УДК 004.02

В.В. Войцеховський
О. В. Бісікало

МЕТОДИ ТА ПІДХОДИ У РОЗРОБЦІ ТЕХНОЛОГІЇ ІНДЕКСАЦІЇ ДОКУМЕНТІВ ТА ЇХ ЗАСТОСУВАННЯ В CLOUD ТЕХНОЛОГІЯХ

Вінницький національний технічний університет

Анотація

У даній статті проведено дослідження і опис методів індексування документів. Наведено опис різних підходів застосування даної технології та доведено її актуальність на прикладах практичного застосування.

Ключові слова:

Програмне забезпечення, індексація, хмарні технології, пошук, методи пошуку, оптимізація пошуку, робота з текстовими даними.

Abstract

In this article, research and description of document indexing methods is carried out. A description of various approaches to the application of this technology is provided and its relevance is proved by examples of practical application.

Keywords:

Software, indexing, cloud technologies, search, search methods, search optimization, work with text data.

Вступ

Індексування документів – це технологія керування інформацією, яка ідентифікує та реєструє певні атрибути документа, щоб зробити його пошук швидшим і простішим. Іншими словами, якісна підтримка індексування документів покращує можливості пошуку та сортування колекцій документів у системі керування документами.

Залежно від варіанту використання атрибути даних або параметри індексування можуть включати широкий спектр описової інформації та метаданих. Наприклад, документи в бухгалтерії можуть бути проіндексовані за номерами рахунків-фактур, іменами постачальників, датою видачі тощо. Подібним чином файли відділу кадрів організації можуть бути проіндексовані за іменем співробітника, номером соціального страхування та іншою подібною відповідною інформацією. Вибір атрибутів даних для індексації зазвичай визначається ймовірністю пошукових запитів, створених кінцевим користувачем.

Актуальність застосування технології

Завдяки розвитку сучасних технологій відбувається глобальна цифровізація даних. Сучасні носії дозволяють зберігати дані у великих кількостях та відтворювати інформацію швидко та якісно. Більшість документації у сучасному світі переноситься та зберігається у електронному форматі, тому виникає потреба ефективного пошуку серед великих масивів даних. Індексація документів пропонує вирішення труднощів зберігання, доступу, організації та захисту електронних документів. Ця технологія є важливою для пошуку інформації, оскільки вона забезпечує швидкість і точність, з якою потрібну інформацію та відповідні

документи можна отримати. Без належної індексації документів пошук інформації буде трудомістким і дорогим. Фактично, згідно з нещодавніми дослідженнями, працівники галузі інформаційних технологій витрачають більше години щодня на те, щоб знайти потрібні документи. Ось чому важливо використовувати індексацію документів – це дозволяє заощадити час на пошук документів, ресурси на сховища, оптимізувати потоки внутрішньої та зовнішньої інформації [1].

Застосування індексації у хмарних технологіях

Хмарні технології – це парадигма, що передбачає віддалену обробку та зберігання даних. Ця технологія надає користувачам мережі Інтернет доступ до комп'ютерних ресурсів сервера і використання програмного забезпечення як онлайн-сервісу. Тобто, якщо є підключення до Інтернету, то можна виконувати складні обчислення, опрацьовувати дані, використовуючи потужності віддаленого сервера. Хмарні технології мають здатність швидко масштабувати сховища даних, що дуже корисно, якщо навантаження на запити можуть змінюватися з часом. Серед компаній що пропонують послуги хмарного зберігання – AWS, GCP, Azure, якими користується більша частина розробників програмного забезпечення. Ці служби є сховищем об’єктів, де кожен об'єкт ідентифікується за назвою. Щоб завантажити дані у хмарне сховище та вивантажити їх з нього, використовуються надані постачальником засоби програмування, зазвичай у вигляді API. Такі API надсилають запити через мережу, щоб отримати доступ до даних [2].

Популярні хмарні інструменти представляють собою онлайн сервіс, для якого найбільш важлива робота пов’язана з пошуковими системами. Для останніх критичним вважається час отримання першого байту інформації, тобто швидке отримання даних. При повнотекстовому пошуку даних зі сховища час передачі зростає відносно їх розміру. Графік відповідної залежності представлено на рисунку 1.

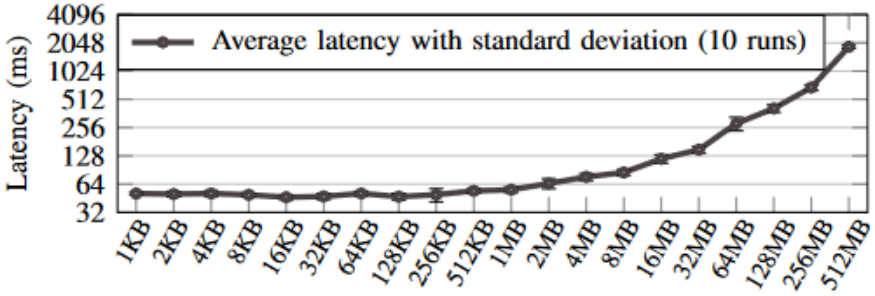

Рисунок 1 – Кореляція між часом затримки та розміром даних

Така тенденція значно погіршує ефективність пошуку. Однак, цей недолік можна виправити, використовуючи різноманітні методи індексації.

Методи індексування

Процес індексування документів передбачає ручне або автоматичне сканування оцифрованих документів для ідентифікації попередньо визначених ключових фраз. Традиційна індексація документів вручну є не тільки трудомісткою, але й зазвичай має значну ймовірність допустити помилку. Тому найбільш вигідний сучасний спосіб полягає в автоматизації цього процесу, що можна забезпечити за допомогою багатьох відомих методів [3].

1. Метод повнотекстового пошуку. Повнотекстовий пошук – це комплексний метод пошуку, який порівнює кожне слово запиту пошуку з кожним словом у документі чи базі даних. У повнотекстовому варіанті пошукова машина аналізує всі слова в кожному збереженому документі та намагається знаходити найближчі відповідності до таких слів-критеріїв, що були визначені користувачем у пошуковому запиті. Метод повнотекстового пошуку став популярним ще у 1990-х роках, коли Інтернет тільки почав входити у повсякденне використання. При роботі з невеликою кількістю документів цілком можливо в процесі повнотекстового пошуку перевірити вміст всіх документів для кожного запиту. Така стратегія отримала назву «послідовного сканування». Проте при використанні такої технології на великих базах даних пошук займав би дуже багато часу, а в інтернеті був би взагалі неможливим. Сучасні алгоритми заздалегідь формують для пошуку так званий повнотекстовий індекс – спеціальний словник, в якому перераховані всі слова і зазначено, в яких місцях вони зустрічаються. За наявності такого індексу достатньо здійснити пошук потрібних слів у

ньому, і тоді одразу буде отримано список документів, в яких вони зустрічаються – такий підхід значно скорочує час виконання пошуку за даним методом.

2. Подвійне індексування ключів. Найбільш традиційним способом індексування є індексування подвійним ключем, яке іноді називають подвійною сліпою перевіркою. Два окремі оператори вводять однакові значення індексування, тоді як програмне забезпечення для введення даних порівнює інформацію безпосередньо в процесі. Якщо є невідповідність, другий оператор, відомий як оператор перевірки, вимагає від них повторного введення значень, доки не буде збігу. Обидва записи мають бути введені однаково, щоб для перевірки було можливо отримати чотири відповідні записи.

Інша стратегія подвійної індексації також включає в себе два оператори, які вводять однакові значення індексації. Після того, як обидва оператори завершать введення ключа, ці два отриманих значення порівнюються і, якщо є невідповідність, повідомлення надсилається третьому й останньому оператору – арбітру. Арбітр переглядає зображення і ключові значення та затверджує коректні ключові значення або повторно вводить значення правильно, якщо неправильними є обидва.

Третя стратегія – напівавтоматизована. Перший прохід збору даних виконується за допомогою програмного забезпечення зонального оптичного розпізнавання символів (OCR), яке автоматично витягує машинописний/машинодрукований текст із відсканованого зображення документа замість того, щоб оператор вводив весь текст вручну. Другий прохід виконує оператор, який потім вручну порівнює індекси так, щоб забезпечити точність даних [4].

3. Індексування змінного пошуку (VLI) – це розширений варіант індексування, який спрощує цей складний і трудомісткий процес. VLI відрізняється від стандартних процесів індексування через заповнення полів індексу поєднанням рівнів автоматизації, що включає пошук бази даних із процесами перегляду вручну. VLI пропонують лише найдосвідченіші компанії зі сканування документів, такі як MetaSource. Винятки трапляються, коли не знайдено відповідності між полем індексу та базою даних, наприклад, має місце спроба зіставити чиєсь ім’я з номером соціального страхування. Це може призвести до ручного дослідження та обробки цих записів. VLI мінімізує винятки, використовуючи кілька баз даних для заповнення полів індексу. Чим нижча частота винятків, тим повнішими та точнішими будуть індексовані поля, що забезпечує швидший і надійніший пошук відсканованих документів.

4. Метадані. За визначенням метадані описуються як «дані, які містять інформацію про інші дані». Іншими словами, це «інформація про дані». Метадані містять інформацію, необхідну для розуміння та ефективного використання даних для різноманітних додатків [5], зокрема застосовуються для підвищення якості пошуку. Пошукові запити, що використовують метадані, можуть врятувати користувача від зайвої ручної роботи з фільтрації. Найважливіші системи на ринку пропонують кілька форм метаданих документів, які допомагають упорядковувати, класифікувати та якомога швидше знаходити документи [6]. Йде мова про:

- Поля або властивості: точний тип інформації, який автор документа зможе ввести на основі типів документів або поля шаблонів, що допомагають надавати додаткову інформацію для цілей упорядкування та розміщення інформації. Прикладами полів є імена, дати, числа та валюта. Вони можуть бути встановлені як обов’язкові та унікальні.
- Шаблони: набір полів, розташованих у певній послідовності для легкого налаштування, видалення та редагування шаблонів документів.
- Інформація про версію: пов’язана з версією інформація, наприклад номер версії, редакція, дата створення, дата зміни тощо, допомагає контролювати номер версії, редакцію, дату створення і т.д. Такі дані мають важливе значення для цілей аудиту.
- Коментарі: враховані системою коментарі до документа допомагають у процесі співпраці. Топові системи дозволяють додавати коментарі до кожної версії. Коментар може бути текстом із вкладеннями, і працівники можуть позначати один одного тегами для запрошень до співпраці та сповіщень. Пошук у коментарях також може допомогти знайти документи, коли вони знадобляться.
- Теги: зазвичай служать методом категоризації документів. Теги часто вказуються на рівні документа, що означає, що їх не можна змінювати для кожної версії.

Висновок

У даній статті було розглянуто актуальність застосування відомих технології індексування текстової інформації. Детально описано популярні методи індексації, способи їх використання. Показано доцільність та перспективність використання даної технології у хмарних сховищах.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. What is Document Indexing? [Електронний ресурс] : [Веб-сайт]. – https://theecmconsultant.com/what-is-document-indexing/ – Назва з екрана.

2. Document Indexing: What Is It, How It Works, and More [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://www.mhcautomation.com/blog/document-indexing-basics/ – Назва з екрана.

3. The Definitive Guide to Document Indexing [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://nanonets.com/blog/document-indexing/ – Назва з екрана.

4. What is METADATA: Why Is It Extremely Important? [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://theecmconsultant.com/metadata/ – Назва з екрана.

5. Ільїн В.В. Дидактичні та технологічні вимоги до програми-оболонки для підготовки та використання електронних навчальних посібників / Ільїн В.В., Теплюк В.М., Бісікало О.В. – Київ: Аграрна освіта, 2004. – 20 с.

6. AIRPHANT: Cloud-oriented Document Indexing [Електронний ресурс] : [Веб-сайт]. – Режим доступу: https://arxiv.org/pdf/2112.13323.pdf – Назва з екрана.

Науковий керівник – Бісікало Олег Володимирович *– професор, в. о. завідувача кафедри автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: obisikalo@vntu.edu.ua*

Войцеховський Вільям Вільямович *– студент групи 1ІСТ-22М, кафедра автоматизації та інтелектуальних інформаційних технологій, Факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: fkca.1akit18.VVV@gmail.com*

Supervisor – Bisikalo Oleh Volodymyrovych *– professor, head of the Department of Automation and Intellectual Informational Technologies, Faculty of Intellectual Informational Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua*

Voitsekhovskyi Viliam Viliyamovych *– student of 1IST-22M group, Department of Automation and Intellectual Informational Technologies, Faculty of Intellectual Informational Systems and Automatics, Vinnytsia National Technical University, Vinnytsia, e-mail: fkca.1akit18.VVV@gmail.com*