# Лекція 11. Пошукові системи

## Загальні відомості про пошукові системи

Пошукова система - це складний програмно-апаратний комплекс, що призначений для здійснення пошуку ресурсів в Інтернет, збереження відомостей про них в своїх базах і надання користувачу переліку посилань відповідно до його пошукового запиту.

На даний час основними пошуковими системами вважають такі, що мають власні бази і алгоритми пошуку - Google, Yahoo, MSN, Яндекс. Більшість інших пошукових систем використовує їх бази. Наприклад, пошук AOL (search.aol.com) використовує базу Google, а AltaVista, Lycos і AllTheWeb - базу Yahoo.

Пошук інформації в Інтернеті складається з двох частин (рис. 1).

1. Пошукові системи обходять різноманітні ресурси по всьому Інтернету, викачують, обробляють і зберігають цю інформацію в базах на своїх серверах.
2. Користувач через головну сторінку пошуковика формує запит і отримує відповідь з серверів даної пошукової системи.

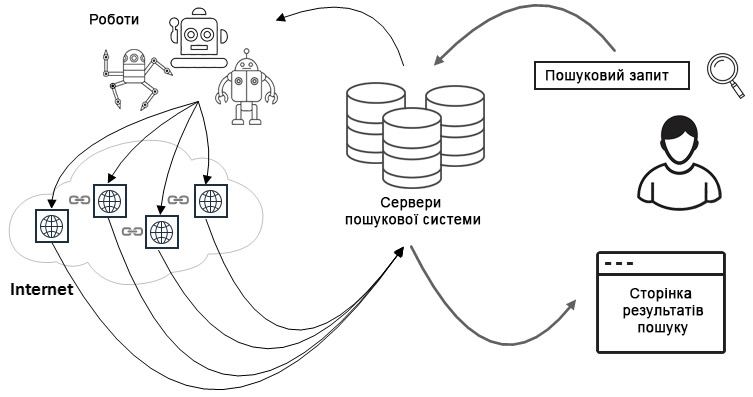


Рис. 1. Загальна схема пошуку інформації в Інтернеті

Головним завданням пошукової системи є здатність надавати користувачам саме ту інформацію, яку вони шукають. Навчити користувачів робити «правильні» запити до пошукової системи, які відповідають її принципам роботи неможливо. Тому, розробники створюють такі алгоритми і принципи роботи пошукових систем, які найкраще пристосовані до поведінки і ходу думок пересічного користувача. Пошукова система повинна діяти так само, як діє користувач при пошуку інформації і надавати за його запитом інформацію максимально швидко і просто.

Користувач оцінює роботу системи за кількома основними критеріями.

* Чи знайшов він те, що шукав і чи була потрібна інформація серед перших результатів пошуку? Наскільки швидко пошукова машина обробляла запит?
* Якщо не знайшов, то скільки разів йому довелося перефразувати запит, щоб знайти потрібне? Як багато непотрібної інформації було знайдено нарівні з корисною?
* Наскільки актуальною є надана інформація і чи зручно представлено результати пошуку?

Для того, щоб задовольнити зростаючим потребам користувачів, розробники пошукових машин постійно вдосконалюють алгоритми і принципи пошуку, додають нові функції і можливості, всіляко намагаються пришвидшити роботу системи.

## Принципи роботи пошукової системи

Практично всі великі пошукові системи мають власні структури, що різняться 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.

Логічну структуру пошукової системи можна представити у вигляді трьох модулів:

1. **Модуль індексування інформації.** Містять спеціальні програми – роботи (боти), які надсилають запити до навколишніх серверів, де зберігаються сайти і завантажують їх вміст. В роботів є спеціальний розклад, згідно з яким вони здійснюють свій обхід.
2. Сторінки сайту, що завантажені роботом, спеціальним чином обробляються і надсилаються на зберігання до **бази даних**. Це гігантські обсяги інформації (мільярди сторінок), і для їх збереження використовуються спеціальні сервери.
3. **Сервер** пошукової системи відповідає за обробку запитів користувачів і видачу результатів пошуку. Щодня пошукові системи обробляють сотні мільйонів запитів. Для того, щоб видавати на кожен запит релевантні результати, система шукає в своїх базах всі документи, що відповідають запиту і видає ті, які мають найкращу відповідність. На сторінці видачі пошукової системи сайти сортуються за спаданням їх релевантності до запиту користувача.

### Індексування інформації

Розміри сучасного Інтернету обчислюються в екзабайтах, тобто в мільярдах мільярдів байтів. При запиті користувача пошуковик звертається до бази даних (пошукового індексу), де для всіх слів, які є на відомих для пошуку сайтах, зазначено їх місце знаходження - адреса сторінки і місце на ній. Індекс можна порівняти з предметним покажчиком в книзі або адресному довіднику. На відміну від звичайного предметного покажчика, індекс містить не лише терміни, а всі слова. На відміну від адресного довідника, в кожного слова є не одне, а дуже багато «місць прописки».

Підготовка даних, за якими шукає пошукова машина, називається індексуванням. Спеціальні програми - пошукові роботи регулярно обходять Інтернет, викачують документи і обробляють їх. Створюється свого роду зліпок Інтернету, який зберігається на серверах пошуковика і оновлюється при кожному новому обході роботів.

Модуль індексування містить основні види програм :

**Scheduler (Планувальник)** - вибудовує маршрут та встановлює черговість обходу документів. При цьому планувальник враховує важливі для пошукової системи характеристики сайтів, такі як, наприклад, популярність, рейтинг та частота оновлення документів. Після створення маршруту планувальник віддає його до іншого пошукового робота Spider (Павука).

**Spider (Павук)** - програма, що призначена для завантажування сторінок з навколишніх серверів відповідно до заданого переліку адрес. Робот отримує від пошукової системи початковий перелік адрес документів (сторінок), які він має відвідати, скопіювати вміст і надіслати його до подальшого опрацювання до пошукової системи. Spider регулярно обходить документи за заданим маршрутом. Якщо сайт працює і доступний, павук викачує заплановані в маршруті документи. Він визначає тип завантаженого документа (html, pdf, swf і т.п.), кодування та мову, а потім надсилає дані до сховища. Робот передає до навколишніх серверів HTTP-запит і у відповідь отримує текстовий потік, що містить службову інформацію і безпосередньо сам документ:

* + URL сторінки.
  + Дата, коли сторінку завантажено на сервер.
  + Тіло сторінки (HTML-код).

Якщо при обході робот-павук бачить на вже відомих сайтах нові посилання, він передає їх до іншої програми Crawler.

**Crawler (Мандрівний павук)** - програма, яка автоматично проходить по всіх посиланнях, які зазначено на сторінці і здійснює індексацію нових документів, які до того не були занесені до баз пошукової системи.

Результати роботи павуків передаються до програми Indexer.

**Indexer (Індексатор)** - програма, яка аналізує вміст сторінки, що завантажили павуки. Індексатор обробляє документ: очищає від html-розмітки, залишаючи чистий текст, виділяє дані про місце розташування кожного слова, розбирає сторінку на складові частини і аналізує їх, застосовуючи власні лексичні і морфологічні алгоритми. Аналізу піддаються різні елементи сторінки, такі як текст, заголовки, посилання, структурні та стильові особливості, спеціальні службові html-теги тощо.

Індексатор додає документ в розібраному та початковому виді до пошукового індексу (проміжне сховище даних), де він зберігається у сховищі до наступного обходу роботів-павуків. Завдяки цьому користувачі можуть знайти і подивитися документи, навіть якщо сайт тимчасово недоступний. Якщо сайт закрився або документ був видалений чи оновлений, пошуковик зберігає копію в архів або заміняє її на нову.

Отже, модуль індексування дозволяє обходити за посиланнями задану множину ресурсів, завантажувати сторінки, витягувати з отриманих документів посилання на нові сторінки та здійснювати повний аналіз цих документів (рис.2).

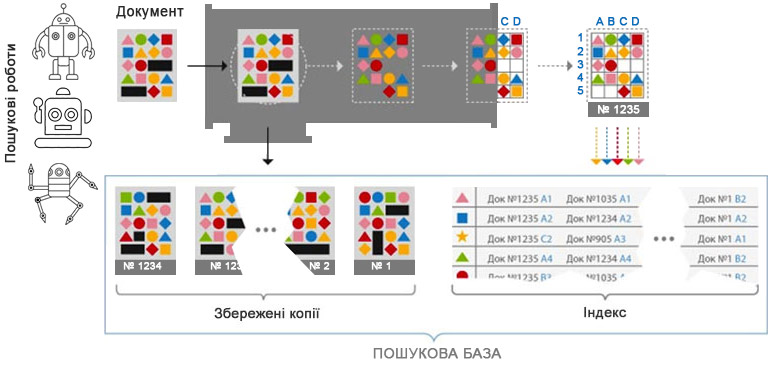


Рис. 2. Функціонування модуля індексування

Пошуковий індекс, дані про тип документів, кодування, мову і збережені копії документів разом складають пошукову базу. Вона оновлюється постійно, але, щоб це оновлення стало доступне користувачам, її потрібно перенести на «базовий пошук».

**Базовий пошук** - сервери, які відповідають користувачам на запити. Туди переноситься не вся пошукова база, а лише її корисна частина - без спаму, дублікатів сайтів (дзеркал) та інших непотрібних документів.

Оновлення пошукової бази зі сховища основного робота потрапляє в пошук «пакетами» - раз у кілька днів. Цей процес створює додаткове навантаження на сервери, тому проводиться вночі, коли до пошуковиків звертаються на порядок менше користувачів. Спочатку нові частини бази поміщаються поруч із такими ж частинами з минулого обходу. Потім вони перевіряються за цілою низкою факторів, щоб оновлення не погіршило якість пошуку. Якщо перевірка пройшла успішно, нова частина бази замінює собою стару.

В пошукових системах існують два типи пошукових роботів - основні і швидкі. Основні роботи індексують Інтернет в цілому, а швидкі відповідають за те, щоб у пошуку можна було знайти найсвіжіші документи, які з'явилися хвилини або навіть секунди тому. В кожного робота є список адрес документів, які потрібно відвідати.

Швидкий робот призначений для пошуку в реальному часі. Його планувальник і павук налаштовані так, щоб знаходити нові документи і вибирати з величезної їх кількості все, що є цікавим. Кожен такий документ швидкий робот відразу обробляє і викладає на базовий пошук. Термінових документів не дуже багато у порівнянні з загальним обсягом Інтернету, тому оновлення бази в реальному часі можна робити і при денних навантаженнях на сервер.

### Обробка запиту

Для розуміння запиту користувача пошукова система здійснює лінгвістичний аналіз запиту. Саме при аналізі запиту система вирішує, за якими словами і словоформами потрібно шукати. Аналізуючи запит користувача, система визначає мову, проводить морфологічний розбір кожного слова, вибирає потрібні для пошуку словоформи і прибирає зайві. На весь аналіз запиту - визначення мови, розбір слів, пошук синонімів - витрачаються долі секунди.

### Морфологічний розбір запиту

Аналіз запиту починається з визначення мови. При визначенні мови пошукова система звертає увагу на регіон користувача і мову інтерфейсу. Наприклад, якщо людина ставить запитання з України і використовує інтерфейс українською мовою, це буде додатковим фактором, щоб вважати запит україномовним і надавати більшість документів українською мовою.

Визначивши мову запиту, відбувається морфологічний розбір. Знання морфології дозволяє знаходити документи, що містять різні форми одних і тих же слів. За словоформою, яка є в запиті, не завжди можна точно сказати, яке слово мав на увазі користувач.

Щоб вибрати для пошуку найбільш ймовірний список форм, система звертається до статистики спільного поєднання слів і граматичних ознак. Для збору статистики пошукові системи використовують власні корпуси, де зібрано величезну кількість текстів.

Для ефективного пошуку визначається, як часто це слово зустрічається з іншими словами запиту - і в питаннях користувачів, і взагалі в текстах. Однокореневі слова і синоніми система бере з відповідних довідників і словників.

Аналізуючи запит, пошукова система виділяє в ньому різні об'єкти - географічні назви, імена людей, назви організацій і т.д. Для виділення стійких фраз і об'єктів теж використовуються різні довідники - наприклад, словник топонімів (географічних назв), словник імен і прізвищ, довідник організацій, словник стійких словосполучень. Отримавши запит, система кожен раз перевіряє за довідниками, чи є в ньому стійкі словосполучення.

Аналізуючи запит, пошукова система завжди перевіряє його на грамотність, оскільки певна частка запитів містять помилки. Це можуть бути орфографічні помилки або набір символів, що утворюється при неправильній розкладці клавіатури. Тому, пошукова система пропонує користувачеві грамотно написаний варіант і перепитує чи шукати за виправленим запитом.

На роботу з помилками і весь лінгвістичний аналіз йдуть частки секунди. За цей час система встигає визначити мову запиту, розібрати кожне слово, знайти синоніми, стійкі поєднання і вирішити, які документи з якими словами потрібно шукати.

### Архітектура відповіді на запитання

Кожен день користувачі задають сотні мільйонів запитів і щоб обробляти такі обсяги даних і робити це швидко, пошукова система використовує тисячі серверів. Сервери об'єднані в кластери і навіть в кластери кластерів. Архітектура пошукової системи влаштована так, що до вже існуючих серверів можна легко додавати нові сервери для нових даних з постійно зростаючого Інтернету.

Всі запити користувачів спочатку потрапляють в систему **метапошуку**. Метапошук обробляє кожен запит в реальному часі - з'ясовує всі необхідні дані про запит (з якого регіону надійшов, до якого класу належить тощо), проводить лінгвістичну обробку. Далі метапошук перевіряє, чи формувалися останнім часом результати пошуку для цього запиту. Результати пошуку за часто вживаними запитаннями деякий час зберігаються в пам'яті метапошуку, а не формуються щораз заново. І якщо новий запит виявився популярним, метапошук покаже користувачеві заздалегідь збережені результати (рис.3).

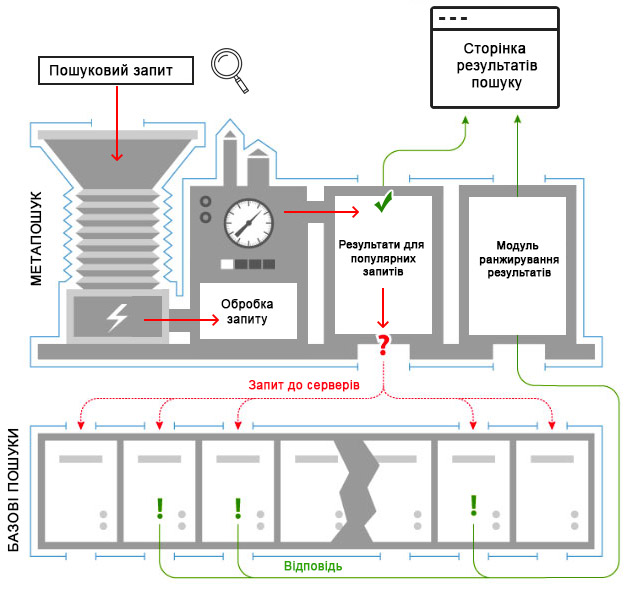


Рис. 3. Оброблення пошукового запиту

Якщо відповіді в пам'яті немає, то метапошук передає запит до серверів іншої комп'ютерної системи - «базового пошуку». На базовому пошуку зберігається зліпок Інтернету - пошукова база. Вона розділена на частини, які зберігаються на різних серверах - шукати відповідь одночасно у кількох частинах бази даних швидше, ніж у всій базі цілком. Крім того, в кожного сервера є кілька копій. Це дозволяє розподіляти навантаження і не втрачати дані - якщо один із серверів не зможе своєчасно відповісти, інформація знайдеться на дублюючих серверах. З тисяч серверів базового пошуку метапошук вибирає найменш завантажені - таким чином, щоб разом вони містили цілу пошукову базу.

Кожен з серверів повертає до метапошуку перелік документів, в яких є слова із запиту. Там вони об'єднуються, ранжируються за допомогою спеціальних технологій і потрапляють на сторінку *результатів пошуку*. Завдяки такій організації пошук може відповідати користувачеві за частки секунди.

### Оновлення інформації в базі пошуковій системі

Частота відвідування сайтів встановлюється логікою пошукових систем з індексації документів (сторінок):

* Після знаходження і індексації нової сторінки, робот відвідує її наступного дня.
* Після порівняння вмісту сторінки з тим, що було вчора і не знайшовши відмінностей робот пошукової системи зайде на неї за три дні.
* Якщо і цього разу на даній сторінці нічого не змінитися, то робот навідується за тиждень і т.д.

З часом, частота відвідування пошукового робота до сторінки наближається до частоти її оновлення. Час повторного заходу робота пошукових систем може вимірюватися для різних сторінок як в хвилинах, так і в роках. Розумні пошукові системи встановлюють індивідуальний графік відвідування для різних сторінок сайтів.

## Алгоритми роботи пошукових систем

#### Алгоритм прямого пошуку

Це метод простого перебору всіх сторінок (документів), що зберігаються в базі даних пошукової системи. Цей метод дозволяє напевно знайти потрібну інформацію не пропустивши нічого важливого, але він є не доречним для роботи з великими обсягами даних, оскільки пошук буде займати багато часу.

#### Алгоритм зворотного пошуку

Для ефективного пошуку у великих обсягах даних всі потужні пошукові системи використовують алгоритм зворотних індексів. За цим алгоритмом пошукові системи перетворюють документи в текстові файли, що містять перелік всіх наявних в документі слів. Слова в таких списках (індекс-файлах) розташовуються в алфавітному порядку і поряд з кожним словом зазначено його координати в документі та параметри, що визначають його статус в документі.

Для формування сторінки видачі результатів пошуку пошукові системи шукають інформацію саме в зворотних індексах оброблених документів. Прямі індекси (оригінальний текст документів) пошуковики теж використовують, наприклад для складання фрагментів опису знайденого документу.

### Математична модель пошуку

Основним завданням математичної моделі будь-якої пошукової системи є пошук документів у базі зворотних індексів відповідно до даного пошукового запиту і сортування їх у порядку зменшення релевантності до пошукового запиту. Використання простої логічної математичної моделі, яка знаходить документ, якщо в ньому зустрічається шукана фраза, не підходить, в силу величезної кількості таких документів.

Математичні моделі пошукових систем відносяться до класу векторних моделей, де використовується поняття ваги документа по відношенню до запиту користувача. Вага документа за заданим пошуковим запитом обчислюється за двома основними параметрами:

* **TF (Term Frequency)**. Частота, з якою зустрічається дане слово в аналізованому документі.
* **IDF (Inverse Document Frequency).** Частота, наскільки рідко це слово зустрічається у всіх інших документах колекції пошукової системи.

Під колекцією пошукової системи розуміють всю сукупність документів, які відомі для пошукової системи. Перемноживши ці два параметри, отримується вага документа за заданим пошуковим запитом.

Природно, що різні пошукові системи, окрім параметрів TF і IDF, використовують багато інших коефіцієнтів для обчислення ваги документа за заданим пошуковим запитом, але суть залишається незмінною: вага документа буде тим більше, чим частіше слово з пошукового запиту зустрічається в документі (до певних меж, після яких документ може бути визнано спамом) і чим рідше зустрічається це слово у всіх інших документах, проіндексованих пошуковою системою.

### Алгоритм ранжирування

Алгоритм ранжирування - це система математичних формул для оцінки певних факторів, на базі якої пошукова система привласнює сайту (сторінці) певний рейтинг.

Неможливо написати для пошукової системи таку програму, в якій передбачено кожен запит, і для кожного запиту передбачити кращу відповідь. Пошукова система повинна вміти приймати рішення самостійно і сама вибирати з мільйонів документів той, який найкраще відповідає користувачеві.

Процес впорядкування результатів пошуку відповідно до запиту користувача - називається **ранжируванням**. Властивості запиту і сторінки, які важливі для ранжирування і які можна виміряти числами, називаються факторами ранжирування. Для точного пошуку важливо враховувати багато різних чинників.

* Статистичні фактори.
* Динамічні фактори.
* Запитальні фактори.

Основними факторами, що характеризують документ:

1. Вік сайту.
2. URL сторінки (доменна адреса сайту).
3. Мова сайту.
4. Число сторінок сайту.
5. Популярність тематики сайту.
6. Загальний обсяг (вага) сайту і кожної окремо взятої сторінки сайту.
7. Обсяг текстової інформації на сайті, а також на кожній сторінці сайту.
8. Загальна кількість ключових слів на сайті і на окремо взятій сторінці.
9. Співвідношення загального числа слів на сайті/сторінці до числа ключових слів на сайті/сторінці.
10. Індекс цитування на інших сайтах.
11. Кількість запитів за конкретним ключовим словом за заданий період часу.
12. Періодичність оновлення інформації на сайті, дата останнього оновлення сторінок сайту.
13. Відстань розташування ключових слів від початку сторінки.
14. Наявність і аналіз мета-тегів.
15. Географічне розташування сайту.
16. Відповідність ключових слів сайту до того розділу каталогу пошукової системи, в якому зареєстрований цей сайт.
17. Загальна кількість гіперпосилань сайту, число внутрішніх посилань, число зовнішніх посилань сайту.
18. Інші важливі фактори.

Пошук відповіді ведеться одночасно на тисячах серверів. Кожен сервер шукає у своїй частині індексу і формує список найкращих результатів. В нього гарантовано потрапляють всі самі релевантні до запиту сторінки.

Далі з цих списків складається один загальний, і сторінки, що потрапили туди, впорядковуються за формулою ранжирування з врахуванням всіх факторів та їх комбінацій. Таким чином, нагорі пошукової видачі виявляються самі релевантні сайти - і користувач майже миттєво отримує відповідь на своє питання.

### Машинне навчання

Формування видачі пошукових систем з певного запиту здійснюється автоматично за математичною моделлю без участі людини. Проте, жодна модель не може працювати ідеально, особливо на початку, тому, за роботою математичної моделі потрібно здійснювати контроль.

Цей контроль здійснюють фахівці - **асесори**, які переглядають видачу пошукових систем і оцінюють якість роботи моделі. Вони беруть пошукові запити і документи, які пошук знаходить за цими запитами, і оцінюють, наскільки добре знайдений документ відповідає на поставлене запитання. Із запитів та хороших відповідей складається навчальна вибірка. Вона повинна містити самі різні запити, причому в тих же пропорціях, в яких їх задають користувачі.

На навчальній виборці пошукова система встановлює залежність між сторінками, які асесори вважають релевантними до запитів, і властивостями цих сторінок. Після цього вона може підібрати оптимальну формулу ранжирування - яка показує релевантні запиту сайти серед перших результатів пошуку.

Всі зауваження асесорів враховуються розробниками, які відповідають за налаштування моделі пошуку. У формули моделі вносяться зміни або доповнення, в результаті чого якість роботи пошукової системи підвищується. Асесори виконують роль своєрідного зворотного зв'язку між розробниками пошукової системи та її користувачами, який необхідний для покращення якості роботи пошуковиків.

### Оцінка якості роботи математичної моделі пошукової системи

Основними критеріями для оцінювання є:

1. **Точність видачі пошукової системи** - відсоток релевантних документів, відповідних до пошукового запиту. Чим точніше пошук, тим швидше користувач знайде документи, що відповідають запиту і тим менше різного роду «сміття» серед них буде зустрічатися.
2. **Повнота пошукової видачі** - процентне відношення релевантних документів в пошуковій видачі до загальної кількості релевантних документів, наявних у всій колекції пошукової системи. Чим повніше пошук, тим більше ймовірність, що користувач знайде потрібний документ.
3. **Актуальність пошукової видачі** - характеризується часом з моменту публікації документів в Інтернет, до їх занесення до бази пошукової системи. Актуальність видачі пошукової системи залежить від частоти сканування роботами документів і поновлення інформації в базах. Також, це ступінь відповідності реального документа в Інтернеті, до того що про нього написано в пошуковій видачі. Наприклад, документ може бути видаленим або сильно зміненим, але в пошуковій видачі за заданим запитом він буде присутнім, незважаючи на його фізичну відсутність за вказаною адресою або ж на його поточну невідповідність до даного пошукового запиту.
4. Ш**видкість пошуку** тісно пов'язана з стійкістю системи до навантажень. В робочі години до пошукових систем може надходити сотні запитів в секунду. Така завантаженість вимагає скорочення часу обробки окремого запиту. Пошукова машина повинна обробити запит максимально оперативно, щоб не гальмувати обчислення наступних запитів.
5. **Наочність представлення результатів.** За запитом користувачеві видається багато ресурсів. Орієнтуватися в результатах пошуку допомагають різні елементи у сформованих відповідях.

## Результати пошуку

Сторінка результатів пошуку (SERP Search Engine Results Page) щодня формується десятками мільйонів разів. Результатами пошуку є невеликі блоки інформації - **сніпети**, що містять посилання на знайдені документи з короткою інформацією про них (рис. 4). Інформація підбирається так, щоб допомогти користувачеві зрозуміти - яка з відповідей підходить йому найкраще.

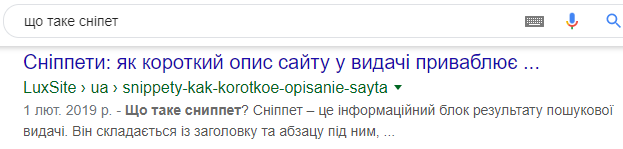


Рис. 4. Сніпет до знайденого документа

Для заголовку сніпету найчастіше використовується заголовок самого документа. Якщо він занадто довгий, обирається фрагмент, який найбільше підходить за змістом. Для формування опису, програма вибирає всі фрагменти тексту документа зі словами із запиту. Кожен з таких фрагментів поділяється ще на кілька частин - наприклад, зі словами із запиту на початку, в кінці і в середині, так можна повніше описати зміст сторінки (рис.3). Потім програма порівнює їх все між собою і вибирає кращі - вони і потрапляють в сніпет. Для різних пошукових запитів один документ буде мати різні сніпети (рис. 5).

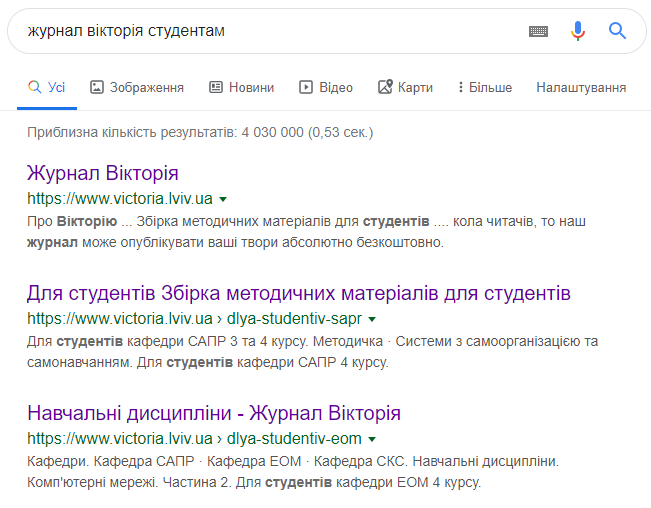


Рис. 5. Фрагмент сторінки видачі Google на запит «журнал вікторія студентам»

Результат пошуку оформляється так, щоб користувачеві було легше його сприймати. Заголовки традиційно виділено синім кольором і підкреслено. Впізнати знайомий ресурс допомагає фавіконка - невеликий фірмовий значок сайту, що може знаходитися поруч.

Для зручності слова із запиту в результатах пошуку виділено жирним шрифтом, а також деякі слова, яких немає в запиті. Це відбувається при відповіді на загальні, багатозначні запити. Спеціальна програма стежить за тим, як користувачі уточнюють свої запити, і обчислює значимість таких уточнень, що використовується при формуванні сніпету.

Важливим є, щоб користувачі могли швидко знайти відповідь - іноді навіть відразу на сторінці результатів пошуку. Для різних відповідей потрібна різна додаткова інформація. Наприклад, якщо людина задає в запиті назву організації, можливо, їй потрібно довідатися, де вона знаходиться або як з нею зв'язатися. Щоб не довелося витрачати час на пошуки сторінки з контактами на сайті організації, пошукова система додає телефон і фізичну адресу з посиланням на карту до сніпету (рис.6).

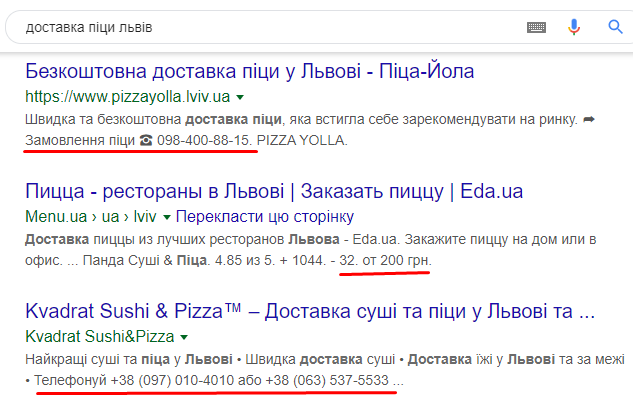


Рис. 6. Сніпети до пошукового запиту «Доставка піци Львів»

Якщо пошуковій системі відомо про структуру сайту, вона показує її користувачеві (рис.7). Над текстом сніпету сайту з'являються посилання на його найбільш відвідувані сторінки - щоб за бажанням користувач міг перейти в потрібний розділ, витрачаючи менше кліків і трафіку.

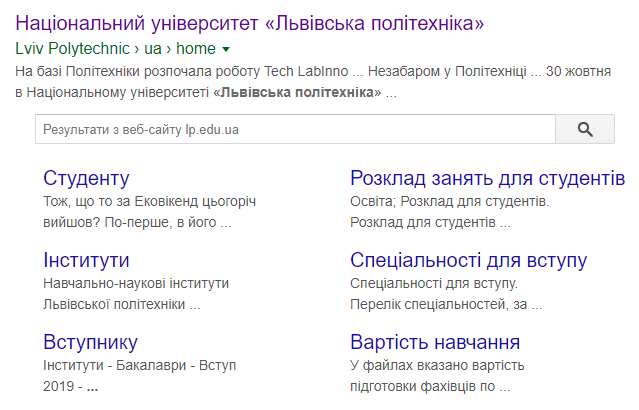


Рис. 7. Сніпет до запиту «Львівська політехніка»

Для деяких предметних областей створюються спеціальні сніпети (рис.8). Наприклад, для сторінок з описами товарів або для сайтів готелів, ресторанів, кінотеатрів. Основна інформація, що з'являється в сніпетах - ціна товару, «зірковість» готелю, кухня ресторану, кількість залів кінотеатру. Завдяки таким спеціальним сніпетам користувач економить час і трафік, а організація отримує відвідувача сайту, зацікавленого саме в її послугах.

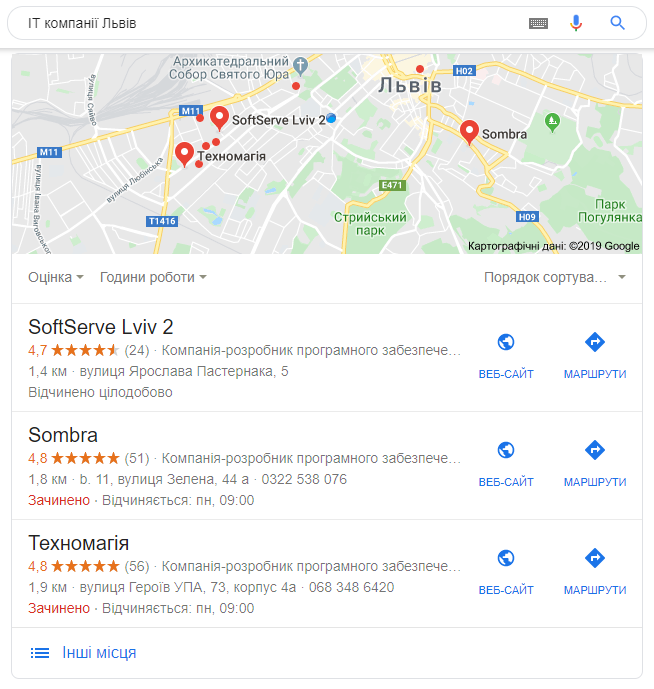


Рис. 8. Інформація до запиту «ІТ компанії Львів»

### Пошукові підказки

Коли користувач починає вводити запит у пошуковому рядку, пошуковик показує кілька найпопулярніших запитів, що починаються на вже введені літери, - це пошукові підказки. Пошукові підказки допомагають заощадити час - можна не друкувати запит цілком. Пошуковик розуміє, які показати підказки, навіть якщо користувач забув змінити розкладку клавіатури або допустив граматичні помилки.

Список, з якого беруться пошукові підказки, формується після фільтрації всього потоку запитів користувачів. Запити проходять множину фільтрів, кожний з яких відсіває запити за кількома умовами. Наприклад, прибирає дуже рідкісні запити або запити, що містять ненормативну лексику. Разом з фільтрацією запитів в них виправляються помилки. У підсумку залишаються мільйони запитів.

Серед решти запитів шукаються схожі, щоб об'єднати їх в групи. До угруповання запитів підходять дуже акуратно - запити, які здаються схожими для машини, не завжди схожі для людини. Щоб два запити об'єдналися в одну підказку, вони повинні не лише мало відрізнятися за написанням, але і вести на однакові результати пошуку. Для кожного регіону є свій список пошукових підказок, що заснований на запитах з цього регіону.

Оскільки постійно з'являються нові популярні запити, список пошукових підказок регулярно оновлюється - не рідше, ніж раз на день. Запити, що втратили актуальність, видаляються.

Для запитів про події та інциденти, що сталися щойно і зацікавили велику кількість користувачів, пошукова система має «швидкий» список підказок. Він оновлюється раз на півгодини. Запити для цього списку відбираються за складною формулою, яка враховує, наскільки різко зріс пошуковий інтерес до теми, як багато з'явилося новинних повідомлень і постів у блогах.

### Персональний пошук

На одне запитання може бути кілька правильних відповідей - залежно від потреб користувача. Персональний пошук вміє враховувати особисті інтереси та переваги користувача і вибирати найбільш доречну для нього відповідь.

В основі персонального пошуку лежить аналіз пошукової поведінки - наприклад, які запити задає людина, на які сайти переходить, якою мовою спілкується з пошуковою системою. Аналізуючи питання і переходи користувача, пошукова система робить висновок про те, що йому зараз важливо і цікаво. Ці дані враховуються як при виборі підказок, так і при пошуку персональної відповіді на запитання.

Для підбору релевантних відповідей використовується спеціальна формула ранжирування, яка підлаштовується під кожного користувача з врахуванням його інтересів і мовних уподобань. За допомогою персональної формули визначається, наскільки кожен із знайдених за запитом документів підходить для конкретної людини. Всі документи отримують свою оцінку і розташовуються в результатах пошуку у відповідному порядку. Два різних користувача з різною пошукової історією на однаковий запит отримують різні відповіді

Формула ранжирування враховує як постійні (довгострокові і середньострокові), так і сьогочасні інтереси людини. Всі вони різною мірою впливають на якість відповіді на питання. Так, постійні інтереси відображають мовні переваги користувача, його звички і близькі для нього теми. А одноразові випадкові - показують, що важливо людині просто зараз. Наприклад, користувач, який зазвичай запитує про кіно і музику, може несподівано зацікавитися комп'ютерними іграми, а потім надовго про них забути. Таких інтересів, які швидко виникають і так само швидко сходять нанівець, більше половини.

Щоб бути в курсі постійних інтересів користувача, пошукова система аналізує його пошукову історію за останні два місяці і за тиждень. А щоб стежити за миттєвими інтересами - обробляє дані пошукових сесій в режимі реального часу. Це дозволяє зрозуміти, що важливо людині прямо зараз, і підлаштуватися під його інтереси.

Щоб миттєво реагувати на зміни в пошуковій поведінці людей, застосовується система обробки даних в режимі реального часу. Завдяки цій системі запити та кліки, які користувач зробив щойно, починають впливати на результати пошуку вже за кілька секунд.

Вибираючи для людини підказки, пошукова система намагається вгадати його наступні запити. При цьому пошукова система спирається на питання людей з схожою пошуковою поведінкою.

Практика показує, що приблизно чверть своїх запитів користувачі задають повторно і часто клацають по одних і тих же результатах. Таким чином вони переходять до улюблених сайтів і документів. Щоб полегшити їм це завдання, пошукова система додає до підказок нещодавні запити користувача та його улюблені сайти. Відбираючи персональні підказки також враховується, які запити характерні для цієї пошукової сесії.

За замовченням персональний пошук працює для всіх користувачів пошуку. І чим більше запитів задає користувач, тим краще пошук його розуміє. Втім, кожен користувач може сам вирішувати, чи потрібні йому відповіді з врахуванням інтересів. Персональний пошук можна увімкнути або вимкнути в налаштуваннях сервісу.

## Висновки

Архітектурно сучасні пошукові системи представляють складні комплекси з багатьох серверів. В такій системі основне навантаження покладається на пошук, оскільки щосекунди надходять десятки і сотні запитів. Для того, щоб впоратися з цією проблемою, індекс розбивають на частини і розкладають по десяткам, сотням і навіть тисячам комп'ютерів.

Для пришвидшення пошуку застосовують спеціальні техніки. Наприклад, ешелонування полягає в поділі індексу на свідомо більш релевантну і менш релевантну частини. Пошук спочатку виконується в першій частині, а потім, якщо нічого не знайдено, або знайдено мало, пошукова система звертається до другої частини індексу. Техніка відсікання (скорочення) полягає в тому, щоб динамічно припиняти обробку запиту після накопичення достатньої кількості релевантної інформації.

Великої уваги приділяють організації безперебійної роботи багато комп’ютерних комплексів, безшовне оновлення індексу, стійкість до збоїв і затримок з відповідями від окремих компонентів. Для спілкування між пошуковими серверами і серверами, які збирають відгуки і формують сторінку видачі розробляються спеціальні протоколи.

Вирішальне значення набуває продумування архітектури всього комплексу з самого початку, оскільки будь-які зміни, наприклад, додавання незвичайного фактору при ранжируванні або складного джерела даних стає виключно складною процедурою.

## Контрольні питання

1. Які пошукові системи на сьогодні є самими популярними в світі, в Україні?
2. Які основні програми-роботи використовують для збору та підготовки даних? Які функції вони виконують?
3. Яка різниця між основним та швидким роботом?
4. Які дії відбуваються для обробки запиту користувача? Для чого здійснюється лінгвістичний аналіз та морфологічний розбір запиту?
5. Яким чином формується відповідь на запит користувача?
6. Яким чином працює алгоритм зворотних індексів, в чому його особливість та ефективність?
7. На які фактори орієнтується процес ранжирування знайдених відповідей? Який фактор є самим визначальним для відображення ресурсу в топ -видачі?
8. Назвіть основні критерії в оцінці якості роботи математичної моделі пошукових систем.
9. Які інструменти, окрім запиту можуть покращити пошук конкретного зображення?
10. Для чого здійснюється зворотний пошук зображень?

## Використані джерела

1. Документація Google Пошук - <https://support.google.com/>
2. Як працює Google Пошук - <https://habr.com/ru/company/ua-hosting/blog/277819/>
3. Як працюють пошукові системи - <https://habr.com/ru/company/yandex/blog/464375/>
4. Історія розвитку пошукових алгоритмів Google - <https://blog.promopult.ru/seo/istoriya-razvitiya-poiskovyx-algoritmov-google.html>
5. Енциклопедія пошукового просування - <http://www.sembook.ru/book/poiskovye_sistemy/kratkoe_vvedenie_v_poiskovye_sistemy/>
6. Пошукові системи світу. Статистика, рейтинги - <https://marketer.ua/search-engine-stat-2018/>
7. Альтернативні пошукові системи - <https://ahrefs.com/blog/ru/alternative-search-engines/>
8. Відео курс машинного навчання від розробників Google - <https://proglib.io/p/google-ml-recipes/>
9. Як фільтрувати результати пошуку - <https://support.google.com/websearch/answer/142143?hl=ru>
10. Важливі SEO-тренди на 2020 рік - <https://webrost.ru/blog/10-vazhnykh-seo-trendov-na-2020-god/>