УДК 004.94 https://doi.org/10.15407/pp2019.04.053

О. Захарова

КЛАСИФІКАЦІЯ МЕТАДАНИХ ВЕЛИКИХ ДАНИХ

Насьогодні накопичені величезні обсяги даних різної структури (або в загалі не структуровані) та походження, їх обсяги зростають експоненційно. Проблема полягає у тому, що існуюче програмне та апаратне забезпечення, яке не може обробити таку кількість різноманітних типів даних, що створюються з великою швидкістю. Великі дані стали надто складними та динамічними, щоб їх можна було обробляти, зберігати, аналізувати та управляти ними за допомогою традиційних засобів. Це обумовило виникнення нових платформ та підходів для роботи з даними, а разом з цим чітке розуміння того, що для вирішення задач великих даних ці необроблені дані повинні бути доповнені метаданими. Насьогодні метадані є засобом класифікації, впорядкування та характеристики даних, та їх вмісту. Їх найважливішою особливістю є впорядкована структура. Завдяки структурованому вигляду, метадані доступні для читання не лише для людини, але й для комп'ютера. Таким чином, вони можуть оброблятися автоматизовано та використовуватися для різних цілей: для індексації, пошуку, об'єднання, автоматичної обробки, класифікації великих даних тощо. Побудова ефективних систем керування метаданими, перш за все, вимагає їх узгодженої загальної класифікації з урахуванням типів джерел (способів отримання) даних, що формують контент задач, що мають вирішуватися на різних етапах життєвого циклу, існуючих форматів представлення відповідних даних, принципів розумної ефективності, так як часто розміри метаданих значно перевищують обсяг самих даних (навіть великих), які вони описують. Тому, мета даної роботи полягає в аналізі існуючих джерел великих даних, способів створення та обробки відповідних метаданих, а також програмних засобів, що дозволяють опрацьовувати метадані певним чином, та побудові класифікації метаданих на основі проведеного аналізу.

Ключові слова: джерела великих даних, керування метаданими, Hadoop, класифікація метаданих, аналіз метаданих, сервіси обробки метаданих, створення метаданих, перегляд метаданих, редагування метаданих, метадані зображень, метадані аудіо- файлів, метадані відео-файлів, метадані сховищ, метадані в соціальних мережах.

Вступ

При впровадженні кожного нового проекту великих даних має бути можливість їх ідентифікувати. Важлива можливість для розробки та розвитку сервісів обробки великих даних – створення комплексної програми керування метаданими. Метадані можуть значно спростити та вдосконалити процеси збору, інтеграції та аналізу великих даних. За відсутністю метаданих підприємства можуть втратити глибоке розуміння того, що саме можуть дати великі дані. Метадані можуть керувати всім життєвим циклом даних, процесами, процедурами, а також клієнтами або користувачами, які впливають на певну бізнес-інформацію. Вони є основою для збору величезних обсягів даних з розрізнених джерел та інформаційних сховищ, перш ніж вони стануть некерованими.

Слід також зазначити, що зростання ролі метаданих значною мірою обумовлене розвитком Веб. Це стосується не лише виникнення самих великих даних. Веб звільняє від обмежень фізичного світу, а саме контент, що формують великі дані, вже не повинен знаходитись в один момент часу в одному місці, як стілець чи шафа. Веб контент може існувати в багатьох місцях одночасно. Метадані дозволяють помітити контент термінами з таксономії, що описує його мету. Для цього можуть бути використані півдюжини різних словників. Деякі можуть описувати семантику контенту, інші – функції/задачі контенту та аудиторію, для якої він призначений, формат контенту, його структуру, API платформу чи компоненти, що використовуються та інше. Потім ці метадані можуть використовуватися багатьма організаціями відповідно до різних цілей. Правила відповідної системи керування сайтом (CMS) визначають, де буде з'являтися контент.

Метадані можуть бути застосовані як для опису контенту, так й бути представлені як фасети у результатах пошуку, щоб визначити фільтрацію результату.

Окрім цього, можна створювати складні фільтри, які комбінуватимуть метадані різними способами, наприклад, якщо потрібно знайти весь контент, що відповідає (інструменту) ACME API та використовується для налагодження віджетів (функція), а також призначений для розробників (персонал). Інформацію можна отримати миттєво за допомогою фасетних фільтрів або завчасно побудованих запитів.

Звільнення контенту від статичної, єдиної позиції у змісті, ймовірно, є найбільш актуальною та явною перевагою, яка з'явилася в результаті переміщення контенту до Інтернету. Але доки контент не збагачений метаданими та семантичними мітками, його складно знаходити, витягувати, просувати та обробляти різними способами. Це вимагає спеціальних інструментів для створення та анотування контенту метаданими. Слід зазначити, що контент має бути таким, що легко переноситься. Звісно, що використання сторонніх схем метаданих вимагає додаткових зусиль щодо оформлення контенту, але це може забезпечити кращу взаємодію з користувачами на різних платформах у пристроях у сьогоденні та майбутньому. Тому, створення добре структурованого контенту та забезпечення його гнучкості, та можливості легкого перенесення є ключем для вирішення багатьох задач.

Впорядкована структура саме є найважливішою особливістю метаданих. Інформація категоризована та має визначену форму/формат. Наприклад, категорію *час створення* можна визначити лише за допомогою формата запису дати та часу. Завдяки структурованому вигляду, метадані доступні для читання не лише для людини, а й для комп'ютера. Таким чином, вони можуть оброблятися автоматизовано та використовуватися для різних цілей: для індексації, пошуку, об'єднання, автоматичної обробки тощо.

Побудова ефективних систем управління метаданими, перш за все, вимагає їх узгодженої загальної класифікації з урахуванням типів джерел (способів отримання) даних, що формують контент, задач, що мають вирішуватися на різних етапах життєвого циклу, існуючих форматів представлення відповідних даних, принципів розумної ефективності, так як часто розміри метаданих значно перевищують обсяг самих даних (навіть великих), які вони описують.

Джерела великих даних

Класифікація метаданих неможлива без усвідомлення типів джерел надходження великих даних, різновидів самих великих даних, потреб користувачів у їх використанні та способів обробки.

На найвищому рівні відповідно до способів отримання даних можна виділити внутрішні та зовнішні джерела (рис. 1).

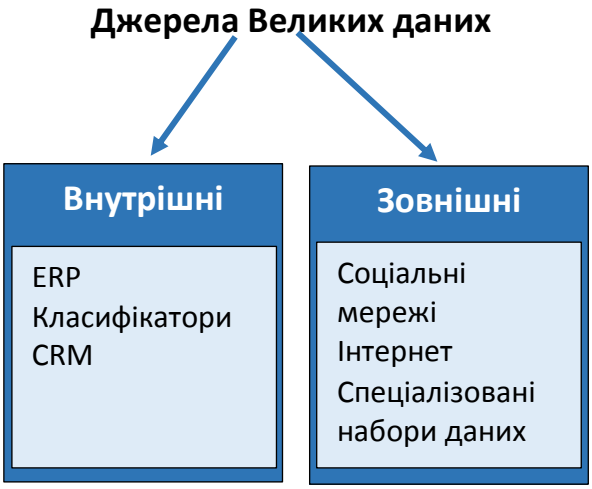

Рис. 1. Внутрішні та зовнішні джерела даних

Найбільш традиційним типом джерел даних є бази даних (БД). Існує багато різних БД, які мають власну архітектуру та властивості. Але використання транзакційних БД насьогодні вже не є оптимальним рішенням для вирішення задач у бізнес-аналізі. На це існує багато причин, зокрема, вимоги щодо попередньої оптимізації даних для звітності та аналізу, вимоги до структурованості контента, певні обмеження обсягів контенту, низька швидкість виконання запитів до даних тощо. У деяких випадках компанії використовують ETL засоби для збору даних з транзакційних БД, перетворення їх певним чином, щоб вони були оптимізовані для бізнес-аналізу, та завантаження їх у сховище та інші вітрини даних. Але, насьогодні, в умовах відкритого світу та інформаційного

буму, треба мати можливість обробляти дані з не структурованих або слабко структурованих інформаційних джерел. На рис. 2 показано порівняльну характеристику можливих джерел інформації різного рівня структурованості. Аналіз джерел проводиться за основними властивостями великих даних, а саме 3V характеристиками: швидкість (Velocity), різноманітність (Variety) та обсяг (Volume).

Як основні джерела отримання інформації можна виділити.

1. *Архіви відсканованих документів*, заяв, страхових форм, медичних записів, кореспонденції, архіви паперових документів, друковані файли потоку, які містять вихідні системи записів між організаціями та їх користувачами.

2. *Документи*: файли різних форматів xls, word, html, html 5, pdf, csv, ppt, txt, xml, json тощо.

3. *Сховища даних* (SQL або NoSQL), файлові системи тощо.

Сьогодні підприємства вважають за краще використовувати як традиційні, так й сучасні БД (разом), щоб отримати відповідні великі дані. Ця інтеграція відкриває шлях для гібридної моделі даних і вимагає низьких інвестиційних витрат та витрат на IT-інфраструктуру. Крім того, такі гібридні БД розгортаються також й одночасно для кількох цілей бізнес-аналітики, та потім можуть забезпечувати вилучення інформації, яка використовується для отримання прибутку. Процес вилучення та аналізу даних з джерел великих даних є складним процесом. Ці проблеми можна вирішити, якщо організації охоплюють всі необхідні міркування великих даних, беруть до уваги відповідні джерела даних і розгортають їх у спосіб, який добре налаштований на цілі організації.

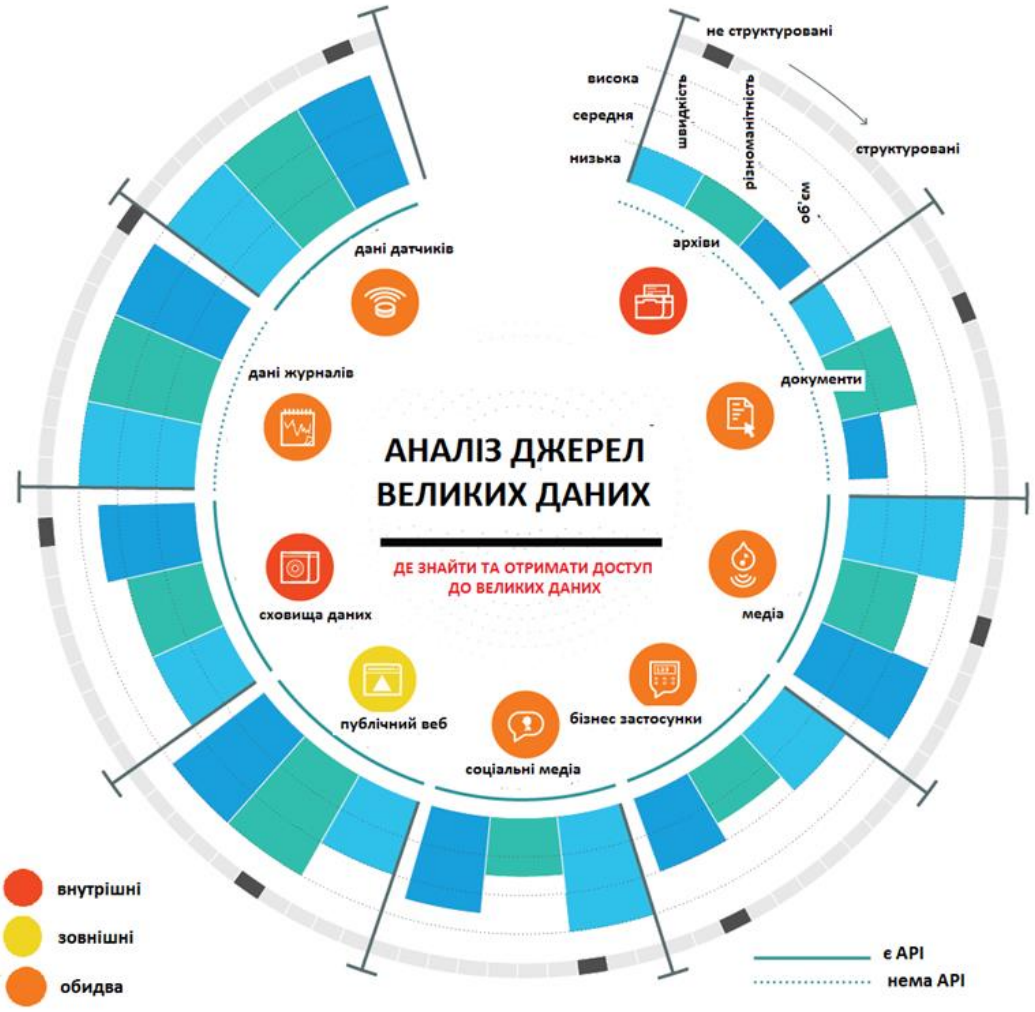

Рис. 2. Аналіз джерел великих даних

4. *Бізнес-застосунки*: системи керування проектами, системи автоматизації маркетингу, продуктивності, CRM системи, керування ERP контентом, HR, керування талантами, системи закупівель, керування витратами, портали, Інтернет системи тощо.

5. *Публічний веб*. Цей веб містить дуже поширені та легко доступні великі дані. Дані в Інтернеті є загально доступними як для фізичних осіб, так і для компаній. Різноманітні веб сервіси (сервіси охорони здоров'я, загальнодоступні фінансові сервіси, Вікіпедія, сервіс перепису населення, економічні сервіси, встановлення відповідності тощо) надають всім безкоштовну та швидку інформацію. Величезні розміри мережі забезпечують зручність її використання у різних аспектах. Веб є особливо корисним для стартапів та малих і середніх підприємств, оскільки звільняє їх від необхідності розробки власної інфраструктури великих даних й сховища даних, перш ніж вони зможуть використовувати ці великі дані.

6. *Засоби масової інформації* (ЗМІ) є найпопулярнішим джерелом великих даних, оскільки надають цінну інформацію про переваги споживачів і змінюють тенденції. Оскільки це інформація, що транслюється самостійно і перетинає всі фізичні та демографічні бар'єри, медіа джерела є для підприємств найшвидшим способом отримати глибокий огляд своєї цільової аудиторії, побудувати закономірності, зробити висновки та підвищити рівень прийняття рішень. ЗМІ включають соціальні медіа та інтерактивні платформи, такі як: Google, Facebook, Twitter, YouTube, Instagram, а також загальні засоби масової інформації (медіа), такі як: зображення, відео, аудіо, записи прямих ефірів та підкасти (цифрові аудіо файли в Інтернеті, готові до завантаження), які забезпечують кількісні та якісні висновки щодо кожного аспекту взаємодії з користувачем.

7. *Дані журналів*: дані серверів, журнали подій, журнали бізнес-процесів, журнали прикладних систем, дані потокових каналів, детальні записи викликів, мобільні локації, використання мобільних застосунків тощо.

8. *Автоматично генерований контент* (IOT) – це значиме джерело великих даних. Це дані, які зазвичай генеруються датчиками, що пов'язані з електронними пристроями. Це не лише інформація, що генерується комп'ютерами та смартфонами, але й дані з медичних пристроїв, датчиків автомобілів, розумних електричних лічильників, дорожніх камер, супутників, пристроїв запису дорожнього руху, процесорів у транспортних засобах, відеоігр тощо.

Метадані різних типів даних

Контент, що описується метаданими, може бути представлений у різних форматах. І, в загальному випадку, кожний тип файлу має власний стандарт для визначення метаданих. Але, насправді існує не так вже й багато схем, протоколів та форматів представлення метаданих. Розглянемо далі можливі схеми метаданих відповідно до типів джерел великих даних, які вони описують.

Метадані зображень. Зображення, як й будь-який інший контент, також мають різні формати та *схеми метаданих* [1]. У випадку графічних файлів, таких як фотографії з цифрової камери або смартфону (формати JPEG, TIFF та RAW), ми переважно маємо справу з метаданими формату Exif, які є досить розвиненими порівняно з метаданими інших типів файлів.

EXIF (Exchangeable Image File Format) – стандарт зберігання метаданих у зображенні, що використовується цифровими камерами для зберігання інформації про витримку, діафрагму та інші параметри зйомки. Метадані у форматі EXIF можуть зберігатися у файлах форматів JPEG, TIFF та RIFF WAV.

EXIF формат був розроблений з виникненням цифрових камер та, перш за все, для опису зроблених ними зображень, і, тому, ці дані є у кожній фотографії, не залежно від того, з якого пристрою вона була зроблена. EXIF – це не лише параметри фотоапарата/смартфона, з якого було зроблено фотографію, але й багато іншого: дата створення, геолокація, інформація про власника кадра тощо. За стандартом з

користувацьких описових метаданих в EXIF може зберігатися лише опис (тег Description) та коментар (тег User Comment), але Windows Explorer використовує також декілька додаткових тегів (XPTitle, XPSubject, XPAuthor, XPComment, XPKeywords). Windows Explorer ігнорує тег XPTitle за наявності стандартного тега Description. Таким чином, це в основному відомості про характеристики, структуру, розміщення зображення, умови та способи його отримання, автора, дату та час його змінення, та дуже мало справжньої семантики. З повним переліком елементів метаданих можна ознайомитись у специфікації стандарту [2].

EXIF є частиною більш широкого стандарта DCF (Design rule for Camera File system). DCF [3] дозволяє визначити метадані не лише для файла зображення, але й пов'язаного з ним файлу, наприклад, аудіофайлу. DCF також містить досить детальну, але технічну інформацію.

Якщо виникає необхідність визначення більш детального опису змісту фотографії, Exif може бути розширений набором *метаданих у стандарті IPTC*, який, окрім полів, що пов'язані з темою фотографії, має розділ для контактних даних фотографа. Це стандартний додаток графічних файлів, що доступні через банки фотографій.

IPTC (International Press Telecommunications Council) – це скоріше назва організації, що розробила даний стандарт, Міжнародна Рада з питань преси та телекомунікації. Назва самого стандарту – IIM (Information Interchange Model). На відміну від EXIF, який спрямований на технічну інформацію, IIM дозволяє зберігати різну детальну інформацію. У метаданих даного стандарту можуть зберігатися такі описові поля, як ObjectName (заголовок), Key words (ключові слова), Caption (опис, існує декілька варіацій тега).

У початкових версіях стандарту метадані зберігалися таким чином, що програмне забезпечення (ПЗ), яке не знало про існування IPTC, не могло працювати з файлами зображень з такими метаданими. Але, згодом стандарт був розширений Adobe, та метадані були перенесені до блоку APP13 JPEG-файлу, що дозволило такому ПЗ успішно читати JPEG-файл, ігноруючи невідомі метадані. IPTC інформацію підтримують фотобанки, пошукові сервіси тощо.

XMP (eXtensible Metadata Platform) [4] – стандарт, розроблений Adobe, який, починаючи з 2012 року, є також стандартом ISO. Метадані зберігаються у моделі RDF, яка представлена в XML форматі. Це дозволяє включити будь-яку необхідну інформацію до файлу зображення. Це стандарт з відкритим кодом. XMP метадані можуть бути додані до графічних файлів багатьох різних типів.

Властивості метаданих групуються в схеми. Кожна схема ідентифікується унікальним URI простору імен та містить довільну кількість властивостей. Хоча URI просторів імен виглядає досить схожим на веб-адреси (в дійсності, вони часто виглядають однаково), важливо зазначити, що вони не ідентифікують конкретну веб сторінку. Це просто унікальні ідентифікатори для деякої сутності, що використовується в XMP. Остання специфікація включає більше десятка передвизначених схем з сотнями властивостей для документа загалом та характеристик зображення. Більшість таких схем називається Дублінським ядром (DC) [5] та включає загальні властивості такі як: Назва, Тема, Автор та Опис. Окрім передвизначених схем, можуть бути визначені спеціальні (власні) схеми, щоб задовільнити специфічні вимоги до метаданих організації. Власні формати метаданих можна імпортувати та розповсюджувати разом з іншими XML файлами.

Все це дозволяє розробникам досить просто адоптувати специфікацію стандарту до стороннього ПЗ. Дана технологія забезпечує також обмін метаданими між прикладними системами. Метадані, що визначені в інших форматах, таких як Exif, IPTC (IIM), GPS, та TIFF, можна синхронізувати з XMP, що полегшує їх використання та керування ними. Даний стандарт, реалізований у всіх Adobe продуктах підтримується десятками незалежних розробників ПЗ та груп користувачів.

Насьогодні специфікація XMP

складається з трьох частин:

– *«Модель даних, серіалізація та базові властивості»* [6] охоплює модель представлення основних метаданих, що є базисом формату XMP стандарту. Модель даних прописує, як можуть бути організовані метадані XMP, незалежно від формату файлу чи специфіки використання. Модель серіалізації визначає, як модель даних представляється в XML, зокрема, в RDF.

– Друга частина [7] забезпечує детальний перелік властивостей для схем метаданих XMP стандарта; вони включають схеми загального призначення, такі як: Дублінське ядро, та схеми спеціального призначення для застосунків Adobe, таких як Photoshop. Також забезпечує інформацію з розширення існуючих схем та створення нових.

– *«Зберігання у файлах»* забезпечує інформацію про те, як серіалізовані XMP метадані пакуються у XMP пакети та вбудовуються у файли різних форматів. Включає інформацію щодо відношення XMP з іншими форматами метаданих, та узгоджує значення, що представлені в інших форматах.

Існує також стандарт ISO 16684-2:2014, *Graphic technology – Extensible metadata platform (XMP) – Part 2: Description of XMP schemas using RELAX NG* (опис схеми XMP за допомогою RELAX NG), який специфікує використання RELAX NG для опису серіалізованих XMP метаданих. Стандарт визначає, як відповідні схеми можуть використовувати функції RELAX NG.

Метадані аудіофайлів. Завдяки приєднанню метаданих до звукових файлів виникає можливість доповнити звук (чи музику) будь-якою необхідною інформацією. Інформація може бути дуже різноманітною, виходячи з вимог та бажань автора чи власника запису. Метадані – це не лише назва трека чи рік випуску музичного альбому, але й ім’я композитора, автора аранжировки, тексту слів до пісні, адреси сайтів, електронні адреси, все, що пов’язане з художнім оформленням пісні або альбому, якщо це музичний файл та багато іншого. У сучасних мультимедійних пристроях пошук композицій здійснюється не за назвою файлів чи папок, а за метаданими, які там містяться, та лише за умови їх відсутності, за назвою файлів. Виробники мобільних пристроїв таких відомих марок, як наприклад, Nokia, Sony Ericsson, iPod, і та інші слідують цьому ж принципу.

Першим запропонував додавати до MP3-файла невеличкий блок з даними програміст Эрик Кемп (проект «Studio3»). Цей блок був названий ID3tag (tag – ярлик, мітка, ID3 – Identification Data for Studio3). В наслідок цього назва TAG надійно закріпилася за метаданими інших форматів, таких як: WMA, OGG, MP4 та інших.

Щоб не викликати не сумісності з плеєрами, тег розміщувався наприкінці файлу, в результаті чого, міг просто ігноруватися без будь-яких наслідків. Це дало можливість додавати потрібну текстову інформацію до будь-якого MP3 файлу.

Серед існуючих насьогодні форматів метаданих аудіофайлів можна виділити: ID3tag, Lyrics3 tag, APE tag, WM metadata (частина стандарту Windows Media), Vorbis comments, MP4/iTunes metadata, ATRAC metadata.

Перелік можливих властивостей, що визначаються метаданими, досить великий [8], тому, наведемо деякі з них на прикладі фреймів стандарту ID3v2 [9]:

– назва альбому, фільму або шоу, якому належить уривок;
– головний виконавець;
– група, оркестр, супровід;
– кількість ударів в секунду;
– коментар;
– композитор;
– уточнення до назви;
– інформація про авторські права;
– ім’я людини, що закодувала файл;
– жанр музики;
– мова тексту;
– назва твору;
– номер твору в альбомі;
– текст;
– синхронізований текст;
– рік;

– обкладинка;

– точки початку/кінцівки;

– мітки синхронізації з аудіопотоком для тексту пісні.

Повний перелік фреймів стандарту ID3v2 та їх значень можна знайти на офіційному сайті [10]. Окрім фреймів, наданих в цій специфікації, користувачі можуть створювати власні фрейми з власною структурою. Інші відомі стандарти метаданих аудіо – iTunes, XSPF [11], також містять в основному інформацію про альбом, трек, авторів, тривалість запису і т. і. Це дуже корисна інформація, яка може бути використана для пошуку, але вона не розкриває семантичної сутності самого запису. Єдиний елемент, що може характеризувати зміст запису, є назва (Title).

Метадані відеоматеріалів. Як і зображення, відео також містять додаткову описову інформацію – метадані про технічні подробиці зйомки, характеристики камери, місце зйомки, назву та опис відео тощо. Теги Назва (Title) та Опис (Description) дозволяють задати семантичну інформацію, яка може бути використана для пошуку. Окрім цього, як правило, можна задати додаткові теги, або метатеги, де визначити слова або фрази, що описують відео.

Теги відрізняються від ключових слів тим, що вони не відображаються користувачам у результатах пошуку. Але, пошукові системи впливають на них при певній мірі індексації сторінки чи відео. Як правило, метадані асоціюють з текстовою інформацією на сторінці (назви, описи та ін.), але вони можуть використовуватися й для надання більш глибокої інформації. Наприклад, "прямі тимчасові метадані" або метадані на основі часу – це інформація, яка прив'язана до часової шкали в межах відео. Вдалим прикладом однієї з форм часових метаданих є закриті підписи. Приєднаний текст може також включати інформацію про те, коли та яка музика відтворюється, сцену, коли є сміх, хто говорить тощо.

Різні засоби відеозйомки можуть генерувати власний набір метаданих, залежно від його цілей, призначення тощо. Так, IBM® Intelligent Video Analytics [12] аналізує будь-який об’єкт, який рухається у відеопотоці, та створює метадані для опису ідентифікованих дій та подій. Компонент Smart Surveillance Engine (SSE) для Intelligent Video Analytics генерує метадані при обробці відеофідів. Компонент Middleware for Large Scale Surveillance витягує й зберігає дані, та керує ними.

Кожний екземпляр метаданих, що згенеровані Intelligent Video Analytics:

– описує єдиний об’єкт відеоспостережень;

– включає у себе відмітку часу, що використовується для перегляду відео;

– містить ключовий кадр, який надає зведення події або сповіщення;

– містить ідентифікатор представлення, що використовується для посилання на камеру цього відеозапису.

Метадані, які отримані у Intelligent Video Analytics, залежать від аналітичного профіля, що обраний при створенні цього каналу. Операційний аналітичний механізм може генерувати різні формати метаданих. Ці формати описані визначеннями типу контенту (CTD), що сконфігуровані в межах аналітичного профілю.

До складу кожної події або сповіщення включається декілька повідомлень метаданих, які містять різні атрибути для виявлення сегменту змінення або руху у джерелі відеосигналу. За потребою до Intelligent Video Analytics можна додати потрібні типи контенту, включивши аналітичні механізми для відправлення метаданих датчиків та подій. Можливе також виконання перехресної кореляції метаданих від усіх аналітичних механізмів, дозволяючи користувачам пошук за модальностями. Такі розширені можливості індексації забезпечують унікальний та потужний диференціатор порівняно з іншими існуючими рішеннями для спостережень.

Ступенем деталізації метаданих можна керувати. Більша рухливість об’єктів у полі зору призводить до більшої кількості метаданих, що зберігаються в БД системи для обраного каналу. Існує також кореляція між об’ємом метаданих та продуктивністю.

Враховуючи особливості складу та використання відео файлів, окрім технічних та загальних описових властивостей, що визначаються у метаданих, та спеціальних описових властивостей, притаманних відповідній предметній області, цілям бізнесу, чи застосунку, можна виділити ще дві групи елементів метаданих (які, зокрема, генеруються й Intelligent Video Analytics): метадані подій та метадані оповіщень.

Метадані подій:

– описують об'єкти або дії, що спостерігаються у відеоматеріалі;

– містять атрибути, які описують об'єкти. Атрибути змінюються залежно від аналітичного профілю;

– зазвичай використовуються для криміналістичного пошуку або статистичного аналізу.

Метадані оповіщень:

– ініціюються, коли для об'єкта на відео виконані попередньо визначені умови;

– базуються на поведінці та атрибутах;

– використовуються для моніторингу в реальному часі та можуть бути використані також для статистичного аналізу.

– За замовченням метадані оповіщень включають наступні атрибути:

– час, коли воно було ініційоване;

– канал, по якому це відбулося;

– передвизначені ім'я та тип оповіщення;

– пріоритет оповіщення;

– затримка перед відтворенням (час в секундах перед ініціацією оповіщення).

Технічні метадані можуть визначати інформацію про файл (назва та розмір, формат, тривалість запису, формат аудіо даних, роздільна здатність тощо) у CVS-документі чи таблиці [13]. Щодо стандартів метаданих, кожний формат відеофайлу має власний метод зберігання метаданих, але метадані можна також зберігати у зовнішньому файлі (чи БД) або використовувати XML/ XSD комбінацію.

MPEG7 [14] – стандарт опису мультимедійного вмісту, який набув статусу стандарту в ISO/IEC 15938 («*Інтерфейс опису мультимедійного вмісту*»). Для формалізації/зберігання метаданих він використовує XML та може бути приєднаний до тимчасового коду, наприклад, щоб синхронізувати текст з піснею. Але стандарт не набув широкого розповсюдження.

Досить вдала реалізація намірів щодо узагальнення структурованої інформації для опису великих даних була зроблена спільними зусиллями компаній Google, Microsoft, Yahoo та Yandex, які створили словниковий ресурс схем для структурованих даних Schema.org [15], який детально, розглядатимемо далі. Разом з іншими, даний словник визначає перелік характеристик для опису відеооб'єктів.

Окрім цього, набори метаданих для декількох передвизначених типів контенту, таких як: відеокліпи, аудіофайли, вебсторінки і т. і. були запропоновані розробниками Open Graph, що також, розглядатимемо далі.

Словники Schema.org – спільна робота спільноти з місією створення, підтримки та просування схем для структурованих даних в Інтернеті, на веб-сторінках, у повідомленнях електронної пошти та за її межами. Словник Schema.org можна використовувати з багатьма різними кодуваннями, включаючи RDFs, Microdata та JSON-LD. Ці словники охоплюють сутності, відносини між сутностями та діями і можуть бути легко розширені за допомогою добре документованої моделі розширення. Понад 10 мільйонів сайтів використовують Schema.org для розмітки своїх веб-сторінок і повідомлень електронної пошти. Багато додатків від Google, Microsoft, Pinterest, Yandex та інших вже використовують ці словники для енергійного, розширюваного досвіду.

Заснована компаніями Google, Microsoft, Yahoo і Yandex, словники Schema.org розробляються в рамках спільноти, використовуючи список розсилки public-schemaorg@w3.org і через GitHub. Спільний словник (колекція словників) спрощує веб-майстрам та розробникам вирішення

питань щодо схеми метаданих та максимізує ефективність їх роботи.

Так, для відеоданих у словнику пропонується наступна схема метаданих, яка приведена у таблиці.

Таблиця

Елементи *VideoObject*		
actor	Person	Актор, наприклад, у TV, радіо, фільмі, відео грі тощо, або в події. Актори пов'язані з індивідуальними елементами або серією, епізодом, кліпом
caption	MediaObjectабо Text	Підпис об'єкта (наприклад, субтитри). Вказує формат кодування
director	Person	Директор, наприклад, у TV, радіо, фільмі, відео грі або події. Директори пов'язані з індивідуальними елементами або серією, епізодом, кліпом
musicBy	MusicGroup or Person	Композитор звукового треку
thumbnail	ImageObject	Мініатюра для зображення або відео
transcript	Text	Чи цей MediaObject є AudioObject або VideoObject, транскрипція цього об'єкту
videoFrameSize	Text	Розмір фрейму
videoQuality	Text	Якість відео

Метадані документів та веб-контенту. Поняття метаданих для документів визначене у стандарті ISO 15489-2001 «*Інформація та документація. Управління документами*» [16]. Відповідно до цього стандарту документ окрім свого інформаційного змісту повинен містити метадані або бути постійно зв'язаним з метаданими, необхідними для виконання дій з документами. Це потрібно для забезпечення незмінності структури документа (його формату та взаємозв'язків між складовими елементами) при виконанні різних операцій з електронним документом. Наявність метаданих робить зрозумілими обставини, за яких документ був створений, отриманий та використаний, – діловий контекст (включаючи відомості про те, частиною якого ділового процесу є виконана операція, про дату, час та учасників ділової операції).

Міжнародний стандарт ISO 15836-2003 (*«Information and documentation – The Dublin Core metadata element set»*) визначає універсальний набір метаданих для будь-яких інформаційних ресурсів. Стандарт покладений в основу практично всіх наборів метаданих, що описують текстові контенти.

Будь-який документ характеризується змістом (контентом) та структурою.

Контент – інформація документа, що фіксує управлінську діяльність.

Структура – зовнішній вигляд та розміщення частин контенту (наприклад, носій, формат, організація даних, розміщення реквізитів, шрифти, примітки і т. і.), та наявність у документі зв'язків з іншими документами (гіперпосилань).

На базі Дублінського ядра було розроблено ГОСТ 7.70-2003 *«Опис баз даних та машиночитаємих інформаційних масивів. Склад та позначення характеристик»,* який вводить нескладну систему опису інформаційних, зокрема, мережевих, ресурсів. Для опису інформаційного ресурсу стандарт пропонує використовувати 29 характеристик, обов'язковими з яких є для всіх типів ресурсів:

1) ідентифікатор інформаційного ресурсу (посилання на ресурс);
2) назва ресурсу;
3) власник;
4) опис (зміст ресурсу, влючаючи анотацію або реферат, або опис контенту візуальних, аудіо- або мультимедійних ресурсів);
5) коди рубрикатора – тематика ресурсу, що виражена кодами стандартного переліку тематичних рубрик;
6) ключові слова;
7) мова (для текстового ресурсу);
8) період оновлення;
9) фінансування – форма фінансування при створенні та введенні ресурсу;
10) дата останнього оновлення вмісту або дата створення.

Умовно обов'язкові:

1) дата останнього оновлення метаданих інформаційного ресурсу;

2) адреса в мережі;

3) консультант – особа, до якої треба звертатися за додатковою інформацією;

4) дата реєстрації інформаційного ресурсу;

5) служба, що реєструє інформаційний ресурс.

В стандарті визначені також факультативні реквізити, які можуть надати додаткову інформацію про інформаційний ресурс, зробити його зручнішим: «Творець», «Учасник», «Права», «Ресурс-джерело», «Обмеження доступу», «Об'єм» тощо.

Більшість комп'ютерних програм автоматично створюють деякі метадані та асоціюють їх з файлами. Наприклад, кожний документ MS Word має певний набір властивостей файлу (назва, автор, розмір файлу тощо), який визначається автоматично або вручну. Автоматично формується особливий тип метаданих, так звані системні аудиторські перевірки, які фіксують дії, які виконуються з окремим документом (наприклад, дату дії, вид дії, особа, що її ініціювала).

Великий відсоток веб-контента генерується спеціальними ПЗ, наприклад, генераторами сайтів. Такі програмні засоби дозволяють генерувати пости, тобто вебконтент, фактично, дані, які супроводжуються метаданими. Кожний такий «генератор» може мати власні вимоги до метаданих та використовувати власні механізми створення метаданих та протоколи їх зберігання. Так, у Wintersmith [17] обов'язковими вважають три види метаданих: *template* визначає шаблон, що використовується для рендерінга, title – назву поста, date – його дату. Але існує можливість додавання будь-яких метаданих.

Метадані, як правило, мають власний формат представлення (окремий від формату даних). Одним з найбільш розповсюджених є YAML формат [18, 19]. Широке використання YAML отримав завдяки тому, що він орієнтований на зручність введення/виведення типових структур даних багатьох мов програмування, тобто даних багатьох форматів, включаючи Json об'єкти. Він є дружнім до людини форматом серіалізації даних, концептуально близький до мов розмітки. Дозволяє визначити будь-які метадані відповідно до вимог бізнес – застосунку. Json об'єкти, в свою чергу, довели свою зручність та доцільність з розвитком великих даних та мережевих технологій, як стандартний текстовий формат для представлення структурованих даних на основі синтаксису об'єкта JavaScript, що, зазвичай, використовується для передачі даних до вебзастосунків.

Метадані у соціальних мережах

Найрозповсюдженішими схемами метаданих, що використовуються у соціальних мережах, є Twitter Card, розроблений Twitter та Open Graph (OG), запропонований Facabook. Ці протоколи метаданих виконують одну й ту саму функцію – забезпечення найкращого досвіду взаємодії з користувачем при розповсюдженні інформації через соціальні платформи.

Twitter Cards та OG [20] це два окремі набори метаданих. Багато соціальних платформ зчитують їх частину та відображають у відповідному вигляді. У Twitter можна побачити результат у вигляді анотацій до новин чи зображень у стрічці. Результат використання OG можна побачити, наприклад, у Facebook у вигляді блока-анотації з заголовком, описом та зображенням, що характеризують інформаційний блок.

За основу протоколу OG, при його створенні, було покладено стандарт Dublin Core Metadata Element Set. Спочатку Facebook створив цей стандарт для власного використання. Тому, він є досить складним та містить функції, які потрібні лише Facebook. Але, інші платформи, такі як LinkedIn та Google+, також успішно використовують метадані OG.

Twitter Cards служить лише одній меті: наповненню інформаційних блоків, які відображаються у Twitter та застосунках, що працюють з цим сервісом. Невеличкі відмінності між Twitter Cards та OG обумовлені лише специфічними особливо-

стями Twitter. Унікальність Twitter як соціальної платформи полягає у розповсюдженні інформації через ретвіти, що створює певні проблеми. Обмеження твіту у 140 символів не завжди дозволяє додати коментар або інформацію про автора чи власника контенту. Twitter Cards дозволяє офіційно визначити у метаданих інформацію про автора та правовласника. Twitter досить вміло сформували власну схему метаданих, щоб не відставати від інших соціальних платформ. Автори та власники сайтів, які реалізували підтримку схеми Twitter Cards, тепер можуть надавати більше інформації про контент, що розповсюджується через Twitter.

Будь-які застосунки, що забезпечують функції соціальної мережі, підтримують відправлення повідомлень наступних типів [21]: текстові повідомлення; текстові повідомлення з приєднаним зображенням, аудіо, відео чи просто будь-яким файлом; повідомлення з приєднаним шаблоном; повідомлення з прикріпленим відступом. Відповідно, для кожного повідомлення зберігаються загальні властивості (відправник, отримувач, властивості, що ідентифікують повідомлення) та властивості, специфічні для обраного типу повідомлення (текст повідомлення, тип приєднаного файлу, посилання на нього тощо).

Обидві схеми також спрямовані на надання інформації про контенти найпопулярніших типів:

– для статей формується текстова анотація і зображення для попереднього перегляду;
– для аудіо-записів додається аудіоплеєр;
– для відеокліпів – відеоплеєр;
– для зображення – можливість їх попереднього перегляду.

Протоколи Twitter Cards та OG є досить схожими та мають специфічні властивості, що визначаються для передвизначених типів контенту. Розробники визначили декілька типів контенту, таких як відеокліпи, аудіофайли, веб-сторінки та інші, та запропонували для кожного з них власний набір мета-властивостей. Коли користувач ділиться аудіофайлом або відеокліпом, для їх відтворення на основі метаданих обирається відповідний плеєр.

Далі наводиться перелік мета-тегів, що надають детальну інформацію про статтю в OG та Twitter Cards.

OG:

```
        <meta property="og:type"
content="article">
        <meta property="og:url"
content="URL об’єкта">
        <meta property="og:site_name"
content="Назва ресурсу, де розміщена
стаття">
        <meta property="og:image"
content="URL зображення для статті">
        <meta property="og:title"
content="Заголовок статті">
        <meta
property="og:description"
content="Опис статті">
        <meta
property="article:author"
content="URL сторінки автора статті">
        <meta
property="article:section"
content="Розділ, до якого відноситься
стаття">
        <meta property="article:tag"
content="Ключові слова">
```

Twitter Cards:

```
<meta name="twitter:card"
content="summary">
        <meta name="twitter:url"
content="URL статті">
        <meta name="twitter:title"
content="Заголовок статті">
        <meta
name="twitter:description"
content="Опис статті">
        <meta name="twitter:image"
content="URL зображення для статті">
```

Наступні опціональні елементи Twitter Card дозволяють визначити ідентифікатор автора або організації автора контента у Twitter:

```
<meta name="twitter:site"
content="@username">
        <meta name="twitter:site:id"
content="Twitter ID">
        <meta name="twitter:creator"
content="@username">
        <meta
name="twitter:creator:id"
content="Twitter ID">
```

Протокол Open Graph [22]. Визначає метадані для опису веб-сторінок, та перетворює їх у вузли соціальної мережі. Хоча існує багато різних технологій і схем, які можна було б об'єднати разом, не існує єдиної технології, яка надає достатньо інформації для повномасштабного представлення будь-якої веб-сторінки у соціальній мережі. Протокол OG побудований на цих існуючих технологіях. Його головною метою є простота для розробників. Протокол містить базові елементи метаданих, опціональні елементи метаданих, структуровані властивості для аудіо, відео та зображень, спеціальні властивості для музичних записів та визначення типів об'єктів.

До базових елементів метаданих відносяться:

– назва об'єкта;
– тип об'єкта;
– URL на зображення, що представляє об'єкт у графі соціальної мережі;
– URL на об'єкт, що використовується як ідентифікатор;
– опціональні та загально рекомендовані елементи метаданих:
– URL на аудіофайл, що супроводжує об'єкт;
– стислий опис об'єкта;
– слово, що виводиться перед назвою об'єкта (за замовченням пусте);
– місцева мова;
– масив інших доступних на сторінці мов;
– назва сайту, якому належить об'єкт;
– URL на відеофайл, що доповнює даний об'єкт.

Деякі властивості також можуть мати додаткові метадані [23].

Протокол Twitter Card (TC). Зміст протоколу визначається, перш за все, потребами користувачів Twitter. Користувачі Twitter організовують людей, яким вони слідують, у списки, постять текст та відео, використовують хеш-теги для визначення коментарів у твіті та пов'язують їх один з одним, ретвітять чужий контент з/без коментарів, виділяють "улюблені" твіти, керують властивостями, такими як «список популярних тем». Глибина даних про суспільство, що представлена контентом у Twitter, та його метадані призвели у 2010 році до заключення угоди про те, що Бібліотека Конгресу США буде архівувати цей цінний матеріал для досліджень [24].

Перелік елементів метаданих ТС за складом дуже близький до OG, що обумовлюється, як було вже зазначено вище, єдиним набором типів контентів, які вони намагаються структурувати та описати. Існує 5 різних типів карт, відповідно до різних об'єктів, які вони описують [25]:

– зведена карта;
– зведена карта з великим зображенням;
– карта програвача;
– карта застосунку;
– карта Lead Generation.

Зведена карта може використовуватися постами блогів, новинами, сторінками продуктів та іншими бізнес-новинами. Їх мета полягає у наданні такої інформації, як назва, опис та зображення, що супроводжує даний пост.

Зведена карта з великим зображенням відрізняється лише допустимими розмірами зображення.

Карта застосунку може використовуватися мобільними застосунками. Вона надає інформацію про назву, опис, іконку, рейтинг та ціну. Щоб витягнути цю інформацію Twitter може використовувати ідентифікатори застосунків.

Карта програвача була створена, щоб рекламувати мультимедійні потокові мультимедіа, такі як аудіо- чи відео- програвач у Twitter. Містить такі дані як: опис, зображення, програвач.

Карта *Lead Generation* підходить для спілкування з потенційними клієнтами. Вона дозволяє зібрати інформацію про перспективи. Дані потенційних клієнтів (ім'я, контакти), зазвичай, вже введені та їх не треба заповнювати ще раз. Карта містить таку інформацію як: назва, повідомлення посту, зображення, заклик до дії.

Щоб глибше усвідомити значимість та обсяги метаданих у соціальних мережах, розглянемо невеличкий приклад [26] метаданих Twitter (рис. 3), що пов'язані з твітом лише у 140 символів.

Останні не представляють великих обсягів даних, однак об'єми їх вибухають, якщо зв'язати твіт з усіма метаданими, що необхідні для розуміння цих 140 символів у контексті розмови.

Наведений далі приклад (рис. 3) демонструє підмножину повного переліку елементів метаданих протоколу TC, а саме містить наступні елементи:

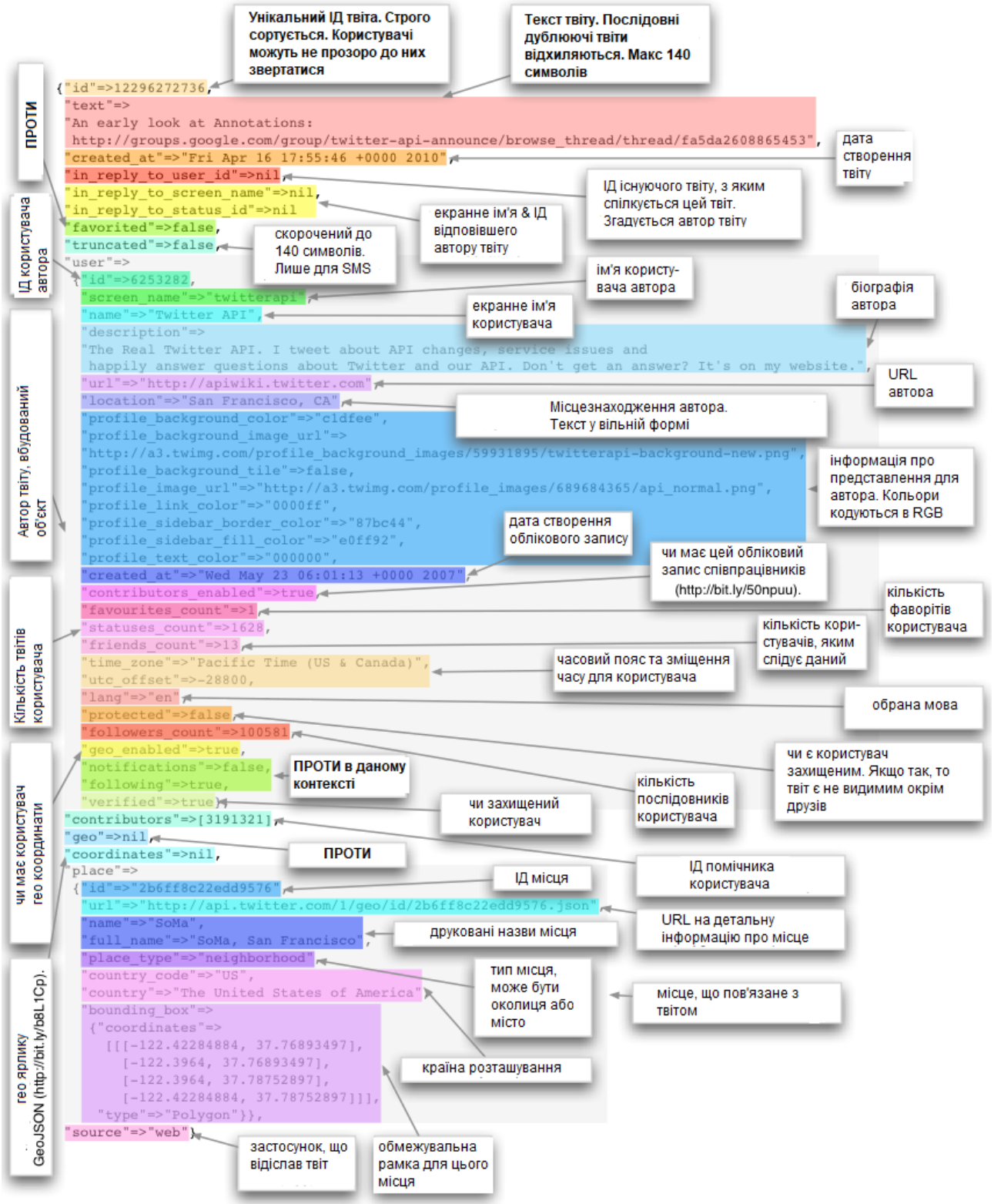

Рис. 3. Метадані, пов'язані з твітом

– ім'я та ідентифікатор користувача, що відповів автору твіта;
– дата та час створення твіту;
– ім'я автора;
– ім'я користувача;
– біографія автора;
– URL автора;
– місцезнаходження автора;
– надання інформації для автора;
– дата створення облікового запису;
– кількість обраних, що має користувач;
– кількість користувачів, на яких підписаний даний користувач;
– часовий пояс та зміщення часу для даного користувача;
– мова, обрана користувачем;
– чи є користувач захищеним;
– кількість підписників користувача;
– ідентифікатор місця;
– друкована назва цього місця;
– тип місця;
– країна;
– застосунок, який відіслав твіт.

Особливістю даного прикладу є те, що всі метадані у деякій мірі структуровані. Вони збираються таким чином, щоб виконувати корисну задачу, та сортуються за відомими категоріями. Саме це поняття структури перетворює необроблену інформацію у реальні метадані.

Метадані сховищ даних

Склад та зміст цього класу метаданих визначається цілями їх використання у сховищах даних. А саме, метадані, у даному випадку, повинні сприяти мінімізації робіт з розробки та адміністрування сховища даних та забезпечувати більш ефективне витягнення інформації із сховища. Метадані містять всю інформацію, що необхідна для витягнення, перетворення та завантаження даних з вихідних систем, а також для наступного використання та інтерпретації вмісту сховища.

Ральф Кимболл виділив наступні типи метаданих:

– метадані вихідної системи:
  o специфікації джерел даних, таких як репозиторії;
  o описова інформація (частота оновлення, юридичні обмеження, методи доступу);
  o інформація про процеси, наприклад, графік завдань та коди витягнення.
– метадані перетворення даних:
  o інформація про отримання даних (планування передачі даних, відомості про використання файлів);
  o керування таблицями вимірювань, (визначення вимірювань та присвоєння сурогатних ключів);
  o перетворення та агрегування, (розширення та відображення даних, програми завантаження СКБД, визначення агрегатів даних);
  o документування перевірок, робіт та журналів (журналів перетворення даних й записів стеження за походженням даних.
– метадані СКБД:
  o зміст системних таблиць СКБД;
  o рекомендації з обробки.

Ці метадані можуть використовуватися трьома способами:

– *пасивно*, забезпечуючи чітку документацію щодо структури, процесу розробки та використання системи Сховища даних. Документація необхідна як кінцевим користувачам, так і системним адміністраторам та розробникам застосунків;
– *активно*, шляхом збереження конкретних аспектів (наприклад, правил перетворення) у вигляді метаданих, які можна інтерпретувати та використовувати під час виконання. У даному випадку, метадані керують процесами сховища даних. Як наслідок, код (активні метадані) та додаткова документація погоджено та уніфіковано керуються в одному репозиторії, при цьому зростає актуальність документації;
– *напівактивно*, за рахунок зберігання статичної інформації (наприклад, визначень структур, специфікацій конфігурацій), яку буде зчитувати інший програмний компонент під час виконання.

Наприклад, обробникам запитів необхідні метадані для перевірки існування атрибутів. На відміну від активного використання, тут метадані лише читаються, але не виконуються.

Метадані сховищ дозволяють вирішувати наступні задачі відповідно до вище визначених цілей [27]:

– *підтримка інтеграції систем*. Схеми та інтеграція даних залежать від метаданих, що описують структуру та сенс окремих джерел даних та цільових систем. Правила перетворення можна застосувати до вихідних даних та зберігати як метадані. Більше того, інтеграція різних інструментів можлива лише, якщо вони розділяють «дані», які є метаданими системи сховища;

– *підтримка аналізу та проектування нових застосунків*. Метадані підвищують контрольованість та надійність процесу розробки застосунків, забезпечуючи інформацію про зміст даних, їх структуру та джерела. Більше того, метадані, що стосуються рішень з проектування застосунків, можна використовувати повторно;

– *підвищення гнучкості системи та можливості повторного використання існуючих програмних модулів*. Це можливо лише для активного та напівактивного використання даних. Семантичні аспекти, які швидко змінюються, зберігаються явним чином у вигляді метаданих прикладних програм. Тому, підтримка є суттєво простішою. Систему можна розширити та адаптувати без будь-яких труднощів. Даний підхід дає можливість повторного використання «фрагментів коду»;

– *автоматизація адміністративних процесів*. Метадані керують запуском різних процесів Сховища даних (наприклад, завантаження та оновлення). Інформація про їх виконання (журнали доступу, кількість записів та ін.) також міститься у репозиторії, доступному адміністратору;

– *підсилення механізмів безпеки*. Метадані повинні забезпечити правила доступу та права користувачів для всієї системи сховища. Керування доступом у сховищі іноді вимагає застосування складних методів. Наприклад, оперативне джерело може містити нешкідливу інформацію про окремі показники роботи компанії, але сумарні значення у сховищі іноді виявляються важливим секретом. З іншого боку, персональні доходи кожного співробітника є тайною, але при цьому підсумок грошової одиниці у сховищі може взагалі не бути критичною інформацією;

– *підвищення якості даних*. Якість даних визначається наступними характеристиками:

- погодженість (однорідність представлення, відсутність дублікатів, даних з визначеннями, що конфліктують);
- повнота (всі дані присутні)
- точність (співпадання збережених та фактичних значень);
- своєчасність (актуальність значення, що зберігається);
- правила перевірки якості даних необхідно визначати, зберігати у вигляді метаданих та перевіряти при кожному оновленні сховища. Окрім цього, висока якість вимагає підтримки контроля даних. Метадані забезпечують інформацію про час створення, про автора даних, про джерело, значення даних в момент отримання (про спадковість даних), й про подальший шлях від джерела до поточного місцезнаходження (походження даних). Таким чином, користувачі можуть відновити ланцюг проходження даних за час перетворення, та перевірити точність інформації, що повертається;

– *покращення взаємодії в системі сховища*. Взаємодія відбувається як через виконання простих запитів та звітних застосунків, так й з використанням складних аналітичних інструментів. Метадані забезпечують відомості про значення даних, термінологію та бізнес-концепції підприємства, а також їх зв’язок з даними. Тому, метадані підвищують якість запитів, що виконуються, за рахунок більш точного та строгого формулювання, а також скорочують витрати на користувачів, яким потрібний доступ, оцінку та застосування відповідної інформації;

– *покращення аналізу даних*. Методи аналізу даних представлені широко – починаючи від простих застосунків звітності та OLAP, та, закінчуючи склад-

ними застосунками data mining. В цьому напрямку метадані необхідні для розуміння предметної області та її представлення у сховищі, для адекватного застосування та інтерпретації результатів;

– *застосуванню загальної термінології та мови взаємодії у корпорації*. Доступність метаданих як унікального джерела документації для користувачів має й інші переваги. Вона гарантує погоджені дії та інтерпретації інформації із сховища, а також запобігає двозначності та забезпечує погодженість відомостей у компанії, дозволяє розділити знання та досвід.

Метадані системи сховища містяться в репозиторії – структурованій системі зберігання та витягнення, що реалізована на основі СКБД. Для інтерпретації метаданих необхідно зберігати структуру репозиторія (тобто схему метаданих) та їх семантику.

Існують різні способи визначення та зберігання метаданих у сховищі, один з яких є використання технології XML, що завдяки своїм перевагам, таким як зрозумілість, відкритість технології, гнучкість, є дуже зручним засобом опису. Він дозволяє публікувати метадані, що використовуються будь-якою програмою або БД, та забезпечує зв'язок між структурованою базою та неструктурованим контентом. Якщо є ПЗ, яке здатне прочитати та розшифрувати XML-файли, то метадані у будь-якому сховищі можна представити у вигляді звичайного XML-файлу, що створений на основі загального DTD (опис типу документу), а переваги XML представлення є очевидними. Але якщо буде забагато XML файлів, написаних відповідно до різних стандартів, то це зробить їх обробку досить складною. Тому, ефективна обробка метаданих, представлених таким чином, вимагає вирішення наступних проблем XML-середовища:

– погоджений стиль тегів;

– несуперечливі погодження про іменування;

– сумісне визначення тегів;

– керування XML об'єктами для їх наступного повторного використання;

– інструменти для динамічної перевірки DTD та схем;

– документовані набори кодів;

– чітко визначені простори імен бізнес-моделі;

Метадані в екосистемі Hadoop

Hadoop використовує метадані для ефективного керування даними. Метадані вбудовуються у самі дані при проходженні ними різних систем підприємства (робиться це за допомогою визначення спеціальних тегів). Більше того, метадані розширюються та включають додаткову інформацію окрім звичайних атрибутів, таких як: розмір файлу, роздільна здатність, дати модифікації тощо. Наприклад, до метаданих можуть бути включені відомості про бізнес, що може допомогти визначити корисність даних у конкретній моделі. Нарешті, на відміну від самих корпоративних даних метадані можуть бути централізовані на єдиній платформі.

HDFS здатна присвоювати розширені атрибути, що також дозволяє збагатити метадані. Але це не завжди підходить для великих даних. Тому, виникають альтернативні підходи, як, зокрема, визначення тегів (Apache Atlas) та створення централізованого сховища метаданих, а користувачі дружніх до Hadoop систем витягування даних (наприклад, Hive та Spark SQL) можуть визначати теги самостійно.

Задачі роботи з метаданими і сервіси для аналізу та обробки метаданих

Досить швидкий розвиток мережевих технологій та зростання обсягів даних різних типів обумовлюють нагальні вимоги як до самих метаданих, так й до інструментів, що дозволяють якось ними користуватися. Метадані не мають сенсу, якщо вони не є «корисними», тобто немає можливості їх ефективного використання для вирішення задач користувачів. Тобто, необхідно мати розвинені методи та засоби, що дозволять:

– створювати метадані;

– читати метадані;

– редагувати метадані;

– організовувати ефективне зберігання метаданих;
– керувати метаданими;
– аналізувати метадані;
– оптимізувати метадані;
– використовувати метадані при пошуку інформації.

Насьогодні розроблено велику кількість доданків чи окремих застосунків, що намагаються вирішити перелічені задачі. Деякі з них просто допомагають переглянути чи вивести інформацію метаданих, інші додають інструменти для створення та редагування, а деякі є потужними інструментами аналізу метаданих, що дозволяють робити висновки про достовірність контенту тим чи іншим чином та навіть будувати статистичні звіти та прогнози, на основі аналізу метаданих.

Керування метаданими

Найважливішою особливістю метаданих є впорядкована структура. Інформація категоризована та має конкретну форму/формат. Завдяки структурованому вигляду, метадані є доступними для читання не лише людиною, але й комп'ютерами. Таким чином, метадані можуть бути оброблені автоматизовано та використані для різних цілей: індексація, пошук, об'єднання або автоматична обробка. Зокрема, одним з прикладів використання метаданих є пошук зображення на основі Exif значення поля. Це, наприклад, можна зробити в програмі Google Picasa, де передбачені спеціальні команди, які вводять значення в поле пошуку у застосунку. В Інтернеті, у галереях зображень стандартом є визначення метаданих поряд з фотографією, що переглядається. Сайти активно використовують дані з EXIF, наприклад, для «прив'язки» фотографії до географічної мапи. Так, у галереї Google, Google+ або Picasa Web Albums поряд з фото з'явиться мапа з попереднім переглядом.

Створення метаданих, перегляд, читання та редагування

Базовий набір метаданих визначається автоматично застосунком при створенні файлу даних конкретного типу. Так, для файлів фотографій за це відповідає ПЗ, яке встановлене на цифровій камері.

Щоб переглянути метадані, перш за все, необхідно витягнути їх з файлу. Отримати доступ до метаданих веб-об'єктів можна за допомогою доданків до веб-браузера або спеціальних сервісів. Деякі з інструментів, що беруть на себе такі функції, підтримують лише один формат файлу (наприклад, JPEG), інші – багато форматів. Окрім цього, різні програми можуть підтримувати різні типи метаданих.

Серед прикладів інструментів витягування метаданих можна перелічити [28].

1. *Exiv2* [29] – ПЗ з відкритим кодом, яке декодує EXIF, IPTC та XMP метадані.

2. *ExifTool* [30] – один з найбільш потужних засобів витягування метаданих (через командну строку). Підтримує сотні різних форматів файлів та метаданих (PDF, Djvu, JPEG, AVI, MOV, MP3 тощо), включаючи дуже специфічні. Написаний на Perl, добре документований, добре працює під Linux та Mac, але може бути проблемним для користувачів Windows, у яких не встановлений Perl.

3. *Adobe Photoshop* – комерційний застосунок, що включає XMP переглядач. Хоча він не такий потужний або повний як Exiv2 або ExifTool, забезпечує можливість декодування XMP, IPTC, Exif та інші типи метаданих у графічному інтерфейсі.

4. *PreviewInspector*. За замовчуванням Apple Mac OS X для перегляду зображень використовує ПЗ Preview, яке містить спеціальний 'Inspector' для перегляду метаданих. Інструмент відображає невеличку частину доступних метаданих та може забезпечувати не достовірні результати аналізу, тому не рекомендований для використання в офіційних розробках.

Зручним інструментом для читання метаданих графічних файлів також є браузер графічних файлів *IrfanView*, за умови, що встановлено плагін, який включає бібліотеку для декодування Exif. Хоча там не вистачає можливості редагування Exif, IrfanView дозволяє створювати опис у форматі IPTC. З метаданими також чудово ладнають всі ПЗ для обробки цифрових фотографій. Для не професійного викорис-

тання, можна рекомендувати програму *Google Picasa*, яка має панель, що дозволяє перевірити всі дані фотографії. Нажаль, можливості редагування метаданих обмежена. Тому, у випадку більших вимог, слід звернути увагу на *Adobe Lightroom* [31], що містить дуже складні інструменти для перегляду та редагування метаданих зображень. Його використовують професійні фотографи. Програма містить також спеціальні інструменти, що використовують метадані для автоматичної корекції кольорів зображень. Lightroom дозволяє визначати тип датчика, тип об'єктива, фокусну відстань та на цій основі покращити геометрію зображення.

Розвинені можливості щодо створення та редагування метаданих надає відома платформа для побудови клієнтських застосунків WPF (Windows Presentation Foundation). Вона дозволяє обробляти XMP, IPTC та EXIF метадані та використовує для цього спеціальні класи. При цьому, відповідні поля у метаданих різних стандартів знаходяться у наступному порядку: XMP, IPTC та EXIF. Запис тегів метаданих виконується у XMP форматі. Окрім цього, для читання та запису метаданих можуть використовуватися функції GetQuery/SetQuery, які працюють з ієрархічними іменами тегів метаданих. Існують спеціальні класи, що дозволяють працювати з конкретними форматами зображень. А клас InPlaceMetadataWriter дозволяє змінювати метадані на місці, без перекодування файла.

Досить потужним інструментом для редагування тегів аудіофайлів є *TagScanner*. Він вміє редагувати у пакетному режимі теги більшості сучасних аудіоформатів. Підтримуються теги ID3v1 та ID3v2, Vorbis Comments, APEv2, WMA та MP4 (iTunes). Застосунок дозволяє змінювати назву файла за інформацією з тегів, генерувати тег за назвою файла/директорії або виконувати будь-які перетворення та змінення тексту в тегах та іменах файлів. Програма має розвинені можливості для отримання інформації та роботи з музичним альбомом чи архівом.

Розвинені можливості керування метаданими в XMP форматі надає застосунок *Premiere Pro* [32]. Він відкриває загальний доступ до метаданих, а також дозволяє їх переглядати, створювати, видаляти, редагувати та шукати.

Аналіз метаданих

Для перевірки наявності метаданих та їх аналізу у фотографіях (наприклад. визначення місця та часу, коли воно було зроблене, чи було воно відредаговане) можна використовувати спеціальні он-лайн ресурси [33]. Одним з таких ресурсів є *Jeffray's Exif Viewer*, розроблений та викладений у відкритий доступ американським програмистом. Він відображає всю доступну інформацію з метаданих. Аналогічно працює інший схожий ресурс для перевірки метаданих *FindEXIF.com*. Але в ньому відсутня можливість завантаження фотографії. Сервіс працює лише з посиланнями. Фотографії з конкретних географічних місць можна також шукати за допомогою *Panoramio*. Цей сервіс використовує EXIF-дані для публікації фотографій на мапі. Дозволяє при розміщенні фотографії визначити її координати. Розібратися в локаціях допомагають також сервіси Google Maps та Wikimapia.

Сервіс *FotoForensics* дозволяє визначити, чи було фото відредаговане. Сервіс працює як з завантаженими фотографіями, так й з посиланнями на них. Окрім того, що сервіс виводить доступні метадані (дату створення, дату редагування тощо), він пропонує ELA (Error Level Alysis) рівень стискання файлів. Це свого роду сканер, який показує маніпуляції з зображенням, навіть, якщо їх не видно на перший погляд. Знаючи специфіку цих даних, можна ефективно визначати масштаби та тип редагування знімка, наприклад, чи був використаний фотомонтаж при редагуванні зображення.

Інструменти *Google Search by Image* та *TinEye* забезпечують можливості зворотнього пошуку зображень, тобто користувач може завантажити до сервісу фото та знайти його оригінальне джерело й подивитись, де воно ще публікувалося.

Програмний застосунок *JPEGSnoop* дозволяє дивитися метадані не лише зображень, а й форматів AVI, DNG, PDF, THM,

але працює тільки для Windows. Дозволяє перевірити, чи було зображення редаговане, виявити помилки у пошкодженому файлі тощо.

Сервіс *Pipl.com* призначений для пошуку «Інтернет-сліду» користувача, допомагає його ідентифікувати та знайти його контент (фотографії, файли тощо). Програма проводить пошук у всіх соціальних американських мережах (Facebook, LinkedIn, MySpace) – для цього потрібно ввести ім'я та прізвище латиницею. Особливість програми в тому, що вона веде пошук по «глибокому Інтернету», який ігнорується звичайними пошуковими системами та недоступний для користувачів.

Ресурс *WebMii* шукає посилання з визначеним ім'ям людини, дає рейтинг «веб-видимості», за допомогою якого можна встановити фейкові акаунти. Завдяки інструменту кожен може знайти згадування власного імені на іноземних ресурсах.

Застосунок *Geofeedia* є «куратором соціальних мереж», який агрегує результати не за ключовими словами чи хеш-тегами, а за заданим місцем розташування. Сервіс обробляє повідомлення з Twitter, Flickr, Youtube, Instagram та Picasa, надіслані з використанням GPS, і потім представляє їх у вигляді колажу. І хоча значну кількість повідомлень він не охоплює, надає загальну картину. Сервіс платний, безкоштовною надається лише демоверсія.

Окремої уваги заслуговує сервіс *Wolfram Alpha* [34]. Це навіть не пошукова система, а база знань з науковим ухилом, інтелектуальний робот, який може відповідати на будь-які питання. Але він орієнтується лише в темах, які стосуються точної, більш енциклопедичної інформації, а не поточних подій. Він не надає посилань на інші сайти, а видає вже готовий варіант відповіді. Спочатку він був рекомендований для перевірки погоди, в тому числі використовувався для перевірки достовірності зображень через перевірку погоди за датою та місцем зображення. Але, насправді ця система може порівнювати відомі світові компанії за безліччю показників, а також міста, країни, відомих осіб, будівлі. Програма також містить багато економічної інформації, у тому числі може вираховувати прогнози, наприклад, ціна на газ та нафту в довгостроковій перспективі; може вирішувати складні математичні приклади, давати детальну інформацію про поживну цінність різних продуктів, показувати карту зоряного неба для різних точок земної кулі. Окрім того, вираховує індекс маси тіла (потрібно ввести свої дані) та ризики захворюваності; визначає час, потрібний для читання чи написання певної кількості слів та багато іншого. Нарешті, одна з найцікавіших можливостей інструменту – це аналіз статистики користувача у Facebook. Програма формує повний звіт із інфографікою, який показує активність користувача від моменту реєстрації, кількість завантажених лінків та фото, також показує активність упродовж доби й аналізує середню довжину постів. Крім того, можна подивитися статистику по друзях за різними показниками – вік, стать, сімейний стан, місце проживання на карті та мережеві зв'язки друзів між собою.

Питання оптимізації метаданих

Як було продемонстровано прикладом вище, обсяги метаданих часто можуть перевищувати обсяги самих (великих) даних. Тому, якщо документ містить велику кількість метаданих, доцільно зберігати метадані в різних файлах та використовувати дві URL-адреси – окремо для метаданих та для самих даних. Ці дві сторінки можна зв'язати за допомогою вказівників. Це дозволить суттєво підвищити продуктивність роботи з даними. Цей варіант ідеально підійде для адаптивних сайтів: матеріали однієї й тієї самої сторінки можна показувати як у браузері комп'ютера, так і на мобільному пристрої.

Facebook пропонує [35] також для застосунків, що використовують піддомени з метою розміщення додаткових версій, які оптимізовані для мобільних пристроїв, запобігати додавання додаткових даних до представлення сторінок для мобільних пристроїв. Для цього пропонується використовувати канонічні URL, що вказують на представлення цих сторінок для перегляду на комп'ютері.

Висновки

Проведені дослідження продемонстрували нагальність та масштаби проблем у галузі обробки різнорідних не структурованих даних великих обсягів – величезні, безперервно зростаючі обсяги різноманітної, слабкоструктурованої (чи взагалі не структурованої) інформації, велика кількість розрізнених сервісів для обробки їх метаданих, що, як правило, описують лише технічні характеристики даних, відсутність семантичних описів даних та єдиної ефективної системи керування метаданими.

Головною перевагою метаданих є їх структурованість, що забезпечує можливість їх автоматизованої обробки та використання для роботи з великими даними, а саме: для індексації, пошуку, об'єднання, автоматичної обробки та ін. Це набуває особливого значення при роботі з не структурованими даними та сприяє вирішенню складних бізнес-задач.

Але створення ефективної системи керування метаданими, вимагає їх узгодженої загальної класифікації, в першу чергу, з урахуванням походження самих даних, їх цілей, форматів представлення, задач, у вирішенні яких ці дані використовуються. Тому, мета даного дослідження – це аналіз можливих джерел великих даних та визначення притаманних ним характеристик.

В результаті проведеного аналізу було визначено 8 типів джерел, а саме:

- архіви відсканованих документів;
- документи-файли різних форматів;
- сховища даних;
- бізнес – застосунки;
- публічний веб;
- засоби масової інформації;
- дані журналів;
- автоматично генерований контент.

Аналіз перелічених джерел дозволив визначити наступні основні типи контенту відповідних даних: зображення, аудіо-, відео- файли, документи, повідомлення, дані з датчиків, веб-сторінки. Для представлення вказаного контенту можуть бути використані різні формати, що може впливати на вибір протоколу метаданих, що використовується. Насьогодні, розроблено велику кількість протоколів метаданих, що підтримують визначення метаданих для одного чи кілька форматів представлення даних одного чи декількох типів, однак не існує єдиного підходу чи протоколу, який надавав би загальну таксономію характеристик. Окрім цього, більшість таких протоколів визначають досить детальні, але технічні характеристики даних. Опис семантики даних обмежується, зазвичай, двома елементами: Title та Description. Найбільш гнучкими є протоколи, що надають XML формат представлення метаданих, як XMP специфікація. Такий формат представлення дозволяє включити до контенту будь-яку необхідну метаінформацію та використовувати для визначення семантики контенту цільові онтології. Слід зазначити, що на відміну від технічних характеристик, що зазвичай генеруються автоматично, визначення семантичних характеристик неможливе без участі спеціалістів або експертів, а їх застосування при вирішенні задач вимагає спеціальних програмних засобів, які будуть вміти їх розпізнавати та використовувати.

Розуміючи величезну роль метаданих для забезпечення семантичного визначення контенту великих даних, потрібно пам'ятати, що часто розміри метаданих значно перевищують обсяг самих даних (навіть великих), та дотримуватися принципів розумної ефективності при визначенні елементів метаданих та організації їх зберігання та обробки.

Література

1. https://habr.com/ru/post/93119/
2. https://www.exif.org/category/specifications
3. http://exif.org/dcf.PDF
4. https://helpx.adobe.com/after-effects/using/xmp-metadata.html
5. https://www.dublincore.org/specifications/dublin-core/dces/
6. ISO 16684-1:2012, Graphic technology – Extensible metadata platform (XMP) specification – Part 1: Data model, serialization and core properties
7. https://www.adobe.com/devnet/xmp.html

8. https://forum.allnokia.ru/viewtopic.php?t=51934
9. https://habr.com/ru/post/103635/
10. http://id3.org/id3v2.3.0
11. http://www.xspf.org/xspf-v0.html
12. https://www.ibm.com/support/knowledgecenter/ru/SS88XH_1.6.0/iva/ov_metadata.html
13. https://mediaarea.net/AVIMetaEdit/tech_view_help
14. https://stackoverflow.com/questions/2075175/is-there-a-standard-schema-for-video-metadata
15. https://schema.org/
16. https://studref.com/379466/menedzhment/metadannye_dokumentov
17. http://prgssr.ru/development/posty-dannye-i-metadannye-v-wintersmith.html#heading-section-1
18. https://symfony.com.ua/doc/current/components/yaml/yaml_format.html
19. https://uk.wikipedia.org/wiki/YAML
20. https://frontender.info/like-able-content-spread-your-message-with-third-party-metadata/
21. https://www.ixbt.com/soft/audio-tag-editors.shtml
22. http://ogp.me/
23. https://hostenko.com/wpcafe/plugins/kak-nastroit-open-graph-i-twitter-karty-dlja-wordpress/
24. https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/
25. https://www.oncrawl.com/oncrawl-seo-thoughts/a-complete-guide-to-twitter-cards/
26. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world
27. http://iso.ru/ru/press-center/journal/2122.phtml
28. https://fotoforensics.com/tutorial-meta.php
29. https://www.exif.org/Exif2-2.PDF
30. http://www.belursus.info/soft/i.php?c=exiftool
31. https://webznam.ru/blog/metadannye_fajlov_fotografij/2015-04-01-135
32. https://helpx.adobe.com/ru/premiere-pro/using/metadata.html
33. https://www.stopfake.org/metadannye-nevidimaya-informatsiya-o-fotografii/
34. https://ms.detector.media/mediaprosvita/how_to/13_onlayninstrumentiv_dlya_perevirki_kontentu
35. https://developers.facebook.com/docs/sharing/webmasters/ optimizing?locale=ru_RU

References

1. https://habr.com/ru/post/93119/
2. https://www.exif.org/category/specifications
3. http://exif.org/dcf.PDF
4. https://helpx.adobe.com/after-effects/using/xmp-metadata.html
5. https://www.dublincore.org/specifications/dublin-core/dces/
6. ISO 16684-1:2012, Graphic technology – Extensible metadata platform (XMP) specification – Part 1: Data model, serialization and core properties
7. https://www.adobe.com/devnet/xmp.html
8. https://forum.allnokia.ru/viewtopic.php?t=51934
9. https://habr.com/ru/post/103635/
10. http://id3.org/id3v2.3.0
11. http://www.xspf.org/xspf-v0.html
12. https://www.ibm.com/support/knowledgecenter/ru/SS88XH_1.6.0/iva/ov_metadata.html
13. https://mediaarea.net/AVIMetaEdit/tech_view_help
14. https://stackoverflow.com/questions/2075175/is-there-a-standard-schema-for-video-metadata
15. https://schema.org/
16. https://studref.com/379466/menedzhment/metadannye_dokumentov
17. http://prgssr.ru/development/posty-dannye-i-metadannye-v-wintersmith.html#heading-section-1
18. https://symfony.com.ua/doc/current/components/yaml/yaml_format.html
19. https://uk.wikipedia.org/wiki/YAML
20. https://frontender.info/like-able-content-spread-your-message-with-third-party-metadata/
21. https://www.ixbt.com/soft/audio-tag-editors.shtml
22. http://ogp.me/
23. https://hostenko.com/wpcafe/plugins/kak-nastroit-open-graph-i-twitter-karty-dlja-wordpress/
24. https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/
25. https://www.oncrawl.com/oncrawl-seo-thoughts/a-complete-guide-to-twitter-cards/
26. https://www.datasciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world
27. http://iso.ru/ru/press-center/journal/2122.phtml
28. https://fotoforensics.com/tutorial-meta.php
29. https://www.exif.org/Exif2-2.PDF

30. http://www.belursus.info/soft/i.php?c=exiftool
31. https://webznam.ru/blog/metadannye_fajlov_fotografij/2015-04-01-135
32. https://helpx.adobe.com/ru/premiere-pro/using/metadata.html
33. https://www.stopfake.org/metadannye-nevidimaya-informatsiya-o-fotografii/
34. https://ms.detector.media/mediaprosvita/how_to/13_onlayninstrumentiv_dlya_perevirki_kontentu
35. https://developers.facebook.com/docs/sharing/webmasters/ optimizing?locale=ru_RU

Одержано 17.10.2019

Про автора:

Захарова Ольга Вікторівна,
кандидат технічних наук,
старший науковий співробітник.
Кількість наукових публікацій в українських виданнях – 30.
http://orcid.org/0000-0002-9579-2973.

Місце роботи автора:

Інститут програмних систем
НАН України,
проспект Академіка Глушкова, 40.
Тел.: 526 5139.
E-mail: ozakharova68@gmail.com