5. СТРУКТУРА ТА ПРИНЦИПИ ФУНКЦІОНУВАННЯ ПОШУКОВОЇ СИСТЕМИ У ВЕБ

Практично всі великі пошукові системи мають свою власну структуру, відмінну від інших. Однак можна виділити загальні для всіх пошукових машин основні компоненти. Відмінності в структурі можуть бути лише у вигляді реалізації механізмів взаємодії цих компонентів. Розглянемо типову структуру пошукової системи для WWW, запропоновану Budi Yuwono, Dik L.Lee у статті “Search and Ranking Algorims for Locating Resources on the World Wide Web”.

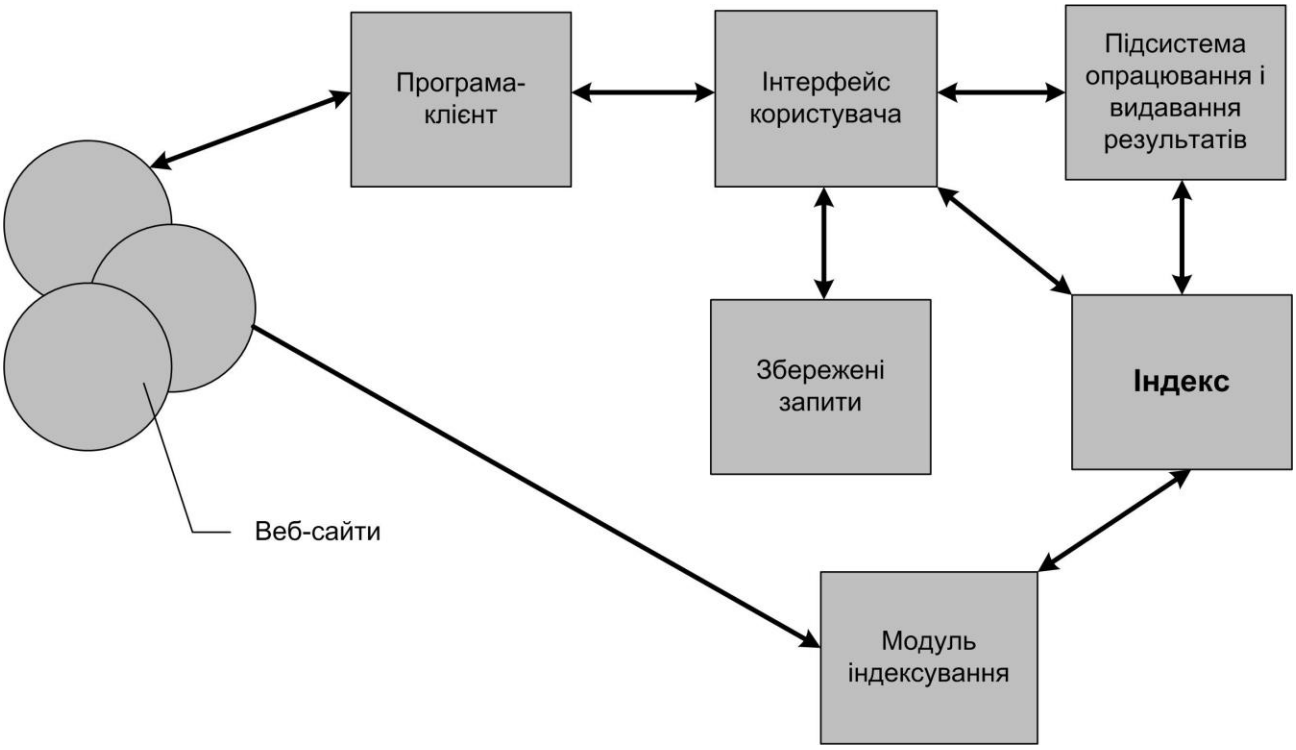

Рис. 1. Типова структура пошукової системи для WWW (Budi Yuwono, Dik L.Lee)

Розглянемо кожну складову окремо.

1. **Модуль індексування.** Він служить для постійного сканування мережі Інтернет та підтримування бази даних індексу в актуальному стані. Цей модуль є основним джерелом інформації про стан інформаційних ресурсів мережі. Він складається з трьох допоміжних програм (роботів):

- **Spider** (павук) - програма, яка призначена для скачування веб-сторінок. «Павук» забезпечує скачування сторінки і витягує всі внутрішні посилання з цієї сторінки, тобто викачується html-код кожної сторінки. Для

скачування сторінок роботи використовують протоколи HTTP. Павук працює за такою схемою: він передає на сервер запит “get/path/document” та деякі інші команди HTTP-запиту. У відповідь робот отримує текстовий потік, що містить службову інформацію і безпосередньо сам документ. Посилання витягуються з тегів <a>, <area>, <base>, <frame>, <frameset> тощо. Разом з посиланнями, багато роботів обробляють редіректи (перенаправлення). Кожна скачана сторінка зберігається в наступному форматі:

- URL-адреса сторінки;
- дата, коли сторінка була завантажена;
- http - заголовок відповіді сервера;
- тіло сторінки (html -код).

Отже, павук робить запит вмісту сторінок так само, як це робить звичайний інтернет браузер, відправляючи на сервер HTTP запит і отримуючи від нього відповідь. Після того, як вміст сторінки завантажено, він надсилається до кроулера та індексатора.

- **Crawler** (“мандрівний” павук) – програма, яка автоматично проходить за всіма посиланнями, що знайдені на сторінці. Краулер аналізує шляхи, що ведуть з поточної сторінки на інші розділи сайту, або на сторінки зовнішніх інтернет-ресурсів, і визначає подальший порядок обходу павуком ниток всесвітньої павутини. Саме кроулер знаходить нові для пошукової системи сторінки і передає їх павуку. Його завдання – визначити, куди далі повинен йти павук, ґрунтуючись на посиланнях або виходячи із заздалегідь заданого списку адрес.
- **Indexer** (робот-індексатор) – програма, яка аналізує веб-сторінки, які завантажені павуками. Індексатор розбирає сторінку на складові частини і аналізує їх, застосовуючи власні лексичні та морфологічні алгоритми. Індексатор здійснює первинний аналіз вмісту завантаженої сторінки, виділяє основні частини (назва сторінки, опис, посилання, заголовки тощо) і розкладає їх у відповідні розділи пошукової бази даних - поміщає в індекс пошукової системи. Цей процес називають **індексуванням інтернет-**

ресурсів, звідси і назва самої підсистеми. На основі результатів первинного аналізу, індексатор також може прийняти рішення, що сторінка взагалі “недостойна” знаходитися в індексі. Причини такого рішення можуть бути різними, наприклад: сторінка не має назви, вона є точною копією іншої, що вже наявна в індексі, або містить посилання на заборонені законодавством ресурси.

2. **Індекс пошукової системи** (index database) – це база даних, яка зберігається на пошуковому сервері і яка містить посилання на проіндексовані ресурси і стислі копії веб-сторінок. В індексі пошукова система зберігає свій “словниковий запас”, тобто набір слів і словосполучень, які зустрічаються на інтернет-сторінках. Він реалізований у формі інвертованого файлу, в якому кожне слово чи словосполучення поєднане з адресами тих веб-сторінок, на яких вони зустрічаються. Стисла копія веб-сторінки зберігається у формі переліку слів, які наявні у тексті сторінки, для кожного з яких перераховані позиції, в яких воно зустрічається на цій сторінці. При цьому відкидаються стоп-слова, а інші слова можуть приводитися до вихідної форми. Індекс використовується системою для пошуку сторінок з входженням ключових слів, які задані запитом користувача. Індекс постійно поповнюється новою інформацією, яку збирає павук пошукової системи. Для того, щоб сайт з'являвся в списку видачі пошукової системи за певними запитами, він, або хоча б певна частина його сторінок, повинні бути внесені в індекс відповідної пошукової системи. Павук пошукової системи може дізнатися про новий сайт одним із двох способів – через звернення власника сайту або через наявність посилань з проіндексованих сайтів на цей сайт.

3. **Підсистема опрацювання та видавання результатів** (Search Engine and Results Engine). Це найважливіша частина будь-якої пошукової машини. Алгоритми роботи цієї підсистеми компанії-розробники зберігають у суворій секретності, оскільки вони є комерційною таємницею. Саме ця частина пошукової машини відповідає за адекватність відповіді пошукової системи на запит користувача. Вона служить для трансляції запиту користувача з інформаційно-пошукової мови (ІПМ) у формальний запит системи, пошуку

посилань на інформаційні ресурси Інтернету і видавання результатів цього пошуку користувачеві. У ній можна виділити два основних компоненти:

- **Підсистема рангування**. Рангування – це сортування сторінок інтернет-сайтів відповідно до їх релевантності певному запиту. Релевантність сторінки – це ступінь відповідності змісту сторінки змісту запиту і цю величину пошукова машина визначає самостійно, виходячи з величезної кількості параметрів. На рангування сторінки, крім її структури та вмісту (контенту) також впливають: кількість і якість посилань, що ведуть на дану сторінку з інших сайтів; вік домену самого сайту; характер поведінки користувачів, які переглядають сторінку і багато інших чинників.
- **Підсистема видавання результатів**. У завдання цієї підсистеми входить інтерпретація користувацького запиту, його переклад на мову структурованих запитів до індекса і формування сторінок результатів пошуку. Крім тексту самого запиту, пошукова машина може також враховувати :
  - **Контекст запиту**, що формується виходячи зі змісту раніше здійснених користувачем запитів. Наприклад, якщо користувач часто відвідує сайти спортивних новин, то на запит зі словом “Дніпро” або “Карпати” він, ймовірно, прагне отримати інформацію про ці футбольні клуби, а не про однойменну річку чи гірську систему. Це називається персоналізованим пошуком, тому результати на один і той самий запит для різних користувачів може істотно відрізнятися.
  - **Користувацькі уподобання**, про які пошукова машина може “здогадуватися”, аналізуючи посилання, які користувач обирає на сторінках результатів пошуку. Це ще один спосіб скоригувати контекст запиту: користувач своїми діями ніби підказує машині, що саме він хоче знайти. Як правило, до результатів пошуку пошукові машини намагаються додавати сторінки, які релевантні запиту, але які відносяться до різних сфер життя. Припустимо, що користувач цікавиться музикою і тому часто вибирає посилання на сторінки про

музичні гурти і їх творчість, навіть якщо ці сторінки не зовсім релевантні вихідному запиту. При формуванні відповіді на наступний запит цього користувача, система може надавати перевагу сторінкам, що пов'язані з музикою, у назві яких зустрічаються слова з тексту запиту.

- **Регіон**, що є особливо важливим при обробці комерційних запитів, пов'язаних з придбанням товарів і послуг у місцевих постачальників. Якщо користувач проживає у Львові і хоче придбати телевізор, то, найймовірніше, що цю особу не цікавлять ціни на телевізори, наприклад, у Запоріжжі, якщо це явно не вказано в тексті запиту. Очевидно, що у результатах повинні найперше з'явитися ціни на телевізори у Львові. Тому сучасні пошукові машини поділяють запити на геозалежні і геонезалежні. Тому, якщо пошукова система вирішує, що запит користувача є геозалежний, то вона автоматично додає до нього ознаку регіону, яку визначає за інформацією від інтернет-провайдера цього користувача.
- **Час**. Пошукові машини часто аналізують, коли відбувалися події, які описані на сторінці. Оскільки інформація постійно застаріває, а користувач потребує найперше найсвіжіші новини, актуальні прогнози, анонси та інформацію про події, які зараз відбуваються. Тому пошуковій системі необхідно зрозуміти, що актуальність сторінки залежить від часу, і зіставити її з моментом виконання запиту.

4. **Інтерфейс користувача** (user interface) – це спосіб спілкування користувача з пошуковим апаратом системи, тобто з системою формування запитів і переглядів результатів пошуку. Це HTML-форма, яка відкривається за допомогою програми-клієнта, такої як Internet Explorer, Mozilla Firefox, Opera тощо, в яку користувач вводить запит і натискає відповідну кнопку для здійснення пошуку.

5. **Збережені запити** (saved queries) – це запити, які надходять до пошукової системи від користувачів. Вони зберігаються в базі даних і дають

можливість пошуковій системі прогнозувати можливі варіанти запиту і пропонувати “підказку” користувачам на основі попередніх збережених запитів.

6. **Програма-клієнт** (програма перегляду) (client) – це засіб перегляду інформаційних ресурсів у WWW. За допомогою програми-клієнта здійснюється доступ, зокрема, до інтерфейсу користувача пошукової системи

7. **Веб-сайти (WWW sites)** – це ті інформаційні ресурси WWW, перегляд яких забезпечується програмами-клієнтами.

Розглянемо принципи взаємодії елементів архітектури пошукової системи.