Моделі і методи зберігання даних

Модель даних - це абстракція над реальними даними, за допомогою якої можна виділити об’єкти що на пряму стосуються програми та встановити зв’язки між ними.

Існує багато різних типів моделів даних, основними з яких є

- реляційні - логічні моделі даних, використовуються у реляційних базах даних
- ієрархічні - представляють дані у вигляді деревовидної структури, зазвичай будується на основі формату xml
- мережеві (графові) - логічна модель даних, яка є розширенням ієрархічної. На відміну від ієрархічної дані представлені у вигляді графа та можуть мати багатонаправлені зв’язки (кожен запис може мати багато дітей та багато батьків)
- документно-орієнтована - дані представлені у вигляді документу, наприклад у форматі JSON чи XML
- об’єктно-орієнтована - дані представлені у вигляді об’єктів, що мають свої методи та властивості

Вибір методу зберігання даних залежить від потреб системи та від обраного типу даних. Наприклад для реляційних моделів використовують реляційні бази даних

Класифікація інформаційних систем і місце серед них інформаційно-пошукових систем

Інформаційні системи можна класифікувати за багатьма параметрами, найчастіше використовують функціональність, масштаб та область застосування.

Класифікація за функціональністю:

- оперативно-інформаційні системи - для обробки транзакцій, реєстрації операцій
- стратегічно-інформаційні системи - для стратегічного планування
- кадрові-інформаційні системи - для управління кадрами
- інформаційно-пошукові системи - для збору, організації та пошуку даних

Класифікація за масштабом:

- малі - призначені для особистого користування чи користування невеликою групою осіб
- середні - для середніх організацій з відносно невеликим обсягом даних

- великі - для великих корпорацій з великим обсягом даних

Класифікація за областю застосування -

- фінансові - для ведення фінансового обліку
- логістичні - для управління ланцюгом постачання та складною логістикою
- медичні
- комерційні - для автоматизації процесів у сфері торгівлі

Отже, інформаційно-пошукові системи є підтипом інформаційних систем, які використовуються для збору, організації та пошуку даних. Прикладами таких систем є бази даних, пошукові рушії або системи індексації

Організація пошуку. Пошукові машини.

Організація пошуку - це процес планування та виконання пошукових операцій з метою ефективного та систематичного знаходження інформації. Зазвичай пошук інформації використовується за допомогою пошукових машин.

Пошукова машина - це сервер на якому встановлено пошуковий рушій. Вона забезпечує можливість користувачів швидко знаходити інформацію в інтернеті за допомогою ключових слів . Основними компонентами пошукової машини є:

Сканування — пошукова машина використовує спеціальних веб-роботів або “веб-павуків”, що ходять по сторінкам, збирають дані та індексують веб-сторінки.

Індексація — зібрана інформація про веб-сторінки зберігається у вигляді індексу, який потім використовують для ефективного пошуку інформації.

Обробка інформації — пошукова машина використовує спеціальні алгоритми ранжування для того щоб визначити наскільки релевантні веб-сторінки для конкретного запиту користувача.

Видача результатів — після запиту користувача система за допомогою індексів та алгоритмів ранжування знаходить необхідні результати та представляє їх у вигляді списка посилань.

Створення і типи індексів

Індекс - це структура даних, призначена для пришвидшення пошуку інформації в базах даних, пошукових рушіях та інших системах.

Процес створення індексу складається з наступних етапів

1. вибір даних для індексації
2. створення ключів, що вказують на розташування елементу

3. організація даних - ключі групуються у структуру даних, яка дозволяє ефективно шукати дані

Існує багато видів індексів, у пошукових рушіях найчастіші використовують наступні індекси:

- текстовий індекс - індексує текстовий вміст веб-сторінок, а саме ключові слова та фрази для пошуку
- метатеговий індекс - індексує заголовки, мета-описи, ключові слова та інші метадані сторінки
- індекс зображень - індексує метадані зображення, що дозволяє користувачам шукати зображення за назвою або ключовими словами в інтернеті
- індекс відео - індексує метадані відео, що дозволяє користувачам шукати відео за назвою або ключовими словами в інтернеті
- географічний індекс - індексує дані, що пов’язані із місцезнаходженням елементів
- соціальний - індексує соціальну активність користувачів, таку як перегляди, лайки, репости в соціальних мережах для визначення релевантності контенту

Проблеми індексування

Найбільшими проблемами індексування є

- розмір індексів - чим більший обсяг даних, тим більше необхідно ресурсів для індексування
- оновлення даних - при оновлені чи видалені даних необхідно також оновлювати й індекс, що може займати доволі багато часу, особливо при великої кількості даних
- неправильний формат даних - пошукові системи здатні лише обробляти статичні html сторінки, що робить індексування даних у таких форматах як pdf або json неможливим. Також неможливо індексувати дані що було завантажено за допомогою JavaScript
- забруднення даних - при індексуванні можна випадково захопити неочікувані дані, які були завантажені спеціально для того щоб потрапити до індексу
- безпека даних - необхідно забезпечити надійний захист індексованих даних, оскільки вони можуть містити конфіденційні дані

Запити до пошукових машин

Запит до пошукових машин - це текст, який вводить користувач у спеціально призначений для цього рядок з метою пошуку інформації.

Існує багато типів різних типів запитів до пошукових машин, основними з яких є:

- пошук за ключовими словами
- пошук за фразою
- виключення слів
- пошук в діапазоні дат або чисел
- географічний пошук
- пошук за типом файлу

Якість роботи пошукачів

Основними критеріями якості роботи пошукачів є:

- повнота пошуку - визначає наскільки велика частка релевантної інформації була знайдена
- точність пошуку - визначає наскільки знайдена інформація відповідає потребам користувача
- швидкість пошуку - визначає наскільки швидко пошукова система повертає інформацію
- конфіденційність даних - пошукова система повинна дотримуватись стандартів конфіденційності та захисту даних користувача

Посилальне ранжування (Page Rank)

Page Rank - це алгоритм ранжування, який було розроблено для пошукової системи Google. Він полягає у тому, що важливість сторінки визначається на основі кількості та якості посилань на сторінку.

Для розрахунку важливості сторінок інтернет представляється у вигляді графу, в якому веб-сторінки є вузлами, а посилання - ребрами. Кожна сторінка має спеціальний коефіцієнт важливості - PageRank. Чим більше знайдено посилань на сторінку тим більшим є її PageRank, також важливим аспектом алгоритму є те що сторінки з більшим PageRank дають більшу оцінку іншим сторінкам коли посилаються на неї.

Поняття інформації як категорії, дані і знання

Дані - це необроблена структура, яка не є інтерпретована. Даними можуть бути цифри або факти.

Інформація - це оброблені дані або дані що мають визначений контекст та значення.

Знання - це певні висновки отримані з інформації.

Програмне та апаратне забезпечення для організації пошуку інформації в мережі інтернет

Основним програмним забезпеченням для пошуку інформації в мережі інтернет є:

- Пошукові рушії - програмні засоби, що індексують та аналізують веб сторінки для їх подальшого пошуку користувачами
- Веб-браузери - служать для простої взаємодії користувача з пошуковими двигунами

Основним апаратним забезпеченням для пошуку інформації в мережі інтернет є:

- Сервери - обчислювальні потужності для збереження даних та забезпечення індексації
- Мережеве обладнання - обладнання, призначене для передачі даних між користувачами та серверами по мережі