1. Моделі і методи зберігання даних

Моделі даних є абстракцією предметів, з якими працює певна система, і відображають властивості, поведінку та взаємозв’язки цих предметів. Існують різні методи зберігання даних:

* Файлове зберігання даних, з використанням спеціальних форматів, таких як csv, xml, json тощо.
* Реляційні бази даних, які зберігають дані у вигляді таблиць, між якими є зв’язки.
* Об’єктно-орієнтовані бази даних, що зберігають дані як об’єкти, що мають атрибути та методи.
* NoSQL (документі, графові, ключ-значення тощо) бази даних.

З фізичної точки зору можна виділити наступні методи зберігання даних:

* Зберігання даних на диску пристрою.
* Хмарне зберігання даних на віддалених серверах.
* Зберігання в оперативній пам’яті пристрою, наприклад, для кешування.
* Розподілене зберігання на декількох серверах чи пристроях для покращення доступності та відновлюваності даних.

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

Інформаційні системи, це, як правило, комп’ютерні системи, що збирають, зберігають, оброблюють та розповсюджують інформацію. Такі системи використовуються практично в усіх сферах людської діяльності. Класифікуючи інформаційні системи за призначенням можна виділити наступні:

* системи обробки транзакцій;
* інформаційні системи управління;
* експертні системи;
* бізнес системи;
* інформаційно-пошукові системи.

Варто зазначити, що окрім окремої ролі, інформаційно-пошукові системи можуть бути складовою частиною інших інформаційних систем. Вони можуть використовуватися в будь-яких сферах людської діяльності, наприклад бізнес, освіта, державне управління, мистецтво, наука, розваги тощо. Також вони можуть працювати з різними формами передавання інформації, як текстовою, так і мультимедійною.

3. Організація пошуку. Пошукові машини.

Пошукові машини здійснюють процес пошуку, що складається з наступних частин:

* Індексування: аналіз вмісту документів і створення індексу, який зіставляє ключові слова або терміни з розташуванням цих термінів у документах.
* Власне пошук: визначення за певним алгоритмом документів, актуальних до запиту користувача.
* Ранжування: процес визначення порядку відображення результатів пошуку на основі їх релевантності запиту користувача.
* “Crawling”: пошукові роботи систематично проходять і завантажують нові документи в системі для індексування їх вмісту.

Такі пошукові машини, як Google, Bing, Yahoo здійснюють процес пошуку вебсторінок в мережі Інтернет.

4. Створення і типи індексів

Індекс — це структура даних в інформаційно-пошукових системах, яка пов’язує ключові слова або фрази з документами, які їх містять. Це дозволяє швидко отримувати релевантні результати, коли користувачі виконують пошукові запити.

На прикладі веб пошукових систем, створення індексів відбувається таким чином:

1. Пошуковий робот проходиться по вебсторінкам.
2. Завантажує HTML зміст кожної сторінки та аналізує його.
3. Отримує ключові слова зі структури, тексту та метаданих вебсторінки та додає їх до індексу.
4. Проходиться по посиланням всередині сторінок.
5. Процес оновлення індексів періодично повторюється, щоб вони залишалися актуальними.

Існують наступні типи індексів:

* кластеризовані: індексовані дані і записи індексу зберігаються разом, що часто призводить до покращення продуктивності запитів;
* некластеризований: індекс має зберігається окремо від фактичних даних, і містить покажчики на рядки даних.

За типом даних індекси можуть бути:

* бітмап індекси: використовують бітмап вектори для представлення наявності чи відсутності значень у наборі даних;
* хешеві: використовують хеш-функцію для зіставлення ключів із записами індексу, забезпечуючи швидкий пошук, але не підтримують запити з діапазоном;
* текстові: спеціалізовані на ефективному текстовому пошуку;
* просторові індекси: використовуються для ефективного зберігання та пошуку просторових даних, таких як географічні координати або геометричні дані.

5. Проблеми індексування

При індексуванні інформаційно-пошукові системи можуть стикатися з наступними проблемами:

* Непідходящий тип індексу для даних;
* Застарілість індексів;
* Неструктурованість документів, що ускладнює процес аналізу та отримання ключових слів;
* Дублювання документів;
* Необхідність фільтрувати “сміттєві” документи, що не несуть в собі корисної інформації.

6. Запити до пошукових машин

Запити до пошуковим машин надаються їх користувачами для отримання ними необхідної інформації, найчастіше в формі тексту. Часто використовуються для пошуку в мережі Інтернет за допомогою таких пошукових машин як Google, Bing тощо.

Це можуть бути:

* прості запити по ключовим словам;
* фразові запити;
* запити, що включають логічні оператори ТАК, АБО, НЕ тощо;
* запити до конкретних вебсайтів;
* географічні запити;
* запити з конкретним видом файлу (pdf, docx тощо);
* запити по пошуку зображення чи іншої нетекстової інформації.

7. Якість роботи пошукачів

Оцінками якості роботи пошукачів є швидкість отримання результатів, їх релевантність та правильне ранжування, ефективність індексування.

Конкретними оцінками релевантності є відповідність результатів пошуку до запиту користувача, максимальна гранична релевантність, тобто оцінка того, скільки нової інформації приносить документ враховуючи попередні результати пошуку та актуальність – співвідношення кількості корисної інформації до загальної кількості інформації, отриманої під час інформаційного пошуку.

8. Посилальне ранжування (Page Rank)

Посилальне ранжування або PageRank — це алгоритм, розроблений співзасновниками Google Ларрі Пейджем і Сергієм Бріном, названий на честь Ларрі Пейджа. Він був частиною оригінального алгоритму пошукової системи Google і використовувався для визначення рейтингу веб-сторінок у результатах пошуку. Основна концепція PageRank полягає в оцінці важливості вебсторінки на основі структури посилань у World wide web.

Кожна вебсторінка отримує своє значення важливості, на основі кількості і важливості посилань, що ведуть на дану сторінку. Це ітеративний процес, що відбувається поки значення не стабілізуються. Чим більше сторінок з чим більшим значенням PageRank посилаються на певну сторінку, тим вище її значення PageRank і так далі. По завершенні цього процесу його значення використовується для ранжування результатів пошуку.

Варто зазначити, що цей метод ранжування має свої обмеження і може бути вразливим для маніпуляцій. Тому зараз це не єдиний метод, що використовується Google.

9. Поняття інформації як категорії, дані і знання

Поняття інформації є одним з найбільш дискусійних в науці та філософії, йому важко дати визначення, що не посилається саме на себе. Між даними, інформацією та знаннями можна виділити наступні зв’язки.

Дані – це певні факти, символи та показники, отримані за допомогою органів чуттів чи апаратури. “Сирі” дані є необробленими людиною, і потребують контексту для розуміння і інтерпретування.

Інформація – це оброблені людиною дані, яким було надано контекст, структуровано та інтерпретовано. Інформація надається в формі, що може розумітися людиною.

Знання отримуються на основі великої кількості інформації відносно певних предметів чи аспектів якогось явища. Поєднуючи інформацію з різних сфер людина отримує знання, на основі котрих провадить власну діяльність.

Для ілюстрації можна привести наступний приклад: “20 км” є даними, що не мають контексту для розуміння людиною, “До місця призначення 20 км” є інформацією, яку людина може інтерпретувати. На основі цієї інформації та на основі інформації про громадський транспорт, людина може отримати знання про те, що найоптимальніший спосіб дістатися до місця призначення – на метро, та діяти для досягнення свої мети на основі цього знання.

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

Програмним забезпеченням для організації пошуку інформації в мережі інтернет є, по-перше, веббраузери (Chrome, Firefox, Safari, Edge тощо), які надають користувачам інтерфейс для доступу до Інтернету і можливість введення запитів, отримання, фільтрування та доступ до результатів пошуку.

Браузери додають доступ до використання пошукових систем (машин), таких як Google, Bing, Yahoo тощо. Ці системи індексують, ранжують і повертають вебсторінки в залежності від відповідності запиту користувача.

Пошукові системи мають власне програмне забезпечення для індексування документів, таке як Apache Lucene, Elasticsearch.

Пошукові роботи пошукових машин, такі як Googlebot, Bingbot проходяться по новим стіореним вебсторінкам та індексують їх для пошуку користувачів.

З точки зору апаратного забезпечення, для організації пошуку інформації в мережі Інтернет найчастіше використовують хмарні обчислення, з розподіленими по різним точкам світу датацентрам, що містять велику кількість серверів для зберігання та оброблення великої кількості даних. Ці датацентри пов’язані між собою спеціалізованою та відновлюваною мережевою інфраструктурою. Такий підхід забезпечує належну швидкість, відновлюваність та доступність.