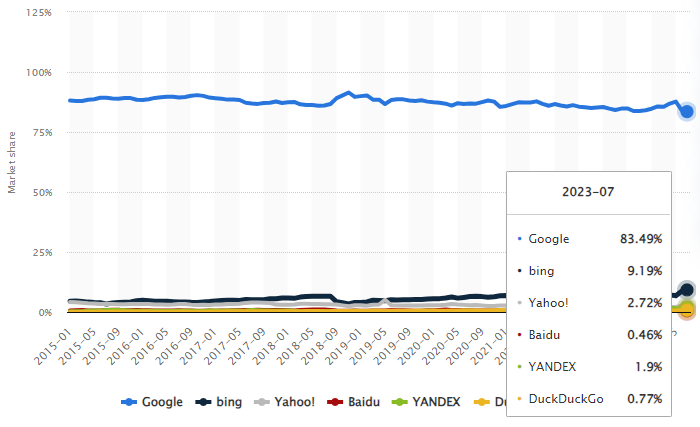
# 5.1. Пошукові системи

## Загальні відомості про пошукові системи

За оцінками аналітичної компанії International Data Corporation (https://www.idc.com/), обсяг інформації в Інтернеті у 2023 році становить близько 500 екзабайтів (500 мільярдів гігабайтів). Це означає, що щороку в Інтернеті створюється близько 2,5 ексабайтів інформації. Пошукові системи значно покращують і прискорюють процес пошуку інформації в Інтернеті, забезпечуючи користувачам широкий і зручний доступ до величезного обсягу контенту.

**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.

За даними статистики ([statista.com](https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/)) на жовтень 2023 року перші місця на світовому ринку займають наступні пошукові системи (рис.1).



###### Рис.1. Популярність пошукових систем у світі

Провідні пошукові системи мають власні бази даних, де зберігається інформація про веб-сайти, включаючи їх URL-адреси, контент, метадані та інші дані. Пошукові системи використовують ці бази даних для індексування веб-сайтів та відображення результатів пошуку користувачам. Менші пошукові системи зазвичай орендують бази даних у великих компаній.

Пошукові системи, які мають власні бази даних:

* Google (<https://www.google.com.ua/>) американська публічна транснаціональна корпорація. Найпопулярніший пошуковий сервіс станом на 2023 рік.
* Bing (<https://www.bing.com/>) пошукова система, що належить компанії Microsoft.
* Yahoo! (<https://www.yahoo.com/>) пошуковик, який належить компанії Yahoo!.
* Baidu (<https://www.baidu.com/>) лідер серед китайських пошукових систем.
* Yandex (<https://www.yandex.com/>) російська ІТ-компанія, що володіє однойменною пошуковою системою та інтернет-порталом.
* DuckDuckGo (<https://duckduckgo.com/>) пошукова система з відкритим початковим кодом, що використовує інформацію з багатьох джерел для надання точних та різноманітних результатів. Застосовує неперсоналізовий пошук

Функціонування пошукових систем складається з двох частин (рис.2).

1. Збір, обробка та збереження інформації. Пошукові системи обходять різноманітні ресурси по всьому Інтернету, викачують, обробляють і зберігають інформацію в базах на своїх серверах.
2. Обробка пошукового запиту і формування сторінки видачі результатів.

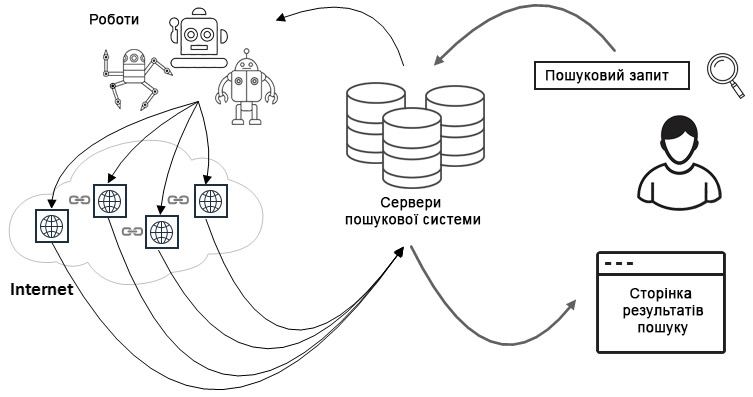


Рис.2. Загальна схема функціонування пошукової системи

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.

Для надання допомоги користувачам у пошуку потрібної інформації пошукові системи виконують наступні етапи.

* **Індексація інформації.** Пошукові системи сканують мільярди веб-сторінок, збираючи інформацію з кожної з них. Ця інформація обробляється (індексується) і зберігається в пошуковому індексі. Пошуковий індекс це база даних, що містить інформацію про сайти – зліпок інтернету, що дозволяє пошуковим системам швидко і ефективно витягувати релевантні відповіді.
* **Пошук за ключовими словами.** Користувачі вводять запити, що складаються з ключових слів або фраз в пошуковому рядку. Пошукова система аналізує ці запити і шукає відповідні ключові слова в пошуковому індексі.
* **Алгоритми ранжування.** Пошукові системи використовують складні алгоритми ранжування, щоб визначити релевантність кожної сторінки до запиту користувача. Алгоритми враховують багато факторів, таких як ключові слова, структуру сторінки, авторитетність сайту та інші фактори.
* **Фільтрація результатів.** Після виконання пошуку у пошуковому індексі система надає результати, відсортовані за релевантністю. Для уточнення результатів пошуку можуть бути використані різні фільтри та параметри, такі як дані, тип контенту, країни та інші.
* **Персоналізований підхід.** Багато пошукових систем враховують попередню активність користувача та надають персоналізовані результати. Це адаптація результатів пошуку до переваг користувача, актуальні оголошення, рекомендації ТОЩО.
* **Контекстні підказки.** Щоб допомогти користувачам уточнити свої запити та отримати більш точні результати пошукові системи надають контекстні підказки та виправлення запитів,.
* **Універсальний доступ до різних типів контенту.** Пошукові системи інтегрують різні типи контенту в результати пошуку, такі як текст, зображення, відео, новини та різноманітні сервіси. Це дозволяє користувачам передавати інформацію в різних форматах.
* **Швидкий доступ до інформації.** Завдяки вдосконаленню алгоритмів пошуку та ранжування користувачі можуть миттєво отримувати доступ до потрібної інформації, мінімізуючи час і зусилля, які раніше вимагалися для пошуку вручну.

## 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.** Важлива ланка у пошуковій системі, що забезпечує ефективний та структурований пошук за величезним обсягом інформації в інтернеті. Для сканування і обробки сторінок використовуються 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. Оброблені сторінки з модуля індексування надсилаються на зберігання до **пошукового індексу**. 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

#### Сканування інформації

Сканування сайтів у пошукових системах є важливим кроком у процесі індексації та забезпечує обхід та збирання інформації з веб-сторінок для подальшого використання при формуванні результатів пошуку. Для сканування сайтів пошукові системи використовують програмні інструменти - роботи для обходу та збору інформації з веб-сторінок сайтів.

#### Визначення початкової точки.

За це відповідає робот Scheduler (Планувальник). Планувальники використовують різні алгоритми та стратегії для визначення пріоритетів сканування сторінок. Планувальник бере до уваги такі фактори:

* Розмір сайту. Великі сайти скануються рідше за невеликі веб-сайти, оскільки сканування великих сайтів потребує більше часу та ресурсів.
* Актуальність сайту. Сайти, які часто оновлюються, скануються частіше, ніж сайти, де інформація на сторінках оновлюються рідко.
* Якість сайту. Сайти з якісним контентом та чіткою структурою скануються частіше.

#### Запуск роботів

Пошукова система запускає роботів, які за розкладом Планувальника автоматично переходять на сторінки сайтів. Роботи починають з початкової точки і просуваються до інших сторінок, створюючи граф зв'язків між ними.

Для сканування та індексування інформації застосовуються різні типи роботів.

* Загальні павуки (Spiders, Crawlers) є основним типом роботів, що автоматично обходять веб-сторінки, дотримуючись посилань і збирають інформацію.
* Мобільні павуки. Зі збільшенням використання мобільних пристроїв пошукові системи також використовують мобільних павуків, які сканують та індексують мобільно-оптимізовані версії веб-сторінок. Вони імітують поведінку мобільних браузерів та перевіряють, наскільки добре сторінки оптимізовано для мобільних пристроїв.
* Глибокі веб-павуки (Deep Crawlers) використовуються для сканування та індексування вмісту, що знаходиться за захистом авторизацією. Вони можуть збирати інформацію, яка зазвичай недоступна для загальних павуків.
* Павуки для зображень спеціалізуються на скануванні та індексуванні зображень на сторінках. Вони аналізують атрибути зображень, такі як назва файлу, заголовок title, альтернативний текст alt та текстове оточення навколо зображення.
* Павуки для відео сканують та індексують відео-контент на сторінках. Аналізуються метадані відео, такі як заголовки, описи, ключові слова та прев'ю, щоб визначити вміст відео та зв'язати його з відповідними запитами пошуку.
* Веб-павуки для новин спеціалізуються на скануванні та індексуванні новинних веб-сайтів та статей. Вони можуть обробляти різні метадані, такі як заголовки, автори, дата публікації та категорії новин.

Це лише деякі з типів роботів, які використовуються в пошукових системах. Конкретні пошукові системи можуть використовувати власні спеціалізовані роботи в залежності від своїх потреб і особливостей індексації та пошуку контенту.

#### Запит, завантаження сторінок та витягування інформації

Роботи Spiders надсилають запити HTTP до серверів, на яких зберігаються веб-сторінки, і отримують відповіді у вигляді HTML-коду сторінок. Павук визначає та фільтрує однаковий контент на сторінках, щоб запобігти небажаному дублювання в індексі пошукової системи. Враховується файл robots.txt на сайті, який містить інструкції для пошукових роботів щодо доступу до певних сторінок або каталогів. Також враховується мета-тег noindex, який може бути встановлений у HTML-коді сторінки.

Для подальшої обробки, отримана HTML-сторінка надсилається до робота Індексатора.

#### Слідування за посиланнями.

Якщо на сторінці є посилання, то їх дослідженням займаються роботи Crawlers (Краулери) слідують за посиланнями, що знайдені на сторінках, щоб перейти на інші сторінки сайту або на зовнішні сайти. Це дозволяє краулерам обійти більше сторінок та розширити охоплення сканування.

#### Обробка даних.

Зібрана інформація обробляється роботом Indexer (Індексатор), що використовує такі методи для індексування інформації:

* Створення індексу сторінок, що містить інформацію про кожну сторінку, включаючи її URL-адресу, контент, метадані та посилання на інші сторінки.
* Створення індексу ключових слів, що містить інформацію про ключові слова, які зустрічаються на сторінках.
* Створення індексу посилань, що містить інформацію про посилання між сторінками.

Інформація у пошуковому індексі зберігається у різних формах, залежно від типу інформації.

**Інформація про сторінки** зберігається у вигляді документів. Документ є набір даних, який описує сторінку і містить такі дані:

* URL-адреса сторінки - унікальна адреса сторінки в Інтернеті.
* Контент сторінки - текст, зображення, відео та інші дані, які містяться на сторінці.
* Метадані сторінки - додаткові дані про сторінку, такі як заголовки сторінок та ключові слова.
* Посилання на інші сторінки, які трапляються на сторінці.

**Інформація про ключові слова** зберігається у вигляді індексу ключових слів. Індекс ключових слів є таблицею, яка містить інформацію про ключові слова, які зустрічаються на сторінках і містить такі дані:

* Ключове слово - слово чи фраза, що зустрічаються на сторінках.
* Кількість разів вживання ключового слова на сторінці.
* Вага ключового слова, яка визначає, наскільки релевантною є сторінка для пошукового запиту.

**Інформація про посилання** зберігається у вигляді індексу посилань. Індекс посилань є таблицею, яка містить інформацію про посилання між сторінками і містить такі дані:

* URL-адреса сторінки, на яку вказує посилання.
* URL-адреса сторінки, з якою вказує посилання.
* Тип посилання - внутрішнє або зовнішнє.

Після обробки інформація про сторінки надсилається до пошукового індексу, який вже використовується пошуковою системою для пошуку релевантних результатів відповідно до запитів користувачів.

#### Регулярне оновлення індексу.

Процес сканування та індексації відбувається з певною періодичністю, щоб виявити нові сторінки, оновити існуючі та видалити видалені або недоступні сторінки. Це дозволяє пошуковим системам відстежувати зміни Інтернету та підтримувати актуальність своїх результатів пошуку. Пошукові системи можуть встановити обмеження на частоту та обсяг сканування, щоб не навантажувати веб-сервери та забезпечити справедливий розподіл ресурсів.

#### Фактори, які можуть вплинути на частоту оновлення пошукового індексу:

* **Розмір пошукової системи.** Великі пошукові системи, такі як Google, обробляють більше інформації, ніж маленькі пошукові системи. Тому їм необхідно оновлювати свої пошукові бази частіше, щоб забезпечити користувачам доступ до найактуальнішої інформації.
* **Кількість нової інформації.** Якщо в Інтернеті публікується багато нової інформації, пошукові системи повинні оновлювати пошукові бази частіше, щоб включити цю інформацію в індекс.
* **Алгоритми індексування.** Пошукові системи постійно вдосконалюють свої алгоритми індексування. Це може призвести до збільшення або зменшення частоти оновлення пошукової бази.

Пошукові системи постійно працюють над покращенням процесу оновлення пошукової бази. Вони прагнуть забезпечити користувачам швидкий та ефективний доступ до найактуальнішої інформації.

### Зберігання інформації в пошуковому індексі

Пошукові системи використовують різні способи зберігання інформації в пошуковому індексі, які можуть відрізнятися залежно від конкретної системи та її архітектури. Однак, загальними методами, які часто використовуються для зберігання інформації в пошукових індексах, є:

* **Інвертований Індекс.** Це один із основних методів індексації текстових даних. Для кожного унікального терміну (слова) у колекції документів створюється список документів, в яких цей термін зустрічається із зазначенням позицій у кожному документі. Таким чином, інвертований індекс забезпечує швидкий доступ до документів, що містять певні терміни.
* **Хешування.** Хешування можна використовувати для швидкого доступу до даних. Хеш-функції можуть перетворювати унікальні ідентифікатори документів або термінів на хешовані значення, які потім використовуються як індекси для швидкого пошуку.
* **Кластеризація та Сегментація.** Індекс може бути поділений на сегменти або кластери, що спрощує керування та оновлення інформації. Це особливо корисно для великих обсягів даних.
* **Компресія даних.** Для економії місця та прискорення доступу дані в індексі можуть бути стиснуті. Різні алгоритми стиснення можуть використовуватись залежно від характеру даних.
* **Кешування.** Для прискорення доступу до даних, що часто використовуються в індексі, можуть використовуватися закешовані сторінки з результатами до популярних запитів.
* **Багатомірні Індекси.** Для пошуку за кількома параметрами (наприклад, за датою та ключовими словами) може використовуватися багатовимірний індекс.
* **Дублювання і розподіл.** При роботі з великими обсягами даних використовуються методи реплікації (дублювання даних) та шарингу (розподіл даних по різних вузлах) для забезпечення масштабованості.

Ці методи можуть комбінуватися в різних пошукових системах залежно від їх вимог до продуктивності, масштабованості та ефективності.

### Архітектура пошукового індексу

Пошуковий індекс є основною складовою пошукової системи та забезпечує ефективне зберігання та організацію цих даних для швидкого доступу. Там зберігаються оброблена інформація, зв'язки між словами або токенами та відповідними сторінками.

Основні функції пошукового індексу

* **Пошук та витягування даних** на основі запитів користувачів. Виконуються операції пошуку за індексами та повертаються відповідні результати, що задовольняють заданим критеріям.
* **Оновлення та синхронізація даних** під час виявлення нових або змінених сторінок, щоб забезпечити свіжі результати пошуку.
* **Резервне копіювання та відновлення даних** для забезпечення безпеки та можливості відновлення у разі збоїв або втрати даних.
* **Керування пам'яттю та оптимізація** доступу до даних для підвищення продуктивності та швидкодії системи. Застосовуються різні техніки оптимізації, такі як кешування, попереднє завантаження даних та індексування.

Це лише загальні функції та модулі пошукового індексу. Різні пошукові системи можуть мати свої власні особливості та додаткові модулі, а також застосовувати різні алгоритми та методи для забезпечення ефективного пошуку та доступу до даних.

## Обробка запиту

Обробка запиту користувача проходить кілька етапів для отримання релевантних результатів.

* **Введення запиту.** Користувач вводить запит, який містить ключові слова або фрази, що описують те, що він шукає.
* **Морфологічний розбір запиту.** Для розуміння запиту користувача пошукова система здійснює попередню обробку запиту, щоб вирішити, за якими словами і словоформами потрібно шукати. Визначається мова запиту, проводяться процеси нормалізації та попередньої обробки.
* **Побудова запиту.** На цьому етапі запит користувача перетворюється на структуровану форму, яка буде використовуватися для пошуку. Запит розділяється на окремі слова або токени, і використовуються різні алгоритми для його аналізу та інтерпретації.
* **Пошук та витягування.** Використовуючи оброблений запит виконується пошук в пошуковому індексі. Шукаються відповідні документи, які можуть містити інформацію, яка цікавить користувача. У цьому процесі застосовуються багато алгоритмів і методів визначення релевантності документів стосовно запиту.
* **Ранжування та сортування.** Знайдені документи ранжуються за рівнем їхньої відповідності до запиту користувача. Ранжування ґрунтується на різних факторах, таких як частота народження ключових слів, місце входження ключових слів, наявність синонімів та інші аспекти. Документи можуть бути відсортовані за релевантністю або іншими критеріями, такими як дата публікації або популярність.
* **Подання результатів.** Формується сторінка видачі результатів, де відповідні документи представлені у вигляді сніпетів: короткі фрагменти інформації про веб-сторінку, що вмістять URL-адресу, заголовок та короткий опис.
* **Уточнення запиту.** Якщо результати не задовольняють користувача, він може скорегувати або переформулювати запит, додати нові слова або використати інші параметри пошуку.

Кожна пошукова система може мати власні нюанси та алгоритми, але загальні етапи обробки запиту, описані вище, є типовими для більшості пошукових систем.

### Морфологічний розбір запиту

Аналіз запиту починається з визначення мови. До уваги береться регіон користувача і мова 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.

Щоб забезпечити точний пошук всі літери слів запиту перетворюються на нижній регістр. Наприклад, слова "Собака", "собака" та "СОБАКА" будуть розглядатися як однакові при пошуку. Видаляються зайві пробіли, символи пунктуації, спеціальні символи та стоп-слова (прийменники, сполучники тощо), які не несуть змістовного навантаження.

Для приведення слів до їх основної форми використовується лематизація та стемінг. Наприклад, слова "бігати", "бігає" та "пробіг" приводяться до загальної форми "бігти". Це допомагає врахувати різні форми слова під час пошуку та покращує точність результатів.

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. Для підвищення точності та релевантності пошуку застосовуються алгоритми виправлення помилок. Вони дозволяють автоматично коригувати неправильно введені або опечатані слова у запиті. Пошукова система пропонує користувачеві грамотно написаний варіант і перепитує чи шукати за виправленим запитом.

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. визначається, як часто це слово зустрічається з іншими словами запиту - і в питаннях користувачів, і взагалі в текстах. Для збору статистики пошукові системи використовують власні корпуси, де зібрано величезну кількість текстів.

Деякі пошукові системи можуть проводити обробку синонімів, щоб врахувати різні варіанти виразів, які мають схоже значення. Наприклад, запит "автомобіль" може бути розширений за допомогою синоніму "машина" для більш повних результатів. Однокореневі слова і синоніми система бере з відповідних довідників і словників.

Аналізуючи запит, виділяються різні об'єкти - географічні назви, імена людей, назви організацій, абревіатури і скорочення. Для виділення стійких фраз і об'єктів використовуються різні довідники і словники: топонімів (географічних назв), імен і прізвищ, організацій, стійких словосполучень, абревіатур. Отримавши запит, система кожен раз перевіряє за довідниками, чи є в ньому стійкі словосполучення.

На роботу з помилками і весь морфологічний 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.

#### Переваги морфологічного аналізу запиту:

* Морфологічний аналіз допомагає врахувати контекст запиту і визначити про що конкретно запитує користувач.
* Морфологічний аналіз дозволяє розширити запит. Це сприяє знаходженню сторінок, які відповідають запиту користувача, але не містять всіх слів із запиту.

Пошукові системи постійно вдосконалюють свої методи морфологічного аналізу запиту. Вони прагнуть забезпечити користувачам швидкий та ефективний доступ до інформації.

### 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.2).

Якщо сформованої сторінки у кеші немає, то запит передається до серверів іншої комп'ютерної системи - «базового пошуку». На базовому пошуку зберігається зліпок Інтернету – пошукові індекси, 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.

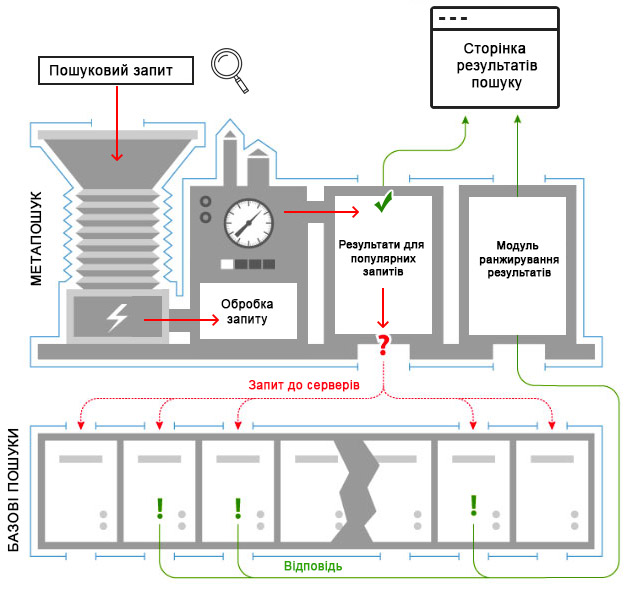


Рис.11.3. Оброблення пошукового запиту

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 сторінка з відповідями формується за частки секунди.

### Чинники, що впливають на видачу пошукової системи

* **Релевантність.** Це рівень відповідності контенту на сторінці запиту користувача. Пошукова система прагне розміщувати на перших позиціях ті сторінки, які найбільш релевантні до запиту користувача.
* **Трастовість**. Це довіра пошукової системи до сайту. Ступінь трастовості залежить від віку домену (чим він старший, тим більше довіра), розміру трафіку, оригінальності вмісту (як тексту, так і графічного контенту). Ступінь довіри також залежить від трастовості інших ресурсів, на яких розміщено зворотні посилання на конкретний сайт. Чим вища трастовість ресурсу, тим вище пошукова система розмістить його на сторінці видачі результатів.
* **Поведінкові показники.** Пошукові системи вміють визначати, які сторінки переглядали і скільки часу, на які кнопки натискали, з якої сторінки поверталися назад у видачу тощо. Наприклад, якщо користувачі залишали сторінку протягом кількох секунд, значить сайт нецікавий людям (незручний інтерфейс, банер на пів-екрану, повільне завантаження).
* **Комерційні показники**. Насамперед, фактори з цієї групи дуже важливі для інтернет-магазинів та сайтів компаній. Пошукова система зацікавлена, щоб користувачі могли отримувати товар або послугу без обману компанії. Тому, на сайті обов'язково мають бути контактні дані, якісні фото товарів, відгуки та інше.
* **Технічні показники.** В інтересах пошукових систем забезпечити відвідувачам зручність користування сайтом. Тому, програма враховує технічні моменти: швидкість завантаження сторінок, розміри рекламних банерів, відсутність помилок у коді тощо. Особлива увага приділяється наявності зручної мобільної версії. У сайтів із низькими технічними показниками надто мало шансів потрапити на першу сторінку видачі.

Ще один важливий фактор, який дуже впливає на формування видачі — оновлення алгоритмів пошукових систем. У такому випадку пошукова система починає інакше оцінювати веб-сторінки, через що принцип формування видачі сильно змінюється.

### Алгоритм ранжирування

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**. Алгоритм ранжирування - це система математичних формул для оцінки певних факторів, на базі якої пошукова система привласнює сайту (сторінці) певний рейтинг. Факторами ранжирування називаються властивості запиту і сторінки, які важливі для ранжирування і які можна виміряти числами. Для точного пошуку важливо враховувати багато різних чинників.

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

Основні фактори, які враховуються під час ранжирування:

* **Відповідність ключовим словам (Keyword Relevance)**. Пошукові системи оцінюють, наскільки запит користувача відповідає змісту веб-сторінки. Відбувається аналіз використання ключових слів і фраз на сторінці.
* **Семантичний контекст (Semantic Context)**. Пошук враховує як ключові слова, а й семантичний контекст і пов'язані поняття.
* **Якість контенту (Content Quality).** Якість тексту, його оригінальність та корисність для користувача важливі. Плагіат та низькоякісний контент можуть негативно вплинути на ранг сторінки.
* **Заголовки та метадані (Headings and Metadata)**. Наявність на сторінці заголовків, мета-тегів та описів (заголовки H1, H2 і так далі, а також мета-опис).
* **Оптимізація зображень (Image Optimization)**. Використання оптимізованих зображень та підписів до них.
* **Анкорні тексти (Anchor Texts)**. Тексти, які використовуються у посиланнях на сторінку.
* **Соціальні сигнали**. Активність сторінки в соціальних мережах. Наявність на сторінці кнопок вподобання, поширення, система коментарів.
* **Геолокація**. Пошукові системи можуть враховувати розташування користувача та надавати релевантні результати.
* **Авторитетність (Page Authority)**. Пошукові системи враховують авторитетність веб-сторінки, що визначається, наприклад, через алгоритм PageRank. Посилання на сторінку та їх якість також впливають на авторитет.
* **Зворотні посилання (Backlinks)**. Кількість та якість зовнішніх посилань, що вказують на сторінку.
* **Адаптивна структура URL (URL Structure)**. Змістовні URL, що стосуються теми сторінки.
* **Адаптивність**. Здатність сторінки адаптуватися під різні пристрої та браузери важлива для користувачів.
* **Мобільна дружність (Mobile Friendliness)**. З врахуванням зростання кількості користувачів з мобільними пристроями, мобільна дружність сторінки стала критичним фактором.
* **Швидкість завантаження (Page Load Speed)**. Швидке завантаження сторінки є важливим для задоволення користувачів.
* **Метрики поведінки користувачів (User Engagement Metrics).** Час знаходження користувача на сторінці, частота відмов та інші метрики взаємодії користувачів.

Пошук відповіді ведеться одночасно на тисячах серверів з пошуковими індексами. Кожен сервер шукає у своїй частині індексу і формує список найкращих результатів з релевантними до запиту сторінками.

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.

## Математичні моделі пошукової системи

Математична модель пошукової системи – це абстрактне представлення роботи пошукової системи з використанням математичних концепцій, методів та формалізму. Вона дозволяє описати різні аспекти роботи пошукової системи та її компонентів, включаючи процес індексації, ранжування результатів та отримання інформації.

Математичні моделі пошукової системи може бути реалізована з використанням різних математичних методів та технік, включаючи теорію ймовірностей, статистику, лінійну алгебру, графові моделі та багато інших. Вона дозволяє досліджувати та аналізувати різні аспекти пошукової системи, а також покращувати її функціональність та продуктивність.

* **Модель ранжування** описує алгоритми та методи ранжирування документів, засновані на різних факторах релевантності, таких як частота ключових слів, авторитетність документа, популярність та інші семантичні аспекти. У моделі можуть бути включені статистичні методи, машинне навчання чи інші фактори.
* **Модель подання результатів** описує методи формування та подання результатів пошуку, включаючи вибір та відображення сніпетів, заголовків, описів та інших елементів, які допомагають користувачам оцінити релевантність та якість документів.
* Модель оцінки якості містить методи оцінки якості пошукової системи, такі як показники точності, повноти, середньої позиції релевантних документів та інші метрики для оцінки ефективності та продуктивності системи.

Формування видачі пошукових систем з певного запиту здійснюється автоматично за математичними моделями без участі людини. Проте, жодна модель не може працювати ідеально, особливо на початку, тому, за роботою моделей 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 можна підібрати оптимальну формулу ранжирування - яка показує релевантні сайти серед перших результатів пошуку.

Всі зауваження асесорів враховуються розробниками, які відповідають за налаштування моделей. У формули 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 пошукових систем.

### 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.

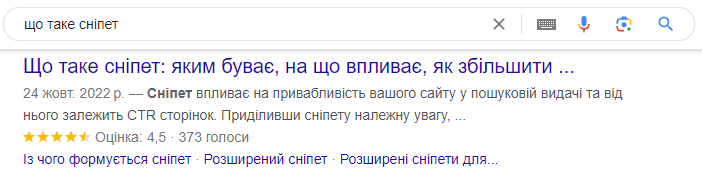


Рис. 4. Сніпет до знайденого документа

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.4).

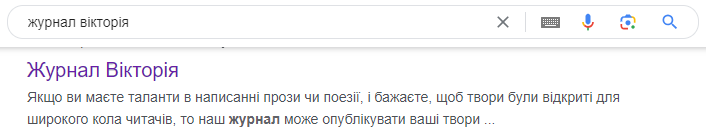


Рис. 4. Фрагмент сторінки видачі Google на запит «журнал вікторія»

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).

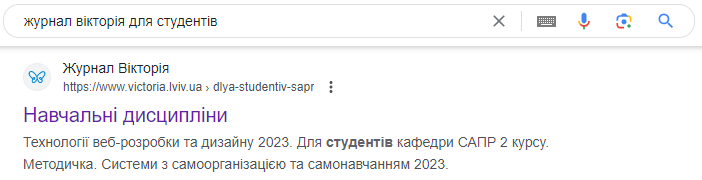


Рис. 5. Фрагмент сторінки видачі Google на запит «журнал вікторія студентам»

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.5).

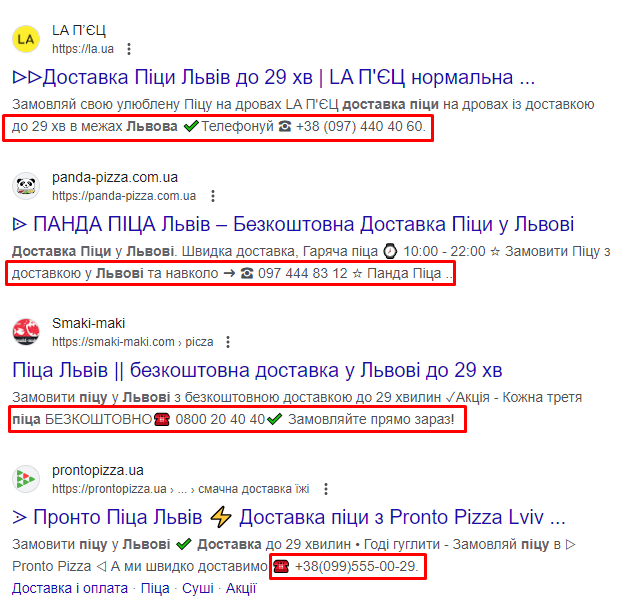


Рис.11.6. Сніпети до пошукового запиту «Доставка піци Львів»

Якщо пошуковій системі відомо про структуру сайту, вона показує її користувачеві (рис.6). 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.

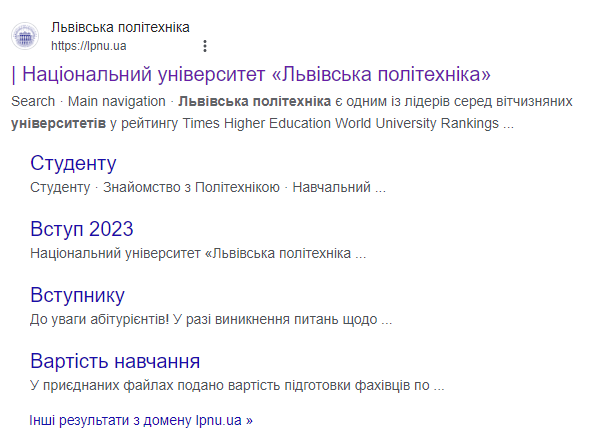


Рис.11.7. Сніпет до запиту «Львівська політехніка»

Для деяких предметних областей створюються спеціальні сніпети (рис.7). 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.

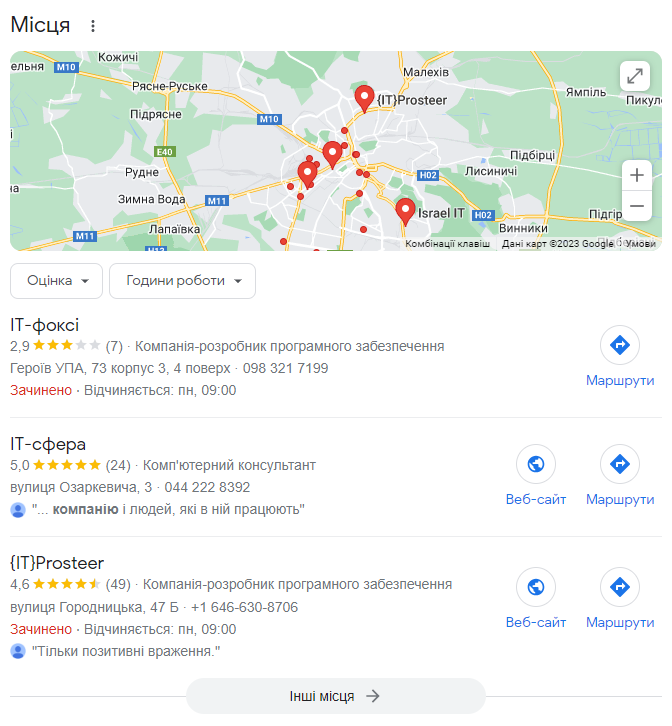


Рис.11.8. Інформація до запиту «ІТ компанії Львів»

### Пошукові підказки

Коли користувач починає вводити запит у пошуковому рядку, пошуковик показує кілька найпопулярніших запитів, що починаються на вже введені літери, - це пошукові підказки, також відомі як "автозаповнення". Підказки прагнуть відповідати поточному запиту користувача, допомагають заощадити час і можна не друкувати запит цілком.

Підказки зазвичай динамічно оновлюються, коли користувач вводить нові символи в пошуковий рядок. Пошуковик розуміє, які показати підказки, навіть якщо користувач забув змінити розкладку клавіатури або допустив граматичні помилки.

Пошукові підказки формуються за допомогою алгоритмів пошукових систем, які постійно збирають інформацію про запити та семантичні зв'язки між ними. Аналізується, наскільки часто певні фрази та запити зустрічаються у їхній базі даних. Популярні запити більш імовірно будуть запропоновані як підказки. Наприклад, якщо користувач почав вводити "погода", система може запропонувати додаткові запити, такі як "погода у Львові", "прогноз погоди" тощо.

Для створення персональних підказок пошукові системи, такі як Google, враховують історію пошукових запитів, попередні інтереси користувача, географічне знаходження та інше. Пошукові системи також враховують помилки та помилки у запитах, надаючи кориговані варіанти.

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

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. 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. Запитів за інтересами, що швидко виникають і пропадають, 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 запити 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.

### Google Knowledge Graph

Google Knowledge Graph – це проект Google, вперше представлений у пошуковій системі Google у 2012 році. Проект розроблено з метою покращення якості та релевантності результатів пошуку, надаючи користувачам більш інформативну та контекстну інформацію про різні сутності та поняття, пов'язані з їхніми запитами.

Google Knowledge Graph є базою даних, в якій сутності (entities) семантично пов'язані між собою. Система розуміє сенс запитів користувачів та може надавати більш інформативні результати. Knowledge Graph здатний розширювати запити користувачів, пропонуючи пов'язані із запитом теми. Це збагачує результати пошуку та дозволяє користувачам досліджувати пов'язані поняття.

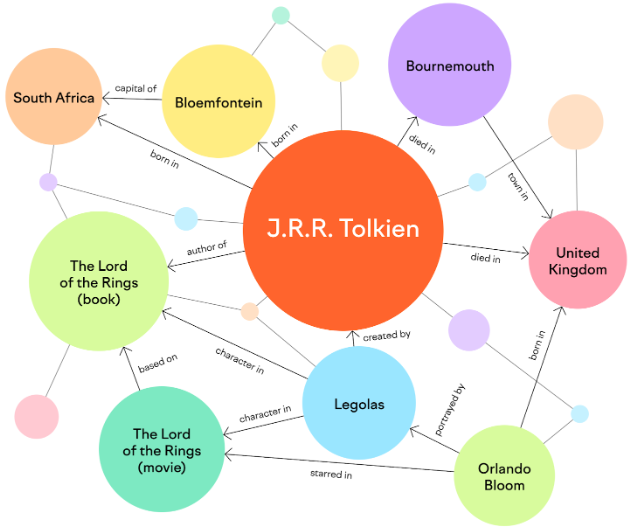


Рис.4. Взаємозв’язок сутностей в мережі знань

Коли користувач робить пошуковий запит про популярну сутність, Knowledge Graph може надавати бічну панель з додатковою інформацією про сутність, пов'язану із запитом. Це дозволяє користувачам швидко отримувати основні відомості без необхідності переходу на інші сайти.

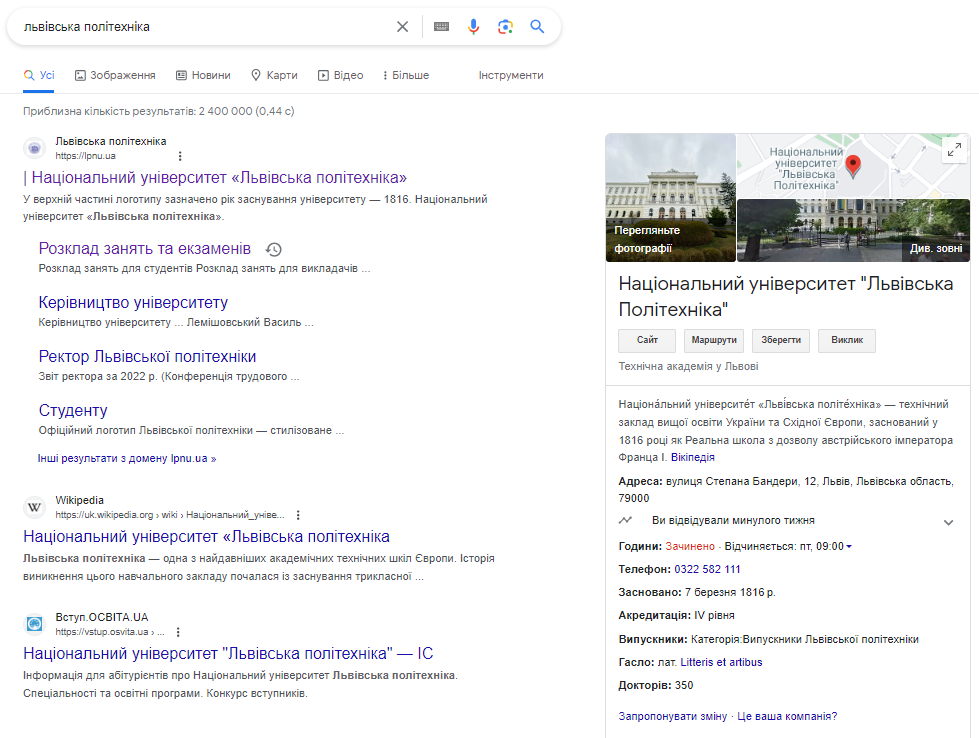


Рис.3. Відображення інформації «Львівська політехніка» у Графі знань

Відображення даних на панелі знань, як і інших результатів пошуку, залежить від цілого ряду чинників, таких як релевантність, відстань і популярність. Наявність даних про компанію не гарантує їх відображення на панелі знань. Рейтинг в сервісах "Пошук", "Карти" і "Мій бізнес" не можна підвищити на замовлення або за гроші. Алгоритм обчислення рейтингу тримається в секреті і забезпечує рівні можливості для всіх. На сьогоднішній день відомо більше 200 чинників, а за деякими даними близько 1000, що впливають на появу в панелі знань.

Панелі знань створюються та оновлюються автоматично, і Google має майже 500 мільярдів фактів про п'ять мільярдів об'єктів у базі даних Google Knowledge Graph. Результати отримано за допомогою автоматизованих алгоритмів Google (рис.5).

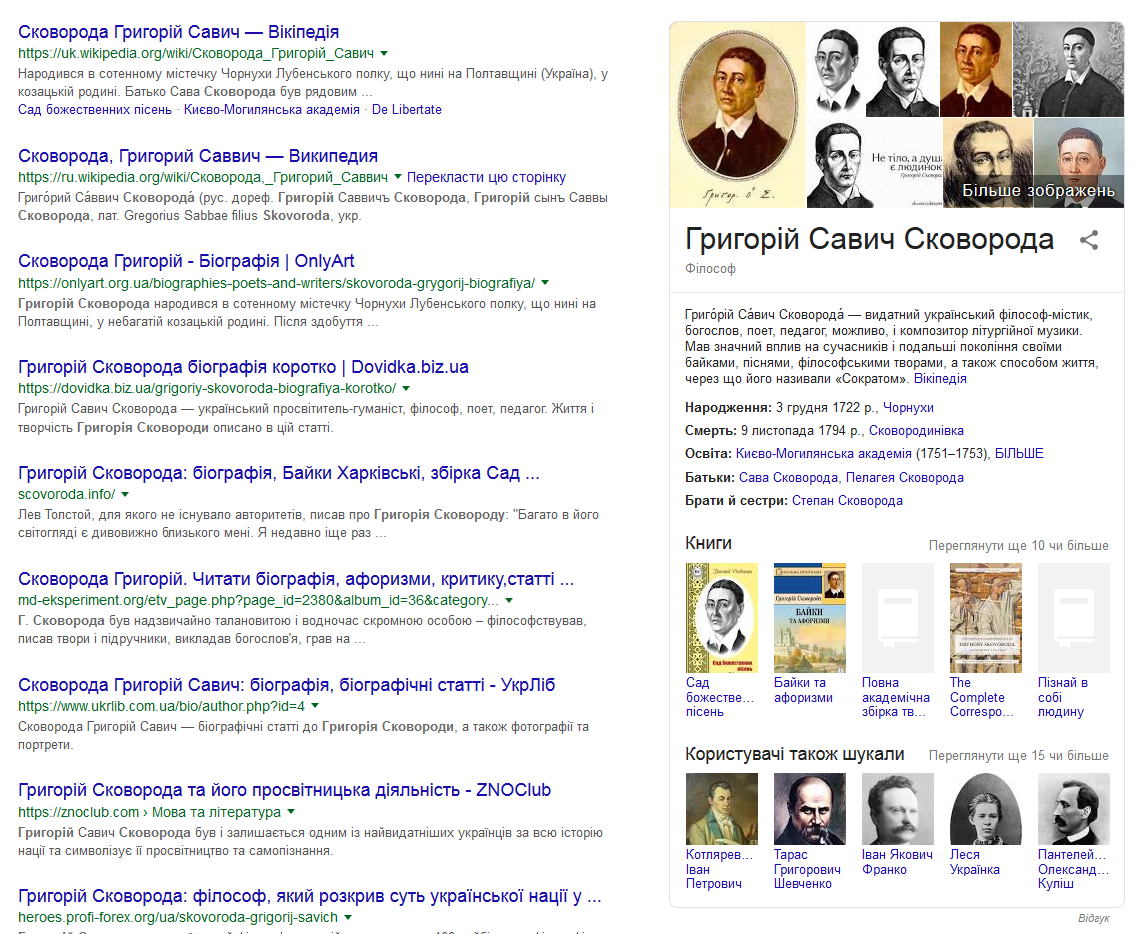


Рис.5. Відображення інформації «Григорій Сковорода» у Графі знань

Формати інформації, які відображаються у панелі знань:

* Заголовок знайденої сутності.
* Короткий, але змістовний опис про цю сутність.
* Одне або кілька зображень людини, місця або події.
* Основні та популярні факти про сутність.
* Особлива інформація, наприклад, популярні цитати, книги чи фільми.

На вкладці "Користувачі також шукають" відображаються об'єкти, пов'язані з даною сутністю.

Google збирає інформацію для Google Knowledge Graph з ліцензійних баз даних та веб-джерел. Формування відповідної сутності може бути в різний спосіб:

* За допомогою аналізу пошукових запитів користувачів.
* За допомогою вивчення та структурування відповідної області.

В граф знань поступово додаються різні відомості не лише шляхом аналізу запитів та інформації, що надходить від користувачів, а й спеціально створюються сектори графу знань. Якщо контент стосується деяких популярних подій, у яких спостерігаються часті зміни, то роботи Google сканують інформацію з більшою частотою, щоб підтримувати актуальність. Що стосується медичного та пов'язаного зі здоров'ям контенту, який потрапляє до мережі знань, Google тісно співпрацює з медичними фахівцями, щоб отримати перевірений контент.

Сприятливі фактори для внесення до мережі знань:

* **Відомості у Вікіпедії або Wikidata.org.** Оскільки більша частина інформації береться з бази даних Вікіпедії, варто мати відповідну сторінку на платформах Wikipedia або Wikidata.org. Якщо компанія є великим і відомим брендом, то потрапити на ці платформи буде не так складно.
* **Профіль у Google Мій бізнес.** Наявність профілю GMB не лише підвищує довіру до бізнесу в Інтернеті, але й приваблює більше клієнтів, не витрачаючи коштів на маркетинг.
* **Наявність посилань на сайтах з високим авторитетом.** Якщо на ресурс посилаються солідні видання, це підвищує шанси потрапити в граф знань.
* **Структуровані дані.** Чітка організація інформації важлива як для користувачів, так і для роботів. Структуровані дані на сайті допоможуть роботам швидше читати та індексувати інформацію.
* **Профілі у соціальних мережах.** Соціальні сигнали мають важливе значення, оскільки демонструють високу довіру до бренду в Інтернеті.
* **Усунення технічних проблем.** Слід виправити всі технічні проблеми сайту, включаючи швидкість завантаження сайту, проблеми зі змішаним контентом, показник відмов тощо.

Графи знань є потужним інструментом, який може бути використаний для покращення якості аналізу запиту в пошукових системах.

* Google Search використовує графік знань для визначення контексту запиту. Це дозволяє Google Search знаходити більш релевантні результати пошуку, навіть якщо запит є неповним або неточним.
* Bing використовує графік знань для розширення запиту. Це дозволяє Bing знаходити більш релевантні результати пошуку, навіть якщо запит є загальним чи абстрактним.

## Великі мовні моделі

Сучасний етап розвитку пошукових систем пов'язаний із застосуванням нейронних мереж, глибокого навчання та великих мовних моделей.

Великі Мовні Моделі (Large Language Models) – це моделі машинного навчання, призначені для вирішення завдань обробки та розуміння природної мови. Вони навчаються на великих обсягах даних, мільярдах параметрів, досягаючи нового рівня якості в обробці природної мови.

Великі мовні моделі – це потужний інструменти, що використовуються в пошукових системах і показують визначні результати у вирішенні різноманітних завдань.

* Розуміння, узагальнення та класифікація інформації у різних формах та шаблонах.
* Розуміння запитів користувача природною мовою та формування відповіді. Великі мовні моделі аналізують контекст запиту та виконують пошук за великим масивом текстів, щоб знайти доречні відповіді.
* Машинний переклад між багатьма мовами.

Великі мовні моделі забезпечують більш точний та ефективний аналіз запиту: обробляти текст, розуміти зміст запиту та визначати частини мови та граматичні ознаки слів. Це призводить до підвищення якості результатів пошуку.

#### Переваги використання LLM для аналізу запиту:

* Точність у визначенні частин мови та граматичних ознак слів, ніж традиційні методи, такі як використання морфологічних словників.
* Ефективність і швидкість обробки запитів, ніж традиційними методами.
* Гнучкість в адаптації до конкретних завдань.

Google використовує кілька великих мовних моделей для своєї пошукової системи.

* LaMDA навчена на величезному наборі даних тексту та коду, спроможна генерувати текст, перекладати мови, створювати різні види творчого контенту та відповідати на запитання інформативним чином.
* Turing NLG для створення природної мови, такої як листи, код, сценарії, музичні твори, електронні листи, листи тощо. Turing NLG використовується для покращення якості результатів пошуку, для створення доречних сніпетів.
* PaLM 2 нова велика мовна модель, яка працюватиме з оновленим чат-інструментом Google Bard, і буде базовою моделлю для більшості нових функцій штучного інтелекту.

Google взялася за розробку нової пошукової системи зі штучним інтелектом, де пошуковик намагається передбачити питання користувача і запропонувати набагато більш персоналізований досвід. Google змінює спосіб подання результатів пошуку, додаючи чат зі штучним інтелектом, а також короткі відео та пости в соціальних мережах. Компанія відходить від традиційного списку результатів, який зробив його домінуючою пошуковою системою.

Нововведення є відповіддю на великі зміни у способах доступу людей до інформації в інтернеті, включаючи появу ботів зі штучним інтелектом, таких як ChatGPT. Google планує зробити свою пошукову систему більш «візуальною, зручною, особистою та людяною» з акцентом на обслуговування молодих людей у всьому світі.