1. Моделі і методи зберігання даних

В контексті інформаційно пошукових систем (ІПС) існує поняття моделі інформаційно пошукової системи. Власне ІПС предаставляє собою набір алгоритмів, які забезпечують відповідність відображуваних документів пошуковим запитам. Простіше кажучи, вона працює для сортування та ранжування документів на основі запитів користувача. Алгоритми ІПС базуються на моделі збереження даних, що використовується у конкретній ІПС. До моделі висувається ряд вимог:

- модель має забезпечити систему перетворення документів у колекції та запитів користувачів
- модель повинна включати функціональність того, як система визначає релевантність документів на основі запиту, наданого користувачем.
- модель також повинна включати логіку ранжування знайдених документів на основі релевантності.

Таким чином модель - це основний інструмент, що визначає ІПС.

Всі моделі можна розділити на 3 класи: класичні моделі ІПС, некласичні та альтернативні.

Класичні моделі - розроблені на основі базових математичних концепцій і є найпоширенішими серед моделей ІПС і можуть бути найлегше реалізовані. До класичних моделей відносять:

- *Булева модель* - це одна з найпростіших і найстаріших моделей пошуку інформації, що базується на булевій логіці та використовує оператори AND, OR і NOT для комбінування умов запиту. Документи представлені як набори термінів, і запит обробляється для визначення документів, які відповідають встановленим умовам. Незважаючи на те, що булева модель ефективна для точного відповідання запиту, вона не здатна ранжувати документи за релевантністю або надавати часткові збіги.
- *Модель векторного простору* - документи та запити представлені векторами в багатовимірному просторі. Кожен вимір відповідає унікальному терміну, а значення в кожному вимірі відображає важливість і частоту терміну в

документі або запиті. Релевантність документів для запиту визначається за допомогою косинусної схожості між вектором запиту і векторами документів. Розроблена частково для вирішення недоліків булевої моделі, модель векторного простору надає ранжовані результати на основі рівня релевантності і широко застосовується в текстовому пошуку.

- *Імовірнісна модель* - модель, що оцінює ймовірність того, що документ релевантний заданому запиту, враховуючи фактори, такі як частота термінів і довжина документа. Вона дуже корисна для обробки великих обсягів даних і надає ранжовані результати на основі ймовірності релевантності. Беручи до уваги вагову статистику, імовірнісна модель ідеально підходить для надання ранжованих результатів.

Некласичні моделі - вони відрізняються від класичних моделей тим, що побудовані на пропозиційній логіці. Пропозиційна логіка (також відома як логіка висловлювань) - це розділ логіки, що вивчає способи комбінування або зміни висловлювань чи пропозицій для формування більш складних висловлювань чи пропозицій.

Альтерантивні моделі - моделі, що використовують принципи класичних моделей та вдосконалюють їх для створення більш функціональних моделей. До альтернативних відносять:

- *Приховане семантичне індексування (LSI)* - LSI використовує декомпозицію сингулярних значень (SVD) для виявлення семантичних зв'язків між термінами та документами. Як і семантичний пошук, семантичне індексування використовує мету і контекст для виявлення концептуально пов'язаних документів, навіть якщо вони не мають точних термінів. Ця ключова можливість робить LSI корисним для вилучення контекстного значення слів у тексті.
- *Okapi BM25* - це важливий варіант імовірнісної моделі, що використовується для ранжування релевантності документів у пошукових системах. Цей метод визначає релевантність документів до пошукового запиту, враховуючи взаємозв'язок термінів у документах. BM25 складається з ряду функцій

оцінювання з різними компонентами та параметрами, а сам термін "найкраща відповідність" вказує на його спрямованість на забезпечення оптимальних результатів у пошуку.

Якщо ж розглядати СУБД, то існують такі моделі зберігаання даних:

- Ієрархічна - в основі упорядкований граф або дерево;
- Мережева - в основі довільний граф;
- Реляційна - грунтується на понятті відношення. Дані описані у вигляді таблиць.

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

Єдино правильного та однохначного способу класифікації існофрмаційних систем (ІС) не існує, тому для їх класифікації зазвичай виділяють ряд ознак.

1) За типом збережених даних:
   a) Фактографічні - призначені для зберігання та обробки структурованих даних у вигляді чисел і текстів. Над такими даними можна виконувати різні операції.
   b) Документальні - у таких системах інформація представлена у вигляді документів, що складаються з найменувань, описів, рефератів і текстів. Пошук по неструктурованим даними здійснюється з використанням семантичних ознак. Відібрані документи надаються користувачеві, а обробка даних в таких системах практично не проводиться.
2) За ступенем автоматизації:
   a) Ручні - відсутність сучасних технічних засобів переробки інформації та виконанням всіх операцій людиною.
   b) Автоматичні - всі опреації виконуються без участі людини;
   c) Автоматизовані - припускають певний відсоток участі людини у процесі обробки інформації. Саме цей клас систем відповідає сучасному уявленню поняття "інформаційна система".

3) За способом використання інформації:
    a) Інформаційно-пошукові системи - призначені для ефективного пошуку та аналізу інформації.
    b) Інформаційно-аналітичні - Системи, призначені для аналізу даних з використанням експертних систем та баз знань;
    c) Інформаційно-вирішальні - ІС, які виконують накопичення та обробку даних з використанням прикладного ПЗ.
4) За архітектурою:
    a) Локальні - ті, що працюють на локальному пристрої без вхаємодії з сервером;
    b) Клієнт-серверні - працюють в локальній чи глобальній мережі з єдиним сервером даних;
    c) Розподілені - мережа компонентів, що взаємодіють між собою (мікросервісна архітектура)
5) За сферою застосування:
    a) Системи організаційного управління – забезпечують автоматизацію функцій управлінського складу.
    b) ІС керування техпроцесами – забезпечують управління пристроями та техпроцесами на автоматизованому виробництві.
    c) Системи наукових досліджень – призначені для наукових вишукувань.
    d) САПР – програмні системи для виробництва проектних робіт з використанням математичних методів.
    e) Навчальні – електронні підручники та довідники.
    f) Інтегровані – забезпечують автоматизацію безлічі функцій підприємств.
    g) Економічні – для роботи з управлінсько-економічною інформацією.

Таким чином, ***інформаційно-пошукові системи*** представляють собою спеціалізовану підгрупу загальних інформаційних систем і спрямовані на забезпечення ефективного пошуку та отримання інформації. Їх можна класифікувати за призначенням, враховуючи їхню спрямованість на ефективний

пошук та аналіз інформації. З точки зору функціональності, інформаційно-пошукові системи можна розглядати як підклас операційних систем, що виконують конкретну функцію - пошук та обробку інформації. Вони використовують різні моделі пошуку, такі як Boolean Model, Vector Space Model, Probabilistic Model, Okapi BM25 та інші, для поліпшення релевантності та ефективності процесу пошуку інформації.

3. Організація пошуку. Пошукові машини.

Організація пошуку - це систематична структура та методи, що використовуються для пошуку інформації в базі даних чи в Інтернеті.

Основні етапи пошуку інформації включають в себе:

1) Визначення ключових слів та формування запитів - процес, що визначає ключові слова, за якими здійснюватиметься пошук та створення запитів на основі ключових слів, які відповідають конкретним потребам користувача.
2) Індексація - створення індексу, який швидко вказує на місце розташування інформації в базі даних чи на вебресурсах. Це важливо для ефективного та швидкого доступу до даних.
3) Використання метаданих та тегів - для докладного опису та класифікації інформації в базі даних чи на вебресурсах.
4) Фільтрація та сортування результатів - застосування різноманітних фільтрів та сортування отриманих результатів для максимального поліпшення їх релевантності.
5) Ранжування результатів пошуку - визначення та використання алгоритмів ранжування для того, щоб користувач отримав найбільш релевантні результати пошуку в першу чергу.
6) Представлення та виведення результатів - Відображення результатів у зручному та зрозумілому форматі для користувача, що забезпечує зручність взаємодії та сприяє розумінню отриманої інформації.

Пошукові системи - це програми або сервіси, які автоматично просуваються по Інтернеті та інших джерелах для збору інформації та створення індексу з

метою надання швидкого та ефективного пошуку. Деталізуємо основні аспекти пошукових систем:

1) Індексація та обхід - пошукові системи проводять сканування веб-сторінок та інших ресурсів для створення індексу, який полегшує швидкий доступ до інформації.
2) Алгоритми ранжування - використання різноманітних алгоритмів для визначення порядку виведення результатів пошуку відповідно до їх релевантності.
3) Ключові слова та метадані - наліз ключових слів та метаданих з метою кращого розуміння тематики та змісту веб-сторінок.
4) Алгоритми фільтрації та класифікації - використання алгоритмів для фільтрації та класифікації інформації з метою відбору та систематизації результатів.
5) Збереження та оновлення індексу - постійне оновлення індексу для врахування нової інформації та змін на веб-сайтах, що гарантує актуальність результатів.
6) Краудсорсинг та коригування результатів - залучення користувачів через краудсорсинг та коригування результатів з використанням їхніх внесків та даних.

Пошукові системи, такі як Google, Bing та інші, використовують ці технології для поліпшення якості та ефективності пошуку інформації в Інтернеті.

4. Створення і типи індексів

Створення індексу з вебсторінок за допомогою пошукової машини включає такі етапи:

1) Конверсія в чистий текст - тексту сторінки перетворюється в чистий текст шляхом вилучення нетекстових елементів, таких як графіка та розмітка HTML.

2) Вибірка слів - всі слова виділяються з тексту для подальшого їх алфавітного упорядкування. Визначення того, що вважати словом, включає в себе літери, числа, буквено-цифрові послідовності та інші критерії.
3) Лінгвістична обробка - застосування алгоритму машинної морфології для приведення слів до їхніх початкових граматичних форм або основ (лематизація, стемінг слів).
4) Складання індексу - об'єднання основ слів у впорядкований за алфавітом словник. Кожна основа має вказівку про номер сторінки та місце входження на цій сторінці.

Цей процес дозволяє створювати ефективний індекс для подальшого поліпшення швидкості та якості пошуку інформації на веб-сторінках.

Індекс пошукової системи - це величезна база даних або бібліотека інформації, яка містить відомості про всі веб-сторінки, які пошукова система просканувала та проаналізувала за певний час, що дозволяє пошуковим системам швидко та ефективно відповідати на пошукові запити користувачів, надаючи список релевантних веб-сторінок. Коли користувач вводить пошуковий запит у пошукову систему, вона шукає у своєму індексі сторінки, які відповідають запиту, а потім ранжує їх на основі факторів свого складного алгоритму, включаючи більш загальні фактори, такі як релевантність, авторитетність і популярність.

Пошукові машини (пошукові системи) використовують такі типи індексів:

- Суфіксне дерево - фігурно структурований як дерево, підтримує пошук за лінійним часом. Будується шляхом зберігання суфіксів слів. Суфіксне дерево є типом префіксного дерева, яке підтримує розширену хеш-функцію, що важливо для індексації пошукових систем. Використовується для пошуку патернів у послідовностях ДНК та кластеризації. Однак його недолік полягає в тому, що зберігання слова в дереві може потребувати більше простору, ніж для самого слова. Альтернативним представленням є масив суфіксів, який вважається меншим за обсягом віртуальної пам'яті і підтримує методи стиснення даних, такі як алгоритм BWT.

- Інвертований індекс - зберігає список входжень кожного атомарного критерію пошуку, зазвичай у формі хеш-таблиці або бінарного дерева. Широко використовується в пошукових системах для відображення термінів до документів, в яких вони зустрічаються. Це дозволяє ефективно витягти документи, що містять конкретні терміни.
- Цитатний індекс - зберігає цитати або гіперпосилання між документами для підтримки аналізу цитувань, що є предметом бібліометрії. Цей тип індексу є важливим для розуміння взаємозв'язків між документами на основі цитат або посилань.
- Індекс n-грам - зберігає послідовності довжини n в даних для підтримки різних видів пошуку чи текстового аналізу. n-грами корисні для завдань, таких як моделювання мови, перевірка правопису та інформаційний пошук.
- Матриця термів документа - використовується в латентному семантичному аналізі, зберігає входження слів в документах у двовимірній розрідженій матриці.

Крім інвертованого індекса пошукові системи також використовують прямий індекс. Він використовується для відображення цитат з виділеними словами запиту. Для цього пошукові машини зберігають тексти всіх проіндексованих сторінок в компактному та стислому вигляді. Цей прямий індекс представляє собою текстову копію всього Інтернету, звідки можна витягти цитати при пошуку.

5. Проблеми індексування

Один із основних викликів індексації полягає в управлінні величезним обсягом даних в Інтернеті. З ростом кількості веб-сторінок індексаційні системи повинні ефективно справлятися зі зберіганням та обробкою цієї інформації для швидкого та точного пошуку.

Постійна зміна інформації на веб-сторінках створює виклик для пошукових систем, оскільки вони повинні систематично оновлювати свої індекси, щоб забезпечити користувачам актуальні результати.

Проблеми виникають із слів, які мають багатозначність. Індексаційні системи повинні враховувати контекст та вдосконалювати алгоритми для правильного та точного розрізнення між різними значеннями.

Використання синонімів може впливати на коректність індексації, і важливо розрізняти схожі терміни для точних та релевантних результатів пошуку.

Індексація інформації в різних мовах вимагає урахування мовних відмінностей та розробки адаптивних стратегій для різноманітних мовних контекстів.

Поява неструктурованих даних на веб-сайтах ускладнює завдання індексації, вимагаючи удосконалення методів виокремлення та індексації цих даних.

Стоп-слова, хоча широко вживані, можуть впливати на точність індексації, тому важливо використовувати ефективні фільтри для їх відсіювання та оптимізації ресурсів індексації.

Вирішення цих проблем вимагає постійного вдосконалення алгоритмів індексації та врахування специфіки джерела інформації для забезпечення високої якості пошукового досвіду.

6. Запити до пошукових машин

Взаємодія користувачів з пошуковими системами є важливою складовою їхньої онлайн активності. У процесі пошуку інформації в Інтернеті користувачі використовують різноманітні стратегії та техніки для максимально ефективного використання пошукових систем.

Однією з основних стратегій є використання ключових слів. Користувачі можуть вводити одне слово, щоб отримати загальну інформацію, або комбінацію слів для точнішого обмеження результатів і знаходження конкретної інформації. Це дозволяє адаптувати пошук до конкретних потреб та очікувань користувача.

Фразові запити, які використовують лапки, стають корисним інструментом для пошуку точних виразів чи фраз. Вони дозволяють зберегти порядок слів і знайти конкретні вирази, що полегшує знаходження необхідної інформації.

Логічні оператори AND, OR і NOT розширюють можливості користувача для точного визначення умов пошуку. Це важливо при потребі включити чи виключити певні терміни з результатів пошуку.

Спеціальні запити, такі як обмеження результатів до конкретного сайту чи фільтрація за типом файлу, роблять пошук більш цільованим. Це особливо корисно, коли користувачі шукають конкретний вид інформації чи ресурс.

Додавання "define:" перед словом для отримання визначення чи введення математичних виразів для обчислень є частиною розширених стратегій пошуку.

Запити з контекстом, такі як вказання періоду часу для отримання актуальних результатів чи додавання географічного контексту, роблять пошук більш зорієнтованим та адаптованим до конкретних умов користувача. Всі ці стратегії сприяють більш точному та ефективному використанню пошукових систем у різних ситуаціях.

7. Якість роботи пошукачів

Якість роботи пошукових машин можна оцінити за різними критеріями. Основними з них є:

- Точність видачі: відсоток релевантних документів, відповідних до пошукового запиту в пошуковій видачі.
- Задоволеність користувача: ступінь відповідності результатів пошуку потребам користувача.
- Швидкість пошуку: час, необхідний для отримання результатів пошуку.
- Доступність: можливість доступу до пошукової машини з різних пристроїв і місць.

Точність видачі є найважливішим критерієм оцінки якості роботи пошукової машини. Вона залежить від того, наскільки добре пошукова машина розуміє сенс пошукового запиту і здатна відшукати релевантні документи.

Задоволеність користувача є комплексним критерієм, що враховує такі фактори, як точність видачі, корисність результатів, зручність використання пошукової машини, тощо.

Швидкість пошуку також є важливим фактором, особливо для користувачів, які шукають інформацію в реальному часі.

Доступність є важливим критерієм для користувачів, які мають обмежений доступ до Інтернету.

Крім цих основних критеріїв, існують також інші, більш специфічні критерії, які можуть бути використані для оцінки якості роботи пошукових машин. Наприклад, можна оцінити:

- Обсяг і різноманітність інформації, доступної через пошукову машину.
- Якість представлення інформації в результатах пошуку.
- Наявність додаткових функцій, які полегшують користувачам пошук інформації.

Оцінка якості роботи пошукових машин може проводитися різними способами. Одним із поширених способів є тестування користувачів. У цьому випадку користувачам пропонується виконати ряд завдань, використовуючи пошукову машину. На основі результатів виконання завдань оцінюється якість роботи пошукової машини.

Іншим способом оцінки якості роботи пошукових машин є статистичний аналіз. У цьому випадку аналізуються дані про результати пошуку, отримані з різних джерел. На основі цих даних оцінюється, наприклад, точність видачі, задоволеність користувачів, тощо.

Оцінка якості роботи пошукових машин є важливим завданням, оскільки вона дозволяє розробникам пошукових машин поліпшувати якість своїх продуктів.

8. Посилальне ранжування (Page Rank)

PageRank - алгоритм ранжування посилань, який оцінює кількість і якість посилань, що ведуть на веб-сторінку. Для кожної сторінки алгоритм обчислює дійсне число від 0 до 10, чим більше число — тим «важливіша» сторінка (тим більше на неї посилань). PageRank виник як академічний підхід до оцінки важливості публікацій автора через їх згадки в бібліографічних посиланнях інших дослідників. Для його адаптації до використання в Інтернеті внесено деякі зміни:

вага кожного посилання розглядається індивідуально та нормується за кількістю посилань на сторінці.

PageRank працює за логарифмічною шкалою, а не за лінійною, як може здатися на перший погляд. Багато фахівців вважають, що шкала має логарифмічну основу, рівну п'яти. Це означає, що кожне додаткове збільшення на шкалі - це п'ятикратне збільшення важливості сторінки. Тобто сторінка з PR4 буде вважатися в 25 разів важливішою, ніж сторінка з PR2.

Формула, за якою обчислюється PageRank наступна (1):

$$PR(A) = (1 - d) + d \sum_{i \in In(A)} \frac{PR(i)}{L(i)}$$

де:

- PR(A) - PageRank сторінки A;
- d - демпфіруючий фактор - визначає ймовірність того, що користувач, перебираючи веб-сторінки, продовжить переходити до іншої сторінки, замість того, щоб завершити свій перегляд. Введення цього фактора дозволяє враховувати характеристику випадкового блукання користувача в мережі. Зазвичай становить 0,85;
- In(A) - множина сторінок, які посиалються на сторінку A;
- PR(i) - PageRank сторінки i;
- L(i) - кількість посилань на сторінці i.

9. Поняття інформації як категорії, дані і знання

Інформація - це складна і багатогранна категорія, яка має різні визначення в різних науках. У загальному розумінні інформація - це відомості про що-небудь, повідомлення, дані, що містять нові знання.

Інформація може бути представлена в різних формах, наприклад, у вигляді тексту, графіків, таблиць, зображень, звуку, відео. Інформація може бути отримана з різних джерел, наприклад, з природи, з інших людей, з технічних пристроїв.

Інформація може бути використана для різних цілей, наприклад, для навчання, роботи, розваг, управління.

Дані - це необроблена інформація, яка не має чіткого значення або значення.

Знання - це інформація, яка пройшла обробку і має чітке значення або значення.

Дані можуть бути перетворені в знання за допомогою таких процесів, як:

- Сортування - відділення важливих даних від неважливих.
- Аналіз - виявлення закономірностей і зв'язків у даних.
- Інтерпретація - надання даних сенсу або значення.

Знання можуть бути використані для різних цілей, наприклад, для прийняття рішень, вирішення проблем, творчості.

Інформація, дані і знання є взаємопов'язаними поняттями. Інформація може бути представлена у вигляді даних, а дані можуть бути перетворені в знання. Інформація може бути використана для отримання даних, а дані можуть бути використані для отримання знань. Наприклад, новина про те, що в Києві стався вибух, є інформацією. Ця інформація може бути представлена у вигляді даних, наприклад, у вигляді тексту, фотографії або відео. Ці дані можуть бути використані для отримання знань, наприклад, для розуміння причин вибуху або для оцінки його наслідків.

Таким чином, інформація, дані і знання є важливими категоріями, які використовуються в різних сферах людської діяльності.

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

Інтернет - це величезний інформаційний простір, який містить безліч веб-сторінок, файлів, зображень, відео та інших видів інформації. Пошук інформації в Інтернеті - це складне завдання, оскільки потрібно знайти потрібну інформацію серед величезної кількості даних.

Для організації пошуку інформації в Інтернеті використовуються спеціальні програмні та апаратні засоби.

Основним програмним забезпеченням для організації пошуку інформації в Інтернеті є пошукові системи. ***Пошукова система*** - це програмно-апаратний комплекс, який забезпечує доступ користувачів до інформації, розміщеної в Інтернеті.

Пошукова система працює за наступним алгоритмом:

1) Павук (*crawler*) переміщується по Інтернету і збирає інформацію з веб-сторінок. Павлук використовує різні методи для виявлення веб-сторінок, наприклад, аналізує посилання на веб-сторінки, які вже були зібрані, або використовує спеціальні алгоритми для пошуку нових веб-сторінок.
2) Індексатор (*indexer*) аналізує інформацію, зібрану павуком, і створює її індекс. Індекс - це база даних, яка містить інформацію про веб-сторінки, наприклад, їх адреси, заголовки, текст, зображення тощо.
3) Пошуковий сервер (*search server*) відповідає на запити користувачів і видає результати пошуку. Пошуковий сервер використовує індекс для пошуку веб-сторінок, які відповідають запиту користувача.

Крім пошукових систем, існує також інше програмне забезпечення, яке може використовуватися для пошуку інформації в Інтернеті. До такого програмного забезпечення відносяться:

- Агенти пошуку (*search agents*) - це програми, які допомагають користувачам знаходити інформацію в Інтернеті, використовуючи різні методи пошуку. Агенти пошуку можуть використовувати пошукові системи, а також інші джерела інформації, наприклад, каталоги веб-сайтів або форуми.
- Метапошукові системи (*metasearch engines*) - це системи, які об'єднують результати пошуку з декількох пошукових систем. Метапошукові системи дозволяють користувачам отримати більш широкий спектр результатів пошуку, ніж при використанні однієї пошукової системи.
- Спільні каталоги (*web directories*) - це каталоги веб-сайтів, які поділяються на категорії. Спільні каталоги можуть бути використані для пошуку інформації за темами або категоріями.

Апаратне забезпечення для організації пошуку інформації в Інтернеті включає в себе:

- Комп'ютери - на яких працюють пошукові системи, агенти пошуку, метапошукові системи та спільні каталоги.
- Сервери - на яких зберігаються інформація, зібрана павуками, а також індекси та результати пошуку.
- Мережеві пристрої - які забезпечують доступ до Інтернету.

Розвиток програмного та апаратного забезпечення для організації пошуку інформації в мережі Інтернет спрямований на підвищення ефективності пошуку, а також на розширення можливостей користувачів.

Одним із перспективних напрямків розвитку є використання штучного інтелекту для розуміння сенсу запитів користувачів і для пошуку інформації, яка є найбільш релевантною для цих запитів.

Іншим перспективним напрямком розвитку є використання мобільних пристроїв для пошуку інформації. Це дозволить користувачам отримувати доступ до інформації в будь-якому місці і в будь-який час.