Моделювання та проєктування інформаційних систем

Тарелкіна Катерина, КП-31мп

1. Моделі і методи зберігання даних

Модель даних – це абстракція, що дозволяє виділити змістовні об’єкти (сутності, entities) та зв’язки між ними (relations), а також операції над цими об’єктами і зв’язками.

Модель даних дозволяє будувати формальне представлення даних для різноманітних предметних галузей.

Прикладами моделей даних є:

- реляційні моделі (об’єкт – відношення, операції над об’єктами – перетин, об'єднання, декартів добуток, проєкція, віднімання, ділення тощо). Ця модель використовується в реляційних базах даних;
- ієрархічні (на основі мов XML);
- графові. Модель даних є графом, сутності можуть мати безліч зв’язків одна з одною. Така модель, наприклад, може відображати контакти в соціальній мережі. Прикладом СУБД, що підтримує роботу з графовими моделями, є Neo4j;

- моделі ключ-значення. У цьому випадку дані зберігаються у вигляді колекції із парами ключ-значення. Перевагою такої моделі є швидкі операції читання та запису. Зауважмо, що це не реляційна модель, тобто вона не підходить до систем із чіткими рамками відношень між об'єктами та чіткою структурою об'єктів (наприклад, система з даними робітників корпорації);

- документі моделі. У цьому випадку дані зберігаються у вигляді документів, здебільшого у форматі JSON або BSON. Модель дозволяє зберігати складні об'єкти (наприклад, вкладені об'єкти або масиви значень).

Також можна розглянути рівні абстракції даних:

- зовнішній – подання для користувача у візуальній формі (таблиці, діаграми тощо);

- концептуальний (логічний) – подання даних в СУБД (схема даних, мова доступу до даних тощо);

- фізичний – представлення даних на носіях (спосіб організації файлів, індексів тощо).

2. Класифікація інформаційних систем і місце серед них інформаційно-пошукових систем

Класифікація ІС за характером логічної організації інформації:

1. Фактографічні. У цьому випадку дані представлено у вигляді т.зв. інформаційних об'єктів, що мають чітку структуру (= набір реквізитів). Наприклад, це може бути система обліку робітників організації. У цьому випадку система міститиме інформаційні об'єкти з ім'ям, прізвищем робітника, його посадою, номером телефону тощо.

2. Документні. У цьому випадку дані не структуровані або мінімально структуровані (наприклад, обов'язковими атрибутами об'єкта можуть бути дата його створення та автор). Крім того, у деяких документних системах можливе встановлення зв'язків між документами.

3. Геоінформаційні. Такі системи застосовуються у галузях, де присутня просторово-географічна компонента (наприклад, маршрути або карти).

Класифікація ІС за функціями:

- довідкові (користувачі можуть отримати певні "класи" об'єктів із визначеною структурою. Прикладом можуть бути електронні довідники, словники тощо);

- пошукові (*інформаційно-пошукові системи*. Користувачам надається можливість пошуку об’єктів у заданому інформаційному просторі – сукупності інформаційних об’єктів і зв’язків між ними);
- розрахункові;
- технологічні.

3. Організація пошуку. Пошукові машини.

Існують три основні способи пошуку.

1. Каталоги. У цьому випадку сторінки (вебсайти) відсортовані по категоріях і користувач може знайти потрібну сторінку вибравши відповідну категорію.
2. Гіперпосилання. За допомогою посилань користувач може переходити з одного вебсайту на інший.
3. Пошук за ключовими словаим. Це основний метод, за якими працює більшість сучасних пошукових машин.

Метод полягає в наступному:

1. Пошукова машина будує початкову колекцію сторінок, з яких вона будуватиме індекс. Для цього використовуються пошуковий павук (crawler), який рекурсивно збирає гіперпосилання зі сторінок, які він відвідує, і додає їх у колекцію адрес.
2. Після цього пошукова машина збирає весь текст із знайдених вебсторінок для подальшої обробки і побудови індексу.
3. Формування індексу. Пошука машина виділяє змістовні слова тексту і розташовує їх в алфавітному порядку, зберігаючи при цьому номер сторінки та іншу службову інформацію.
4. Пошук. Коли користувач набирає пошуковий рядок, пошукова машина звертається до індексу і шукає задані слова в індексі, після чого надає користувачу знайдені сторінки.

4. Створення і типи індексів

Створення індексу полягає в наступних кроках.

1. Попередня обробка тексту. Із тексту видаляються елементи, не мають змістовного значення: теги, графічні компоненти тощо.

2. Виділення слів. Кожна пошукова машина має свої правила, за якими вона визначає, чи є заданий набір символів (цифр, літер, знаків тощо) словом. За цими правилами пошукова машина складає список усіх тексту в алфавітному порядку.

3. Лінгвістична обробка. Ідея цієї обробки полягає в тому, що машина намагається зрозуміти семантику та контекст тексту для надання кращих результатів користувачу. До лінгвістичної обробки, наприклад, входить токенізація (виділення окремих слів), лематизація (зведення слів до їхніх базових форм – лем), частиномовний аналіз (визначення частин мови кожного слова) тощо. Лінгвістичні алгоритми відрізняються залежно від мови та пошукової машини.

4. Складання індексу. В індексі отримані на попередньому кроці токени сортуються в алфавітному порядку. Для кожного токена також зберігається інформація про те, з якої сторінки його взято, та місце входження токена на цій сторінці.

Для підвищення швидкодії структура індексів часто більш складна (наприклад, замість самих токенів використовують їхній номер, а токени тримають в окремій таблиці).

Існують такі типи індексів.

- Координантий. Цей індекс враховує не лише самі сторінки, що містять задане слово, а і місце розташування слова на сторінці. Це дозволяє більш точно вирахувати релевантність сторінки, а також швидко знайти найбільш підходящу цитату з тексту сторінки.

- Інверсний індекс. Пошукова машина у цьому випадку йде від слів до сторінок, де зустрічаються ці слова (а не навпаки, як це звично, наприклад, для людини).

- Прямий індекс. У цьому випадку пошукова машина має стислу (очищену від не змістовних елементів) текстову копію всіх сторінок. Це дозволяє точно показувати користувачу місце входження слова на сторінці, а також “відновлювати” сторінки (показувати сторінки, які було вже видалено).

5. Проблеми індексування

Головною проблемою пошукових машин є те, що вони націлені на HTML-розмітку. Таким чином, текст, що присутній на вебсайті в інших форматах (наприклад, у вигляді файлів DOCX або Excel), або доданий динамічно за допомогою JavaScript, не буде взятий до уваги пошуковиком.

Крім того, пошуковик не має доступу до не статичної інформації на вебсторінці (наприклад такої, що з’являється лише за запитом користувача). Такі дані називають “глибинним інтернетом”, адже навіть існуючи на вебсторінках, вони не доступні пошуковим машинам.

Також пошуковик не використовує всі сторінки вебсайту, що варто брати при розробці сайту для того, щоби найбільш ефективно розмістити дані по сторінках.

Окрім цього, пошуковик оновлює індекс лише з певним проміжком, що може бути критично для вебсайтів зі швидко змінюваними даними (наприклад, новини або курси валют). У цьому випадку пошуковик використовує ”швидкого робота”, що обходить такі сторінки кілька разів на день. До списку такого робота зазвичай потрапляють сайти із високим посилальним рангом та великою кількістю часто оновлювальних сторінок.

6. Запити до пошукових машин

При написанні запитів до пошукової машини користувач може використовувати різноманітні оператори, як-от АБО та І, щоби уточнити або розширити свій запит. Для пошуку словосполучення буквально можуть використовуватися лапки.

Правила вживання операторів формують мову запитів, що різниться серед пошукових машин. Сьогодні, однак, є тенденція відходити від мови запитів та використовувати природну мову, адже це простіше для пересічних користувачів.

Запити до пошукових машин можна поділити на такі типи:

- навігаційні. Це запити, за яких користувач хоче знайти конкретну сторінку в мережі, наприклад, вебсайт онлайн-магазину;
- інформаційні. У цьому випадку користувач шукає певні відомості і не знає, де саме буде знайдено інформацію. Наприклад, це може пошук автора книги або пошук за зображенням;
- транзакційні. Це запити, за яких користувач хоче здійснити певну дію (наприклад, купити, підписатися, замовити тощо).

Пошукові машини використовують різні підходи до обробки кожного із цих запитів.

7. Якість роботи пошукачів

Для оцінки роботи пошукачів використовують в основному два поняття: повнота і точність пошуку.

Повнота пошуку – це міра того, чи знайшов пошукових усі сторінки, що відповідають запиту. Проте оскільки користувача зазвичай не цікавлять усі результати пошуку, а лише перші 10-20 найбільш релевантних, частіше використовують поняття точності пошуку.

Точність пошуку – це відношення кількості релевантних сторінок до кількості усіх сторінок, що видав пошукач на заданий запит. Окрім просто значення точності, важливим є і ранжування, тобто сортування результатів у порядку зменшення їхньої релевантності. Різні пошукачі використовують різні алгоритми ранжування.

8. Посилальне ранжування (Page Rank)

При ранжуванні сторінок крім релевантності варто брати до уваги і авторитетність сторінки. Наприклад, користувачам більш цікаві релевантні вебсайти, які також є більш авторитетними.

Для розрахунку авторитетності сторінки використовують алгоритм посилального ранжування, ідея якого полягає в наступному. Усі сторінки складаються у матрицю, де кожному рядку та стовпцю відповідає сторінка, а кожній комірці – вага посилання сторінки-стовпця та сторінку-рядок (або навпаки). Спочатку усім сторінкам присвоюється одна й та сама вага. При обході матриці ваги перераховуються. “Авторитетність” сторінки пропорційна кількості посилань на неї на інших вебсайтах. Крім того, вагу додає також авторитетність джерела, на якому розміщене посилання. Таким чином, вищий ранг отримують сторінки з більшою кількістю посилань з більш “авторитетних” джерел.

На практиці для отримання достатньо стабільних ваг достатньо лише кілька обходів матриці.

9. Поняття інформації як категорії, дані і знання

Інформація – це міра усунення невизначеності. Інформація є протилежністю до ентропії, що обчислюється за наступною формулою:

$$H(X) = p(X) \log(p(X)),$$

де p(X) – ймовірність події X.

Одиницею інформації в комп’ютерах є біт – це мінімальна інформація для повідомлення про настання однієї з двох рівноймовірних подій.

Загалом, інформація, дані і знання є синонімами. Знання можна окреслити як певним чином систематизовану інформацію. Дані часто визначають як необроблену інформацію, тобто дані самі по собі не несуть корисного змісту.

Наприклад, на основі даних про температуру та погодні умови метеостанція може створити прогноз погоди – прогноз уже буде інформацією.

10. Програмне та апаратне забезпечення для організації пошуку інформації в мережі інтернет

Для організації пошуку інформації в Інтернеті використовується комплексне програмне та апаратне забезпечення. Основні компоненти такого комплексу включають:

- Пошукові сервери, що зберігають індексовану копію вебсторінок та іншого вмісту Інтернету. Вони відповідають на запити користувачів та повертають результати пошуку.
- Алгоритми пошуку, що аналізують запит користувача та визначають сторінки, що найкраще його задовольнятимуть. При цьому враховуються різні фактори, такі як релевантність, авторитетність та популярність сторінок.
- Для забезпечення пошуку використовується механізм індексації, що полягає в створенні індексу, що дозволяє швидкий доступ до релевантних результатів пошуку.
- Ранжування. Це алгоритм сортування сторінок за спаданням їхньої релевантні та авторитетності.
- Алгоритми машинного навчання. Для покращення ефективності пошуку, пошукові машини все частіше використовують штучний інтелект (наприклад, алгоритми машинного навчання аналізують поведінку користувачів та використовують ці дані для прогнозування та вдосконалення рекомендацій).

Використані джерела

1. https://stud.com.ua/53288/informatika/informatika_dlya_ekonomistiv
2. https://aws.amazon.com/nosql/key-value/