ЛЕКЦІЯ 15
ШТУЧНИЙ ІНТЕЛЕКТ У ВЕБ-ДОДАТКАХ
Пошукові системи

ПЛАН

1. Пошукові системи склад, функції, принцип роботи
2. Приклад пошукової системи Яндекса

ПОСИЛАННЯ НА ДЖЕРЕЛА

https://ai.google/
https://www.victoria.lviv.ua/library/students/ai/web.html

ВСТУП

На сьогодні всесвітня мережа Інтернет є певно найбільшим вмістовищем теорії і практики штучного інтелекту. Тому що зараз вона широко застосовується і всіх сферах людського життя і застосовує найновіші досягнення в області штучного інтелекту.

Найбільш поширені приклади застосування ШІ у веб-додатках:
Пошукові системи
Голосові пошукові системи
Пошукові системи за картинкою
Віртуальні співрозмовники і консультанти
Обробка текстів і переклад
І багато-багато іншого….

1. Пошукові системи склад, функції, принцип роботи

Коротка історія розвитку пошукових систем

На початку розвитку Інтернет, число користувачів та обсяг доступної інформації були порівняно невеликим. Доступ до мережі Інтернет мали переважно співробітники науково-дослідницької сфери і завдання пошуку інформації в Інтернеті не була таким актуальним, як тепер.

Одним з перших способів організації доступу до інформаційних ресурсів Інтернет стало створення відкритих каталогів сайтів, посилання на ресурси в яких групувалися згідно до тематики. Першим таким проектом став сайт Yahoo.com, що відкрився навесні 1994 року. Після того, як кількість сайтів в каталозі Yahoo значно збільшилася, було додано можливість пошуку потрібної інформації всередині каталога. В повному розумінні це ще не було пошуковою системою, оскільки пошукову область було обмежено лише ресурсами, присутніми в каталозі, а не всіма Інтернет ресурсами.

Каталоги посилань широко використовувалися раніше, проте практично повністю втратили свою популярність на даний час, бо навіть величезні за своїм обсягом каталоги, містять інформацію лише про мізерно малу частину Інтернет. Найбільший каталог мережі DMOZ (його ще називають Open Directory Project) містить інформацію про 5 мільйонів ресурсів, тоді як база пошукової системи Google складає мільярди документів.

Першою повноцінною пошуковою системою був проект WebCrawler, що вийшов у світ в 1994 році.

У 1995 році з'явилися пошукові системи Lycos і AltaVista.

У 1997 році Сергій Брін і Ларрі Пейдж створили пошукову машину Google як дослідницький проект в Стенфордському університеті. На даний момент Google є найпопулярнішою пошуковою системою в світі!

У 1997 року було офіційно анонсовано пошукову систему Yandex, яка є найпопулярнішою в Рунеті.

На **даний час існують три основні міжнародні пошукові системи - Google, Yahoo і MSN**, що мають **власні бази і алгоритми пошуку**. Більшість інших пошукових систем використовує їх бази. Наприклад, пошук AOL (search.aol.com) використовує базу Google, а AltaVista, Lycos і AllTheWeb - базу Yahoo.

В Рунеті провідними пошуковими системами є Яндекс, Rambler.ru, Aport.ru, Mail.ru.

Призначення та завдання пошукової 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.

Головним завданням пошукової системи є здатність надавати користувачам саме ту інформацію, яку вони шукають. А от 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 правилами**.

Чи знайшов він те, що шукав?

Якщо не знайшов, то скільки разів йому довелося перефразувати запит, щоб знайти потрібне?

Наскільки актуальною є надана інформація?

Наскільки швидко пошукова машина обробляла запит?

Наскільки зручно було представлено результати пошуку?

Чи була потрібна інформація серед перших результатів пошуку?

Як багато непотрібної інформації було знайдено нарівні з корисною?

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.

Основні характеристики пошукової системи

Повнота - це відношення кількості знайдених за запитом документів до загальної кількості документів в Інтернет, що задовольняють даному запиту. Наприклад, якщо в Інтернеті є 100 сторінок, що містять словосполучення «Як вибрати автомобіль», а за відповідним запитом було знайдено всього 60 з них, то повнота пошуку буде 0,6. Очевидно, що чим повніше пошук, тим більше ймовірність, що користувач знайде потрібний документ.

Точність визначається ступенем відповідності знайдених документів до запиту користувача. Наприклад, якщо за запитом «Як вибрати автомобіль» знаходиться 100 документів, у 50 з них міститься словосполучення «Як вибрати автомобіль», а в інших просто наявні ці слова («як правильно вибрати магнітолу і встановити в автомобіль»), то точність пошуку вважається рівної 50/100 (= 0,5). Чим точніше пошук, тим швидше користувач знайде документи, що відповідають запиту і тим менше різного роду «сміття» серед них буде зустрічатися.

Актуальність характеризується часом з моменту публікації документів в Інтернет, до їх занесення до бази пошукової системи. Наприклад, на наступний день після появи цікавої новини, велика кількість користувачів звернеться до пошукових систем з відповідними запитами. Об'єктивно з моменту публікації новинної інформації на цю тему минуло менше доби, однак основні документи вже було проіндексовано і доступно для пошуку, завдяки існуванню у великих пошукових систем так званої «швидкої бази», яка оновлюється кілька разів на день.

Швидкість 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. Тут інтереси користувачів та пошукової системи збігаються: відвідувач бажає отримати результати як можна швидше, а пошукова машина повинна обробити запит максимально оперативно, щоб не гальмувати обчислення наступних запитів.

Наочність представлення результатів є важливим компонентом зручного пошуку. До популярних запитів пошукова машина знаходить сотні, а то й тисячі документів. Внаслідок нечіткості складання запитів або неточності пошуку, навіть перші сторінки видачі не завжди містять лише потрібну інформацію. Це означає, що користувачеві часто доводиться здійснювати додатковий пошук всередині знайденого списку. Орієнтуватися в результатах пошуку допомагають різні елементи сторінки видачі пошукової системи.

Склад і принципи роботи пошукової системи

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.

Модуль індексування

Модуль індексування складається з трьох допоміжних програм (роботів):

Spider (павук) - програма, що призначена для завантажування веб-сторінок з навколишніх веб-серверів до заздалегідь заданого переліку адрес. Робот отримує від пошукової системи початковий список адрес документів (веб-сторінок), які він має відвідати, скопіювати вміст і віддати його на подальшу переробку до пошукової системи (вона перетворює ці документи в зворотні індекси).

Для завантаження сторінок роботи використовують протоколи HTTP. Робот передає на сервер запит "get / path / document" та інші команди HTTP-запиту. У відповідь робот отримує текстовий потік, що містить службову інформацію і безпосередньо сам документ. «Павук» витягує з документа html-код, посилання з відповідних тегів і редиректи (перескерування зі сторінки).

Кожна завантажена сторінка зберігається в базі в наступному форматі (прямий індекс):

- URL сторінки
- Дата, коли сторінка була завантажена на сервер
- HTTP-заголовок відповіді сервера
- Тіло сторінки (HTML-код)

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.

Indexer (робот-індексатор) - програма, яка аналізує вміст веб-сторінки, що завантажили павуки. Індексатор 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.

Таким чином, модуль індексування дозволяє обходити по посиланнях задану множину ресурсів, завантажувати сторінки, витягувати з одержаних документів посилання на нові сторінки та здійснювати повний аналіз цих документів.

База даних

База даних, або індекс пошукової системи - це система зберігання даних, інформаційний масив, в якому зберігаються спеціальним чином перетворені параметри всіх завантажених і оброблених модулем індексування документів.

Пошуковий сервер

Пошуковий сервер є найважливішим елементом всієї системи, оскільки від його алгоритмів функціонування, безпосередньо залежить якість та швидкість пошуку.

Пошуковий сервер працює наступним чином:

• Отриманий від користувача запит піддається морфологічному аналізу. Генерується інформаційне оточення кожного документа, що міститься в базі (зворотній індекс).

• Отримані дані передаються в якості вхідних параметрів до спеціального модулю ранжирування. Відбувається обробка даних по всіх документах, в результаті чого, для кожного документа обчислюється власний рейтинг, що характеризує релевантність запиту, введеного користувачем, і різних складових цього документа, що зберігаються в індексі пошукової системи.

• Залежно від вибору користувача цей рейтинг може бути скориговано додатковими умовами (наприклад, так званий «розширений пошук»).

• Далі генерується *сніппет*, тобто, для кожного знайденого документа з таблиці документів витягуються заголовок, коротка анотація, найбільш відповідна до запиту і посилання на сам документ, причому знайдені слова виділено грубішим шрифтом.

• Отримані результати пошуку передаються користувачеві у вигляді *SERP (Search Engine Result Page)* - сторінки видачі пошукових результатів.

Як видно, всі ці компоненти тісно пов'язані один з одним і працюють у взаємодії, утворюючи чіткий, достатньо складний механізм роботи пошукової системи, що вимагає величезних витрат ресурсів.

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, бо пошук буде займати багато часу.

Алгоритм зворотного пошуку (інвертованих індексів)

Для ефективного пошуку у великих обсягах даних всі потужні пошукові системи використовують алгоритм зворотних (інвертованих) 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 координати його знаходження в документі та параметри, що визначають його статус в документі.

Це подібно до алфавітного покажчика слів в технічних або наукових книгах, де наводиться список використаних слів із зазначенням номерів сторінок, де вони зустрічаються.

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

Для пошуку по зворотних індексах документів, що містяться в базі даних пошукових систем, використовується ***математична модель***, що дозволяє

спростити процес виявлення потрібних документів (за введеним користувачем пошукового запиту) і процес визначення релевантності всіх знайдених документів до цього запиту. Чим більше документ відповідає даному запиту, тим вище він розташований в пошуковій видачі.

Основним завданням математичної моделі будь-якої пошукової системи є пошук документів (сторінок) у своїй базі зворотних індексів відповідних до даного пошукового запиту і сортування цих знайдених документів у порядку зменшення їх 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.

Математична модель, яку використовується пошукові системи, **відноситься до класу векторних математичних моделей**. В ній використовується поняття ваги документа по відношенню до заданого користувачем запиту.

В базовій векторній математичній моделі вага документа за заданим пошуковим запитом обчислюється за двома основними параметрами: частотою, з якою зустрічається дане слово в аналізованому документі (**TF - term frequency**) і частотою, наскільки рідко це слово зустрічається у всіх інших документах колекції пошукової системи (**IDF - inverse document frequency**). **Під колекцією пошукової системи розуміють** всю сукупність документів, які відомі пошуковій системі. Перемноживши ці два параметри, отримується вага документа за заданим пошуковим запитом.

Природно, що різні пошукові системи, крім параметрів TF і IDF, використовують багато різних 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.

Оцінка якості роботи векторної математичної моделі пошукової системи

Формування видачі пошукових систем з тих чи інших запитів здійснюється автоматично за математичною моделлю без участі людини. Проте, жодна модель не може працювати ідеально, особливо на перших порах, тому, за роботою математичної моделі потрібно здійснювати контроль. Цей контроль здійснюють фахівці - **ассесори**, які переглядають видачу пошукових систем і оцінюють якість роботи математичної моделі пошукової системи.

Всі внесені ними зауваження враховуються розробниками, які відповідають за налаштування математичної моделі пошукової системи. У формулу векторної математичної моделі вносяться зміни або доповнення, в результаті чого якість роботи пошукової системи підвищується. Ассесори виконують роль своєрідного зворотного зв'язку між розробниками пошукової системи та її користувачами, який необхідний для поліпшення якості роботи пошуковиків.

Основними критеріями в оцінці якості роботи математичної моделі пошукових систем є:

1. **Точність видачі пошукової системи** - відсоток релевантних документів, відповідних до пошукового запиту в пошуковій видачі.
2. **Повнота пошукової видачі** - процентне відношення релевантних документів в пошуковій видачі до загальної кількості релевантних документів, наявних у всій колекції пошукової системи.
3. **Актуальність пошукової видачі** - ступінь відповідності реального документа в Інтернеті, до того що про нього написано в пошуковій видачі. Наприклад, документ може вже не існувати або бути сильно зміненим, але при цьому в пошуковій видачі по заданому запиту він буде присутнім, незважаючи на його фізичну відсутність за вказаною адресою або ж на його поточну невідповідність до даного пошукового запиту. Актуальність видачі пошукової системи залежить від частоти сканування роботами документів і поновлення інформації в базах.

Сніппет документа

Сніппет в пошуковій видачі розташовується відразу під посиланням на знайдений документ (текст якої береться з тега TITLE документа):

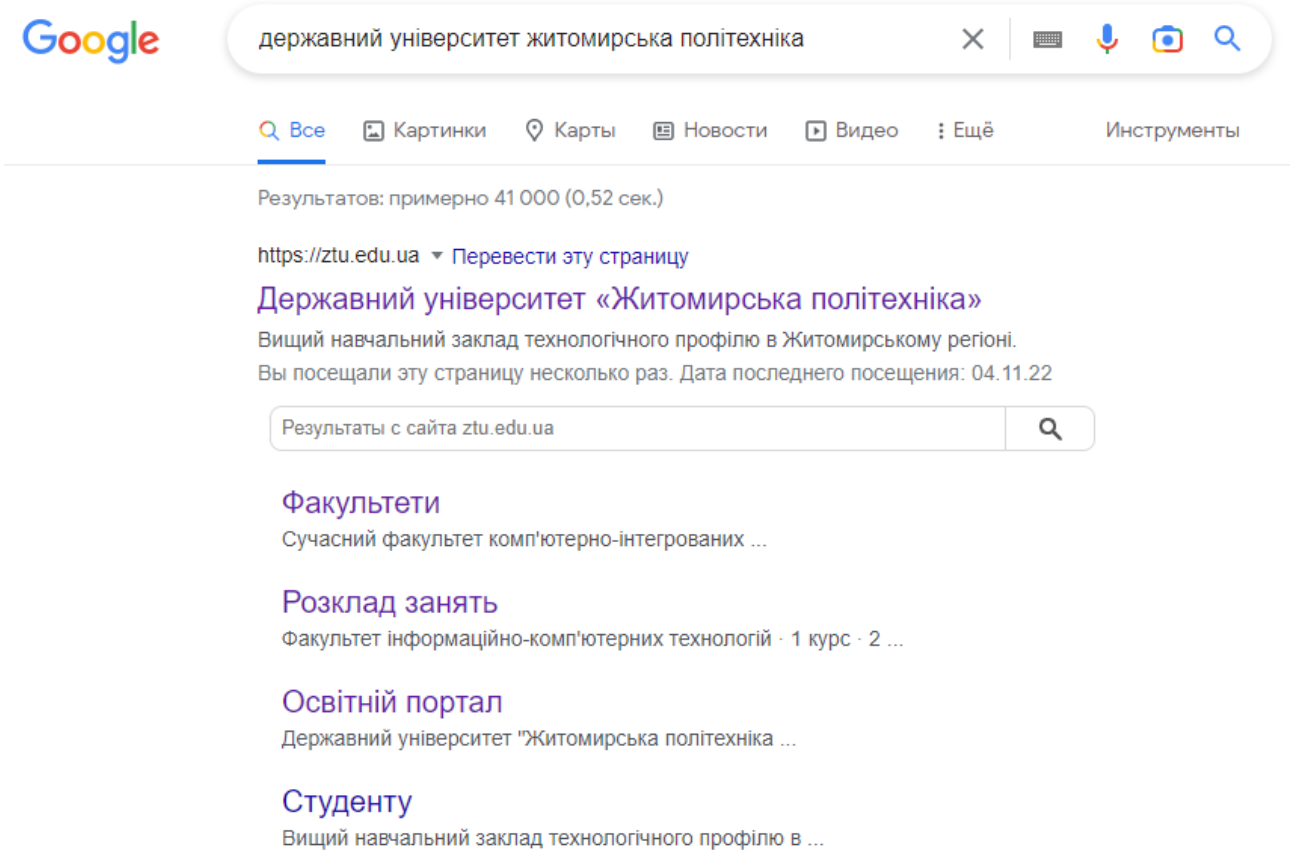


Рис. 1. Фрагмент сторінки видачі Google

Для сніппету використовуються фрагменти тексту з прямого індексу. Ідеальний сніппет має надати користувачеві коротку змістовну інформацію про вміст документа. Сніппет формується автоматично, пошукова система сама

формує фрагменти тексту документа. Для різних пошукових запитів один документ буде мати різні сніппети.

Сніппет не можна отримати з зворотного індексу, оскільки там зберігається інформація лише про використані на сторінці слова та їх розташуванні в тексті. Саме для створення фрагментів одного і того ж документа в різних пошукових видачах (за різними пошуковими запитами) пошуковики, окрім зворотного індексу, зберігають ще і прямий індекс, тобто копію документа, з якої зручно нарізати потрібні сніппети.

Формування сторінки пошукової видачі

В пошуковій видачі за заданим запитом зазвичай міститься лише один (релевантний до запиту) документ з кожного сайту. Пошукові системи зацікавлені в тому, щоб користувач отримував різноманітну інформацію з різних сайтів, а не гортати кілька сторінок пошукової видачі з документами одного сайту. Іноді, як виняток, допускається відображення в пошуковій видачі іншого документа з сайту, якщо цей документ виявиться також доречним.

Частота індексування сайтів

Логіка роботи 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 наблизиться до частоти її оновлення. Час повторного заходу робота пошукових систем може вимірюватися для різних сайтів як в хвилинах, так і в роках. Розумні пошукові системи встановлюють індивідуальний графік відвідування для різних сторінок різних сайтів.

Апдейт пошукової системи

Апдейтом пошукової системи в середовищі оптимізаторів прийнято називати зміну алгоритмів ранжирування того чи іншого пошукача. Оновлення позицій, займаних інтернет-ресурсами, - це і є зовнішній прояв чергового апдейта.

Апдейт індексу пошукової системи - вельми важлива подія в роботі оптимізаторів, в результаті якого той чи інший сайт може не тільки «просісти», а й «піднятися», в результатах видачі пошукових роботів.

Зміна алгоритмів ранжирування таких пошукових систем, як Яндекс і Google - це не тільки зайвий привід похвалитися своїми успіхами або невдачами на форумах. Апдейт - це своєрідний момент істини, який дозволяє говорити про те, яка стратегія просування сайту виявилася успішнішою.

Аналізуючи черговий апдейт, оптимізатори формують деякі його оціночні характеристики : «поганий апдейт » і «хороший апдейт ».

Як часто трапляються апдейти ?

Пошукові роботи обходять, за словами оптимізаторів, сайти безперервно. Разом з тим, результати видачі пошукових систем оновлюються з якоюсь періодичністю. Варто відзначити, що якщо в Google апдейти результатів видачі відбуваються практично щодня, то наприклад в Яндексі - всього лише 1-2 рази на тиждень.

Які ж характеристики сайтів враховуються при апдейті видачі ?

- Різні зміни на сайтах/
- Поява на них різних посилань/
- Санкції, накладені на ті чи інші сайти (песимізація, бан і фільтри)/
- Та інше.

Як би то не було, апдейт індексу пошукової системи - це те, на чому будується робота всіх seo оптимізаторів/

Які бувають апдейти?

Оскільки в Інтернеті нічого не буває незмінним, все постійно оновлюється, то існує і кілька різних видів апдейтів. Насамперед - це **апдейт видачі**.

Що це таке? Це оновлення бази даних і перерахунок позицій, відповідно до інформації, зібраної пошуковою машиною з часу минулого апдейта, або зміною заданого алгоритму оновлень.

У пошуковій системі Яндекса систематично проходить апдейт **ТИЦ (тематичного індексу цитування)**. Цей індекс показує, кількість сайтів, що дають посилання на ваш ресурс. Саме це і враховує пошукова машина Яндекса.

Є ще і такі показники, як **ІЦ (індекс цитування)** і **ВІЦ (зважений індекс цитування)**. При обробці цих показників враховуються не тільки кількість, але і якість сайтів, що дають посилання на вас.

Пошукова машина Google проводить апдейт **PR (PageRank)**. Тут основним критерієм є посилання і відвідуваність сторінки. Оновлюється PR достатньо рідко. Цей критерій дозволяє швидко прикинути, до якої міри сайт розкручений, але не більше того.

Що таке **фавікон**, знає не кожен. Це іконка, розташована в адресному рядку поруч з назвою сайту. **Апдейти фавікона** проводяться досить рідко, та й багато чого тут залежить від браузера.

Пошукова машина Яндекса має у своєму розпорядженні таку корисну програму, як **бистроробот**. Принцип його роботи такий: чим частіше оновлюється сторінка, тим швидше вона індексується. Через два дні вона потрапить в загальну базу. Саме тому, працювати з сайтом треба постійно. Але треба знати, що індексування і зміна позицій сайту бистроботом апдейтом не є.

Як часто відбувається апдейт пошукової бази?

Абсолютно ясно, що пошукові бази повинні регулярно оновлюватися. Адже нова інформація надходить постійно. Це впливає на положення сайту, на його індекс відвідуваності, а, значить, і на доходи, які він приносить. Саме тому апдейт так важливий.

Відповідно, частота, з якою на різних пошукових машинах проходить апдейт, є досить важливим фактором. У різних пошукових машин апдейт відбувається з різною частотою. Найчастіше, приблизно раз на добу, оновлюється Рамблер. Проіндексовані сторінки, які найбільш активно відвідувалися, вже через дві-три години знаходяться в базі даних Рамблера. Принцип індексації в Рамблері взагалі дуже цікавий. Весь Інтернет тут поділений на сектори, пофарбовані в різні кольори. Всі частини бази окремо оновлюються і збираються в одну, відповідно до кольору в строгій черговості. Такий алгоритм і дозволяє пошуковій машині Рамблера отримувати найсвіжішу інформацію.

Пошукова машина Яндекса проводить апдейт двічі на тиждень: найчастіше з вівторка на середу і з п'ятниці на суботу.

ТІЦ (тематичний індекс цитування) зазвичай оновлюється по вівторках. Цікаво ще й те, що в пошуковій системі Яндекса працюють два роботи. Один посилає інформацію відразу в систему, а інший збирає її в «буферній "базі, щоб, накопичивши певну її кількість, провести апдейт.

Один-два рази на місяць оновлює свої бази Google. Цей процес навіть має свою назву: Google Dance. Що стосується такого важливого показника, як PR (Page Rank, то-есть ранг сторінки), то апдейт тут відбувається лише раз на три - чотири місяці.

Навіщо потрібен апдейт?

Для початку, згадаємо, що ж це таке. Апдейт - це заміна застарілих файлів на більш нові. Для того щоб інформація постійно оновлювалася, що не була застарілою і проводяться апдейти.

У різних пошукових системах апдейт проводиться з різною частотою. Відбувається це приблизно таким чином: робот викачує сторінки і постійно змінює застарілі. Коли індексує програма набирає певну кількість таких оновлень, формується нова база даних, в яких, відповідно, змінюється все, від посилань до тексту.

Далі, весь вміст переноситься в постійну базу даних. Ось цей момент і називається апдейтом. Проблема полягає в тому, що під час апдейта можлива некоректна робота пошукових систем, але, зазвичай, через деякий час нормальне функціонування пошукачів відновлюється.

При апдейті змінюються такі показники, як ТІЦ на Яндексі і PR на Google. ТИЦ впливає на позицію сайту в Яндекс-каталозі, так само як і на продаж реклами, розміщеної на сайті. PR ж визначає, як високо буде знаходитися сайт в пошуковику, а, відповідно, дуже важливий для збільшення відвідуваності.

Зазвичай, після апдейта всі сайти, які активно просуваються, впевнено підвищують свої позиції. Звичайно, апдейти відбуваються не тільки в пошукових системах. Постійно оновлюються будь-які програми. У кожному разі, оновлення системи робить її більш досконалою та зручною.

2. Приклад пошукової системи Яндекса

Підготовка до відповіді

Індексування Інтернету

Пошук в Інтернеті складається з двох частин.

1. Пошуковик обходить Інтернет, створюючи його зліпок на своїх серверах.
2. Користувач задає запит і отримує відповідь з серверів пошуковика.

Пошукова машина Яндекса відповідає на питання користувачів, знаходячи потрібні документи в Інтернеті. А розміри сучасного Інтернету обчислюються в екзабайтах, тобто в мільярдах мільярдів байтів. Звичайно ж, Яндекс не обходить весь Інтернет кожен раз, коли йому ставлять питання. Пошукова система, так би мовити, робить домашнє завдання.

Яндекс шукає за пошуковим індексом - базі даних, де для всіх слів, які є на відомих для пошуку сайтах, зазначено їх місцезнаходження - 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 тільки терміни, а взагалі всі слова. А на відміну від адресного довідника, у кожного слова-адресата є не одне, а дуже багато «місць прописки».

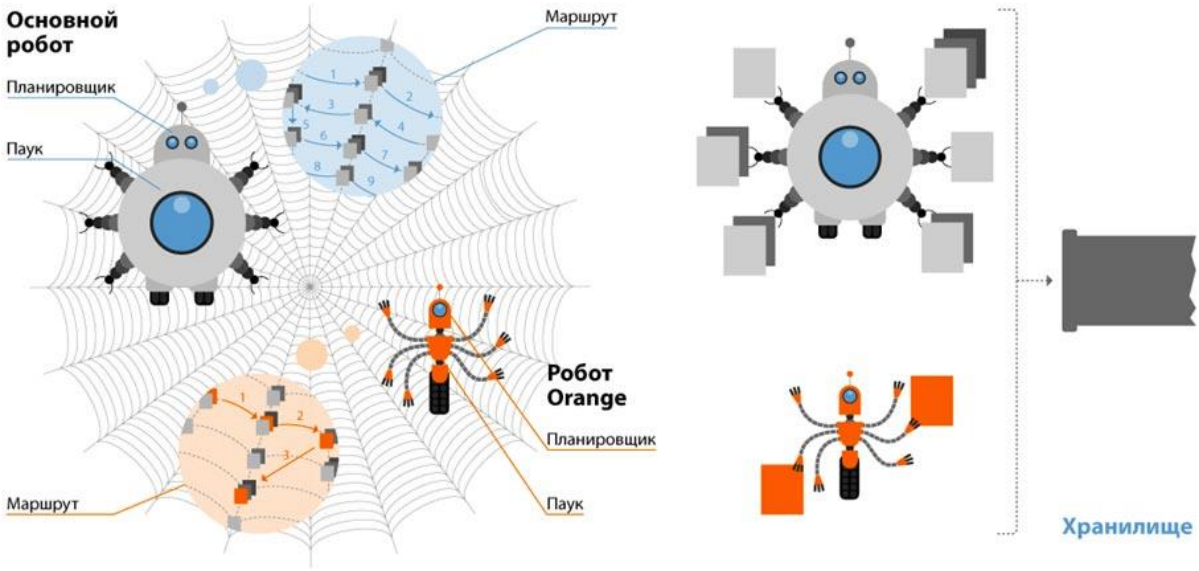
Підготовка до відповідей

Підготовка даних, за якими шукає пошукова машина, називається індексуванням. Спеціальна комп'ютерна система - пошуковий робот - регулярно обходить Інтернет, викачує документи і обробляє їх. Створюється свого роду зліпок Інтернету, який зберігається на серверах пошуковика і оновлюється при кожному новому обході.

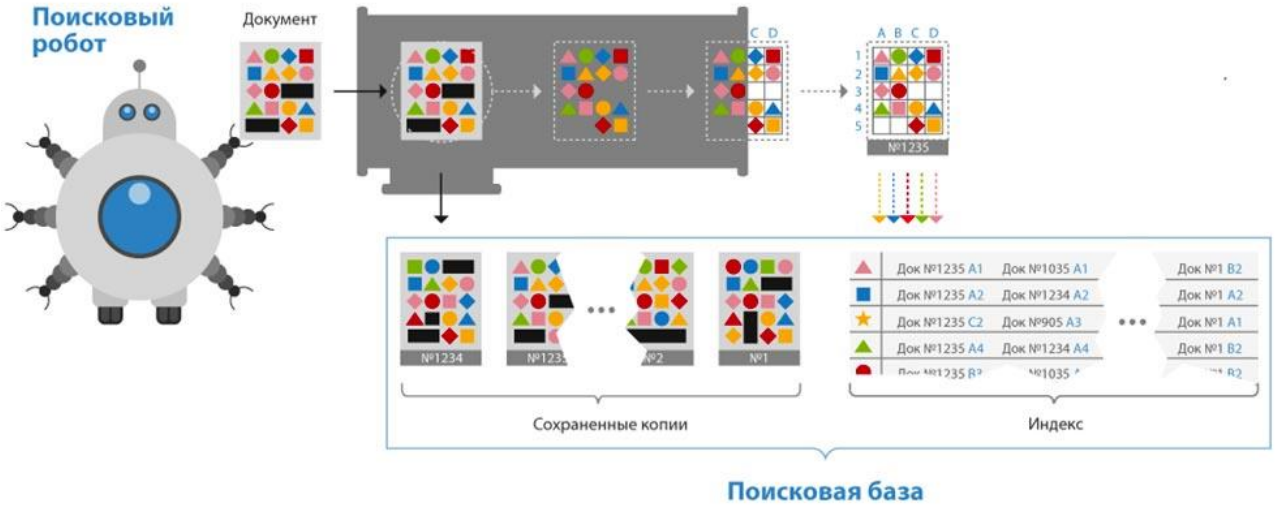
У Яндекса два пошукових робота - основний і швидкий (він називається Orange). Основний робот індексує Інтернет в цілому, а Orange відповідає за те, щоб у пошуку можна було знайти найсвіжіші документи, які з'явилися хвилини або навіть секунди тому. В кожного робота є список адрес документів, які потрібно проіндексувати.

Коли при обході робот бачить на вже відомих сайтах нові посилання, він додає їх до свого списку, збільшуючи кількість індексованих сторінок. Втім, власник сайту сам може допомогти основним роботам Яндекса знайти свій ресурс і підказати, наприклад, як часто оновлюються його сторінки - через сервіс Яндекс.Вебмастер.

Спочатку програма-планувальник вибудовує маршрут - черговість обходу документів. При цьому планувальник враховує важливі для пошукової системи характеристики сайтів, такі як, наприклад, цитованість або частота оновлення документів. Після створення маршруту планувальник віддає його до іншої частини пошукового робота - «павука». Павук регулярно обходить документи за заданим маршрутом. Якщо сайт на місці, тобто 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 відправляє дані в сховищі.



Там програма розбирає документ по цеглинці: очищає від html-розмітки, залишаючи чистий текст, виділяє дані про місцеположення кожного слова і додає їх до індексу. Сам документ у вихідному виді також залишається в сховищі до наступного обходу. Завдяки цьому користувачі можуть знайти в Яндексі і подивитися документи, навіть якщо сайт тимчасово недоступний. Якщо сайт закрився або документ був видалений або оновлений, Яндекс видалить копію зі своїх серверів або замінить її на нову.



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 тільки 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 Яндексу 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.

Робот Orange 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 Orange 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.

Архітектура відповіді на запитання

Архітектура пошуку Яндекса влаштована так, що до вже існуючих серверів можна легко додавати нові сервери для нових даних з постійно зростаючого Інтернету.

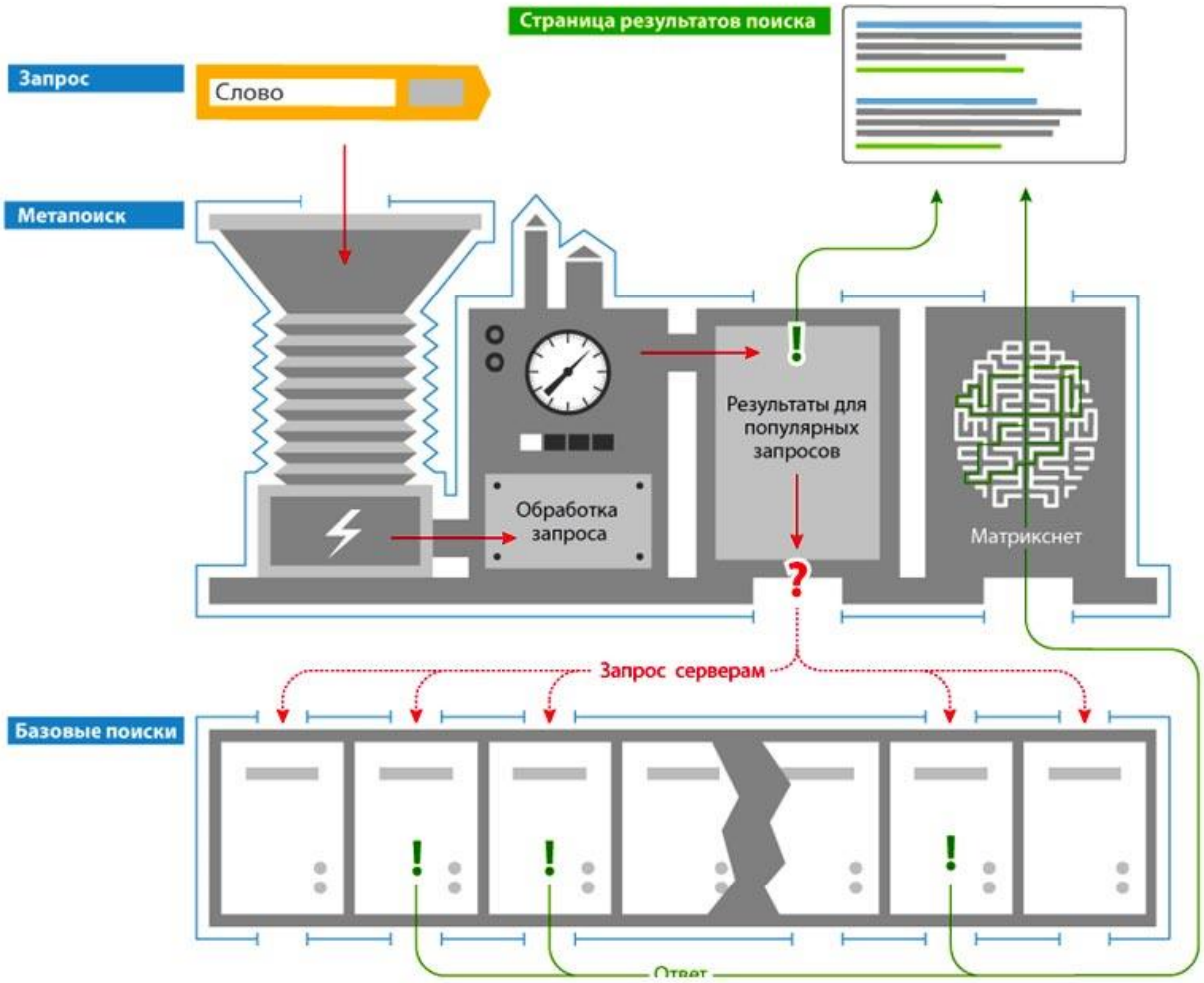
Кожен день користувачі задають Яндексу десятки мільйонів запитів, і пошукова система повинна не тільки точно відповідати, але і швидко обробляти весь цей потік. Для цього Яндекс використовує заздалегідь підготовлені дані - індекс. Безумовно, пошук за допомогою індексу прискорює процес відповіді користувачеві, як, наприклад, предметний покажчик в книзі допомагає швидше знайти потрібне слово. Але розміри самого «предметного покажчика» в пошуку - величезний. Щоб обробляти такі обсяги даних і робити це швидко, Яндекс використовує тисячі серверів. Сервери об'єднані в кластери і навіть в кластери кластерів.

Всі запити користувачів спочатку потрапляють в комп'ютерну систему «метапошуку». Метапошук обробляє кожен запит в реальному часі - з'ясовує всі необхідні дані про запит (з якого регіону він був заданий, до якого класу належить тощо), проводить лінгвістичну обробку. Потім метапошук перевіряє, чи формувалися останнім часом результати пошуку для цього запиту. Результати пошуку по часто заданим запитам деякий час зберігаються в пам'яті метапошуку, а не формуються щораз заново. І якщо новий запит виявився популярним, метапошук покаже користувачеві заздалегідь збережені результати.

Якщо ж відповіді в пам'яті немає, то метапошук передає запит на сервери іншої комп'ютерної системи - «базового пошуку». На базовому пошуку зберігається зліпок Інтернету, за яким шукає Яндекс, - пошукова база. Вона розбита 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 все одно 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 віддає список документів, в яких є слова із запиту, назад до метапошуку. Там вони об'єднуються, ранжуються за допомогою технології Матрікснет [1] і потрапляють на сторінку результатів пошуку [2].

Завдяки такій організації пошук Яндекса може відповідати користувачеві за частки секунди.



[1] Матрікснет - метод машинного навчання, за допомогою якого будується формула ранжирування пошуку Яндекса.

[2] Результати пошуку - посилання на різні веб-сторінки, які користувач бачить у відповідь на свій запит до Яндексу.

Яндекс. Обробка запиту

Щоб вникнути в суть питання, людині потрібно подумати, а пошуковій системі - провести лінгвістичний аналіз запиту. Тільки потім можна приступати до пошуку. Саме при аналізі запиту система вирішує, за якими словами і словоформами потрібно шукати. Наприклад, за запитом «готелі в Іркутську» недостатньо знайти документи з таким поєднанням слів. Хороші відповіді можуть опинитися в документах зі словами «готелі в Іркутську», «іркутські готелі», «Іркутськ готель» і т.д. Аналізуючи питання користувача, система визначає мову запиту, проводить морфологічний розбір кожного слова, вибирає потрібні для пошуку словоформи і відсікає зайві.

На весь аналіз запиту - визначення мови, розбір слів, пошук синонімів - йдуть лише долі секунди.

Визначення мови запиту

Аналіз запиту починається з визначення мови. Наприклад, слово «дружина» в російській мові означає «військова рать», а в українському - і «військова рать», і «жінка». Щоб зрозуміти, що має на увазі користувач, потрібно з'ясувати, якою мовою він спілкується з пошуковою системою. Для цього Яндекс дивиться, який алфавіт використовує людина, які в запиті є характерні поєднання букв і слова. Так, за запитом [дружина князя Игоря] Яндекс буде шукати інформацію про військо, а за запитом [дружина князя Ігоря] - ще й про дружину полководця, княгиню Ольгу.

Крім того, 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 порахувати запит україномовним.

Морфологічний розбір і зняття омонімії

Визначивши мову запиту, Яндекс переходить до морфології. Знання морфології дозволяє знаходити документи, що містять різні форми одних і тих же слів. Наприклад, за запитом [стали для ножів] Яндекс буде шукати документи, в яких є не тільки поєднання «стали для ножів», а й «сталь для ножа», «ножі сталь » і т.д. Аналізуючи запит, Яндекс складає список можливих словоформ для кожного слова.

За словоформою, яка є в запиті, не завжди можна точно сказати, яке слово мав на увазі користувач. Наприклад, у запиті [сталі для ножів] «сталі» - це не тільки іменник «сталь», а й дієслово «стати». І в одному випадку ([сталі для ножів]) потрібно шукати форми іменника, а в іншому ([стали випадати волосся що робити]) - форми дієслова. У такій ситуації потрібно позбутися неоднозначності, тобто зняти омонімією. Омонімія - це збіг слів (словоформ) з різним лексичним значенням.

Щоб вибрати для пошуку найбільш ймовірний список форм, система звертається до статистики спільної зустрічальності слів і граматичних ознак. Наприклад, в морфологічному розборі за запитом [сталі для ножів] система вибере для пошуку слово «сталь». По-перше, тому що за статистикою слово «сталь» частіше зустрічається зі словом «ніж», ніж «стали». А по-друге, тому що іменник в називному відмінку (в даному випадку, «сталь») часто поєднується з іменником у родовому відмінку («ножів»).

Для збору статистики Яндекс використовує Національний корпус російської мови і свої власні корпуси, де зібрано величезну кількість текстів в електронному вигляді.

Розширення запиту

Після зняття омонімії пошукова система вже не шукатиме слова, які користувач точно не мав на увазі. Водночас, якщо обмежити пошук тільки словами із запиту, в поле зору пошукової системи не потраплять багато потрібні документи. Адже для одного і того ж поняття в різних текстах можуть використовуватися різні слова, наприклад на одному сайті може стояти абревіатура, а на іншому - повне найменування.

Для того щоб врахувати всі можливі варіанти, Яндекс розширює запит, додаючи інші формулювання з тим же змістом. Наприклад, разом зі складноскороченим «фізтех» Яндекс буде шукати і офіційне «фізико-технічний інститут», а за запитом «установка скайп» - ще й англійське «skype». Точно так же Яндекс додає в запит різні написання чисел («Петро I» і «Петро Перший»), близькі за змістом однокореневі слова, варіанти написання і синоніми. Так, якщо в запиті є «воронежзький», система може додати до нього однокорінне «воронеж», до [Авто-сервіс міцубіші] - «автосервіс Міцубісі», а до «вітерець» - схоже «бриз».

Вибираючи, яке слово додати, а яке ні, Яндекс дивиться, 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, частину з яких Яндекс сам складає спеціально для таких випадків.

Виділення об'єктів

Аналізуючи запит, пошукова система виділяє в ньому різні об'єкти - географічні назви, імена людей, назви організацій і т.д. Наприклад, якщо пошукова система зрозуміє, що «Сергій Зубов» - це людина, вона не розширюватиме прізвище «зубів» «зубним» або шукати стоматологічні клініки. А якщо в запиті [аптеки біля парку культури] система виявить, що «Парк культури» - це місце, вона врахує це при ранжіруванні: в результатах пошуку перші рядки займуть документи, в яких слова «парк» і «культури» йдуть підряд. Для виділення стійких фраз і об'єктів Яндекс теж становить 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.

Робота над помилками

Аналізуючи запит, пошукова система завжди перевіряє його на грамотність. За статистикою Яндекса, близько 12 % запитів містять помилки. Це можуть бути помилки, орфографічні помилки або абракадабра, яка виходить при неправильній розкладці клавіатури. Якщо шукати те, що зазначено в пошуковому рядку, користувач не отримає потрібну йому відповідь - адже на більшості сайтів слова написані грамотно. Тому, ті слова, в яких часто припускаються помилок («агентство», «вінегрет») або по яких немає хорошої

відповіді на питання, Яндекс відразу ж виправляє і показує відповідь вже на виправлений запит. Зрозуміло, попереджаючи користувача про те, що запит було виправлено.

В деяких випадках складно визначити, помилився користувач чи ні. Наприклад, ресторан «фуджіяма» дуже схожий на вулкан «Фудзіяма», а прізвище футболіста «Массад» на «Моссад» (а також на «масаж» і фортецю «Массада»). В таких випадках, показуючи відповідь на вихідне питання, Яндекс запитує, чи не помилився користувач і чи не хоче він побачити відповідь на виправлений запит. Є ще один варіант - коли система не впевнена, чи було допущено помилку, вона покаже на одній сторінці результатів пошуку відповіді відразу на два питання - на заданий, в якому імовірно є помилка, і на виправлений.

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

Яндекс. Навчання. Ранжирування

Машинне навчання

При нинішньому розвитку Інтернету неможливо передбачити всі запити до пошуку і знайти одну кращу відповідь. Тому пошуковик повинен вміти самостійно визначати, яка відповідь хороша, а яка - ні.

Зараз вже складно придумати такий запит, за яким знайдеться менше десятка сторінок. А по багатьом запитам результатів пошуку - мільйони. І з часом їх стає все більше - Інтернет дуже швидко росте. Тому пошуковій системі вже недостатньо просто показати всі сторінки із словами із запиту - щоб знайти підходящу відповідь, людині доведеться гортати десятки сторінок з результатами пошуку. Пошукова система повинна розташувати знайдені сторінки в потрібному порядку - так, щоб зверху виявилися найбільш підходящі користувачеві (найбільш релевантні).

Процес впорядкування результатів пошуку відповідно до запиту користувача - називається ранжируванням. Саме ранжирування визначає якість пошуку - тобто якість відповіді на питання, задане в пошуковому рядку.

Кожен день Яндекс відповідає на десятки мільйонів запитів. Близько чверті з них - неповторювані. Тому, неможливо написати для пошукової системи таку програму, в якій передбачено кожен запит, і для кожного запиту

відому кращу відповідь. Пошукова система повинна вміти приймати рішення самостійно, тобто, сама вибирати з мільйонів документів той, який найкраще відповідає користувачеві. Для цього потрібно навчити її навчатися.

Пошукова система повинна навчитися будувати правило, яке визначає для кожного запиту, яка сторінка є хорошою відповіддю на нього, а яка - ні. Для цього Пошукова машина аналізує властивості веб-сторінок і пошукових запитів.

У всіх сторінок є особливі ознаки ознаки.

- Статичні - пов'язані з самою сторінкою, наприклад, кількість посилань на цю сторінку в Інтернеті.
- Динамічні - пов'язані одночасно з запитом і сторінкою - наприклад, присутність в тексті сторінки слів запиту, їх кількість і розташування.

У пошукового запиту теж є властивості, наприклад, геозалежні - це означає, що для хорошої відповіді на цей запит потрібно врахувати регіон, з якого він був заданий. Властивості запиту і сторінки, які важливі для ранжирування і які можна виміряти числами, називаються факторами ранжирування. Для точного пошуку важливо враховувати багато різних чинників.

У формулі ранжирування поєднуються різні фактори:

- Статистичні фактори
- Динамічні фактори
- Запитувальні фактори

Асесори

Крім чинників ранжирування пошуковій системі необхідні зразки - запити і сторінки, які люди вважають придатними відповідями на ці запити. Оцінкою того, наскільки та чи інша сторінка підходить для відповіді на той чи інший запит, займаються фахівці - асесори. Вони беруть пошукові запити і документи, які пошук знаходить по цих запитах, і оцінюють, наскільки добре знайдений документ відповідає на поставлене запитання. Із запитів і хороших відповідей складається навчальна вибірка. Вона повинна містити самі різні запити, причому в тих же пропорціях, в яких їх задають користувачі. На навчальній вибірці пошукова система встановлює залежність між сторінками, які асесори порахували релевантними до запитів, і властивостями цих сторінок. Після цього вона може підібрати оптимальну формулу ранжирування - яка показує релевантні запиту сайти серед перших результатів пошуку.

На прикладі це виглядає так. Припустимо, ми хочемо навчити машину вибирати самі смачні яблука. Асесори в цьому випадку отримують ящик яблук, пробують їх і розкладають на дві купи, смачні - в одну, несмачні - в іншу. З різних яблук складається навчальна вибірка.

Машина пробувати яблука не може, але вона може проаналізувати їх властивості. Наприклад - якого вони розміру, якого кольору, скільки цукру містять, тверді або м'які, з листком або без. На навчальній вибірці машина вчиться вибирати найсмачніші яблука - з оптимальним поєднанням розміру, кольору, кислоти і твердості. При цьому можуть виникати певні помилки. Наприклад, оскільки машина нічого не знає про черв'яків, серед обраних яблук

можуть виявитися червиві. Щоб помилок було менше, потрібно враховувати більше ознак яблук.

Перенавчання

В пошукових технологіях машинне навчання застосовується з початку 2000-х років. Різні пошукові системи використовують різні моделі. Одна з проблем, які виникають при машинному навчанні - перенавчання. Перенавчена машина схожа на студента, який перезаймався - наприклад, прочитав дуже багато книжок перед екзаменом з психології. Він мало спілкується з живими людьми і намагається пояснити прості вчинки занадто складними моделями поведінки. І через це поведінка друзів для нього завжди несподівано.

Як це виглядає: коли комп'ютер оперує великою кількістю факторів (у нашому випадку це - ознаки сторінок і запитів), а розмір навчальної вибірки (оцінок асесором) не дуже великий, комп'ютер починає шукати і знаходити неіснуючі закономірності. Наприклад, серед усіх оцінених сторінок можуть виявитися дві з якоюсь складною комбінацією чинників, наприклад, з розміром 2 кб, фоном фіолетового кольору і текстом, який починається на букву «я». І обидві ці сторінки виявляться релевантними до запиту [яблуко]. Комп'ютер почне вважати цю випадкову комбінацію факторів важливою ознакою релевантності до запиту [яблуко]. При цьому всі важливі документи про яблука, які такої комбінації факторів не мають, здадуться йому менш релевантними.

Для побудови формули ранжирування Яндекс використовує власний метод машинного навчання - Матрікснет. Він є стійким до перенавчання.

Матрікснет

Матрікснет - це метод машинного навчання, за допомогою якого будується формула ранжирування Яндекса.

У 2009 році Яндекс впровадив новий метод машинного навчання - Матрікснет. Важлива особливість цього методу - в тому, що він є стійким до перенавчання. Це дозволяє враховувати дуже багато чинників ранжирування - і при цьому не збільшувати кількість оцінок асесором і не побоюватися, що машина знайде неіснуючі закономірності.

За допомогою Матрікснета можна побудувати дуже довгу і складну формулу ранжирування, яка враховує багато різних чинників і їх комбінацій. Інші методи машинного навчання дозволяють або будувати більш прості формули з меншою кількістю факторів, або потребують більшої навчальної вибірки. Матрікснет будує формулу з десятками тисяч коефіцієнтів. Це дозволяє зробити істотно точніший пошук.

Ще одна важлива особливість Матрікснета - в тому, що формулу ранжирування можна настроювати окремо для досить вузьких класів запитів. Наприклад, поліпшити якість пошуку тільки по запитах про музику. При цьому ранжирування за іншим класам запитів не погіршиться. Для прикладу можна уявити собі формулу ранжирування у вигляді складного механізму з великою кількістю ручок. На механізмах, побудованих по інших технологіях, кожна ручка впливає на всі запити. Матрікснет дає можливість налаштувати кожну ручку окремо для свого класу запитів.

Крім того, Матрікснет автоматично вибирає різну чутливість для різних діапазонів значень чинників ранжирування. Це в чомусь схоже на роботу на аеродромі - коли серед постійного шуму злітають літаків потрібно чути й голоси людей. Якщо заткнути вуха, то літаки будуть чутні, а голоси - ні. Співробітники аеропорту працюють у спеціальних навушниках, слабо чутливих до гучного шуму - так можна почути і літаки, і голоси людей.

Як влаштовано ранжирування

Оскільки пошукова система працює з дуже великими обсягами інформації, по кожному запиту їй потрібно перевірити ознаки мільйонів сторінок, визначити їх релевантність і відповідно впорядкувати. Так, щоб зверху виявилися більш відповідні сторінки. Щоб перевірити властивості всіх сторінок по черзі, потрібно або дуже багато серверів, які можуть швидко обробити інформацію про всі сторінках, або дуже багато часу - а пошук повинен працювати швидко, інакше користувачі не дочекаються результатів. Матрікснет дозволяє перевірити дуже багато факторів за короткий час і без істотного збільшення обчислювальних потужностей.

Пошук 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 - тією самою довгою і складною формулою, побудованої за допомогою Матрікснета, з врахуванням всіх факторів і їх комбінацій. Таким чином, нагорі пошукової видачі виявляються всі самі релевантні сайти - і користувач майже миттєво отримує відповідь на своє питання.

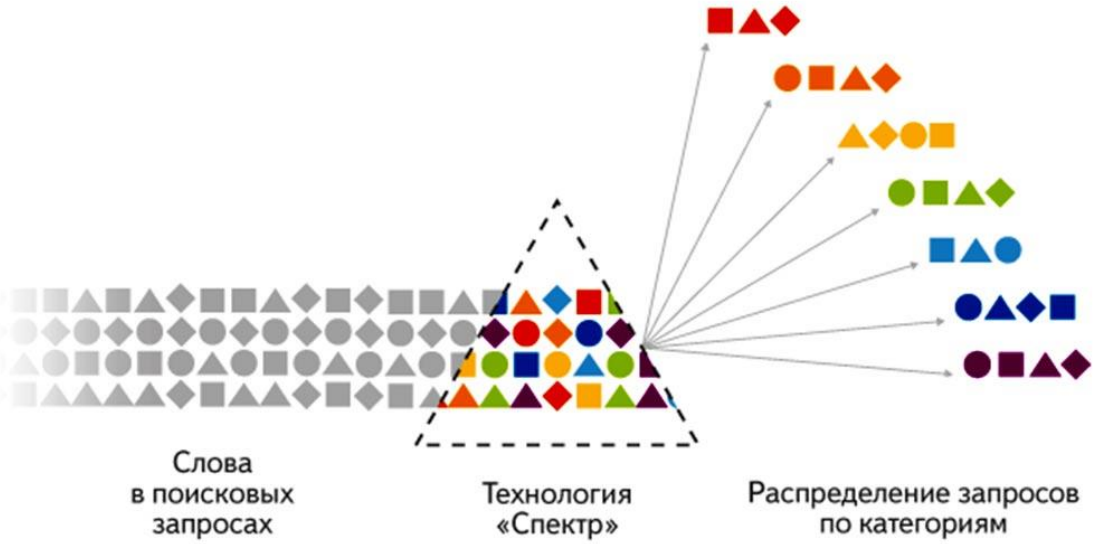
Детальніше: http://company.yandex.ru/technologies/matrixnet/index.xml

Спектр

Спектр - це технологія, яка дозволяє Яндексу враховувати при відповіді різні потреби користувачів.

Коли користувачі задають запити до Яндексу, приблизно в 20% випадків вони формулюють запит неоднозначно. Наприклад, за запитом [наполеон] хтось хоче знайти полководця, а хтось - рецепт торта. А задаючи запит [суші], людина може шукати і ресторан з доставкою додому, і рецепт страви. Спектр можливих цілей може бути дуже широкий - так само, як і спектр можливих відповідей. І якщо користувач не вказав в пошуковому запиті, що він шукає, то зрозуміти це вкрай важко. Технологія Спектр вміє враховувати безліч неявних цілей користувачів та показувати відповідні відповіді.

В основі роботи Спектру лежить статистика пошукових запитів. Система досліджує запити всіх користувачів Яндекса і виділяє в них різні об'єкти - це можуть бути імена людей, назви фільмів і книг, моделі автомобілів тощо. Кожен об'єкт відноситься до однієї або кількох категорій.

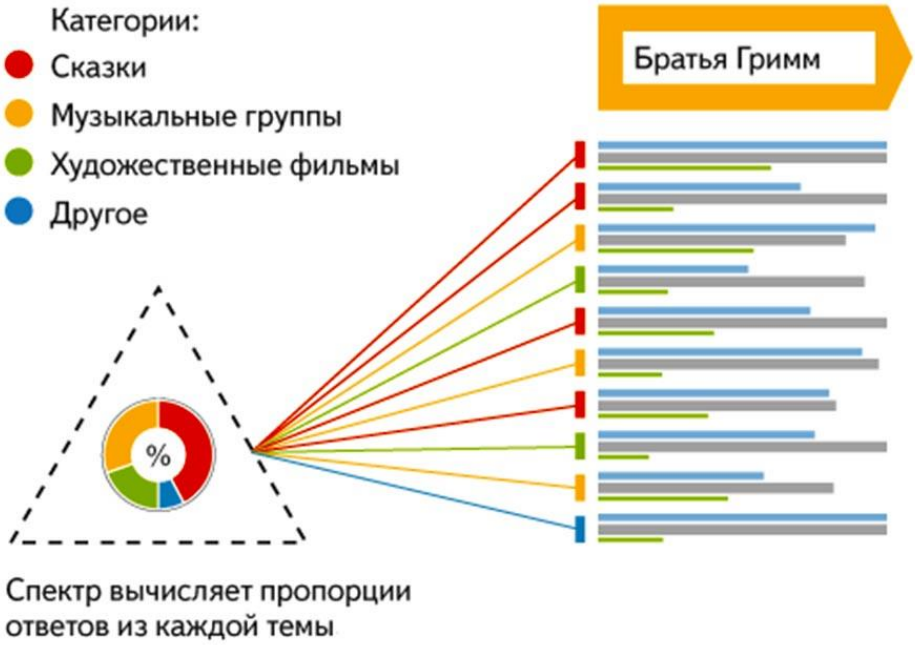


Наприклад, у запиті [колдрекс інструкція] назва ліків «Колдрекс» - об'єкт, який потрапляє в категорію «ліки». А об'єкт «Пушкін» відноситься до двох категорій - «поети» і «міста».

На даний момент Спектр виділяє близько 60 категорій, і ця кількість буде ще рости. Знання категорій дозволяє пошуковій системі розуміти різні значення слів в пошукових запитах.

Крім того, Спектр вміє враховувати при пошуку різні потреби користувачів. В кожній категорії є список можливих потреб - тих намірів, з якими користувачі шукають той чи інший об'єкт. Наприклад, коли люди шукають який-небудь товар, вони, як правило, хочуть купити його або почитати відгуки та огляди. Тобто для категорії «товари» серед потреб будуть «купити», «відгуки» та «огляди». Всього у категорії може бути від двох-трьох до кількох десятків потреб.

З врахуванням того, в які категорії потрапив об'єкт, що люди зазвичай про нього питають, що пишуть в Інтернеті і т.д. Спектр оцінює відсоток людей, які шукають цей об'єкт з кожної з можливих цілей. Ці дані використовуються при ранжируванні результатів пошуку по багатозначним запитам. Використовуючи їх, Спектр обчислює пропорції, в яких відповіді на ту чи іншу тему повинні бути представлені в результатах пошуку. Знайдені сайти впорядковуються таким чином, щоб спектр відповідей відповідав спектру питань. Таким чином, пошук Яндекса максимізує ймовірність того, що людина знайде саме те, що шукав. Навіть якщо він не вказав це явно у своєму запиті, а просто подумав.



Спектр аналізує пошукові запити повністю автоматично - кожного разу розглядається дуже великий масив запитів, більше п'яти мільярдів. Їх обробка відбувається одночасно на кількох сотнях машин. Щоб дані не втрачали актуальність, Спектр запускає процес аналізу кілька разів на тиждень.

Крім статистики запитів, Спектр вміє використовувати дані з довідників та енциклопедій - у тому числі з Вікіпедії. Це допомагає розпізнавати об'єкти, що з'явилися недавно, дізнаватися, які значення об'єктів не вкладаються в жодну з існуючих категорій, і додавати нові.

На базі технології «Спектр» в пошуку Яндекса реалізовано діалогові підказки. Вони з'являються під рядком пошуку у відповідь на неоднозначні запити. Діалогові підказки описують найбільш популярні категорії, в які потрапив запит, і дозволяють в один клік перейти до відповідей тільки з вибраної категорії. Наприклад, за запитом [чорниця] Яндекс пропонує «корисні властивості», «вікіпедія», «калорійність» і «рецепти». За кожним з цих посилань на користувача чекає відповідна сторінка результатів пошуку.

Пошук з врахуванням регіону

Пошук з врахуванням регіону працює для всіх міст Росії, України та Білорусі, де є достатня кількість місцевих ресурсів. Для Казахстану та Туреччини є окремі регіональні формули ранжирування.

Серед всіх запитів до пошуку Яндексу від 15 до 30%, залежно від регіону, становлять ті, у відповідь на які користувач очікує отримати місцеву, регіональну інформацію - наприклад, про послуги або події у своєму місті. На такі запити пошук Яндекса відповідає в різних регіонах по-різному. Наприклад, за запитом [послуги адвоката] жителі Самари знайдуть професійні юридичні послуги в Самарі, а нижньогородці - адвокатів Нижнього Новгорода.

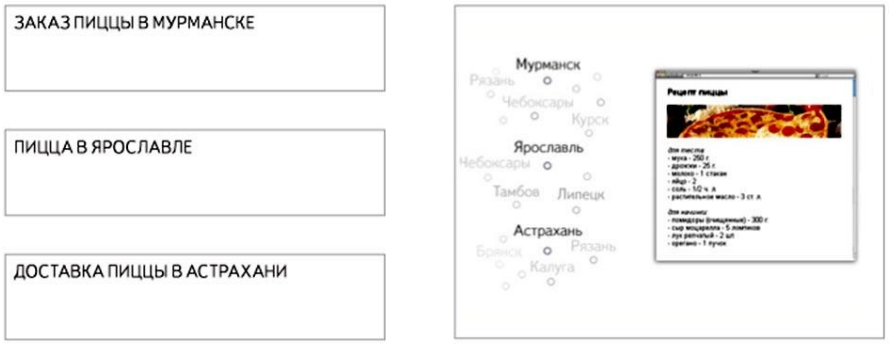
Геозалежні і геоНЕзалежні запити

Відповідь на багато запитів не залежить від регіону, в якому знаходиться користувач. Наприклад, при пошуку художнього твору, рецепту або фізичного

закону враховувати регіон не потрібно - закони фізики скрізь однакові. Але якщо людину цікавить [тренажерний зал] або [замовлення таксі], очевидно, він хоче знайти тренажерний зал або таксі не взагалі, а саме в своєму місті.

Крім того, бувають запити, задаючи які, мешканці різних регіонів мають на увазі різні речі. Найчастіше це прізвища місцевих знаменитостей або назви організацій. Наприклад, за запитом [орбіта] москвичі найчастіше шукають кінотеатр, жителі Ростова-на-Дону - автосалон, а ізраїльтяни - Інтернет-портал.

Вміння розрізняти геозалежні і геонезалежні запити допомагає пошуковій системі краще розуміти запит користувача і давати підходящу відповідь.

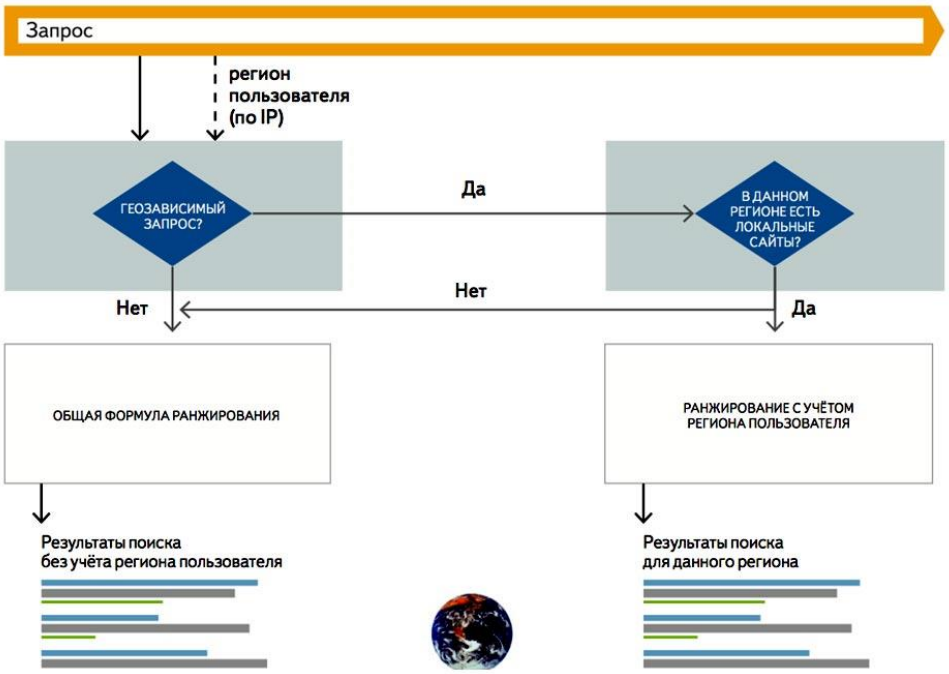


Геозалежні запити визначаються статистично - це запити без вказівки географічних назв, але з такими словами, до яких часто додають географічну назву. Тобто, запит [перевезення вантажів] - геозалежний, тому що поєднання «перевезення вантажів» часто запитують разом з назвами міст.

Як визначається регіон користувача

Регіон користувача визначається, перш за все, за IP-адресою. На ці дані не завжди можна спиратися - наприклад, тому що IP-адреса може присвоюватися провайдером, що працює в кількох регіонах. Яндекс постійно уточнює класифікатор регіону користувача, одержуючи дані від своїх клієнтів, партнерів і безпосередньо від самих користувачів - будь-хто може поміняти свій регіон в налаштуваннях.

Регіон користувача завжди зазначено в правому верхньому куті на сторінці результатів пошуку. Змінити його можна на сторінці http://tune.yandex.ru/region/.



Результати регіонального пошуку

У відповідь на геозалежні запити Яндекс показує різні результати пошуку для різних регіонів. Найбільш релевантні відповіді знаходяться, як правило, на місцевих, регіональних сайтах. Але це не означає, що за геозалежним запитом не можна знайти авторитетний загальноросійський сайт або сайт, що розташований в іншому регіоні. Мова лише про пріоритет для локальних результатів при інших рівних. Яндекс може шукати і виключно по місцевим ресурсам - для цього потрібно відзначити під пошуковим рядком галочку «Шукати у моєму місті». Коли місто зазначено у запиті, сайти цього міста отримують пріоритет в результатах пошуку. Тобто за запитом [готелі Пермі] Яндекс покаже сайти приміських готелів незалежно від того, де перебуває користувач, який загадав цей запит.

Параметри, що дозволяють пошуку дізнатися, для якого регіону потрібно сформувати сторінку результатів пошуку, передаються в адресі сторінки. Так що можна надіслати посилання на неї комусь з іншого регіону - він побачить ту ж саму сторінку.

Щоб відключити локальні результати пошуку за геозалежним запитом, потрібно натиснути на посилання «Пошукати без врахування регіону» під результатами пошуку.

Регіон користувача також враховується в чаклунчиках [4] і пошукових підказках [5]. Так, чаклунчики показують актуальні для регіону користувача погоду, афішу, вакансії, адреси організацій і т.д. - Наприклад, вулиця Зелена буде в кожному регіоні своя.

[4] Чаклунчики - інформаційні блоки в результатах пошуку з відповідями від сервісів Яндекса.

[5] Пошукові підказки - список запитів, один з яких, швидше за все, хотів задати користувач. З'являються під пошуковим рядком при введенні запиту.

Як визначається регіон сайту

Належність сайту до того чи іншого регіону визначається за багатьма ознаками - в тому числі за вказаними на ньому контактами, IP-адресою сайту, регіону, якому присвячена більша частина інформації на ресурсі, і т.д.

Сайти організацій, в яких є офіси в різних регіонах, вважаються місцевими для кожного з них. Однак якщо відділень дуже багато, як у Пошти Росії, то сайт може вважатися загальноросійським, а не регіональним. Те ж саме з сайтами, які створені в одному з регіонів, але розраховані на всю аудиторію інета - наприклад, електронні бібліотеки або поштові служби в мережі. Якщо пошук неправильно визначив регіон того чи іншого сайту, то вебмайстер може виправити його на сервісі Яндекс.Вебмастер.

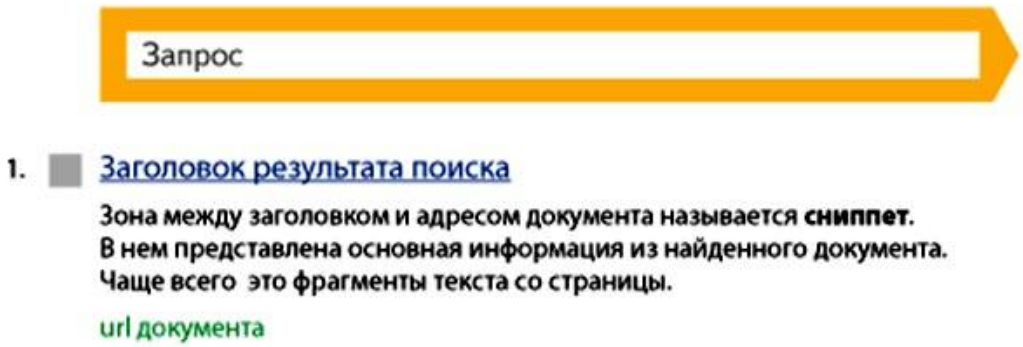
Детальніше: http://company.yandex.ru/technologies/regions/index.xml

Результати пошуку

Сторінка результатів пошуку (SERP) - одна з основних веб-сторінок Яндекса. Кожен день вона формується десятки мільйонів разів.

Сторінка результатів пошуку - це відповідь Яндекса на питання, яке користувач задав в пошуковому рядку. Яндекс знаходить і показує всі відповідні відповіді: чаклунчики своїх сервісів, контекстні оголошення Яндекс.Директа, і, звичайно, самі результати пошуку по Інтернету.

Результати пошуку по Інтернету - це посилання на знайдені документи з короткою інформацією про них. Інформація підбирається так, щоб допомогти користувачеві зрозуміти - яка з відповідей підходить йому найкраще. Яндексу важливо не просто показати релевантну відповідь, але і описати її максимально інформативно.



Формування результатів пошуку

Для заголовка результату пошуку Яндекс найчастіше використовує заголовок самого документа. Якщо він занадто довгий, Яндекс вибирає фрагмент, який найбільше підходить за змістом до заданого запиту.

Буває, що у документа немає заголовка або заголовок не відповідає змісту. Наприклад, назви файлів у форматі doc або pdf часто короткі і малоінформативні. В таких випадках Яндекс створює заголовок самостійно, ґрунтуючись на текстах посилань на документ, заголовках у самому тексті документа та його зміст.

Для формування опису сторінки, яка поміщається в сніпетах, програма вибирає всі фрагменти тексту документа зі словами із запиту. Кожен з таких фрагментів розбивається ще на кілька частин - наприклад, зі словами із запиту

на початку, в кінці і в середині. Потім програма порівнює їх все між собою і вибирає кращі - вони і потрапляють в сніпет.

При виборі програма враховує кілька десятків факторів. Деякі з них підвищують шанси потрапляння фрагмента в сніпет, а деякі - навпаки. Наприклад, якщо слово міститься в довгому реченні, більша ймовірність, що це частина розповіді, а не навігаційне посилання. Це хороший фрагмент для сніпета.

Також в сніпет швидше потраплять фрагменти з різних частин тексту - так можна повніше описати зміст сторінки. А ось фрагмент, схожий з заголовком тексту сторінки, навряд чи потрапить до сніпету - щоб не дублювати інформацію.

Для кожного фактора комп'ютерна система обчислює коефіцієнт. За допомогою машинного навчання система вчиться сама розуміти значимість факторів, ґрунтуючись на даних від фахівців-асесорів (вони дивляться певні набори сніпетів, вручну поділяють їх на хороші і погані і повідомляють ці оцінки системі). Згодом комп'ютерна система вже без допомоги людей будує формулу, за якою створює сніппети.

Оформлення результатів пошуку

Результат пошуку оформляється так, щоб користувачеві було легше його сприймати. Заголовки виділено синім кольором і підкреслено - так на веб-сторінках традиційно виділяються посилання. Впізнати знайомий ресурс допомагає фавіконка - невеликий фірмовий значок сайту - зліва від заголовка результату пошуку. Якщо заголовок або текст опису містить прописні букви, Яндекс намагається зробити їх рядковими - так простіше читати.

А щоб було легше «зачепитися оком», всі слова із запиту в результатах пошуку виділено жирним шрифтом. При цьому Яндекс вміє зіставляти абревіатури та їх розшифровки, повні імена, скорочення та ініціали, числа і їх текстове написання. Наприклад, за запитом [петро 1] Яндекс знайде документи, які містять і «Петро I», і «Петро перший», і виділить у сніпетів різні варіанти написання імені.

Щоб допомогти користувачеві швидше зрозуміти зміст документа, Яндекс може виділити деякі слова, яких немає в запиті. Це відбувається при відповіді на загальні, багатозначні запити. Наприклад, для запиту [сніжна королева] в різних сніпетах будуть додатково виділені слова «мультфільм», «казка», «магазин». Додаткові слова Яндекс дізнається, аналізуючи переформулювання запитів. Спеціальна програма стежить за тим, як користувачі уточнюють свої запити, і обчислює значимість таких уточнень. Потім ці знання використовуються при формуванні сніпета.

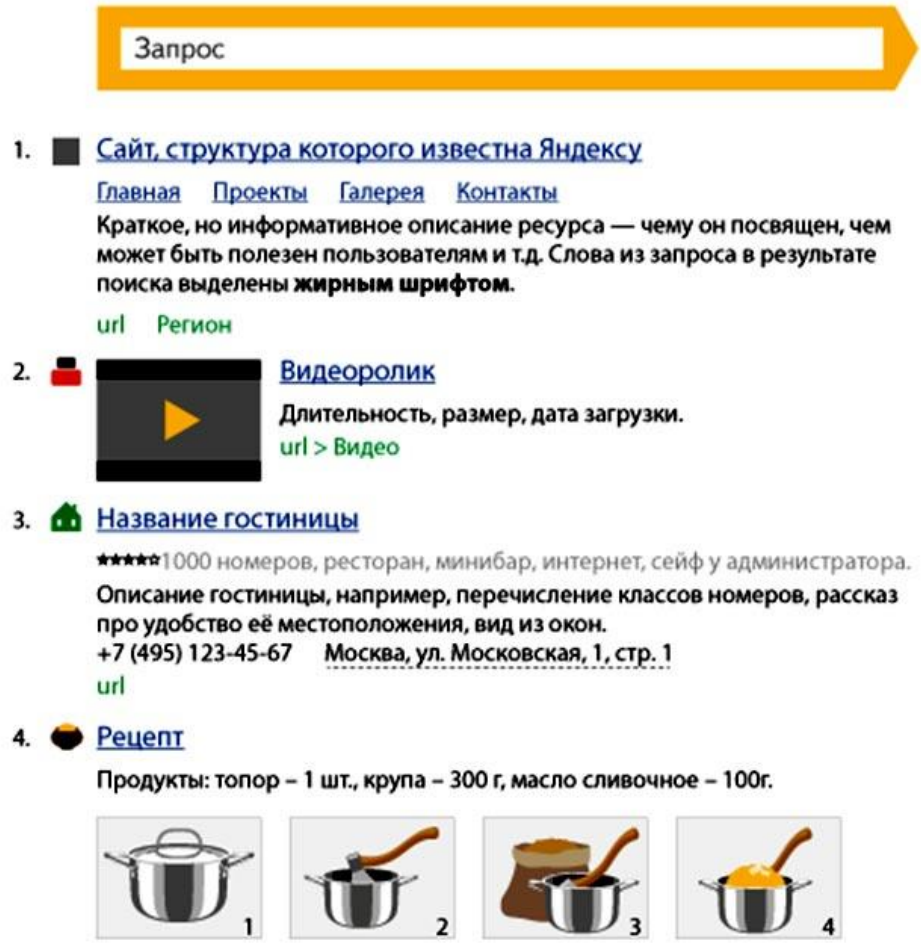
Додаткова інформація в сніпетах

Яндекс намагається зробити так, 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, Яндекс додає телефон і фізичну адресу з посиланням на карту до сніпету.

Якщо Яндексу відома структура сайту, він показує її користувачеві. Над текстом сніпета сайту з'являються посилання на його найбільш відвідувані сторінки - щоб за бажанням користувач міг перейти в потрібний розділ, витрачаючи менше кліків і трафіку. А адресу документа Яндекс перетворює в навігаційний ланцюжок - назви розділів і підрозділів сайту, з яких складається шлях до документа.

Для деяких предметних областей Яндекс створює спеціальні сніпети. 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.



Власники сайтів можуть поліпшити представлення своїх ресурсів в результатах пошуку Яндекса. Багато інструментів для цього є на сервісі Яндекс.Вебмастер.

ВИСНОВКИ

Інтернет з кожним днем все більше нагадує самоорганізований універсум, що еволюціонує з шаленою швидкістю. І хоча ця система ще не має повноцінного штучного інтелекту, зачатки його створення вже починають

з'являтися (наприклад, віртуальний співрозмовник або Акінатор, який читає думки, машинний зір та голосовий інтерфейс пошукових систем). Настане той день, коли тест Тьюринга буде пройдено та Інтернет з функціонального інструменту перетвориться на незамінного помічника, а для когось і друга.

Хто стоїть за всім цим? Безумовно, це спільноти людей. Співтовариства, що об'єднані спільними ідеями, цілями та інтересами, які готові витрачати свій час і ресурси на втілення цих ідей. Тому, з кожним днем в Інтернеті з'являється все більше розумних програм, їх функціонал стає все ширше, а відвідувачі перетворюються зі споживачів в активних творців контенту.