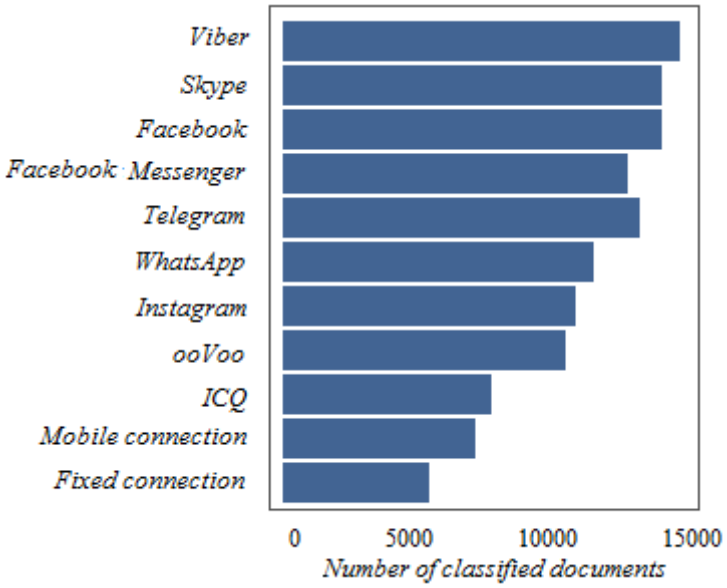

Рисунок 2.2. Кількість знайдених блогів для кожної групи слів

2.3 Моніторинг соціальних медіа

Однією з важливих задач при аналізі соціальних мереж є завдання моніторингу. У загальному випадку, завдання моніторингу полягає в безперервному спостереженні і реєстрації параметрів об'єктів. Систематичний збір і обробка інформації можуть бути використані для поліпшення процесу прийняття рішень, і здійснення організаційних функцій, таких як:

1. виявлення знаходяться в критичному стані об'єктів, для яких надалі може бути вироблений курс подальших дій

2. забезпечення зворотного зв'язку з об'єктами спостереження

3. встановлення відповідності правилам і контрактним зобов'язанням

Моніторинг СМ полягає в безперервному зборі та аналізі інформації про що відбуваються в них події. Залежно від типу соціального медіа та присутньої в ній інформації можна виділити різні типи подій. так для соціальних мереж подіями є: «користувач написав повідомлення», «Користувач А додав в друзі користувача В». У соціальних міді можуть збиратися дані про будь-яких видах активностей користувачів. Зібрана інформація може бути використана для аналізу поточного стану процесів, що відбуваються в СМ.

Моніторинг може допомогти при вирішенні цілого ряду практичних завдань, розв'язуваних в соціальних медіа. Відстеження інтересу до заданих подій або тем, пошук нових подій, що відбуваються в системі, аналіз поширення інформації для мережі або аналіз змін в топологічної структурі мережі - це все завдання які важливі як рішення бізнес завдань, так і для завдань науки. У всіх зазначених задачах важливо швидке виявлення нових подій, що відбуваються в мережі, що дозволяє своєчасно провести аналіз даних і, при необхідності, виробити курс подальших дій.

Виявлення нових подій відбувається шляхом завантаження даних про елементи мережі і порівняння нових даних зі старою версією. Однак великі обсяги даних і жорсткі обмеження на їх завантаження не дозволяють здійснити цю процедуру для всіх елементів мережі. Тому доступні квоти для отримання даних повинні використовуватися ефективно, що дозволить аналізувати більшу кількість елементів мережі.

Одночасно з ефективним розподілом квот, необхідно оптимізувати швидкість виявлення нових подій в системі. Набір алгоритмів вирішальних позначені завдання називаються політиками моніторингу.

Завдання моніторингу соціальних мереж визначається наступним чином. Нехай деяка соціальна мережа - дозволяє завантажувати дані для M елементів за деяку одиницю часу. Не гаючи спільності можна вважати, що одиниця часу дорівнює одним діб. Якщо необхідно здійснювати моніторинг для заданих N елементів мережі, то для кожного з них необхідно отримувати дані про нові повідомлення якомога швидше. Різні способи формалізації слів «як можна швидше» будуть обговорюватися нижче. Це завдання зазвичай вирішується розбиття на дві підзадачі: визначення добової квоти т1

для кожного з N елементів і визначення часу доби, в які необхідно здійснити звернення до елементу 'мережі, щоб виявити нові події як можна швидше.

Наївна реалізація стратегій моніторингу рівномірно розподіляє доступну квоту M між усіма N елементами і робить звернення до кожного з них через рівні проміжки часу. Однак така стратегія не враховує ряд важливих особливостей даних соціальних медіа: різна частота оновлення елементів і наявність ефекту часу доби. Ефективні політики моніторингу повинні враховувати ці ефекти для оптимізації розподілу квот між елементами і найбільш швидкого виявлення нових подій для кожного елемента.

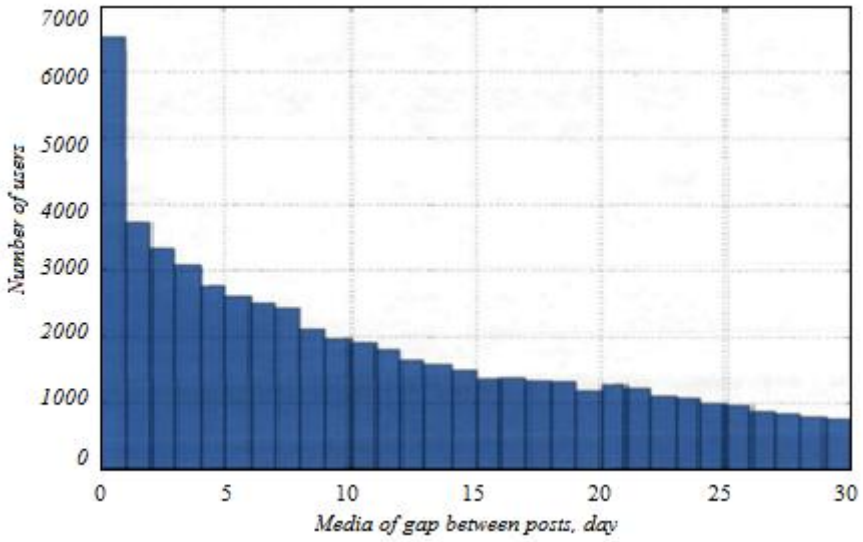

Рисунок 2.3. Гістограма частот поновлення блогів в мережі Facebook

У соціальних мережах різні елементи оновлюються з різною частотою, яка може варіюватися в межах декількох порядків. Так в Facebook присутні блоги, які в середньому оновлюються щодня, так і блоги оновлюються раз на місяць. Ефективний розподіл квот між цими елементами має враховувати цей ефект і для рідко оновлюються елементів виділяти меншу квоту. На рисунку 2.3 представлена гістограма частот поновлення блогів в мережі Facebook. Ефект часу доби полягає в різній інтенсивності появи нових подій в різний час доби. Найбільш яскраво помітна різниця в інтенсивності появи нових подій вдень і вночі. Хоча як показав аналіз, в СМ присутній елементи, для яких характерна скоріше нічна активність, ніж денна. Аналіз так само показав, що для елементів мережі характерний «шаблон поведінки», що описує в який час доби цей елемент мережі оновлюється найбільш ймовірно, і який

досить повільно змінюється в часі. На рис.2.4 представлена залежність появи нових подій в залежності від часу тижні. Видно яскраво виражений ефект дня тижня і часу доби. Графік побудований на основі аналізу даних з Facebook де під подією мається на увазі появу нового запису в блозі.

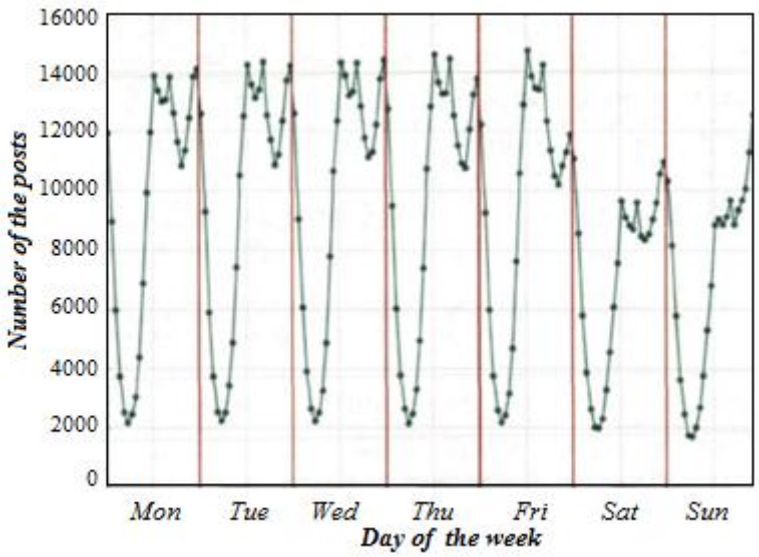

Рисунок 2.4. Число записів в блогах в залежності від часу тижні. побудовано на основі даних Facebook

Деякі соціальні медіа, дозволяють отримати«живий» потік нових подій що відбуваються в системі. Наприклад, на момент написання цієї роботи, Twitter дозволяє спеціальне API, що дозволяє встановити http з'єднання з встановленим прапором keepalive і отримувати список всіх записів, які створюються в системі в режимі реального часу. У таких мережах задача моніторингу спрощується і зводиться в простій завантаженні все нових подій. Однак далеко не всі соціальні міді надають таку можливість, або надають таку можливість тільки для певного типу даних. Незважаючи на те, що Twitter дозволяє отримувати в режимі реального часу список всіх нових повідомлень, що створюються в системі, він не дозволяє в такому ж режимі відслідковувати зміни в структурі мережі - хто з користувачів на кого підписаний.

Тому при вирішенні задач моніторингу структури мережі, як і раніше встає завдання реалізації ефективних політик моніторингу.

Існує кілька метрик для математичної формалізації критерію виявлення нових подій «якомога швидше», які використовуються в різних типах краулерів. При зборі

даних з мережі інтернет найбільш важливі метрики оцінки якості зібраної бази даних є метрики свіжості і віку, детальний аналіз яких наведено в статтях [16] і [17]. Головним завданням моніторингу соціальних медіа є швидке виявлення нових подій, тому нами була використана функція оцінки якості політики моніторингу, яка називається *затримка* (*delay*). Функція затримки показує кількість часу з моменту появи події в мережі до моменту його виявлення. Якщо елемент мережі *E* створив нові події в моменти t_1, t_2, … t_n, - і система моніторингу зробила звернення до цього елементу в моменти часу a_1, a_2, … a_n , то функція затримки для індивідуального подій t_i , визначається як:

$$D(t_i) = a_j - t_i \tag{2.1}$$

Де a_j мінімальний час звернення блогу, для якого виконано ,. На підставі значення затримки, обчисленою для індивідуального події, обчислюється затримка для елемента мережі і для всіх елементів мережі, для яких проводиться моніторинг. Функція затримки для елемента мережі *E* буде дорівнює сумі затримок для всіх його подій:

$$D(E) = \sum_{i=1}^{n} D(t_1) \tag{2.2}$$

Аналогічно визначається функція затримки і для безлічі елементів мережі.

Різні елементи мережі можуть мати різні пріоритети, пораховані виходячи з будь то інших критеріїв. Тому значення затримки для безлічі елементів мережі враховує і пріоритети w_i:

$$D(M) = \sum_{i=1}^{n} w_i D(E_1) \tag{2.3}$$

На підставі цієї метрики порівнюються політики моніторингу. Кожна політика моніторингу обчислює часи звернення до вузлів мережі, на підставі яких

обчислюється значення функції затримки. Політика моніторингу з мінімальним значенням функції затримки вважається найбільш оптимальною для вирішення завдань моніторингу соціальних медіа. Відзначимо, що для обчислення функції затримки, крім часу звернення до вузла, так само необхідно знати час появи події.

Для більшості типів подій оператор соціального медіа надає інформацію про час його скоєння, проте для ряду події така інформація відсутня. В цьому випадку використовують інші метрики оцінки якості політик моніторингу, такі як свіжість, для обчислення яких можна обійтися без знання точного часу здійснення події [18].

Для побудови ефективних політик моніторингу можуть бути використані математичні моделі, які описують появу нових подій в соціальних медіа.

Події мають випадкову природу, оскільки заздалегідь невідомо час, коли воно відбудеться. Тому використовуються імовірнісні математичні моделі, які оцінюють ймовірність появи нового події в той чи інший проміжок часу.

Для визначення оптимальних значень параметрів моделі, вона «навчається» на підставі історії поновлення елементів мережі. Оптимальні значення параметрів моделі, підлаштовують її під навчальну множину таким чином, щоб модель на них пророкувала появу нових подій найбільш точно. Таким чином, процес створення моделі відноситься до методів машинного навчання з учителем.

На підставі моделі, яка описує появу нових подій, створюється політика моніторингу, яка за умови справедливості моделі, визначає квоти елементів і час звернення до них таким чином, щоб мінімізувати очікуване значення функції затримки. Оскільки події мають випадковий характер, то оптимізуються саме очікувані значення.

Численні дослідження показали, що зміна різних об'єктів в Інтернеті, наприклад сайтів, може бути описано за допомогою пуассоновским розподілу. Свідченням цього може бути той факт, що часові інтервали між двома послідовними подіями має експоненціальне розподіл. Хоча гграктіческіе дослідження показали, що часові інтервали між двома послідовними змінами сторінок і не точно повторюють експоненціальне розподіл, воно може бути використано. Крім експоненціального розподілу так само може бути використано гамма розподіл, чиїм окремим випадком є

перший розподіл. Більшість методів може бути використано без змін і для гамма розподілу [18]. Крім опису сайтів в інтернеті Пуасонівский розподіл так само підходить і для опису інших елементів: новинних сайтів, окремих сторінок, даних з соціальних медіа.

Було проаналізовано частота оновлення блогів в Facebook, а саме чи мають часові інтервали між появою двох подій експоненціальний розподіл. У разі істинності цього твердження, для опису зміни блогів може бути використана пуассоновским модель. Функція щільності експоненціального розподілу з параметром λ має наступний вигляд:

$$f(\tau, \lambda) = \lambda * e^{-\lambda\tau} \tag{2.4}$$

де тривалість часового інтервалу. Тоді ймовірність спостерігати до подій на часовому інтервалі т має пуассоновским розподіл:

$$P(\tau, k) = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!} \tag{2.5}$$

Якщо побудувати графік функції щільності експоненціального розподілу (так само як і функцію експоненціального розподілу) на логарифмічних осях, то він повинен мати форму прямої. Для аналізу було вибрано 100000 випадкових блогів з українського сегмента Facebook, для кожного блогу було отримано 25 останніх записів.

Гістограма часових інтервалів між записами для всіх блогів представлена на рисунку 2.5, з якого видно, що розподіл відмінно від експоненціального, оскільки форма лінії гістограми відмінна від прямої лінії. Графік має дуже великим хвостом, який ми не відкидаємо, залишаючи тільки значення часові інтервали, з частотою появи більше 10^{-3}.

З рисунку 2.5 випливає досить очевидне, але важливий висновок, а саме, що зміна всіх блогів нс може бути описано за допомогою єдиного експоненціального

розподілу. Кожен блог повинен бути описаний за допомогою своєї власною моделлю, яка має індивідуальні параметри. Так само було протестовано, підпорядковується зміна кожного індивідуального блогу експоненціального розподілу. Для кожного блогу, аналізуючи його останні 25 записів, було обчислено кількість секунд між появою двох послідовних записів - часові інтервали. В результаті кожен блог описувався 24 значеннями часових інтервалів, для яких було перевірено чи мають вони експоненціальне розподіл. Був використаний рівномірно найбільш потужний критерій для перевірки експоненціального розподілу [19], який показав що з 95% воно не експоненціальне.

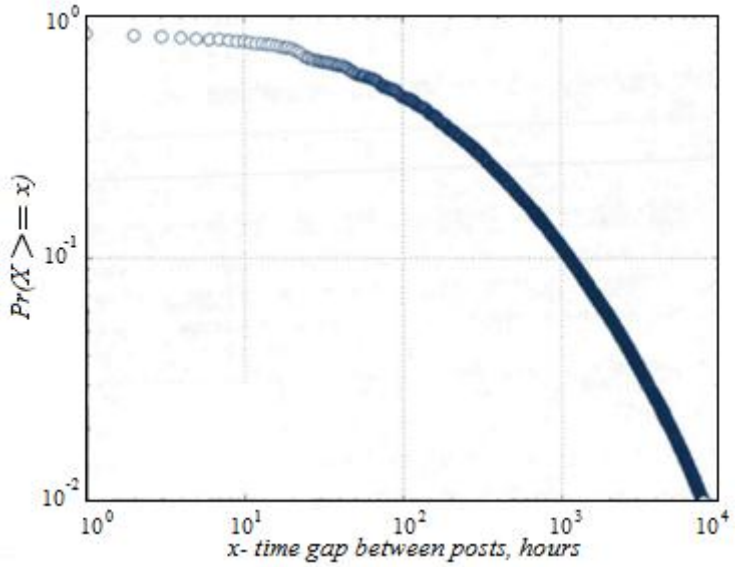

Рисунок 2.5. Гістограма часових інтервалів між появою двох записів у всіх блогах Facebook

Статистичний тест показав, що в разі вимірювання в секундах часових інтервалів між повідомленнями, зміни блогів не підкоряються експоненціальним розподілу. Однак якщо вимірювати часові інтервали в цілому кількості днів, то для багатьох блогів розподіл буде експонентним. Це є наслідком того, що на різних дозволах один і той же процес може описуватися різними законами.

Дослідження показали, що середня кількість повідомлень створюваних в блозі *E* (за один день, не сильно варіюється день у день. Воно змінюється в часі досить повільно і може бути підраховано на підставі історії останньої активності

користувачів. Для новинних сайтів, які дуже інтенсивно оновлюються, досить історії за останні два тижні. Середня кількість повідомлень створюваних в блозі будемо так само називати інтенсивністю оновлення.

Завдяки тому, що інтенсивність оновлення не сильно змінюється день у день, для опису змін блогу може бути використаний однорідний Пуассонівський процес. Пуассонівський процес є безперервним процесом, підраховують кількість подій, що відбулися за певний інтервал часу, і який має властивості відсутності пам'яті, і незалежності від моменту часу.

Пуасонівский процес так само моделює час появи наступної події, вірніше його ймовірність. Однорідний Пуасонівський процес використовує припущення, що нові події виникають з інтенсивністю Я в будь-який момент часу. В однорідному

Пуасонівському процесі часові інтервали між двома подіями мають експоненціальне розподіл описується формулою (2.4), а ймовірність спостерігати до подій протягом часового інтервалу т має пуассоновским розподіл (2.5). Дана модель підходить не тільки для опису появи нових повідомлень в блогах, але і для довільних типів подій, що відбуваються в мережі, однак для спрощення викладу в тексті буде використаний приклад з новими повідомленнями в блогах.

Для моделі однорідного пуассоновским процесу може бути показано, що мінімум функції затримки виконується, коли загальна квота розподілена між елементами мережі пропорційно кореню квадратному з їх інтенсивності оновлення

$$m_i = k\sqrt{\lambda_i w_i} \qquad (2.6)$$

де , - пріоритет елемента мережі , - отримується елементом щоденна квота, а коефіцієнт до - нормувальний коефіцієнт, для якого виконано рівність:

$$\sum_{i=1}^{N} k\sqrt{\lambda_i w_i} = M \qquad (2.7)$$

де M - загальна добова квота на завантаження, яка повинна бути розподілена між N елементами мережі.

У загальному випадку щоденна квота, порахована за формулою (2.6), є дробової величиною і число звернень, які відбуваються до елементів мережі за день, одно . Дрібна частина квоти акумулюється і як тільки вона стає більше одиниці, до елементу мережі відбувається на одне звернення більше. Таким чином, елементи мережі зі значенням щоденної квоти < 1 так само будуть оброблятися.

Інтенсивність оновлення елемента мережі вираховується на підставі недавньої історії активності. Якщо нові події відбулися в елементі мережі за часів . Для них обчислюється часові інтервали між подіями . , де . На підставі часових інтервалів вираховується інтенсивність оновлення блогу , як величина обернена до середнього значення часового інтервалу:

$$\lambda_i = \frac{n-1}{\sum_{i=1}^{n-1} \tau_i} \qquad (2.8)$$

Однак аналіз даних показав, що в соціальних медіа присутні як аномально великі часові інтервали між записами, так і аномально маленькі. Великі інтервали, довгою в кілька місяців або років, можуть бути наслідком того що користувач закинув свій блог, а потім вирішив до нього повернутися. аномально малі часові інтервали, довгою в долі секунд або кілька секунди, можуть бути свідченням того, що даний користувач використовує спеціальні програмні засоби для ведення свого блогу. Найчастіше такі блоги є «ботами» - ведуться не людьми, а програмами, і основною метою яких є поширення реклами в мережі. Подібні дані є шумовими викидами.

Оскільки оцінювання інтенсивності зміни є ключовим для створення ефективних моделей, що описують появу нових подій в елементах мережі, то необхідно фільтрувати шумові викиди. Для зменшення ролі шумових викидів, вони можуть бути ідентифіковані та вилучені під час підрахунку параметра . Також параметр може бути вирахуваний на основі медіанного значення часових інтервалів, яке менш схильне до шумових викидів.

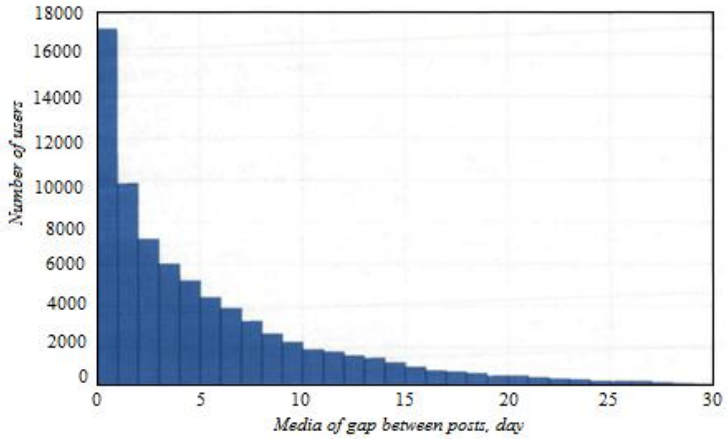

Рисунок 2.6. Розподіл інтенсивностей поновлення блогів. Враховуються блоги зі значенням медіани від одного до 30 днів

На рисунках 2.6 і 2.7 приведені гістограми інтенсивностей оновлень блогів , порахованих через медіанне значення часових інтервалів, а не через середнє арифметичне. На рисунку 2.7 видно велике число користувачів, які писали повідомлення з інтервалом менше 1 години. Ручний аналіз цих користувачів показав, що вони публікували записи з інтервалом в декілька секунд, що дозволяє припустити, що вони є роботами (ботами).

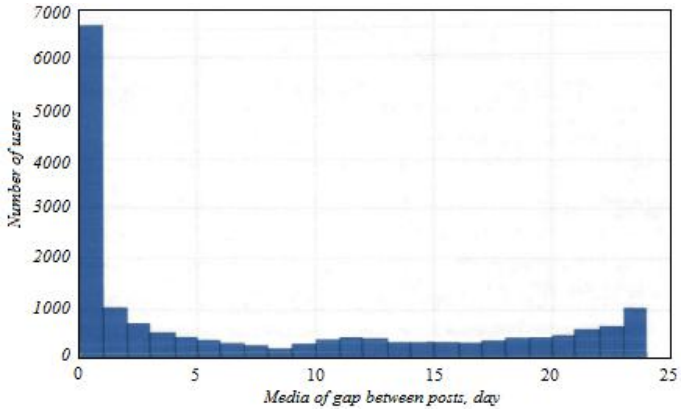

Рисунок 2.7. Розподіл інтенсивностей поновлення блогів. Враховуються блоги зі значенням медіани до одного дня

Описаний в цьому розділі підхід передбачає, що всі дні в тижні і в році є однорідними, з точки зору інтенсивності появи нових подій. Однак для більшої

частини даних з соціальних медіа, присутній ефект «дня тижня», при якому інтенсивність появи подій залежить від дня тижня. Цей ефект не є таким яскравим як ефект часу доби, тому в викладі цієї глави він не враховується. Однак описувані методи можуть бути легко адаптовані і для обліку цього ефекту.

Так само для більшості даних властивий ефект пори року, однак він спостерігається тільки при розгляді великих часових інтервалів. Оскільки оцінка інтенсивності оновлення елементів мережі Л, проводиться за останньою історії активності елемента, невеликої тривалості, то цей параметр буде автоматично враховувати особливості поточної пори року.

Точний час звернення для кожного елемента мережі передбачається на підставі добової квоти m_i. Ми не знає точного часу, коли з'являться нові події, тому, знову, для створення ефективної політики моніторингу ми мінімізує очікуване значення функції затримки. Як і в разі визначення добових квот для кожного блогу ми буде використовувати математичну модель, що описує появу нових подій протягом усього дня. Як було відзначено, в соціальних медіа присутній сильний ефект часу доби, що полягає в тому, що днем з'являється значно більше нових подій, ніж вночі. Щоб врахувати цей ефект використовується періодична неоднорідна пуассоновским модель, яка використовує добовий шаблон поведінки елемента мережі. Відповідно до цієї моделі ймовірність спостерігати до подій на часовому інтервалі [a, b] визначається виразом:

$$P(k) = \frac{e^{-\lambda_{a,b}}(\lambda_{a,b})^k}{k!}, \text{ де } \lambda_{a,b} = \int_a^b \lambda(t)dt \qquad (2.9)$$

Де $\lambda'(t)$ змінюється в часі параметр, що описує інтенсивність появи нових подій в елементі мережі E_i . Якщо використовувати часовий крок рівний одному годині, то $\lambda'(t)$ представлятиме собою 24-х мірний вектор, кожна компонента якого показує інтенсивність появи нових подій в цій годині доби. Параметр $\lambda'(t)$ будемо називати шаблоном поведінки елемента мережі, він є індивідуальним для кожного блогу. Приклади найбільш популярних шаблонів поведінки наведені на рис. 2.8. Якщо

унормувати параметр $\lambda'(t)$, то його можна трактувати як функцію щільності розподілу ймовірності появи події в той чи інший момент часу.

Математично може бути показано, що найменше значення функції затримки за умови використання цієї моделі досягається, якщо звернення до елементам мережі здійснювати за часів $t_1, \dots, t_k$ для яких, для будь-якого $i = 1...k$ до виконується умова

$$\lambda_{t_{i-1},t_1} = \lambda_{t_i,t_{i+1}} \qquad (2.10)$$

Де і , а T - довжина періоду, в даному випадку рівна 24 годинам.

Фактично часи звернення до елементу відбуваються в такі моменти часу $t_1,\dots,t_k$ які ділять шаблон поведінки елемента мережі $\lambda'(t)$ на рівні часові інтервали, з точки зору очікуваного числа подій, які відбуваються на кожному інтервалі. Якщо сприймати $\lambda'(t)$ как функцію щільності розподілу ймовірності, то часи обігу являють собою квантилі цієї функції.

Для визначення часу звернення $t_1,\dots,t_k$ досить визначити два послідовних часу звернення, на підставі яких, використовуючи формулу (2.10), вираховуються всі інші до k - 2 часів. Виходячи з цього, можуть бути запропоновані два підходу до визначення часу звернення. Перший спосіб заснований на визначенні хасу двох звернень, на підставі яких вираховуються часи всіх інших звернень. Пошук оптимального часу двох звернень здійснюється методом перебору: для першого і другого звернення використовується сітка з постійним кроком. Для кожного положення точок на сітці, визначаються часи інших звернень, і вираховується функція затримки. Таким чином, здійснюється пошук часів поводження з мінімальним значенням функції затримки. Однак на практиці цей метод показав погані результати, в першу чергу через складність підбору оптимального кроку для сітки можливих значень. Якщо зробити крок занадто маленьким, то все до звернень будуть здійснені за невеликий проміжок часу. Якщо зробити крок занадто великим, то за добу до елементу буде зроблено менше до звернень.

Другий спосіб полягає в итеративном поліпшення поточного відповіді. Спочатку $t_1,\ldots,t_k$, рівномірно розподіляються по всьому часу доби. на кожній ітерації циклічно розглядаються часи звернення t_i, після розгляду t_k, алгоритм заново дивиться на час t_0. На кожній ітерації відбувається коригування часу звернення t_i, таким чином, щоб для нього виконувалося вираз (2.10), для чого використовуються значення попереднього t_{i-1} і наступного t_{i+1} часів звернення.

Даний метод показав швидку збіжність на практиці і не вимагає для своєї роботи вираховувати значення функції затримки і історії активності елемента мережі. для його роботи потрібно тільки шаблон поведінки $\lambda'(t)$ елемента мережі.

2.4 Визначення шаблону поведінки елемента мережі

Для побудови періодичної неоднорідною пуассоновским моделі необхідний шаблон поведінки елемента мережі $\lambda'(t)$. Визначення шаблону поведінки здійснюється на підставі історії активності елемента, для чого використовуються дані про останні події, що виникли в елементі. Для кожної події обчислюється час його виникнення, і на підставі цього будується гістограма числа подій, що відбулися в той чи інший час доби.

Ми виявили, що в соціальних медіа, для яких не є характерною часте появи нових подій, $\lambda'(t)$ що містить велику кількість нулів. Нулі з'являються через те, що жодна з подій не сталося в той чи інший час доби. неоднорідний

Пуассонівський процес «погано» працює з нульовими можливостями, тому ми використовували різні згладжування для шаблону поведінки, що дозволяють позбутися від нульових ймовірностей і збільшити якість моделей [20]. Найкращі результати ми отримали при використанні згладжування Лапласа [21], яке додає 0.5 при підрахунку кількості подій, що відбулися в той чи інший час доби. При використанні цього згладжування, значення $\lambda'(t)$ на часовому інтервалі [, визначається через число подій, що сталися на часовому інтервалі [,:

$$\lambda_{t_{i-1}}(t) = \frac{n_{t_{i-1},t_i}+0,5}{\sum_{i=1}^{k}(n_{t_{i-1},t_i}+0,5)}$$

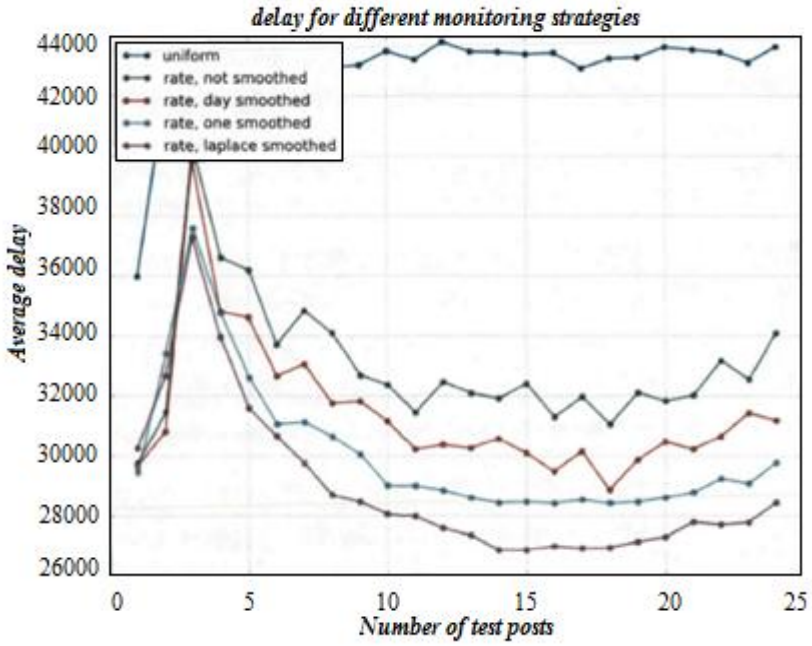

Рисунок 2.8. Порівняння різних методів згладжування шаблону поведінки λ'(*t*)

Результати порівняння згладжування Лапласа з іншими видами адитивних згладжувань наведені на рисунку 2.8. Решта методи згладжування додавати не 0.5, а 1, 1/24 або взагалі не використовують згладжування. Методи були протестовані на даних з мережі Livejournal, для кожного повідомлення було проігноровано дата його створення, враховувалося тільки час доби його появи. Блоги були поділені на 15 груп, в кожній *i* групі блоги описувалися *i* записів, для яких відбувалося в добу *i* звернень. Це дозволило протестувати ефект часу доби як для часто змінюються блогів, так і для рідко змінюються. Решта 25 - *i* записів використовувалися для побудови шаблону поведінки блогу. Результати показали, що використання ефекту часу доби здатне до 30% зменшити значення функції затримки. Це означає, що нові записи можуть бути виявлені до 30% швидше.

Було виявлено, що багато блогів в мережі Facebook володіють схожим шаблоном поведінки - вони створюють нові повідомлення приблизно в один і той же час. Була висловлена гіпотеза про те, що для побудови шаблону поведінки λ(*t*) елемента мережі

можна використовувати шаблон поведінки найбільш схожого на нього кластера елементів.

Незважаючи на те, що ряд досліджень відзначають поліпшення в виявленні нових повідомлень при використанні кластеризації елементів, нами воно не було зафіксовано для випадку блогів Livejournal. Проте, нижче описаний інший спосіб побудови шаблону поведінки $\lambda(t)$.

Для автоматичного пошуку кластерів скористаємося методами кластерного аналізу, для використання яких необхідно визначити метрику близькості між елементами мережі. Метрика схожості буде будуватися на підставі шаблонів поведінки $\lambda(t)$. Кожен кластер буде характеризуватися своїм унікальним шаблоном поведінки, який буде вираховуватися як «усереднений» шаблон всіх елементів входять до цей кластер. Як і раніше шаблон поведінки $\lambda(t)$ являє собою 24-х мірний вектор, кожна компонента якого показує ймовірність появи нового події в той чи інший час доби.

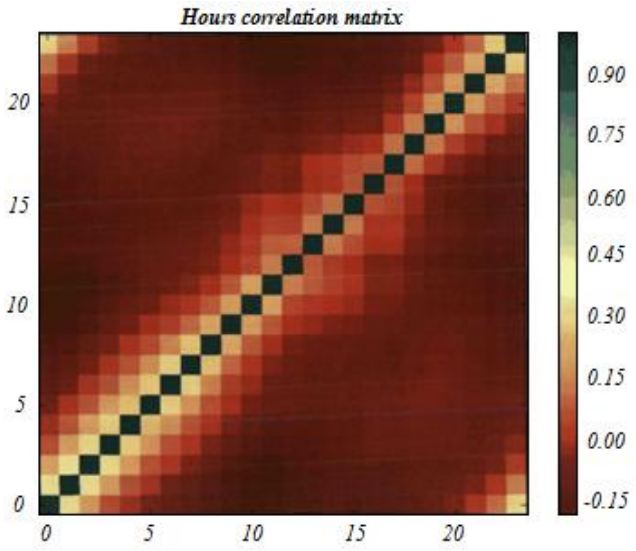

Рисунок 2.9. Матриця кореляції між компонентами

Якщо вважати, що компоненти цього вектора не корелюють один з одним, то можна скористатися евклідовой відстанню для побудови метрики близькості між ними, проте це не так. На рисунку 2.9 представлена матриця кореляції між компонентами шаблону поведінки. З неї випливає, що невеликі кореляції присутні для

сусідніх годин. Щоб врахувати кореляцію між компонентами шаблону поведінки, для побудови метрики близькості між шаблонами пропонується використовувати відстань Махаланобіса [22]:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T s^{-1}(\vec{x}-\vec{y})}$$

де S^{-1} - матриця обернена до матриці кореляції компонент шаблону поведінки.

Використовуючи відстань Махаланобіса блоги Livejournal були розбиті на кластери за допомогою методу K-Meапе [22]. На рисунку 2.10. приведена візуалізація знайдених 4-х кластерів, кожен кластер представлений середнім значенням шаблону поведінки всіх його елементів. З малюнка видно, що кожному кластеру відповідає свій унікальний шаблон поведінки. Деякі користувачі вважають за краще писати в робочий час, тоді як деякі вважають за краще вечірнє.

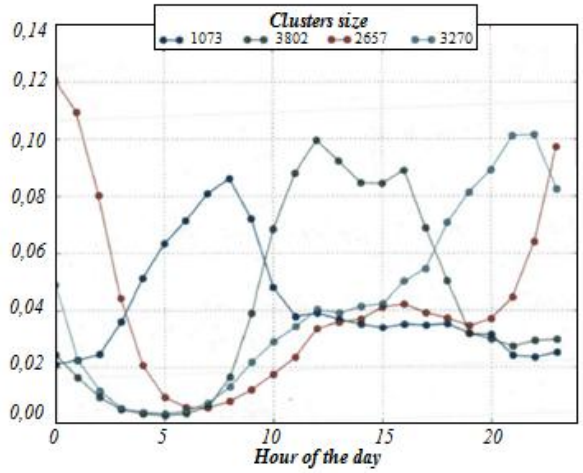

Рисунок 2.10. Візуалізація шаблону поведінки 4-х кластерів

Шаблон поведінки кластера може бути використаний в якості опису шаблону поведінки елемента, оскільки він менш схильний до шумовим викидам. Так само це дозволяє вирішити проблему «холодного старту», при якому для нового елемента мережі ще не накопичено достатню кількість даних, для побудови хорошого шаблону поведінки. Для нового елемента мережі, визначається найбільш близький до нього кластер елементів, який обчислюється на підставі відстані Махаланобіса. Однак як

зазначалося вище, даний метод не показав будь-яких поліпшень при моніторингу блогів в Facebook.

2.5 Порівняння методів

Було проведено порівняння описаних методів на даних про записи з мережі Facebook. Описані методи, що враховують частоту оновлення блогів і шаблон поведінки блогу були зрівняні з наївною реалізацією політики моніторингу, яка рівномірно розподіляє квоти між усіма елементами і робить звернення до мережі через рівні проміжки часу. У порівнянні брали участь чотири стратегії моніторингу:

4. uniform-uniform - рівномірний розподіл квоти між усіма блогами; звернення до блогів відбуваються через рівні проміжки часу.

5. uniform-day - рівномірний розподіл квоти межёду усіма блогами; використовується шаблон поведінки блогу

6. rate-uniform - квота залежить від частоти оновлення блогу; звернення до блогів здійснюються через рівні проміжки часу.

7. rate-day - квота залежить від частоти оновлення блогу; використовується шаблон поведінки блогу.

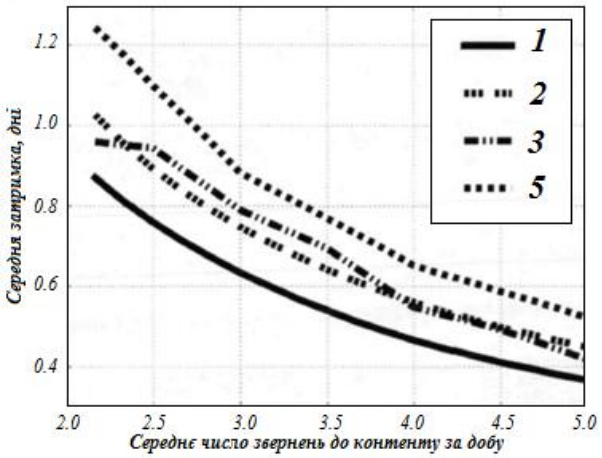

Рисунок 2.11. Порівняння політик моніторингу для різних значень добової квоти

Для тестування методів були використані дані про блогах, що містять мінімум 16 записів, створених за три місяці до моменту завантаження. Дані були очищені від «ботів» - блогів, у яких медіана часового інтервалу між двома записами менше 5 хвилин. Шаблон поведінки блогів був оцінений на підставі більш ранніх записів. У тозі тестової множини складалося з 61152 і 318800 записів. Політик моніторингу були зрівняні для різних значень сумарної квоти на завантаження M. Для кожного значення добової квоти політики моніторингу розподіляли її між блогами і вираховували часи звернень до блогів. Знаючи реальну історію появи повідомлень в блогах, для кожної політики моніторингу обчислювалося значення функції затримки.

Результати порівняння методів представлені на рис. 2.11. По осі абсцис відкладено середнє число звернень здійснюються до одного вузла, воно дорівнює загальній квоті поділеної на число блогів: M / N.

Результати показують, що використання частоти оновлення блогу для розподілу квоти між контентами, дає 13% поліпшень в порівнянні з рівномірним розподілом квот. Використання шаблону поведінки блогу, для обчислення часу звернення, дає 14% поліпшень в порівнянні зі стратегією здійснює звернення до блогу через рівні проміжки часу. Одночасне використання цих двох стратегій дає 25% поліпшень. Покращення вважає за формулою 1 - — де і значення функцій затримок, порахованих для першої і другої стратегії відповідно.

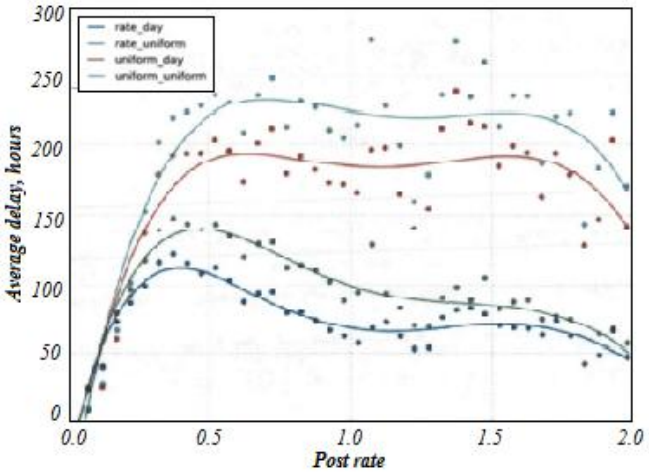

Рисунок 2.12. Порівняння політик моніторингу в залежності від частоти поновлення контенту. Точки апроксимувати поліном 5-ої ступеня

Так само політики моніторингу були протестовані на різних групах контенту.

Контенти були поділені на групи в залежності від частоти їх поновлення, і для кожної групи був вирахувано середнє значення функції затримки. Результати представлені на рисунку 2.12. Результати показують, що стратегії враховують частоту оновлення контентыв дають більше поліпшень для часто оновлюваних контентів. Тоді як для рідко оновлюваних контентів рівномірний розподіл квот дає результати краще.

3. РЕАЛІЗАЦІЯ ТА ОСОБЛИВОСТІ ФУНКЦІОНУВАННЯ ПРОГРАМНОЇ ПЛАТФОРМИ ЗБОРУ ДАНИХ

3.1 Загальна архітектура системи

Архітектура системи збору даних залежить від особливостей сайтів, з яких відбувається збір даних, а гак же вимог, що пред'являються до системи. Особливостями збору даних з соціальних медіа, перерахованими в розділі 2, є.

- збір даних з невеликої кількості сайтів;
- використання API;
- обмеження на кількість запитів;
- великий обсяг даних;
- складний контекст користувача;
- політики обходу різноманітніші.

Вимоги до системи. На підставі особливостей збору даних з соціальних медіа а так же аналізі цілей для яких створюється ця системи, були виділені наступні вимоги до системі. Жорсткі квоти на число запитів і невелика кількість операторів соціальних мереж не вимагають ефективного і швидкого використання мережевого каната.

Однак обсяги оброблюваних даних як і раніше великі, що вимагає реалізації розподіленої системи. Велика розмаїтість типів даних вимагає реалізації гнучкої

системи їх обробки і зберігання. Так само до систем збору даних з соціальних медіа пред'являється ряд вимог, типових для всіх систем збору даних.

- Робота з великими обсягами даних, в тому числі текстовими.

- «Ввічливість» - дотримання квот на завантаження даних з соціальних медіа.

- Модульність - можливість додавати в архітектуру нові реалізації компонент системи.

- Розрахований на багато користувачів режим роботи з системою, при якій кожен клієнт

володіє індивідуальною політикою обходу.

- Відновлювана (безперервна) робота краулер - можливість продовжувати збір даних з останньої точки зупинки.

- Стійкість до помилок - вихід з ладу елементів системи не повинен призводити до втрати даних.

- Якість. В умовах квот накладаються на отримання даних, зібрана база повинна максимально задовольняти вимогам клієнтів системи.

Якість бази даних визначається не стільки реалізацією системи збору, скільки використанням математичного забезпечення, що дозволяє ефективно використовувати квоти для вирішення поставлених завдань. Однак використання математичного забезпечення і гнучка можливість його змінювати накладає ряд вимог до архітектури. Детальніше про це написано в розділі, присвяченому реалізацій політик обходу.

Модель MapReducc для реалізації розподілених обчислень. При реалізації розподіленого краулер виникає ряд завдань, характерних для такого роду систем: розподіл даних між обчислювальними ресурсами, балансування навантаження, координація виконання завдань, відновлення після збоїв, моніторинг виконання завдань. Рішення перерахованих даних завдань можна забезпечити за рахунок стандартних фреймворків, що реалізують стандартні моделі розподілених обчислень, наприклад, MapReduce, яка була створена для паралельної обробки даних об'ємом до декількох петабайт [23] з використанням обчислювального кластера. Для розробки розподіленого batch-краудсра, відповідного наведеним вище вимогам, був використаний фреймворк Apache Hadoop [24], який реалізує більшість завдань,

пов'язаних з організацією розподілених обчислень, на базі моделі MapReduce. Відповідно до цієї моделі, дані зберігаються в розподіленої файлової системи у вигляді пар ключ, значення мають довільний тип і послідовно подаються на вхід операціями Map і Reduce. операція березня здійснює перетворення даних, а операція Reduce виробляє згортку всіх записів, що містять однаковий ключ. Для здійснення операцій Map і Reduce над даними фактично проводиться два типи операцій: послідовне читання і сортування. Сортування необхідна для ефективної реалізації операції Reduce. В моделі MapReduce Паралельна обробка даних забезпечується за рахунок того, що дані зберігаються рівномірними частинами на вузлах кластера та обробляються одночасно і незалежно один від одного. Детальніше з моделлю MapReduce і прикладами її використання можна ознайомитися в роботі [25].

Для аналізу даних, крім коштів Hadoop, можливе використання NoSQL бази даних Apache Hive, що дозволяє здійснювати запити до даних на основі SQL-подібного мови, що значно спрощує і прискорює процес аналізу.

Архітектура системи. У загальному вигляді система збору та обробки даних складається з наступних блоків. Завантажувач даних, парсер даних, модуль керування чергою завантаження. Завантажувач даних витягує з черги завантаження список URL, за якими необхідно завантажити даних з соціальних медіа. Завантажені дані обробляє парсер даних, витягуючи з них структуровані дані про об'єкти соціальних медіа. На її основі формується нова черга завантаження. При цьому для нових URL визначається пріоритет, впливає на порядок обходу сторінок, який може залежати від різних параметрів, наприклад, частоти оновлення і «якісність».

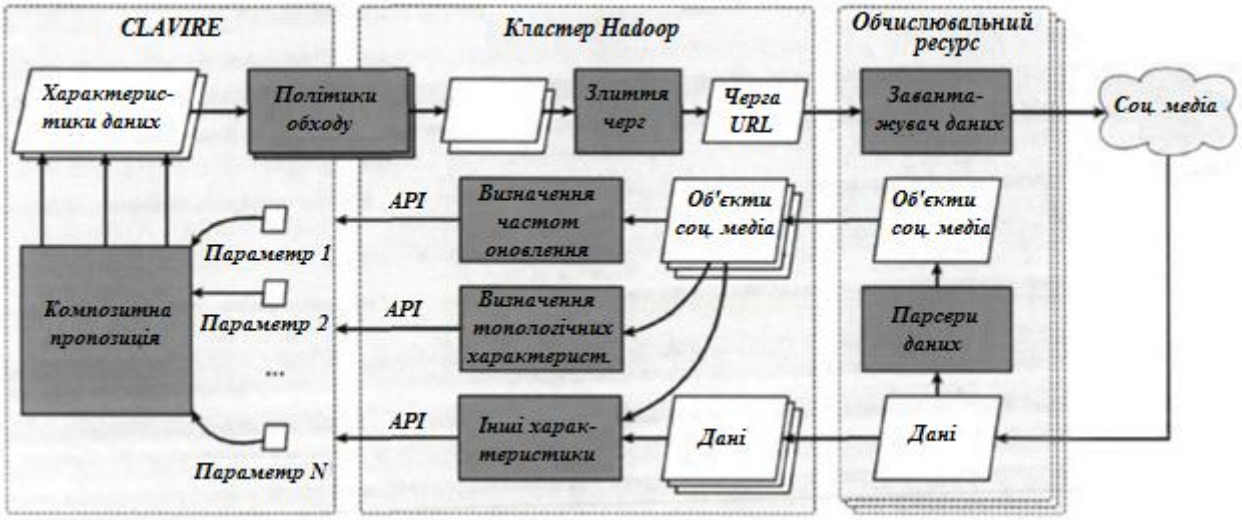

Рисунок 3.1. Загальна архітектура розподіленого краулер соціальних медіа

На рисунку 3.1 приведена загальна архітектура розподіленого batch-краулер, який здійснює однотипні операції над цілим масивом даних. За одну ітерацію обробляється заданий число URL, що дозволяє ефективно працювати з великим об'ємом даних, але збільшує час доступу до знову з отриманими даними. Модулі краулер викликаються послідовно один за одним, але кожен окремий модуль може виконуватися паралельно на декількох обчислювальних ресурсах. Робота краулер є ітераційний процес. Даних зберігаються в сегментах, кожен з яких відповідає певній ітерації роботи краулер. частина модулів реалізовані з використанням фреймворку Hadoop, а частина здатні функціонувати як окремі програми на окремих обчислювальних ресурсах. Система реалізує описану в розділі 2 процедуру збору даних, яка має гнучкість і розширюваністю. Для завантажених даних про об'єкти можуть бути обчислені різні характеристики. Для забезпечення гнучкості і розширюваності, частина функціональності винесена в хмарну середу CLAVIRE, що дозволяє додавати різні сервіси, які можуть бути використані політиками обходу для формування нового списку URL. Самі політики обходу можуть бути реалізовані як в рамках кластера Hadoop, так і в рамках хмарної середовища CLAVIRE.

Архітектуру краулера можна розділити на кілька великих модулів: модуль завантаження даних з соціальних медіа, модуль отримання даних про об'єкти соціальних мереж, модуль здійснює аналіз об'єктів, що обчислює різні характеристики

даних. Так само до складу архітектури входить хмарний сервіс CLAVIRE, на базі якого реалізуються композитні додатки, що обробляють дані про об'єктах соціального медіа і формують різні допоміжні

Характеристики. До складу архітектури комплексу так само входять політики обходу, що визначають об'єкти і відповідні їм URL, для яких так само необхідно завантажити дані. Оскільки система передбачає одночасну роботу з декількома політиками обходу, кожна з яких створює свій список URL, то в системі так само присутній модуль формування єдиної черги URL.

Модуль завантаження даних з соціальних медіа отримує на вхід список URL, по яким необхідно завантажити дані про об'єкти з різних соціальних медіа або мережі Інтернет. Для кожного соціального медіа реалізується компонент дозволяє виробляти в ньому аутентифікацію, і за допомогою використання мегодов API завантажувати необхідну інформацію. Для підвищення продуктивності, робота з різними СМ здійснюється паралельно в кілька потоків, але при цьому дотримується все обмеження. Список URL подавали на вхід цього модулю вже є відсортованим в порядку убування пріоритету URL, тому модуль завантаження не здійснює будь-яка їх ранжування.

З завантажених даних проводиться вилучення інформації про об'єкти СМ.

Для кожного СМ реалізуються парсери даних, що дозволяють витягати структуровані дані про об'єкти з «сирих» даних, отриманих від методів API соціальних мереж. API соціальних медіа надають інформацію в різних форматах, зручних для автоматичного розбору. Найбільш популярними форматами є JSON, XML. Однак в разі, якщо соціальне медіа не надає API для отримання інформації про якомусь типі об'єктів, то виникає необхідність завантажувати дані про ці об'єкти в форматі HTML. Витяг інформації з HTML відбувається на основі побудови відповідного йому DOM-дерева і пошуку в ньому вузлів, містять необхідну інформацію.

Описані два модуля реалізовані як окремі програми, які можуть працювати поза кластера Hadoop. Подальша обробка даних, отриманих краулер на цій ітерації, проводиться на кластері Hadoop, оскільки обробка вимагає отримання агрегованої сгатістікі по всій наявній базі даних. Тоді як для роботи цих модулів агрегування даних не потрібно. окрема реалізація завантажувача даних поза кластера Hadoop і

розміщення його на обчислювальних ресурсах, які можуть перебувати в різних мережах, для деяких соціальних медіа дозволяє ефективніше використовувати мережу. Так само окрема реалізація модулів дозволяє зробити систему більш гнучкою і розширюваної, оскільки дозволяє незалежно збільшувати ресурси для завантаження даних і їх обробки. Однак подібна архітектура вносить необхідність синхронізації між кластером Hadoop і завантажувачами даних, що вирішується за допомогою бібліотеки Apache ZooKeeper. На рисунку 3.2. представлена схема фізичної архітектури системи.

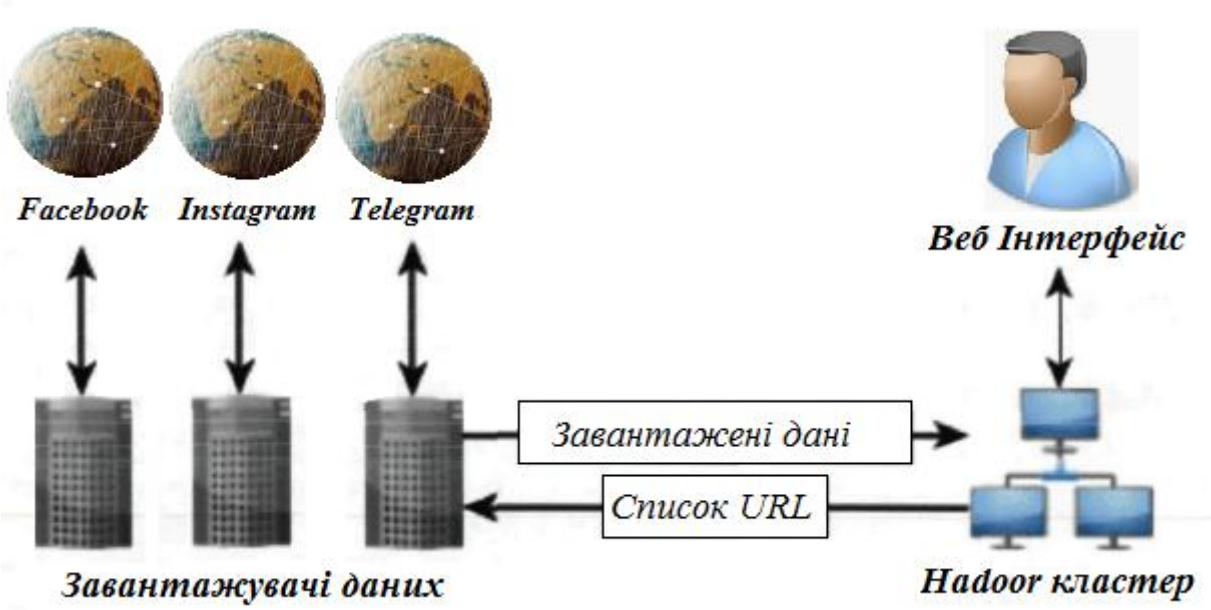

Рисунок 3.2.Компоненти архітектури краулер, без хмарної середовища CLAVIRE

Дані про об'єкти соціальних медіа і «сирі» дані копіюються в розподілену файлову систему HDFS, яка входить у фреймворка Hadoop. на базі цієї файлової системи реалізується база даних, що містить дані про різні об'єктах соціальних медіа, і орієнтована на ефективну обробку великих обсягів даних. База даних кожен тип об'єктів соціальних медіа зберігає в окремому файлі в файлової системі HDFS. Для підвищення продуктивності великі обсяги даних сегментируются, таким чином, щоб кожен сегмент відповідав даним отриманим на одній ітерації і зберігався в окремому файлі. Сегментування дозволяє швидко отримувати деякі характеристики, пораховані за всіма даними. Для цього на кожній ітерації обчислюється характеристики по знову з

отриманими даними, і агрегуються з характеристиками, підрахованими для всіх даних на попередній ітерації.

Прикладами обчислюваних «характеристик», рахованих за всіма даними є визначення всіх відвіданих вузлів мережі, визначення числа посилань ведучих на вузли мережі, помилки при завантаженні даних і т.д. Аналогічним чином можуть бути обчислені частоти оновлення та «шаблони поведінки» вузлів мережі, описані в присвяченому моніторингу розділі 2.Відзначимо, що для об'єктів соціальних медіа може бути пораховано велику кількість різноманітних характеристик, і в процесі розвитку системи їх число може значно збільшуватися.

Тому для поліпшення розширення системи використовується підхід AaaS, згідно з яким доступ до ряду простих характеристик і даних про об'єкти надається у вигляді сервісу за стандартними об'єктів обміну інформації (наприклад, HTTP). В рамках підходу AaaS створюється ряд API, що дозволяють отримати як різні «Прості» характеристики про об'єкти, так і певні дані про самих об'єктах.

Отримані дані можуть бути використані різними композитними додатками, реалізованими в рамках хмарної платформи CLAVIRE. Композитні додатки, орієнтовані на аналіз даних з соціальних медіа, в свою чергу вираховуватимуть болсс «складні» характеристики даних. хмарна платформа дозволяє прозоро для всієї системи збору даних додавати обчислення нових характеристик тим самим підвищуючи розширюваність системи. Відзначимо, що для реалізації обчислень нових характеристик в рамках кластера Hadoop необхідно вносити досить багато низькорівневих змін в роботу системи, в той же час використання хмарної середовища позбавлене цих недоліків.

Характеристики об'єктів соціальних медіа, пораховані в рамках кластера Hadoop або в рамках хмарної середовища обчислення CLAVIRE, необхідні для роботи політик обходу. Політики обходу мережі аналізують дані про вже зібраних об'єктах соціальних медіа та формують новий список об'єктів, для яких так само необхідно завантажити дані. Політики обходу визначають, які саме об'єкти і в якому порядку необхідно «відвідати», тобто завантажити про них дані. Для своєї роботи політики обходу можуть вимагати знання різних характеристик даних, таких як знання зв'язків між

об'єктами, списку вже «відвіданих» об'єктів, різних статистик порахованих за даними. Так само політика обходу для кожного об'єкта визначає URL, за яким можна завантажити необхідну інформацію. Результатом роботи політики обходу є список URL, ранжируваних виходячи зі свого «пріоритету». Політика обходу є «Серцем» системи збору даних. Описані у другому розділі методи підвищення ефективності збору даних реалізуються в політиці обходу. В реалізованої архітектурі політики обходу можуть бути реалізовані як в рамках кластера Hadoop, так і в рамках хмарної середовища CLAVIRE.

Краулер підтримує одночасну роботу з декількома клієнтами одночасно, кожен з яких володіє індивідуальною політикою обходу.

Політики обходу виконуються паралельно і незалежно один від одного. Списки URL, отримані в результаті роботи кожної політики обходу об'єднуються в єдиний список, з якого вибирається безліч URL, за якими на наступній ітерації будуть завантажені дані. Вибір безлічі URL проводиться таким чином, щоб були дотримані квоти на завантаження, що накладаються соціальним оператором і при цьому все клієнти системи, з кожним з яких асоційована політика обходу, отримали частину квоти. Кожному клієнту системи присвоюється пріоритет, пропорційно якому визначається квота клієнта - число URL., які будуть для нього завантажені на наступної ітерації. Для завантаження даних на наступній ітерації роботи вибираються найбільш пріоритетні URL в кількості, що дорівнює квотою клієнта. В результаті роботи цього модуля буде сформовано два списки URL: список URL, за якими буде завантажені дані на наступній ітерації, і список запланованих URL, які будуть оброблені на наступних ітераціях.

3.2 Програмна реалізація системи збору даних та формат зберігання даних

Одним з важливих аспектів роботи з фреймворком Hadoop є визначення формату зберігання даних. Нагадаємо, що всі зібрані дані зберігаються в розподіленої файлової системи HDFS у вигляді пар ключ і значення. На нижчому рівні, вага дані

зберігаються у вигляді послідовності бітів на жорстких дисках. В процесі обробки всі дані перетворюються в Java об'єкти, які подаються на вхід конкретним реалізаціям функцій березня або Reduce. Ключовими моментами є процес перекладу послідовності перекладу бітів в Java об'єкти і назад. Дані процеси називаються процесами сериализации (переклад об'єктів в бітове уявлення) і десеріалізациі (переклад бітового подання до об'єкти). Серіалізация і десеріалізацію використовуються в ситуаціях, коли кільком розподіленим додаткам необхідно обмінюватися один з одним інформацією, або при організації зберігання даних.

Існує безліч різних бібліотек сериализации java об'єктів. Серед найбільш часто використовуваних бібліотек при роботі з Hadoop можна виділити наступні: Apache Thrift, Google Protobuf, Hadoop Writables, Apache Avro. Кожен бібліотека фіксує формат зберігання даних, а так само надає набір методів для сериализации і десеріалізациі. Існує безліч властивостей, за якими бібліотеки сериализации відрізняються один від одного.

- Крос-мовної формат серіалізації або тільки для Java.
- Бінарний або текстовий (JSON, XML, ..) формат представлення даних.
- Ручне або автоматичне відображення об'єктів на дані.
- Простота зміни формату даних.
- Продуктивність: швидкість роботи і обсяг даних.

При роботі з великим об'ємом даних найбільш важливими властивостями є продуктивність, а саме швидкість сериализации і десеріалізациі, а так само простота підтримки даних.

Важливість продуктивності обумовлена роботою з великим об'ємом даних, а так само необхідність їх пересилання по мережі. Оскільки жорсткі диски і процес зчитування і запису на них інформації є найбільш критичними компонентами при роботі з великими даними, продуктивність бібліотек сериализации повинна бути максимальною.

У процесі розвитку системи дані можуть змінюватися змінюються - в них може, як з'являтися нова інформація, так і зникати стара. З точки зору java об'єктів, в них можуть з'являтися і зникати нові поля. Дані зміни відбуваються не тільки внаслідок

розвитку самої системи збору даних, але і внаслідок зміни даних в соціальних медіа. Виникає проблема, при якій дані були серіалізовані в старому форматі, а зчитуються вже в новому форматі. Тому для бібліотеки сериализации особливо важлива простота підтримки даних.

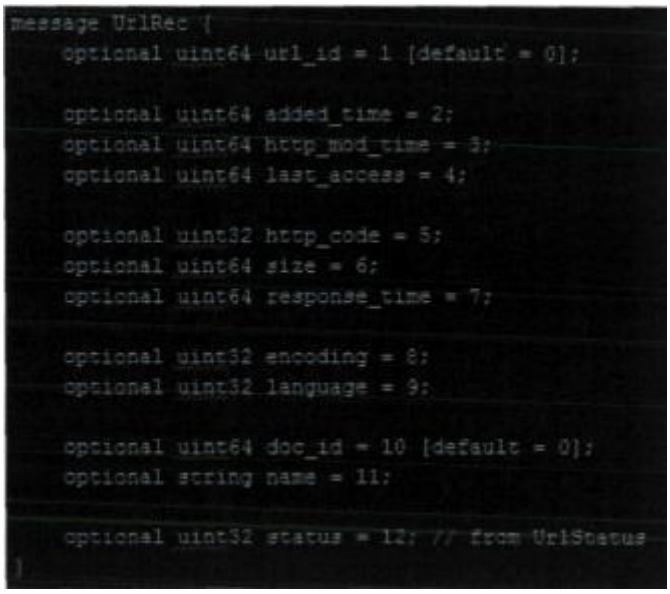

Рисунок 3.3. Приклад опису об'єктів на спеціальній мові для бібліотеки Protobuf

В якості основи для реалізації серіализації і десеріалізациі даних була використана бібліотека Google Protobuf яка володіє необхідним властивостями для роботи з великим обсягом даних.

Дана бібліотека реалізує крос-мовну сериализацию, для чого на спеціальному мовою створиться опис даних рисунок 3.3. На підставі цього опису для безлічі різних мов програмування бібліотекою генерується код, який реалізує сериализацию і десеріалізацію об'єктів. Важливо особливістю є

Простота підтримки коду та кроссплатформенность формату даних.

Для збільшення продуктивності системи все даних, що зберігаються в розподіленої файлової системи IIDFS архівуються. Для стиснення даних використовується бібліотека Google snappy яка прагне не до максимальному стиску даних, а до максимальної швидкості робота з даними.

Реалізований в бібліотеки алгоритм стиснення працює в десятки разів швидше інших популярних алгоритмів, хоча і стислі файли можуть бути більше від 20% до 100%.

Даний алгоритм активно використовується у внутрішніх продуктах компанії Google пов'язаних як із роботою над великими обсягами даних, так і з реалізацією розподілених систем. Швидкість розархівування даних може досягати 500 Мб в секунду, а швидкість стиснення 250 Мб в секунду.

Алгоритм стиснення, який орієнтований на швидкість роботи, особливо на швидкість розпакування, особливо важливий при роботі з великими даними. По перше, великі дані часто записуються один раз і потім багато разів зчитуються. А по-друге, для збільшення продуктивності системи обробної дані, повинен бути дотриманий баланс між навантаженням на процесор і жорсткий диск комп'ютера. Цей баланс і як наслідок максимальна продуктивність, досягаються при роботі з швидшими алгоритмами стиснення.

3.3 Синхронізація між обчислювальними ресурсами і кластером Hadoop

Реалізація всіх операцій в рамках фреймворка Hadoop дозволяє вирішити типові проблеми для розподілених систем. Однак винесення ряду операцій з рамки кластера Hadoop. хоч і дозволяє створити більш гнучку і ефективну систему, але ставить завдання синхронізації між 'обчислювальними ресурсами і кластером Hadoop.

Це завдання вирішується за допомогою використання бібліотеки ZooKccpcr. Zookeeper реалізує розподілене ієрархічне key/value сховище, що володіє можливості:

- Двонаправлений зв'язок між клієнтом і сервером, завдяки якій клієнт може «Підписуватися» на зміни в якомусь із вузлів.
- Можливість створення "тимчасової" пари key/value, яка існує поки клієнт до підключений сервера.
- Стійкість до виходу з ладу кількох машин кластера.

Для виконуваних операцій гарантія виконання запитів FIFO і лінеарізуемость всіх запитів, які змінюють стан системи. Описані властивості дозволяють

використовувати ZooKeeper для здійснення різних координацій в системах розподілених обчислень: розсилка повідомлень між обчислювальними вузлами, реалізація розподіленої пам'яті, сервіс розподілених блокувань, і т.д.

3.4 Швидкий пошук в текстах ключових слів і фраз

Тематичний збір даних заснований на використанні опису теми, яке може складатися з ключових слів або фраз. Опис теми може складатися з десятків і сотень ключових слів і фраз, тому виникає задача їх швидкого пошуку в текстах.

Пошук ключових слів і фраз відбувається в режимі реального часу і здійснюється для всіх нових текстів, які були завантажені на останній ітерації роботи системи.

Так само процес пошуку ключових слів у документі ускладнює наявність «патернів» слів, які прагнуть врахувати різні словоформи слів. Відмітимо, що йод фразами може розумітися як набір слів, що зустрічаються послідовно один за іншому в тексті, так і послідовність слів, які знаходяться на деякій відстані один від одного, а так само слова невпорядковані один щодо одного.

Дане узагальнене поняття фрази так само ускладнює процес їх пошуку в тексті.

Необхідність здійснювати пошук ключових слів у режимі реального часу не дозволяє здійснити «індексацію» документів - процедуру, при якій буде можливо організувати швидкий пошук по слову списку всіх документів, які його містять.

Одним з підходів, який міг би бути використаний для вирішення означеної завдання, є використання регулярних виразів. На підставі списку ключових слів може бути автоматично складено велику регулярний вираз, яке здійснює пошук цих слів в тексті. При цьому можуть бути автоматично враховані різні словоформи ключових слів. Однак на практиці цей підхід виявився не ефективним, оскільки час роботи алгоритму збільшується нелінійно взалежності від числа ключових слів. На темах описуваних десятками ключових слів, час роботи даного алгоритму значно збільшувалася.

Тому для пошуку ключових слів був використаний підхід, заснований на побудові префіксного дерева, що реалізує пошук безлічі подстрок зі словника в цьому рядку [26]. На підставі ключових слів, що описують тему, будується префіксне дерево і реалізується апгорітм Ахо-Корасіка для пошуку подстрок в подається на вхід рядку. Алгоритм дозволяє знайти всі входження в текст всіх ключових слів, що описують тему. Оскільки ключові слова мають безліч словоформ, то префіксне дерево будується виходячи з кореневих форм слів. потім на підставі інформації про становище в тексті цієї кореневої форми, перевіряється, що в тексті дійсно міститься вказане в словнику ключове слово. Таким чином даний алгоритм дозволяє здійснювати швидкий пошук ключових слів у безлічі текстів.

Пошук фраз здійснюється на основі знайдених в тексті ключових слів. В загальному випадку під фразою розуміється модель «мішка слів» (bag of words), при якій не фіксується ні послідовність слів, ні відстань між ними, хоча ці обмеження так само можуть бути враховані. Наприклад, в рамках моделі мішка слів, наявність в тексті фрази «купити машину» означає, що в тексті містяться два ключових слова: «Купити» і «машина». Відзначимо, що порядок розташування слів і відстані між ними так само можуть бути враховані описуваних алгоритмом пошуку фраз.

В рамках моделі мішка слів, кожна фраза описується неврегульованим безліччю ключових слів. Тексти так само описуються неупорядкованими множинами слів. Для перевірки наявності ключової фрази в тексті необхідно перевірити, що безліч всіх ключових слів фрази міститься у великій кількості слів тексту. Оскільки опис теми може складатися з великої кількості фраз, дану процедуру необхідно проводити велику кількість разів.

Для швидкого пошуку всіх ключових фраз в тексті використовуються наступний алгоритм. Всі ключові слова ранжуються виходячи з частоти їх появи в ключових фразах. Слово, що зустрічається в безлічі ключових фраз найчастіше, має мінімальним рангом. Для опису фрази - що входять до неї слова упорядковано виходячи з значення свого рангу. Слова упорядковано для того, щоб можна було виробляти пошук підмножини в безлічі за лінійний час. Потім для кожної фрази вибирається слово представник - слово, яке міститься у фразі і має максимальний рангом. На основі цих

даних створюється асоціативний масив, який за словом дозволяє отримати список всіх фраз, які воно є словом представником.

У тексті проводиться пошук всіх ключових слів, які так само упорядковано виходячи зі свого рангу. Потім ці слова аналізуються в порядку зменшення рангу. Для кожного слова, використовуючи побудований на початку роботи асоціативний масив, виходить список фраз, для яких воно є представником. Для цього невеликого списку фраз перевіряється їх входження в текст. Оскільки списки ключових слів в тексті і у фразі сортовані в однаковому порядку, то перевірка на входження одного безлічі в інше може бути ефективно здійснена за лінійний час від їх розмірів.

Даний підхід виявився ефективним і в разі роботи з тисячами фраз. Після визначення фрази в тексті можуть бути враховані обмеження на порядок слів і відстань між ними. Для цих цілей для кожного знайденого в тексті ключового слова фіксується його позиція в тексті: кількість символів з початку тексту і номер цього слова в тексті. На підставі цих даних можуть бути визначені відносні позиції слів один щодо одного, а так само число слів між ними.

3.5 Реалізація багато режиму

На кожній ітерації роботи краулер формується загальний список URL, по яким необхідно завантажити дані. Виходячи з квот, наданих медійними операторами, на завантаження даних і пріоритетів клієнтів системи для кожного з них визначається індивідуальна квота. Квота клієнта показує число найбільш пріоритетних URL які будуть для нього завантажені на наступній ітерації.

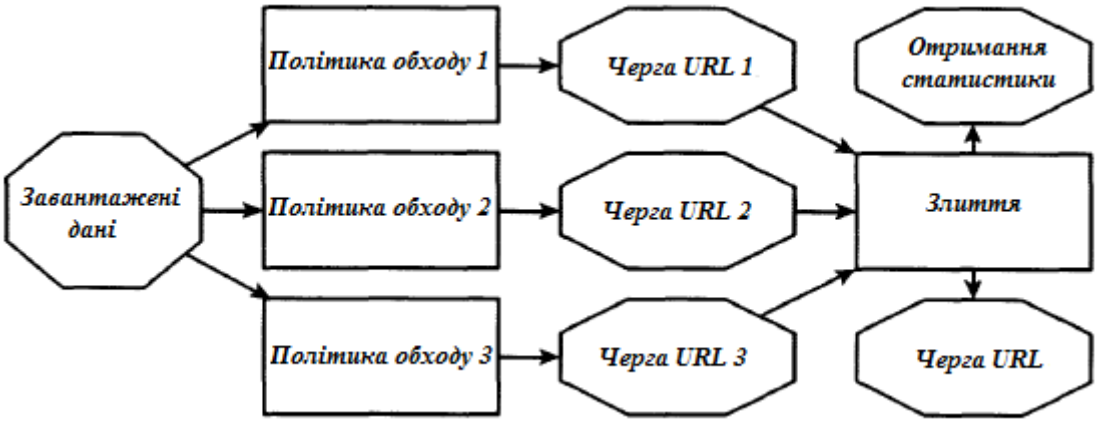

Рисунок 3.4. Процесі формування єдиної черги обходу

Формування єдиної черги і URL дозволяє реалізувати багато користувачів режим роботи з системою, при якому кожен клієнт формує свій список URL. Так ж єдина чергу дозволяє краще контролювати дотримання обмежень, накладаються медійними операторами на завантаження даних. Для цього єдина чергу URL в свою чергу розбивається на дві частини - список URL, за якими будуть завантажені дані на наступній ітерації, і список URL, для яких дані будуть завантажені на подальших ітераціях. При формуванні єдиної черги URL або влаштування її на дві частини порядок проходження URL не змінюється, таким чином, URL обробляються в порядку визначеному політикою обходу. На рис. 3.4. наведено малюнок, який ілюструє процес побудови єдиної черги.

Формально процес визначення квот клієнтів системи описується наступним чином. Для спрощення вважаємо, що всі клієнти працюють тільки з одним соціальним медіа. Так як квоти соціальних медіа незалежні один від одного, то описуваний метод легко розширюється і для випадку роботи з кількома медіа. Нехай на черговий ітерації з системою працює n клієнтів, кожному з яких призначений пріоритет p_i є цілим числом в діапазоні [0,10]. Якщо загальна квота на завантаження даних на наступної ітерації дорівнює M, то кожен клієнт системи на наступній ітерації отримає квоту q_i:

——— (3.1)

Зі списку URL клієнта вибирається q_i найбільш пріоритетних URL за якими будуть завантажені дані на наступній ітерації. Решта URL будуть оброблені на подальших ітераціях. Подібна схема побудови єдиної черги обходу є гнучкою і універсальною, оскільки не залежить від реалізації конкретних політик обходу.

Відзначимо, що реалізовується на даний момент спосіб визначення квот клієнтів не враховує дублікати URL. У разі якщо кілька клієнтів запланували один і той же URL, то при визначенні квот клієнтів цей іЯЬ буде врахований двічі. але при побудові єдиної черги дублікати будуть видалені. Описаний підхід так само дозволяє реалізувати роботу з декількома завантажувачами даних, які працюють паралельно на різних машинах. В цьому випадку для кожного завантажувача даних створюється індивідуальний список URL який буде їм оброблений наступного ітерації.

3.6 Практична реалізація збору даних

Розвиток інформаційних технологій стимулює появу нових методів і напрямків досліджень в різних предметних областях. Зокрема, бурхливе зростання числа користувачів соціальних мереж в Інтернеті забезпечує інформаційну базу, що дозволяє на якісно новому рівні забезпечити дослідження в області социодинамики - розділу соціології, присвяченого кількісним методам моделювання взаємовідносин між індивідами або групами. Традиційно розвиток социодинамики обмежувалося соціометричних фактором – можливістю спостереження (вимірювання) відповідних процесів в суспільстві, оскільки вимірювання і аналіз парних або групових взаємодій індивідів набагато складніші, ніж їх індивідуальних характеристик в рамках вибіркового підходу. Однак в глобальних соціальних мережах подібні взаємини віртуалізуються, формуючи, таким чином, зліпок суспільної структури в просторі Інтернету [27]. Незважаючи на те, що поведінка і характеристики користувачів соціальних мереж в ряді аспектів можуть істотно відрізнятися від реальних, багато в чому ці фактори мають систематичний характер, що дозволяє їх враховувати (шляхом зміщення певних параметрів) при обробці та аналізі відповідних даних. Таким чином, соціальні мережі на даний момент, мабуть, складають основу соціометричних

досліджень нового покоління. У цій статті розглядаються концепція і принципова архітектура спеціалізованого web-орієнтованого виробничо дослідного центру, який надає можливості збору, аналізу та використання в моделях соціометричні даних соціальних мереж в Інтернеті, на основі технологій хмарних обчислень, відповідно до бізнес-моделлю SaaS (Software as a Service, програмне забезпечення як послуга).

Концепція веб-центру. Специфіка виконання кількісних досліджень на основі соціальних мереж пов'язана з рядом особливостей, що обмежують доступність таких даних для широкого кола дослідників; відзначимо деякі з них.

- Доступ до даних глобальних соціальних мереж регламентується політикою оператора мережі та відповідним законодавством в області персональних даних.

Для масштабного збору і аналізу відповідних даних необхідна наявність попередніх угод з оператором.

- Соціальні мережі мають технологічно різні інтерфейси доступу до даними, існують різні принципи обходу мереж. Як наслідок, проведення вимірювань характеристик різних мереж вимагає розробки спеціалізованих засобів збору даних, для окремих дослідників це досить трудомісткий процес.

- Збір даних в соціальних мережах є досить ресурсномісткою операцією і вимагає відповідних виділених обчислювальних ресурсів. Регулярне виконання таких операцій різними користувачами збільшує навантаження на інфраструктуру оператора мережі, що небажано.

- Алгоритми обробки і аналізу таких соціальних мереж у багатьох випадках мають нелінійну складність, оскільки описують взаємини "кожного з кожним ". Тому дослідження мереж досить великого обсягу вимагає застосування

відповідних обчислювальних ресурсів і програмного забезпечення, що допускає ефективне розпаралелювання.

- Візуалізація результатів досліджень на основі соціальних мереж пов'язана з застосуванням досить складних когнітивних алгоритмів, що дозволяють наочно представити різні процеси на непланарних структурах даних великого обсягу, це призводить до необхідності використання спеціалізованого програмного забезпечення.

У сукупності перераховані проблеми перешкоджають розвитку методів і технологій сучасної соціометрії на основі соціальних мереж в Інтернеті, це ускладнюється тим, що потенційні користувачі мереж (фахівці в науках про суспільстві) в більшості випадків не мають спеціалізованими навичками для їх самостійного рішення. Виходом зі сформованої ситуації є експлуатація концепції хмарних обчислень - створення проблемно-орієнтованого середовища, забезпечує доступ до відповідних сервісів (ресурсів, даними і процедурам їх обробки і моделювання) через web-інтерфсйс. В якості основи для розвитку середовища розглядається багатофункціональна інструментально-технологічна платформа CLAVIRE (CLoud Applications VIRtual Environment) забезпечення доступу до сервісів і композитним додатків в середовищі хмарних обчислень.

Проблемно-орієнтоване середовище включає наступні елементи:

- керуюча оболонка (ядро платформи), яка складається з взаємодіючих системних web-сервісів, що функціонують з низькорівневою обчислювальною структурою, що здійснюють підтримку образу хмарної середовища, моніторинг і управління ресурсами, конструювання та виконання сценаріїв, а також підтримку користувацьких інтерфейсів. Керуюча оболонка реалізується на основі концепції iPSE (Intelligent Problem Solving Environment) [27]. Фактично web-центр являє собою відкриту інтелектуальну проблемно-орієнтоване середовище, об'єднуючу розподілені сервіси обчислень і доступу до даних і що дозволяє ефективно управляти паралельними обчислювальними процесами в розподіленої ієрархічної середовищі на основі інтелектуальних технологій [22];

- набір прикладних сервісів обчислень і доступу до даних. У нього входять як різні прикладні пакети для обробки даних і соціодинамічного моделювання, доступні в рамках концепції хмарних обчислень, так і спеціалізовані інструменти для пошуку і вилучення даних збору даних з соціальних мереж - краулери [27].

Всі сервіси будуються на основі предметно-орієнтованих описів пакетів на мовою EasyPackage, що реєструються в базі пакетів керуючої оболонки [28];

- додаткові кошти, що забезпечують підтримку віртуального професійного співтовариства користувачів в рамках концепції web 2.0. вони включають в себе

інтерактивні засоби спілкування, спільного виконання проектів, підтримки єдиного робочого простору, а також коштів, що дозволяють вести дискусії в режимі реального часу з використанням графічних і текстових засобів спілкування. Крім того, є сервіси інтерактивної консультації експертів, передбачається можливість збереження результатів виконаних завдань для подальшого використання іншими членами спільноти або для спільного обговорення та виправлення. В якості окремої завдання розглядаються збір, обробка та аналіз поточної та ретроспсктівной інформації про процеси в віртуальному професійному співтоваристві (включаючи ряд показників індивідуальної та колективної активності користувачів, характеристики затребуваності сервісів і пр.).

Апаратна складова підтримки web-орієнтованого центру формується в складі розподіленої ієрархічної середовища хмарних обчислень, що включає в себе виділені суперкомп'ютери, віртуальні машини в "хмарі" і цільові системи в складі Грід [28,29]. Однаковий спосіб роботи з центром, так само як і оптимізація розподілу обчислювального навантаження, здійснюється засобами керуючої оболонки без залучення користувача.

На рисунку 3.5 представлена схема функціонування web-центру, що демонструє роботу 'сервісів доступу до даних і додатків в області социодинаміки.

Користувач авторизується в проблемно-орієнтованому середовищі через портал провайдера. Через відповідний web-інтерфейс він може вибрати конкретні сервіси або шаблони композитних додатків в формі потоків завдань (workflow, WF), а також отримати (при необхідності) доступ до технічної та експлуатаційної документації. Вибравши необхідні йому сервіси, користувач засобами керуючої оболонки конструює відповідне композитне додаток у формі WF, яке визначає правила збору, обробки та аналізу даних. При цьому може використовуватися як графічна, гак і текстова форма подання композитних додатків. Для підготовленого опису композитного додатки користувач конфігурує умови обчислень: визначає необхідні параметри WF, редагує (при необхідності) його опис, готує і завантажує в сховище середовища вхідні дані для розрахунку. В ряді випадків такі дані (наприклад учбові бази даних фрагментів соціальних мереж) можуть надаватись провайдером web-центра.

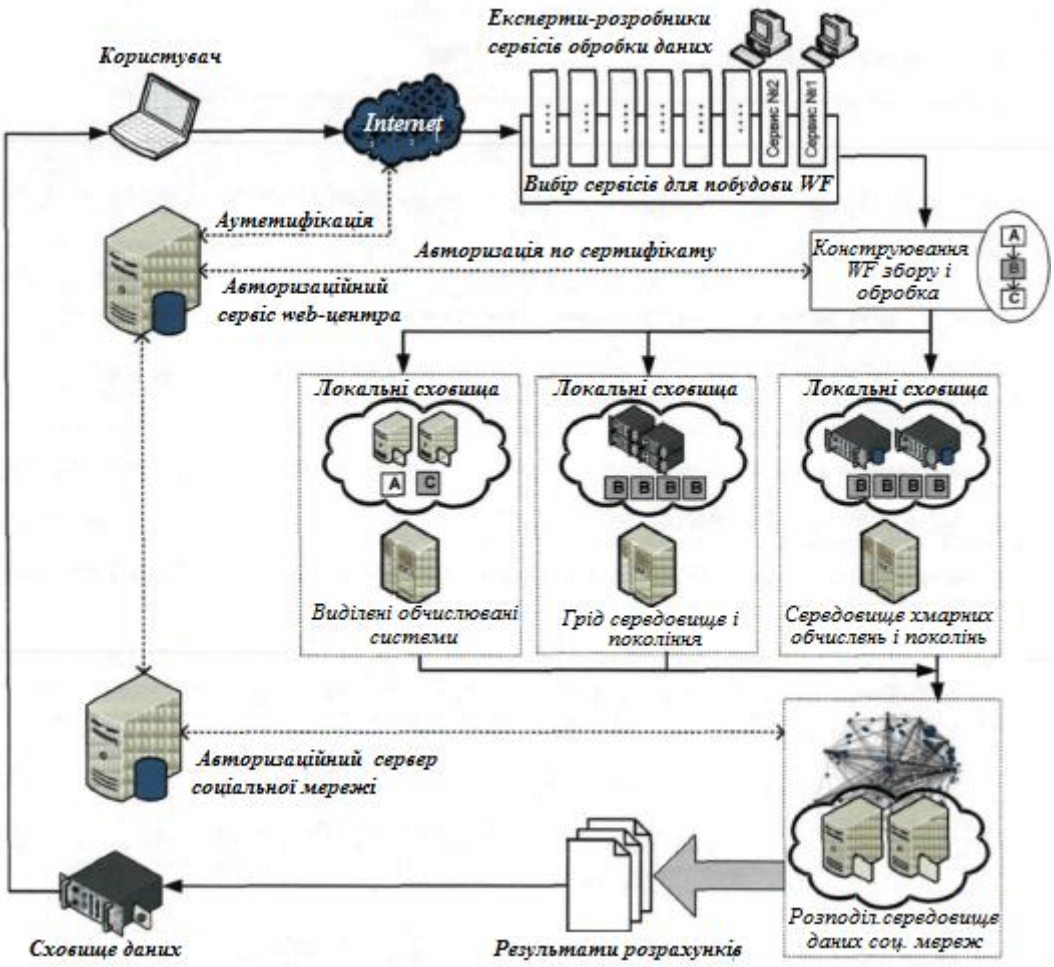

Рисунок 3.5. Схема функціонування web-центру

Потім користувач визначає режим виконання завдання в середовищі (стверджує пропоновані йому варіанти) відповідно до вимог до тимчасових характеристикам розрахунку і правилами доступу до різних джерел даних. При цьому користувачеві пропонуються різні тарифні варіанти, пов'язані з використанням різних соціальних мереж і залученням ресурсів середовищ розподілених обчислень.

Остаточно користувач запускає завдання на виконання в проблемно орієнтованої середовищі. Використання обчислювальних ресурсів і сервісів збору даних в соціальних мережах проводиться з урахуванням єдиного сертифікату, який забезпечує права користувача, делеговані провайдером. В процесі обчислень користувач (при необхідності) здійснює моніторинг процесу виконання, при цьому прогнозується час

завершення обчислень. Коли всі розрахунки завершені, результати поміщаються в сховище даних web-центру; користувачеві надсилається відповідне повідомлення (sms, e-mail та ін.).

Користувач може отримати доступ до результатів розрахунків через інтерфейс проблемно орієнтованого середовища.

Впровадження системи збору даних. Прикладні сервіси можна умовно розділити на чотири функціональні групи: сервіси збору даних в соціальних мережах, сервіси статистичної обробки та аналізу даних, сервіси моделювання сценаріїв, сервіси візуалізації.

Сервіси збору даних в соціальних мережах реалізують різні моделі краулінга (Обхід в глибину, в ширину), з оцінкою спільності відповідно до різних факторів, включаючи семантичний профіль вузлів мережі. В рамках розподіленого середовища ефективним є розпаралелювання мережевого каналу в рамках моделі хмарних обчислень, коли запити до бази відправляються одночасно з різних цільових систем. на кожній цільової системі функціонує робочий агент краулер. Він отримує завдання на перегляд певного безлічі вузлів мережі. Після виконання завдання він передає дані в централізоване сховище. Дії окремих агентів не синхронізуються. Функції керуючого вузла (майстра) полягають в тому, що він визначає порядок, в якому будуть обходитися користувачі мережі, тим самим він реалізує політику обходу краулер. Динамічна архітектура краулер з централізованим управлінням дозволяє динамічно додавати й видаляти агентів, забезпечуючи масштабованість системи в цілому. Крім класичних політик обходу (Обхід в ширину і в глибину) таким чином може бути додатково реалізована політика обходу за ступенем впливу, згідно з якою спочатку відвідуються ті вузли, на які веде найбільше число посилань. Ця евристика дозволяє обходити мережу по топологічним співтовариствам - безлічам тісно пов'язаних один з одним вершин.

Для ефективного збору інформації в соціальних мережах важливо забезпечити високу продуктивність краулер, що досягається за рахунок балансу операцій по перегляду і запису даних в соціальній мережі, і операцій по їх передачі в Інтернет.

Наприклад, в соціальній мережі Instagram за один лінь функціонування краулер обробляє дані близько 700 тисяч користувачів мережі з середньою швидкістю роботи 490 користувачів в хвилину. При цьому виконується близько 270 ітерацій (які відповідають завданням окремим агентам). Аналіз структури витрат часу показав, що найбільш ресурсоємними є операції роботи з базою даних (близько 70%), зокрема - збереження зв'язків між користувачами (18.6%) і списків інтересів користувачів (39.4%). Тимчасові витрати на роботу з мережею не перевищують 27%, що вказує на необхідність оптимізації доступу до бази даних. веб-інтерфейцс доступу до краулера з веб-центра «Социодинаміка» представлено на рисунку 3.6.

Приклад. Динаміка інтересу до заданої теми. Застосування автоматизованих досліджень в соціальних мережах дає нові можливості моніторингу та аналізу реакції громадськості на розроблювані закони, а також моніторингу та аналізу спроб маніпулювання суспільною думкою в політичній сфері.

Можливості розроблених автоматизованих інструментів веб-центру були випробувані в рамках досліджень соціальної мережі.

Другий вивченої темою стали обговорення так званого «Пакету послуг мобільного оператора».

В результаті, стосовно розглянутим темам, були вирішені наступні дослідницькі завдання:

- на основі ключових слів виявлені сукупності користувачів, залучених в обговорення тарифних пакетів послуг або спекуляцій навколо них (в кожній з розглянутих тем виявлено від 400 до 600 обговорюють користувачів, від 500 до 900 повідомлень на обговорювану тему);

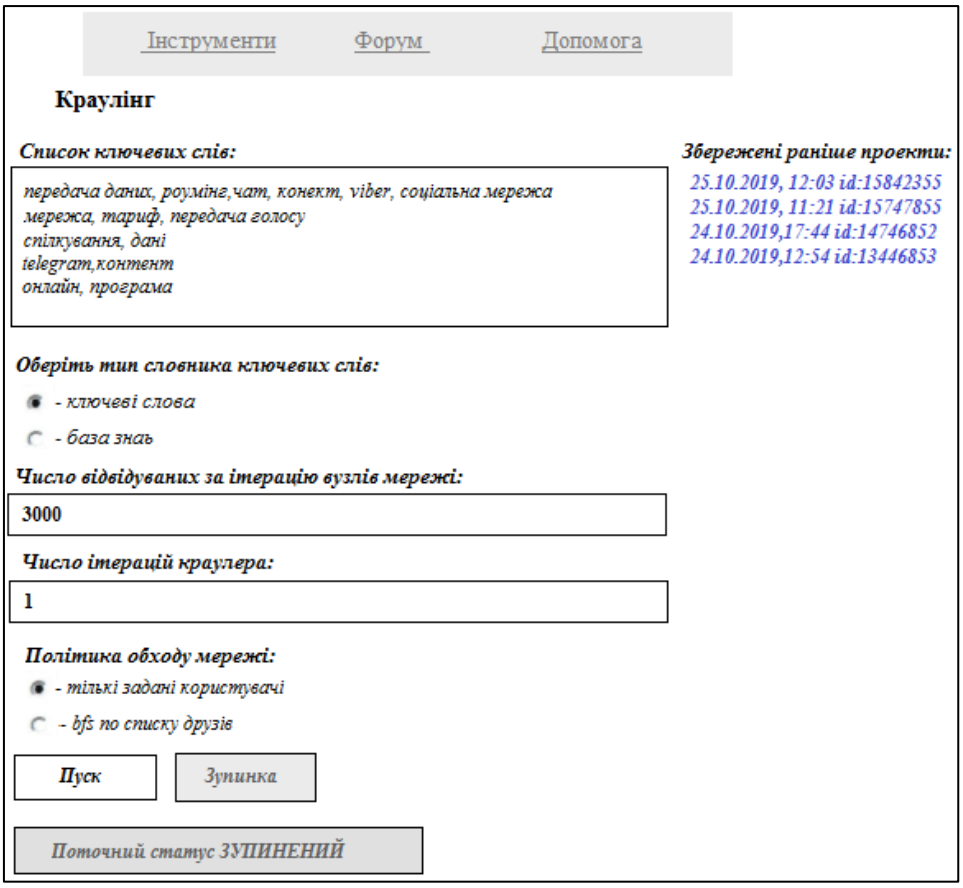

Рисунок 3.6. Інтерфейс доступу до краулер засобами ПАП

- визначено, наскільки серед користувачів, залучених в обговорення розглянутих тем, сильні процеси самоорганізації;

- побудована соціограма спільнот обговорення даних тем (отримані графічні схеми зв'язків між учасниками), обчислені стандартні параметри мереж: щільність мережі, середня кількість зв'язків (передплати), суміжність і т.п .;

- визначені лідери думок;

- введено і проаналізовано індекс активності обговорень по заданих тематиками на основі подобової частоти виникнення повідомлень (виявлена виразна залежність індексу активності обговорення від зовнішніх інформаційних приводів, публікацій ЗМІ і від появи яскравих матеріалів на задану тему в самому співтоваристві);

- в разі спекуляцій навколо теми введення платності середньої освіти виявлені початкові джерела, з яких починається хвиля обговорень.

На закінчення слід зазначити, що поява автоматизованих інструментів дослідження та аналізу віртуальних спільнот на основі платформ хмарних обчислень,

орієнтованим виробничо дослідним центром в області соціодинамики і її додатків, значно розширює дослідницькі можливості операторів зв'язку.

Ці можливості дозволяють на новому рівні застосовувати мережевий підхід до вивчення необхідного контенту, отримуючи цікаві і наочні результати, які не вимагаючи при цьому від дослідника глибоких знань в програмуванні або знання іноземних мов для застосування англомовних програмних пакетів (складових переважна більшість сучасних програмних інструментів).

Розроблені в рамках веб-центру навчально-методичні матеріали дозволяють ефективно і з опорою на практичні приклади навчати студентів гуманітарних напрямків застосування мережевого підходу до аналізу віртуальних спільнот і процесів, що відбуваються в них, що сприятиме більш широкому застосуванню даної методології та інструментарію.

Були проаналізовані та інші теми, викликавши великий резонанс рисунок 3.7.

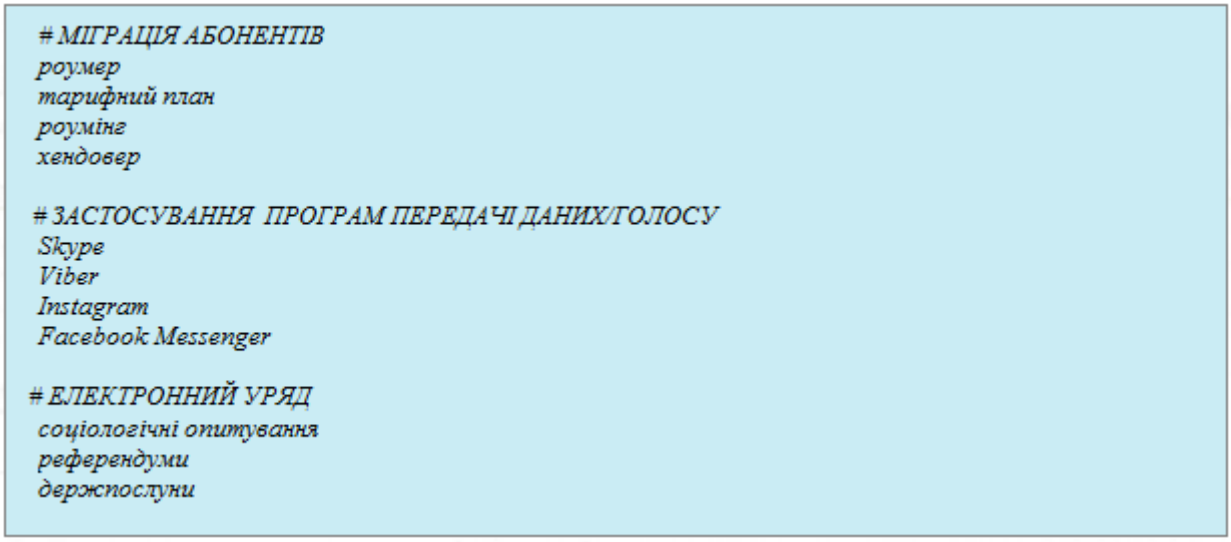

Рисунок 3.7. Опис різних тем, які викликали великий резонанс

На рисунку 3.8 представлена візуалізація інтенсивності обговорення цих тем в різні моменти часу. Кожна точка відповідає числу повідомлень за вказаною темі, які були написані в зазначений тиждень. Для аналізу були використані дані з Українського сегмента Facebook. Для пошуку записів, з яких було розпочато тематичний збір даних, була використана вбудована з сервіс Facebook пошукова система, яка базується на технології «Google. Пошук по блогах».

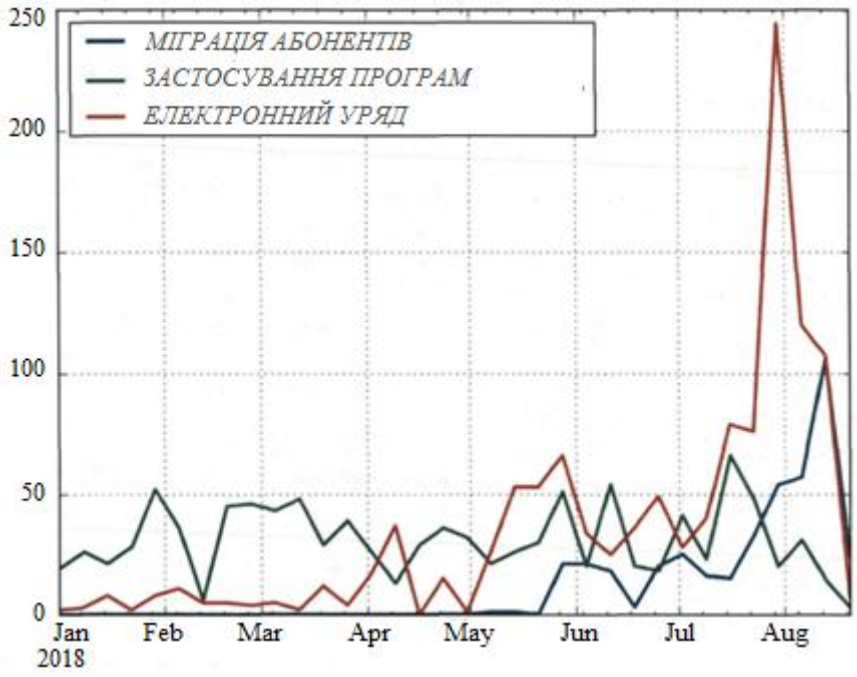

Рисунок 3.8. Інтенсивність запитів користувачів

Зараз абоненти все рідше дзвонять і відправляють SMS і більше користуються інтернетом. У зв'язку з цим більшість великих телеком-операторів розглядають можливість використання технології Big Data для розширення списку послуг. Як приклад може послужити сервіс оператора «Київстар», що аналізує переваги глядача і потім пропонує йому плейлист з фільмів, які, можливо, його зацікавлять.

ВИСНОВКИ

Таким чином, незважаючи на певні складнощі і проблеми розвитку, аналіз показав, що технології Big Data стають одним з найважливіших напрямків

формування нових сервісів, підвищення конкурентоспроможності сервісних підприємств, створення інноваційних маркетингових інструментів просування послуг у різних сферах діяльності людини.

Безсумнівно, великі дані сформують різні ринки – від тих, на яких продають дані лотами різних обсягу й якості, до тих на яких надаються високотехнологічні сервіси з машинним часом суперкомп'ютерів.

Перехід до збору та обробки інформації в обсягах, що перевищують традиційні, може стати хорошим приводом для спеціалізованого або широкого реінжинірингу бізнеспроцесів (і залучених у них бізнес-об'єктів).

При цьому доведеться визнати пріоритет за моделлю інтегрування Великих даних у бізнес-модель по всій структурі й усіх напрямах.

Аналіз принципів збору даних з соціальних медіа виявив ряд важливих відмінностей від збору даних з Інтернету. З точки зору побудови архітектури системи збору даних, найбільш важливою особливістю є робота з контекстом користувача, а не окремими веб-сторінками.

Огляд існуючих систем збору даних показав, що вони орієнтовані на вирішення конкретних завдань і тому досить складно розширюваність. Так само ці системи не орієнтовані на роботу з декількома соціальними медіа і не дозволяють обробляти великі обсяги даних. Тому рішення задач моніторингу та отримання репрезентативних вибірок в них важко. Тому для створення інструментів роботи з даними соціальних мереж доцільно використовувати концепцію і технології Big Data. Завдяки якій можна будувати систему збору даних, абстрагуючись від вирішення завдань, тим самим забезпечуючи універсальність розроблених методів, технологій і програмних засобів.

У третьому розділі розглянута і реалізована архітектура платформи, яка поєднує технології Big Data і хмарних обчислень, за допомогою яких можлива ефективна обробка великих даних. Для організації розподіленого зберігання і обробки даних використовуються технології Big Data , зокрема фреймворк Hadoop, який реалізує модель MapReducc для розподілених обчислень. Це дозволяє уникнути вирішення типових для розподілених програм завдань і сконцентрувати на реалізації системи збору даних.

Запропонована адаптивна процедура ефективного збору даних в соціальній мережі на основі методів передбачення часу виникнення нових подій в соціальній мережі і побудови репрезентативних вибірок з урахуванням топології зв'язків.

Для тематичного збору даних у відсутності навчальної вибірки застосовуються методи інформаційного пошуку, засновані на використанні тематичного словника із ключовими словами по темі, що цікавить. На підставі яких автоматично складаються списки ключових фраз, які використовуються для усунення семантичної неоднозначності. Для організації інформаційного пошуку застосовуються методи методи булевого пошуку або статистичні заходи, які оцінюють важливість терміна в контексті документа.

У практичній частині магістерської роботи виконано дослідження даних, зібраних з соціальної мережі Facebook. Зокрема вивчається питання про аналіз питанню вибору ефективного способу зв’язку соціальних медіа. Аналізується портрет користувачів, залучених в дискусії про кращий програмний спосіб передачі інформації в соціальних медіа, шляхом здійснення тематичного збору даних з Facebook. Здійснюється пошук користувачів, які пишуть про свій вибір програмного забезпечення. Для знайденого спільноти користувачів визначаються характерні для нього інтереси, які зустрічаються в ньому «значно» частіше, ніж в середньому по всій мережі. Так само визначаються користувачі, які мають досить багато інтересів, характерних для користувачів зацікавлених в темі роумінгу.

ПЕРЕІК ПОСИЛАНЬ

1. Viktor Mayer-Schönberger. Big Data: A Revolution that Will Transform how We Live, Work, and Think /Viktor Mayer-Schönberger, Kenneth Cukier. – New York: Houghton Mifflin Harcourt, 2017. – 242 с.

2.Shubham Sharma. Big Data Landscape/ Shubham Sharma.// International Journal of Scientific and Research Publication. – 2016. – №6.

3. White T. Hadoop: The Definite Guide/ Tom White. – Boston: O'Reilly Media, Inc., 2018. – 657 с.

4. Zikopoulos P. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data/ P. Zikopoulos, C. Eaton. – Pennsylvania: McGraw-Hill Osborne Media, 2014. – 176 с.

5. Рижко Б. Розроблення системи автоматизованого збору, оброблення та аналізу даних на основі технологій Big Data/ Б. Рижко, П. Катін // Інфокомунікаційні системи та технології. – 2018. – № 2(2). – С. 37-41.

6.Baburin V. A., Janenko M. E.Big Data technologies in service: new markets, opportunities and challenges// TTPS. № 1 (27). P. 100–105.

7.Analytical review of Big Data market/ Exchange house of Moscow. https://habrahabr.ru/company/moex/blog/25674.

8. Suvorov N. I., Bedenkov A. V. Big data in Russian health care. The time has come// Remedium. 2014 № 6. P. 60–61.

9. Болгова Е.В. автоматизация процесса разработки виртуальных лабораторий на основе облачных вычислений: дис. канд. техн. наук: 05.13.06. 2012.

10. Князьков К.В. технология разработки композитных приложений с использованием предметно-ориентированных программных модулей: дис. Канд. Техн. Наук: 05.13.11.2017. P. 170.

11. Pinkerton B. Finding what people want: Experiences with the WebCrawler// Proc. Second Int. World Wide Web Conf. Chicago, 2003. Vol. 94. P. 17-20.

12.Eiron N., McCurley K.S. Analysis of anchor text for web search // Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. informaion Retr. SIGIR 03. ACM Press, 2013. P. 459.

13.Smirnov I. Overview of Stemming Algorithms//Mech. Transl. DePaul University, 2018.

14. Zhao V. et al. Ontology Classification for Semantic-Web-Bascd Software Engineering. 2014. Vol. 2, № 4. P. 303-317.

15. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval// 2008 /ed. McGraw-Hill. Cambridge University Press, 2018. Vol. 1, № c. P. 496.

16.Cho J., Garcia-molina H. Synchronizing a database to Improve Freshness. 2016. P. 1-30.

17.Cho J., Garcia-Molina H. Effective page refresh policies for Web crawlers//ACM Trans.Database Syst. 2013. Vol. 28, № 4. P. 390-426.

18. Cho J., Garcia-Molina H. Estimating frequency of change // ACM I rans. Internet Technol. 2013. Vol. 3, № 3. P. 256-290.

19.Neyman J., Pearson E.S. Sufficient statistics and uniformly most powerful tests of statistical hypotheses. // Stat. Res. Mem. 2016. Vol. I. P. 113-137.

20.Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics H Language (Baltim). Prentice Hall, 20015

21.Manning C.D., Raghavan P. An Introduction to Information Retrieval // Online / ed. Salas A.C.-B.E. Cambridge University Press, 2018.

22. Tarpcy T. A Parametric k-Mcans Algorithm. // Comput. Stat. 2017. Vol. 22, № 1. P. 71-89.

23.Lammel R. Google’s MapReduce programming model — Revisited // Sci. Comput. Program. 2008. Vol. 70. № 1. P. 1-30.

24.Michael M. et al. Scale-up x Scale-out: A Case Study using Nutch/Lucene // 2007 IEEE Int. Parallel Distrib. Process. Symp. Ieee, 2007. P. 1-8.

25. Lammel R. Google’s MapRcduce programming model — Revisited // Sci. Comput. Program. 2008. Vol. 70, № 1. P. 1-30.

26. McClellan M.T., Minker J., Knuth D.E. The Art of Computer Programming, Vol. 3: Sorting and Searching // Math. Comput. 1974. Vol. 28, № 128. P. 1175.

27. Бухановский А.В., Ковальчук С.В., Марьин С.В. Интеллектуальные высокопроизводительные программные комплексы моделирования сложных систем: концепция, архитектура и примеры реализации // Известия вузов. Приборостроение. 2009. Vol. 52, № 10. P. 5—24.

28.Афанасьев А.П., Волошинов В.В., Посыпкин М.А. С.О.В. Объединение высокоуровневых вычислительных ресурсов в распределённой среде //

Распределенные вычисления и Грид-технологии в науке и образовании Груды 4-й междунар. конф. (Дубна, 28 июня - 3 июля, 2010 г.). 2010. P. 290-296.

29.Афанасьев А. П., Волошинов В. В., Посыпкин М. А., Сигал И.Х. Программный комплекс для решения задач дискретной оптимизации нараспределенных вычислительных системах // Сб. трудов ИСА РАН. 2006. Vol. 25. P. 5-17.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ