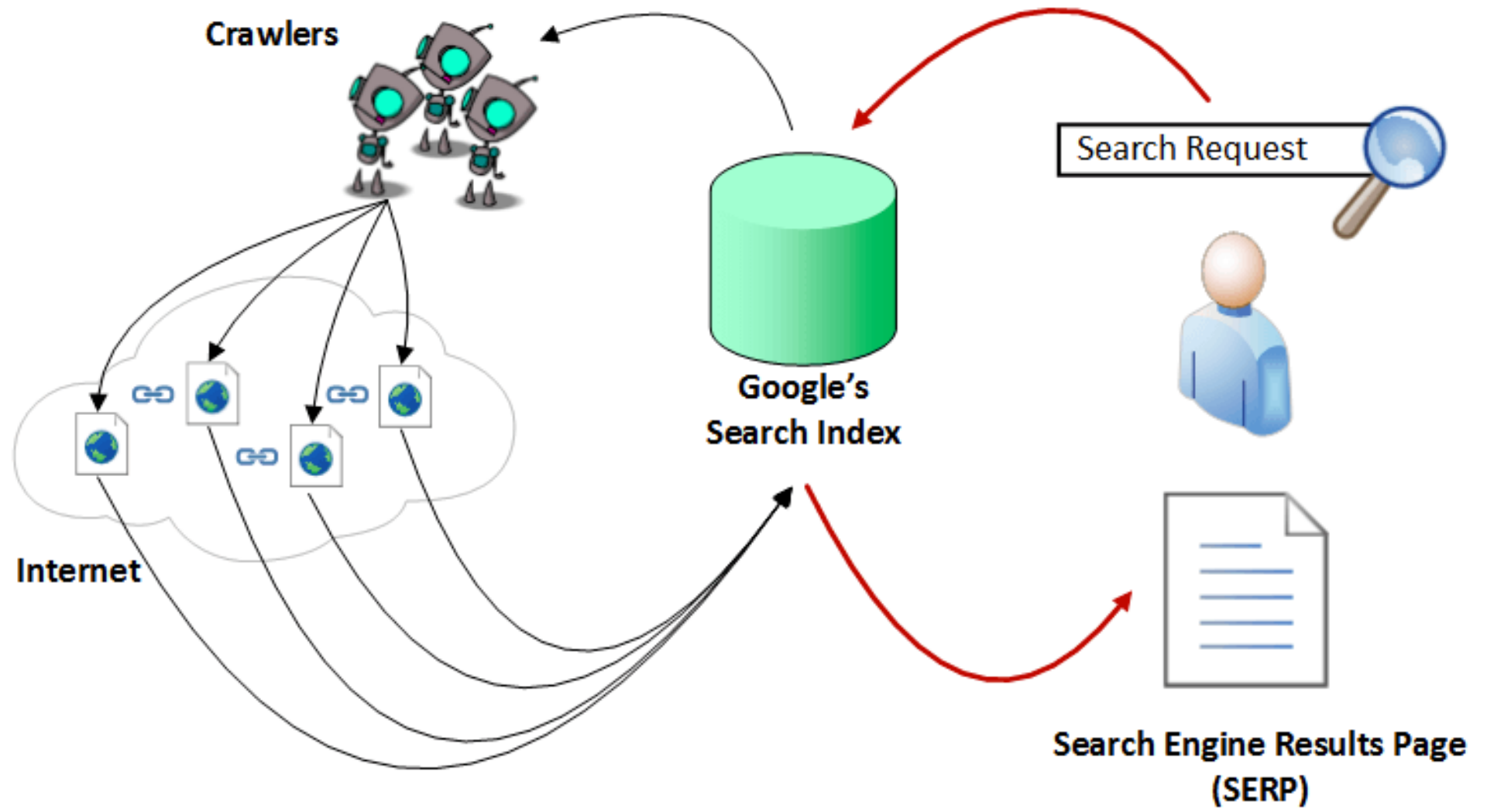
ЛЕКЦІЯ 3
МОДЕЛІ ПОШУКУ.
МЕТАКРАУЛЕРИ.

Курс лекцій
“Робота з інформаційними ресурсами”

Яворський Володимир Антонович
jva@biph.kiev.ua
2019

ПОШУКОВИЙ ПРОЦЕС



КЛАСИЧНІ МОДЕЛІ ПОШУКУ

- Булева модель
- Векторно-просторова модель
- Ймовірнісна модель

Класична концепція «Bag of Words»
розглядаємо документ как множини окремих слів, незалежних одне від одного

БУЛЕВА МОДЕЛЬ ПОШУКУ

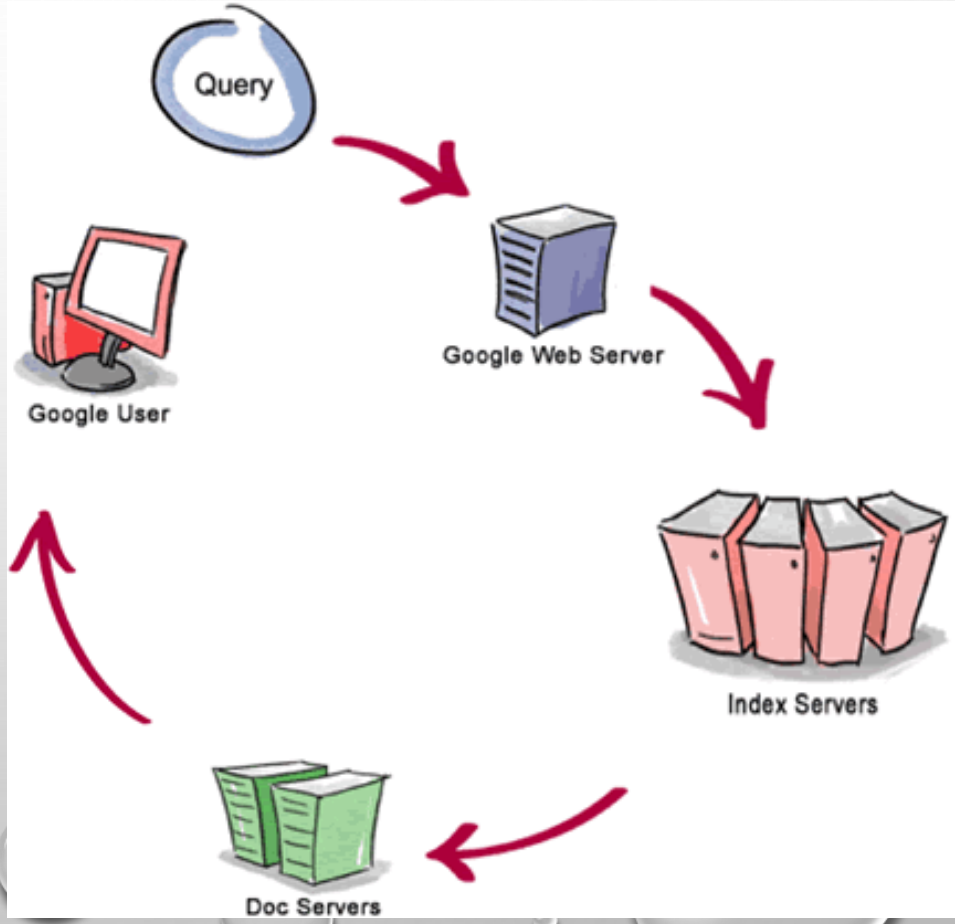
Запит - логічний вираз з операторами (AND, OR, NOT)

Таблиці з інвертованими списками

- тексти
- індекси на тексти
- словник унікальних слів
- інверсна таблиця, що містить списки номерів документів, які відповідають певним словами

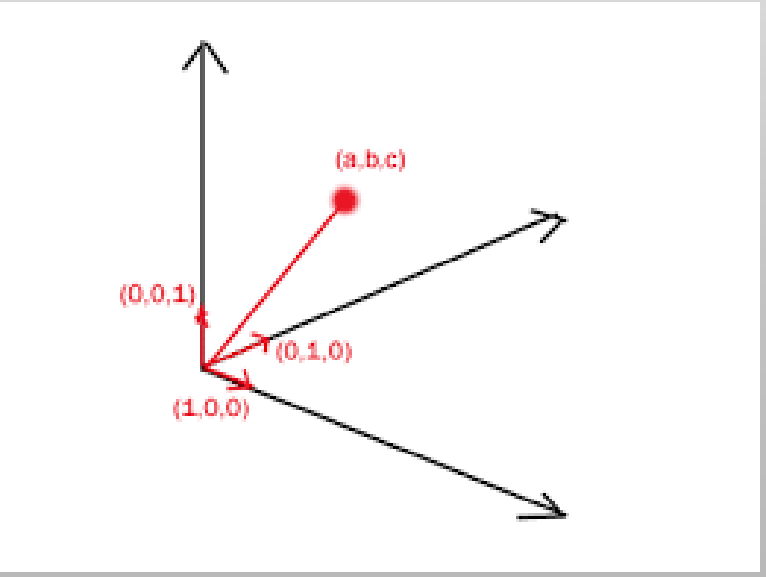
Процес пошуку інформації

- звернення до словника унікальних слів
- звернення до інверсної таблиці
- звернення до покажчиків на тексти
- звернення до текстової таблиці



ВЕКТОРНО-ПРОСТОРОВА МОДЕЛЬ ПОШУКУ

Документ - описується вектором в деякому евклідовому просторі термінів. Кожному терміну зіставляється вага, яка характеризується частотою, місцем розташування, тематикою та інше.



Вага терміну це частота появи терміну в документі, множена на обернену величину кількості документів масиву, де зустрічається термін

$$TF(t_i, d) = \frac{n(t_i, d)}{\sum_k (t_k, d)}$$

Запит також вектор в евклідовому просторі.

Близькість запиту документу скалярний добуток векторів запиту і вектору документу

Модель забезпечує

- Обробку запитів без логічних обмежень їх довжини
- Простоту реалізації режиму ПОСК подібних документів
- Збереження результатів пошуку з можливістю уточнюючого пошуку

ЙМОВІРНІСНА МОДЕЛЬ ПОШУКУ

Релевантність - це ймовірність того, що документ може бути цікавим користувачу (Robertson та Sparck-Jones, 1977)
Пошук будується на гіпотезі, що терми запиту по різному розподілені серед релевантних і нерелевантних документів

- виходячи зі складу термінів, отримуємо апріорні оцінки ймовірності того, що документ є релевантним (релевантність запиту є сумою релевантностей по всім словам)
- отримуємо експертні оцінки користувачів, які визнають документ релевантним або нерелевантним, формуються «учбові набори» релевантних або нерелевантних документів
- Ймовірність нового документу визначаються на основі співвідношенні термів в релевантних і нерелевантних масивів документів
- ітераційно застосовуються експертні оцінки (завдяки зворотному зв'язку) та визначаються документи, зазначених користувачем як такі що задовольняють його інформаційні потреби

МОДЕЛЬ ДОБРЕ ВИЗНАЧАЄ НОВИЙ СПАМ ПО МНОЖИНІ ДОКУМЕНТІВ, ЯКІ ВИЗНАЧЕНІ КОРИСТУВАЧАМИ ЯК СПАМ.

НЕДОЛІКИ КЛАСИЧНИХ МОДЕЛЕЙ

- **Булева модель -** невисока ефективність пошуку, жорсткий набір операторів, неможливість ранжування.
- **Векторно-просторова модель -** пов'язана з розрахунком масивів високої розмірності, малопридатна для обробки великих масивів даних.
- **Ймовірнісна модель -** має низьку обчислювальну масштабованість, пов'язана з необхідністю постійного навчання системи.

GOOGLE ВИКОРИСТОВУЄ ПОНАД 200 ФАКТОРІВ ДЛЯ ВИЗНАЧЕННЯ РЕЛЕВАНТНОСТІ

GOOGLE ІСТОРІЯ

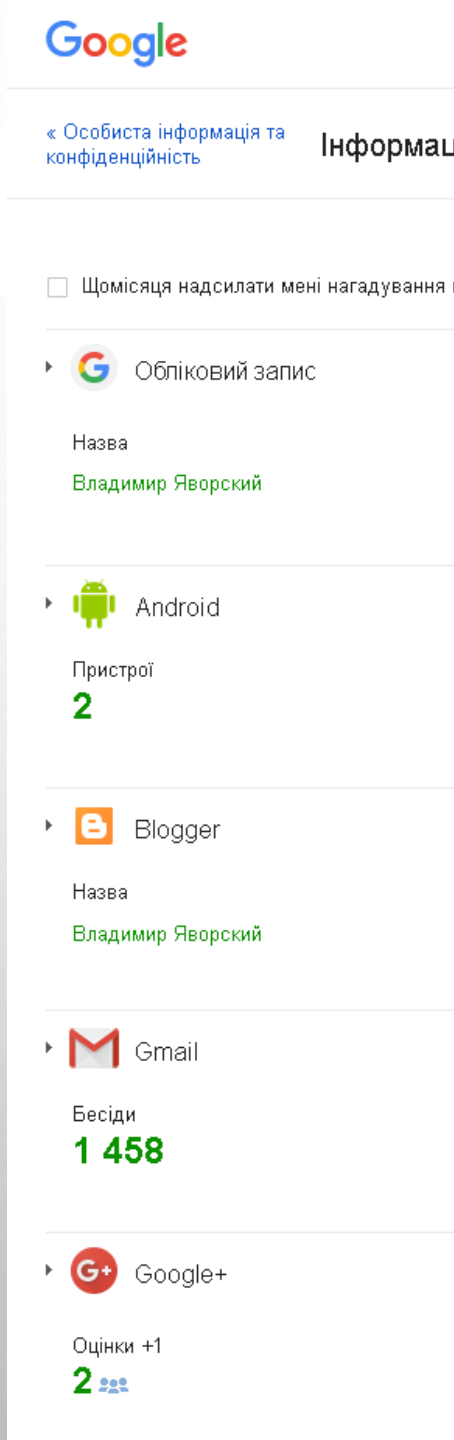
aboutme.google.com

myactivity.google.com/myactivity

myaccount.google.com/dashboard

www.google.com/maps/timeline?hl=uk&pb

adssettings.google.com



Ка н,
Ка
Ка
Ка
Ка
Ка
Ка она,
Ка
Вз
И али
со

YOUTUBE.COM/WATCH?V=H6_1GPRIHKQ

ПОШУКОВИЙ СПАМ «ЧОРНІ» МЕТОДИ

- Сторінка з великою частотою пошукового слова або словосполучення
- Прихований, дрібний і невидимий текст приховані малюнки
- Автоматична переадресація
- Неадекватні ключові слова і опис
- Велика кількість однакових сторінок однієї тематики

ПОШУКОВИЙ СПАМ «ЧОРНІ» МЕТОДИ

- **flood («затоплення» пошукової системи) - індексування сторінки під різними мережевими іменами**
- **перевищення числа сторінок в заявці на індексацію пошуковими краулерами**
- **дорвей (doorway pages) - сторінки, що містять розрізнені набори ключових слів на найрізноманітніші теми, створені для пошукових роботів.**

ПОШУКОВИЙ СПАМ «ЧОРНІ» МЕТОДИ

- **своппінг (swapping) - оптимізація сторінок для досягнення верхніх позицій результатах пошуку з наступною заміною змісту**
- **клоакинг (cloaking) - програмне забезпечення на сервері здатне розпізнавати роботів пошукових систем і підставляти їм не той зміст сторінок, яке побачать відвідувачі**

не помітить пошуковий робот - помітять відвідувачі! І поскаржаться адмінам ...

ПОШУКОВИЙ СПАМ «СІРІ» МЕТОДИ

- Сторінки з швидким оновленням
- Дзеркала сайту
- «Пошуково-активні» вхідні сторінки
- Мережі обміну посиланнями
- Багаторівневий маркетинг (MLM-технології)
- «Стратегічний» спам для індексу цитування

ОНОВЛЕННЯ АЛГОРИТМІВ ПОШУКУ GOOGLE 2011-2015 РОКИ

Panda (лютий 2011)

значне поліпшення алгоритму пошуку, яке спрямоване на підвищення якості контенту веб-сайтів. Оригінальні сайти з авторським контентом в пошуковій системі повинні зайняти місце вище, ніж сторінки з низькою якістю, що повторюють те, що вже і так відомо або ж є копіями інших сайтів.

- **достатній вміст** на сторінці повинен мати обсяг **>1500** слів;
- інформація, представлена на сайті повинна бути оригінальною. Якщо ви просто **копіюєте вміст** інших веб-ресурсів - Google покарає;
- **оригінальність** - для успішного просування контент має бути те, чого немає на інших сайтах;
- текст сайту повинен бути орфографічно і граматично **правильним**, зміст повинен відповідати описаним стандартам.

ОНОВЛЕННЯ АЛГОРИТМІВ ПОШУКУ GOOGLE

Page Layout (січень 2012)

Покарання сайтів, які використовують **занадто багато реклами** *у верхній частині сторінки або роблять її надмірно агресивною, що відволікає від основного змісту (користувачам складно знайти потрібну інформацію і доводилося довго прокручувати сторінку вниз, велика кількість реклами заважає зручності засвоєння інформації).*

Penguin (березень 2012)

Боротьба з **пошуковим спамом***. Сайти, які використовували спам-методи, були значно знижені в рейтингу або зовсім вилучені з нього. Здатність аналізувати кількість посилань.*

Pirate (серпень 2012)

Зниження рейтингу сайтів, які порушують **авторські права** *та інтелектуальну власність. Для оцінки цих порушень, Google використовує систему запитів про порушення авторських прав, засновану на Digital Millenium Copyright Act.*

ОНОВЛЕННЯ АЛГОРИТМІВ ПОШУКУ GOOGLE

Exact Match Domain (EMD, вересень 2012)

Боротьба з доменами, схожими на MFA.
MFA (made-for-adsense) - домен, який створений спеціально для контекстно-медійної системи Google. Зазвичай такий домен призначений для якогось одного запиту (або сімейства запитів) і на ньому встановлений Google Adsense. Користувач, який потрапив на цей домен, не бачить нічого, крім реклами і або закриває сайт, або переходить далі по контекстному оголошенню.

Payday Loan (червень 2013)

Зменшення кількості сторінок, які містять переспамлені запити.
Приклад: вам потрібно купити двері. На запит Google видасть фотографії дверей. З них: 2-3 сторінки, де безпосередньо можна купити двері, 3-4 сайту компаній-виробників дверей і 2-3 сайту про те, як вибрати і поміняти двері. Якби не було оновлення Payday Loan, ви б побачили 15-20 запитів на одну тематику (наприклад, де купити двері).

ОНОВЛЕННЯ АЛГОРИТМІВ ПОШУКУ GOOGLE

Hummingbird (вересень 2013)

Аби повертати точні відповіді на запити із ключовими словами, Google інтерпретує наміри і **контекст пошуку***. Мета полягає в тому, щоб зрозуміти сенс пошукового запиту користувача і повертати відповідні результати. Це означає, що точні співпадіння ключових слів стають менш важливими на користь пошуку наміри.*
Приклад: якщо ви вводите запит «погода», то навряд чи очікуєте отримати повне пояснення самого терміна, натомість отримаєте опис погодних умов.

Pigeon (липень 2014)

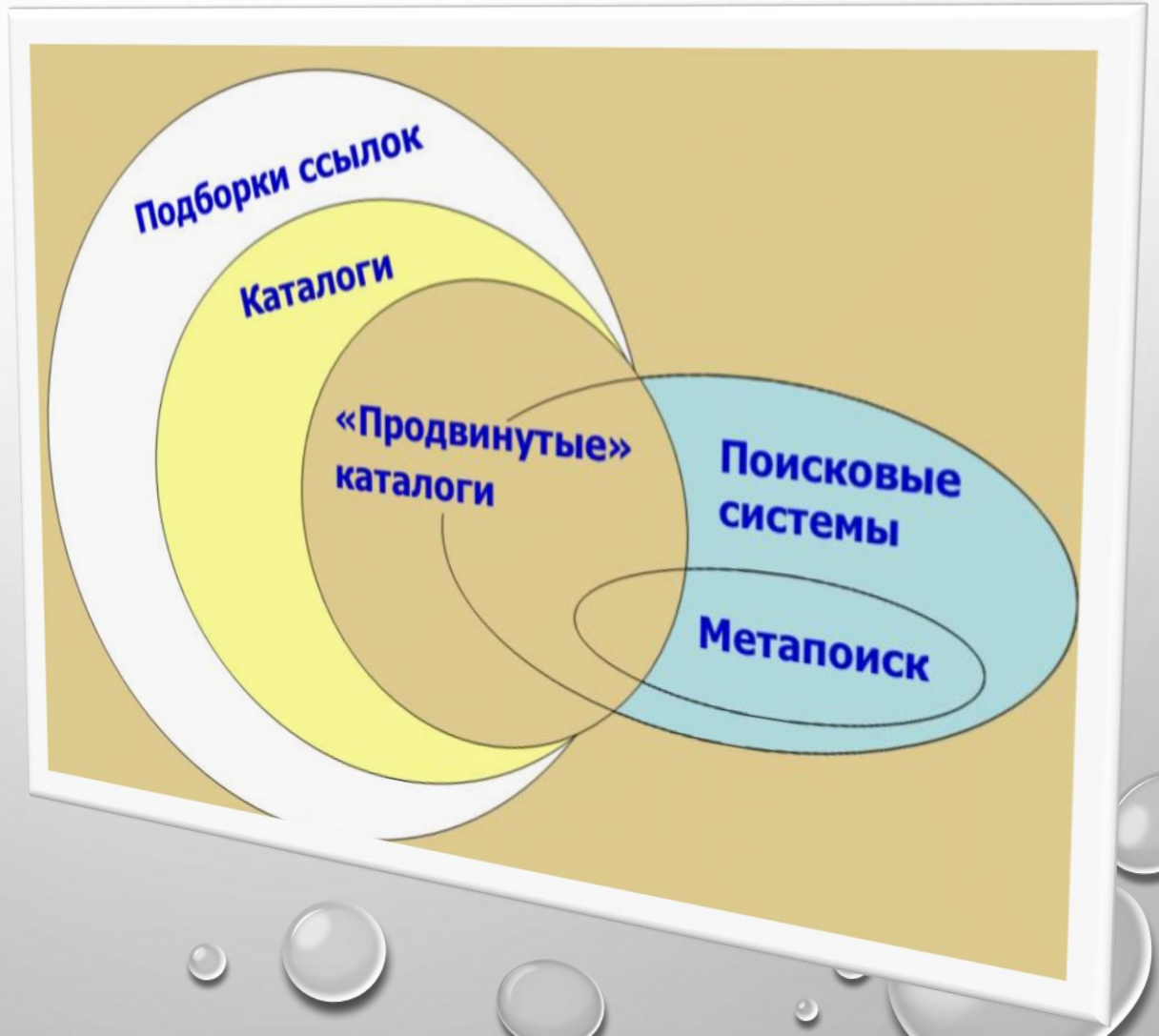
Геозалежний пошук. Відстань і **місце розташування користувача** *є ключовими параметрами ранжування, аби забезпечити точність результату.*

Mobilegeddon (квітень 2015)

Мобільний пошук*. Google дає перевагу сторінкам, дружнім до мобільних пристроїв.*

ТИПИ ПОШУКОВИХ СИСТЕМ

- Системи з пошуковими роботами
- Каталоги ресурсів (керовані людиною)
- Довідникові ресурси
- Локальні програми для пошуку в інтернеті
- Гібридні системи
- Метасистеми



ВЕБ-КАТАЛОГИ

- **Принципово: в наповненні ресурсами беруть участь люди, а не автоматичні пошукові програми.**
- **Рекомендовано для першого знайомства з предметною областю**
- **Рекомендовано для пошуку по нечітким запитам**
- **Недоліки:**

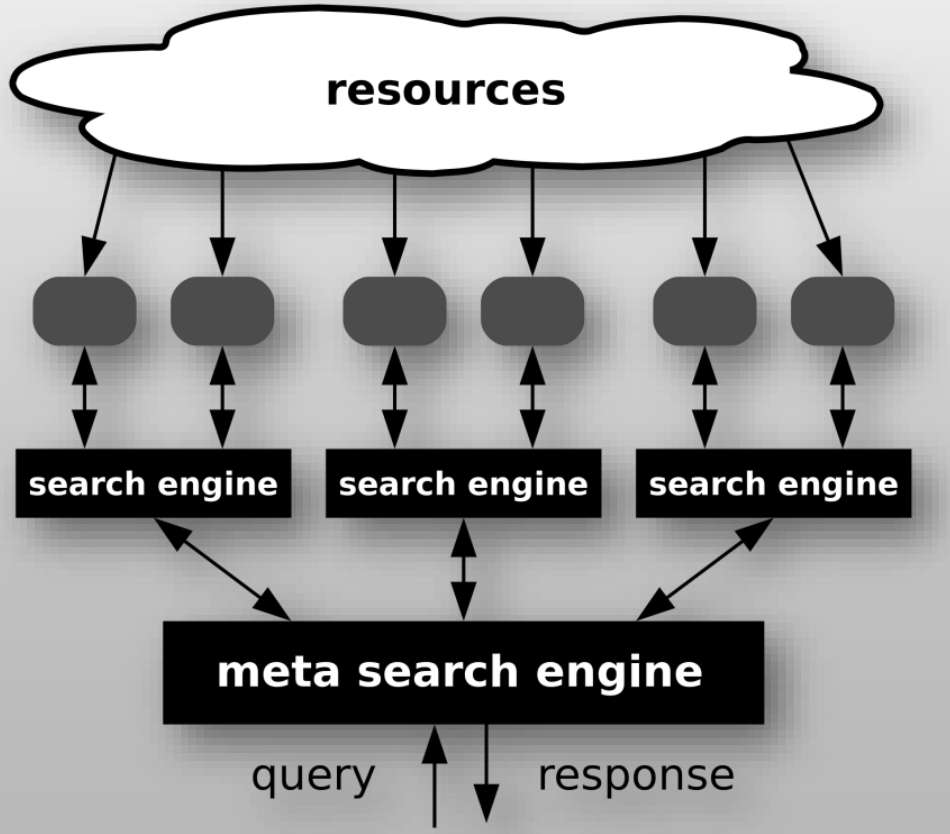
  - Слабка оперативність

  - Замалий розмір баз

  - Відсутня єдина класифікація ресурсів і чіткі критерії віднесення до категорій

МЕТАПОШУКОВІ СИСТЕМИ (SEARCHBOTS, METACRAWLERS)

Замість свого пошуку та індексування інформації, метасистеми надсилають запити багатьом пошуковим ресурсам.



ЗАВДАННЯ САМОКОНТРОЛЮ

- ВИКОНАТИ ПОШУК ГЛОБАЛЬНОГО РЕЙТИНГУ МЕТАПОШУКОВИХ СИСТЕМ

SEARX.ME

searx.me/preferences

searx.me о сайте настройки

Настройки

Общие | Поисковые системы | Плагины | Ответчики | Cookie

общие | файлы | картинки | it | карты | музыка | новости | наука | социальные сети | в

Разрешить	Имя поисковой системы	Сокращение	Выбранный язык	Безопасный поиск	Временной диапазон	Сред врем
○	archive is	ai	не поддерживается	не поддерживается	не поддерживается	0.236
●	wikipedia	wp	поддерживается	не поддерживается	не поддерживается	0.261
●	bing	bi	поддерживается	не поддерживается	не поддерживается	0.409
●	currency	cc	не поддерживается	не поддерживается	не поддерживается	N/A
○	ddg definitions	ddd	поддерживается	не поддерживается	не поддерживается	0.216
○	erowid	ew	не поддерживается	не поддерживается	не поддерживается	1.161
●	wikidata	wd	поддерживается	не поддерживается	не поддерживается	0.775
●	duckduckgo	ddg	поддерживается	не поддерживается	поддерживается	0.662
○	etymonline	et	не поддерживается	не поддерживается	не поддерживается	0.902
○	faroo	fa	не поддерживается	не поддерживается	не поддерживается	N/A
○	gigablast	gb	не поддерживается	поддерживается	не поддерживается	N/A
●	google	go	поддерживается	не поддерживается	поддерживается	0.604

МЕТАПОШУКОВІ СИСТЕМИ

https://www.startpage.com/

- **цільові ресурси з кількох пошукових систем**
- **анонімність серфінгу - комп'ютери клієнтів не зв'язуються з сайтами безпосередньо.**
- **політика конфіденційності Startpage: IP-адреси, відомості про відвідувані порталах, ключові терміни, куки не зберігаються. За це нагороджений першим Європейським Знаком Конфіденційності (EuroPriSe) 14 липня 2008 року**

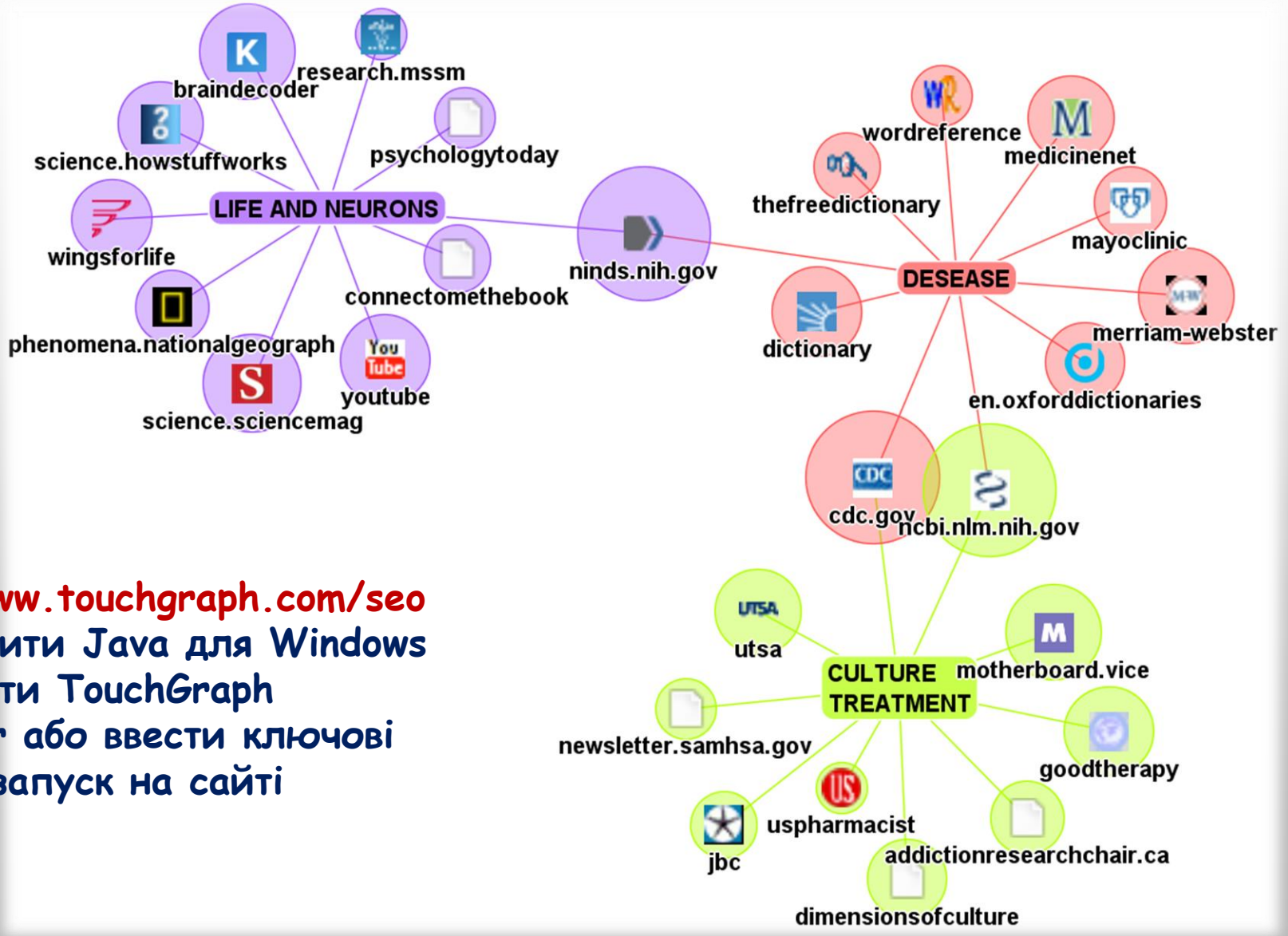
МЕТАПОШУК

- ✓ Не тільки видача результатів різних пошукових систем
- ✓ + кластерізація результатів пошуку

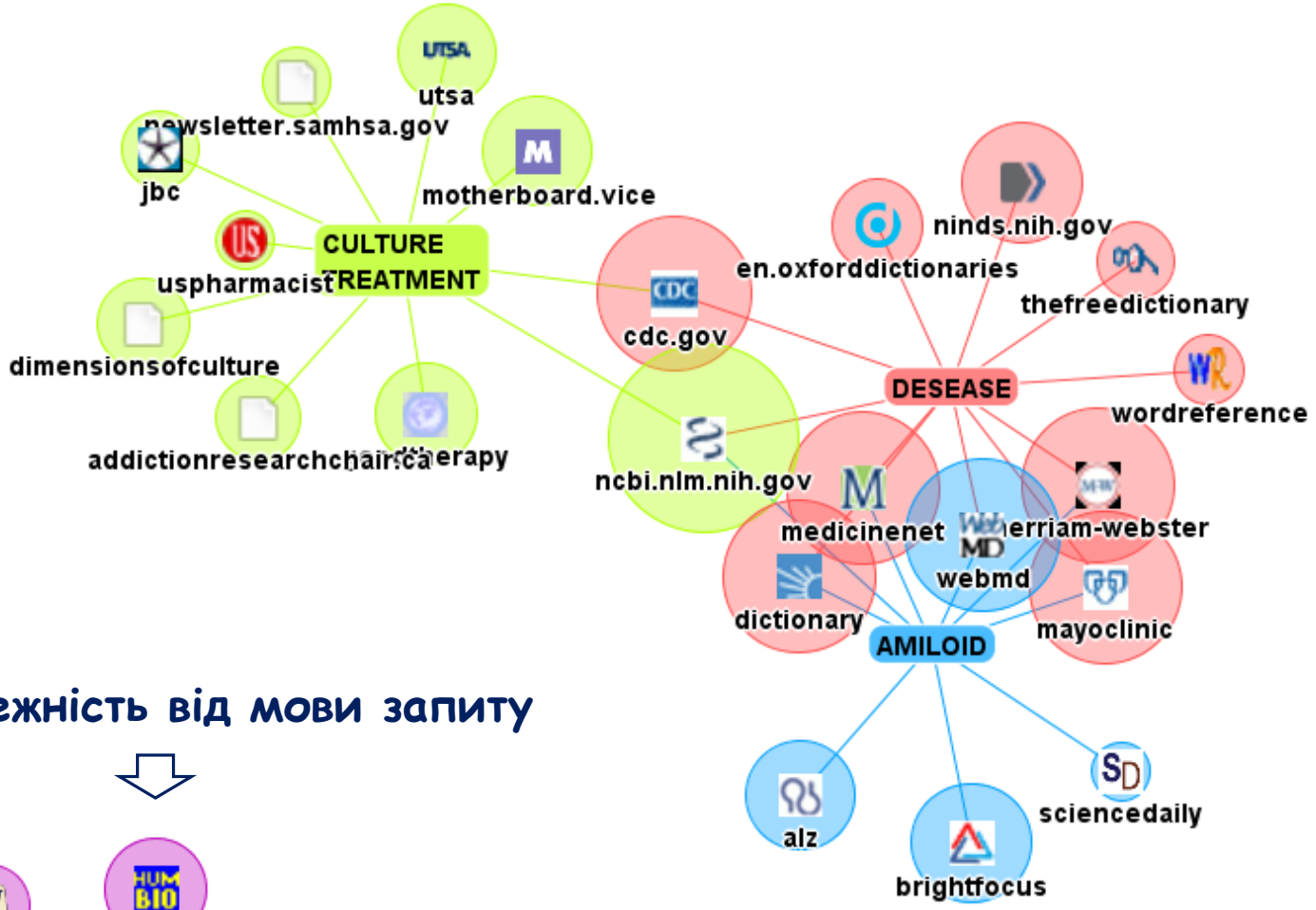
Етапи:

- Пошук веб-сторінок по запиту;
- Аналіз знайдених сторінок, знаходить додаткові ключові слова, які зустрічаються разом з термінами у запиті;
- Підмножини сторінок в результаті аналізу оголошуються кластерами;
- Визначення релевантності посилань і їх позиції в результатах кожного окремого кластера, посилання в межах кластера мають вищу ціну;
- Ранжування виділяє в кластерах корисні ресурси, яким при звичайному пошуку мало що «світить»

ВІЗУАЛЬНІ ПОШУКОВІ СИСТЕМИ



- http://www.touchgraph.com/seo
- Завантажити Java для Windows
- Встановити TouchGraph Navigator або ввести ключові слова + запуск на сайті



Висока залежність від мови запиту

Завдання 4:

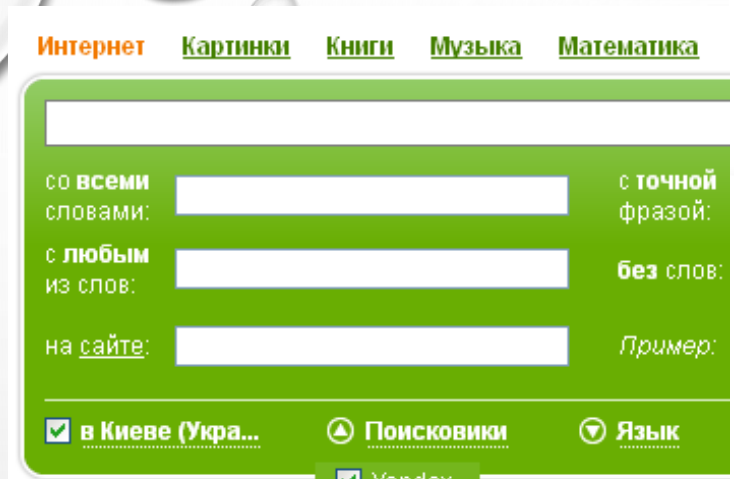
1. **Створити інтелект-карту сайтів, які містять інформацію по Вашій темі дисертаційної роботи**

- Підготувати ключові слова
- Налаштувати візуальну пошукову метасистему
- Провести пошук по кількох варіантам, обрати найкраще
- Зберегти зображення та надіслати jva@biph.kiev.ua

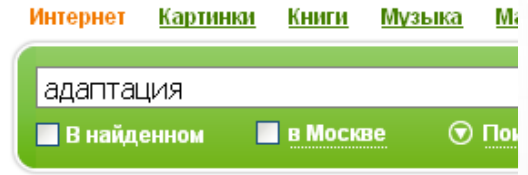
NIGMA.RU

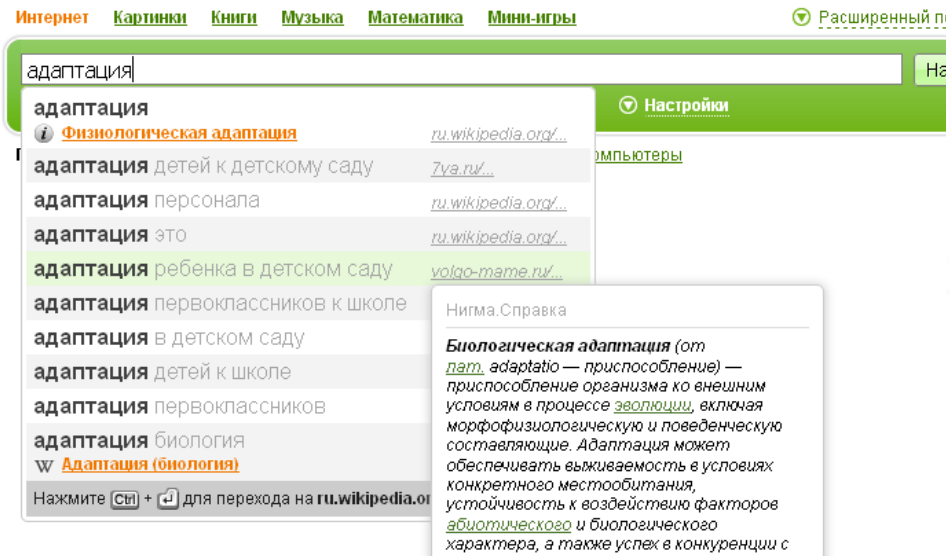
интеллектуальная поисковая система

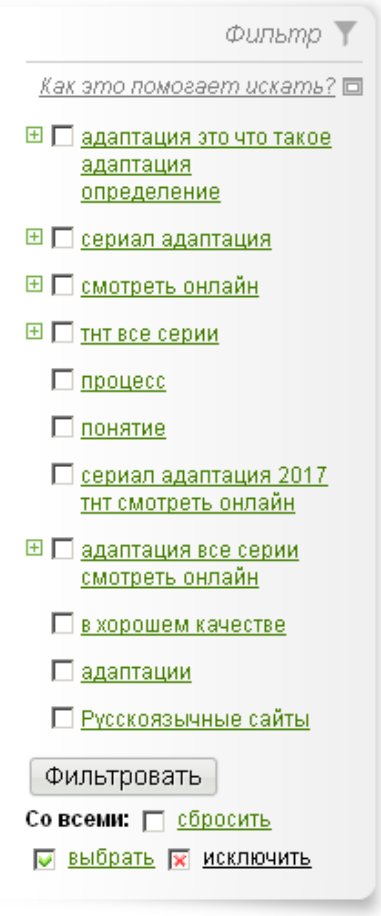
- Власний алгоритм кластеризації результатів
- Врахування специфіки російськомовних запитів
  - Пошук по різним словоформам, синонімам та узагальненим поняттям
  - Потужна система виправлення орфографічних помилок
- Автоматичне доповнення введених запитів та переклад
- Рішення математичних, хімічних рівнянь
- Структурована інформація по запиту

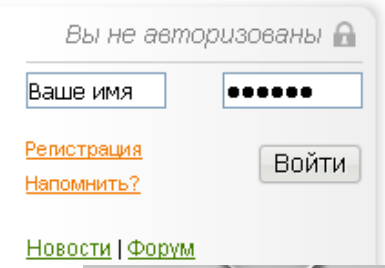

Интернет
Картинки
Книги
Музыка
Математика
со всеми словами:
с точной фразой:
с любым из слов:
без слов:
на сайте:
Пример:
в Киеве (Укра...
Поисковики
Язык
Yandex
Google
Rambler
Bing
Yahoo
Altavista
Nigma

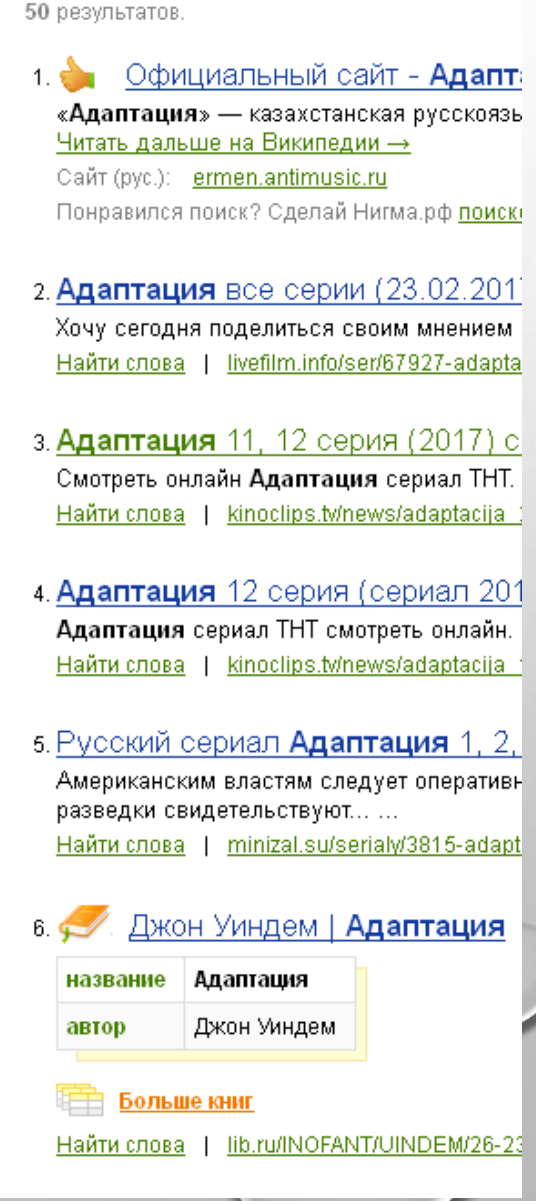

НИГМА.РФ


Интернет
Картинки
Книги
Музыка
адаптация
В найденном
в Москве


Интернет
Картинки
Книги
Музыка
Математика
Мини-игры
адаптация
Настройки
адаптация
Физиологическая адаптация
ru.wikipedia.org/...
адаптация детей к детскому саду
7ya.ru/...
адаптация персонала
ru.wikipedia.org/...
адаптация это
ru.wikipedia.org/...
адаптация ребенка в детском саду
volgo-mame.ru/...
адаптация первоклассников к школе
адаптация в детском саду
адаптация детей к школе
адаптация первоклассников
адаптация биология
Адаптация (биология)
Нажмите Ctrl + для перехода на ru.wikipedia.or
Нигма.Справка
Биологическая адаптация (от лат. adaptatio — приспособление) — приспособление организма ко внешним условиям в процессе эволюции, включая морфофизиологическую и поведенческую составляющие. Адаптация может обеспечивать выживаемость в условиях конкретного местообитания, устойчивость к воздействию факторов абиотического и биологического характера, а также успех в конкуренции с


Фильтр
Как это помогает искать?
адаптация это что такое адаптация определение
сериал адаптация
смотреть онлайн
тнт все серии
процесс
понятие
сериал адаптация 2017 тнт смотреть онлайн
адаптация все серии смотреть онлайн
в хорошем качестве
адаптации
Русскоязычные сайты
Фильтровать
Со всеми:
сбросить
выбрать
исключить


Вы не авторизованы
Ваше имя
Регистрация
Напомнить?
Войти
Новости | Форум


50 результатов.
1. Официальный сайт - Адапт
«Адаптация» — казахстанская русскоязы
Читать дальше на Википедии →
Сайт (рус.): ermen.antimusic.ru
Понравился поиск? Сделай Нигма.рф поиск
2. Адаптация все серии (23.02.201
Хочу сегодня поделиться своим мнением
Найти слова | livefilm.info/ser/67927-adapta
3. Адаптация 11, 12 серия (2017) с
Смотреть онлайн Адаптация сериал ТНТ.
Найти слова | kinoclips.tv/news/adaptacija
4. Адаптация 12 серия (сериал 201
Адаптация сериал ТНТ смотреть онлайн.
Найти слова | kinoclips.tv/news/adaptacija
5. Русский сериал Адаптация 1, 2,
Американским властям следует оперативн
разведки свидетельствуют... ...
Найти слова | minizal.su/serialy/3815-adapt
6. Джон Уиндем | Адаптация
название Адаптация
автор Джон Уиндем
Больше книг
Найти слова | lib.ru/INOFANT/UINDEM/26-23

Интернет Картинки Книги Музыка Математика Мини-игры

C2H5OH + CH3COOH

В найденном | в Киеве (Украи... | Поисковики | Язык | Сортировка | Настройки

54 млн. результатов.

Реакция:

(видео: почему происходят химические реакции)

Реакция образования сложных эфиров

Условия:

В присутствии сильной кислоты. Реакция идёт до конца при наличии водоотнимающего агента (напр.

Больше реакций

Реакция:

$CH_3COOH + C_2H_5OH = CH_3COOC_2H_5 + H_2O$

Реакция:

$CH_3COOC_2H_5 + H_2O = CH_3COOH + C_2H_5OH$

Условия: В присутствии H^+

Список решаемых задач

1. Ответы@Mail.Ru: C2H5OH+CH3COOH= помогите пожалуйста))

C2H5OH+CH3COOH = CH3COOC2H5 (этилацетат) + H2O. Следует указать что условием реакции этерификации является кислая среда! ...

Найти слова | otvet.mail.ru/question/75713978 951 б

Интернет Картинки Книги Музыка Математика Мини-игры

log(x+10)*(20-x)=0

В найденном | в Киеве (Украи... | Поисковики | Язык | Сортировка | Настройки

125 млн. результатов.

Дано:

$$\ln(x+10)\cdot(20-x)=0$$

Скрыть Решение

1 ОДЗ уравнения:

$$x \in (-10, \infty)$$

2 Делаем преобразование левой части уравнения:

$$\ln(x+10)\cdot(20-x) = -(x-20)\cdot\ln(x+10)$$

3 Уравнение после преобразования:

$$-(x-20)\cdot\ln(x+10)=0$$

4 Решаем уравнение:

$$x = 20$$

5 Решаем уравнение:

$$-x = 9$$

6 Возможные решения:

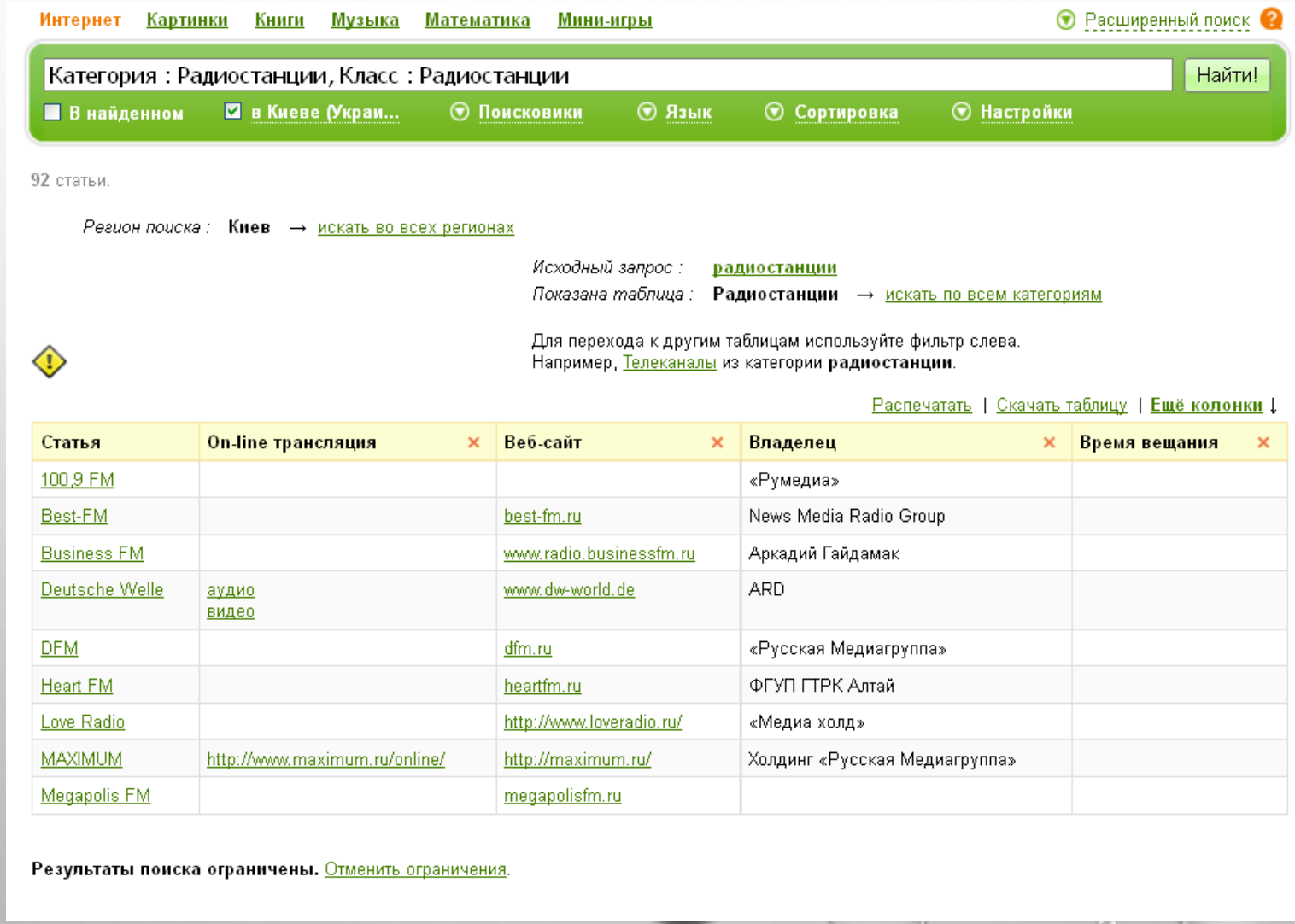
$$-9;$$
$$20;$$

Ответ: (Решение уравнения с учётом ОДЗ)

$$x = -9,$$
$$x = 20$$

Что это такое? Список решаемых задач Пожаловаться

ТАБЛИЧНИЙ NIGMA-ПОШУК



Интернет Картинки Книги Музыка Математика Мини-игры Расширенный поиск

Категория : Радиостанции, Класс : Радиостанции Найти!

В найденном в Киеве (Украи... Поисковики Язык Сортировка Настройки

92 статьи.

Регион поиска : **Киев** → искать во всех регионах

Исходный запрос : радиостанции
Показана таблица : **Радиостанции** → искать по всем категориям

Для перехода к другим таблицам используйте фильтр слева.
Например, Телеканалы из категории **радиостанции**.

Распечатать | Скачать таблицу | Ещё колонки ↓

Статья	On-line трансляция	Веб-сайт	Владелец	Время вещания
100,9 FM			«Румедиа»	
Best-FM		best-fm.ru	News Media Radio Group	
Business FM		www.radio.businessfm.ru	Аркадий Гайдамак	
Deutsche Welle	аудио видео	www.dw-world.de	ARD	
DFM		dfm.ru	«Русская Медиагруппа»	
Heart FM		heartfm.ru	ФГУП ГТРК Алтай	
Love Radio		http://www.loveradio.ru/	«Медиа холд»	
MAXIMUM	http://www.maximum.ru/online/	http://maximum.ru/	Холдинг «Русская Медиагруппа»	
Megapolis FM		megapolisfm.ru		

Результаты поиска ограничены. Отменить ограничения.

ТАБЛИЧНИЙ NIGMA-ПОШУК