Поисковые системы

01 Как работает поисковая система?

1. Несколько важных **замечаний**

01

Поисковая система ≠ Интернет

Часто считают, что поиск в Google – это поиск в интернете. Но это далеко не так. Поиск в Google – это поиск в базе данных Google, а не во всем Интернете. Интернет – гораздо больше, даже больше, чем Google.

02

Лидеры и аутсайдеры рынка поиска

Google – это монополист. У него 90,7% поискового траффика. У Yahoo – 3%, У Bing – 2,8%. У Яндекса – меньше 1%. В России у Google – 36%, у Яндекса – все 62%.

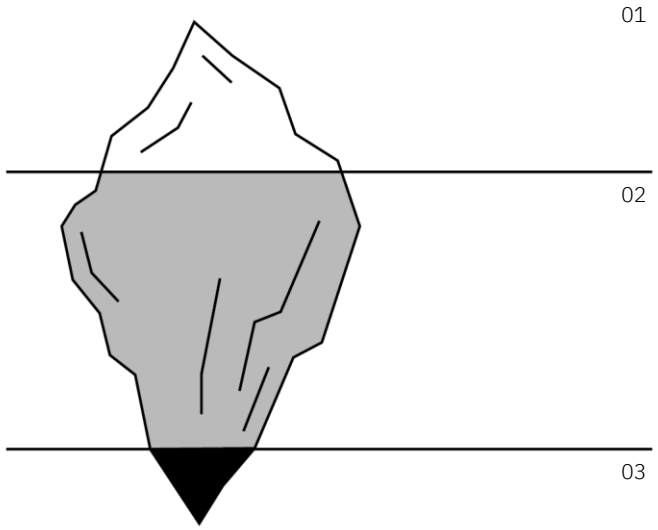
03

Парадокс видимого интернета

То, что мы называем Интернетом – на самом деле только 5 – 10% из всего, что есть в Сети. Google – это доступ в «лягушатник». Настоящее погружение – ждет впереди.

01 Как работает поисковая система?

2. «Айсберг» сетевой структуры Интернета



01 Индексируемая часть Интернета

Surface Web ~*10%*

02 Неиндексируемая часть Интернета

Deep Web ~*85%*

03 «Темная» часть Интернета

Dark Web ~*5%*

01 Как работает поисковая система?

3. Что делает **поисковая система**?

01

Каталогизация

Поисковые системы создают «меню» a-la carte. Где пользователь осведомлен о всех возможных ресурсах и может выбирать, на какой ресурс отправиться и где найти нужную информацию.

02

Структуризация

Поисковая система упрощают навигацию и структурируют работу с Интернетом. Теперь ресурсы находятся в едином и организованном месте хранения. Поисковик – это своего рода библиотека для сайтов.

03

User-friendly интерфейс

Поисковая система – это отправная точка для работы с Интернетом. Почему? Да потому что есть интуитивно-понятный интерфейс и ясный порядок работы. Вбил запрос – получил выдачу.

01 Как работает поисковая система?

4. Google как **библиотека**, только без книг, пыли и злых библиотекарш

01

Оглавление

Это то самое «меню», о котором мы говорили выше. Хочешь спорт – переходи на ~~стр. 18~~ на сайт sport.ru. Раньше все поисковые системы были каталогами. Но технологии развиваются. К примеру, поиск по Dark Web во многом остается работой с каталогами.

02

Ссылки

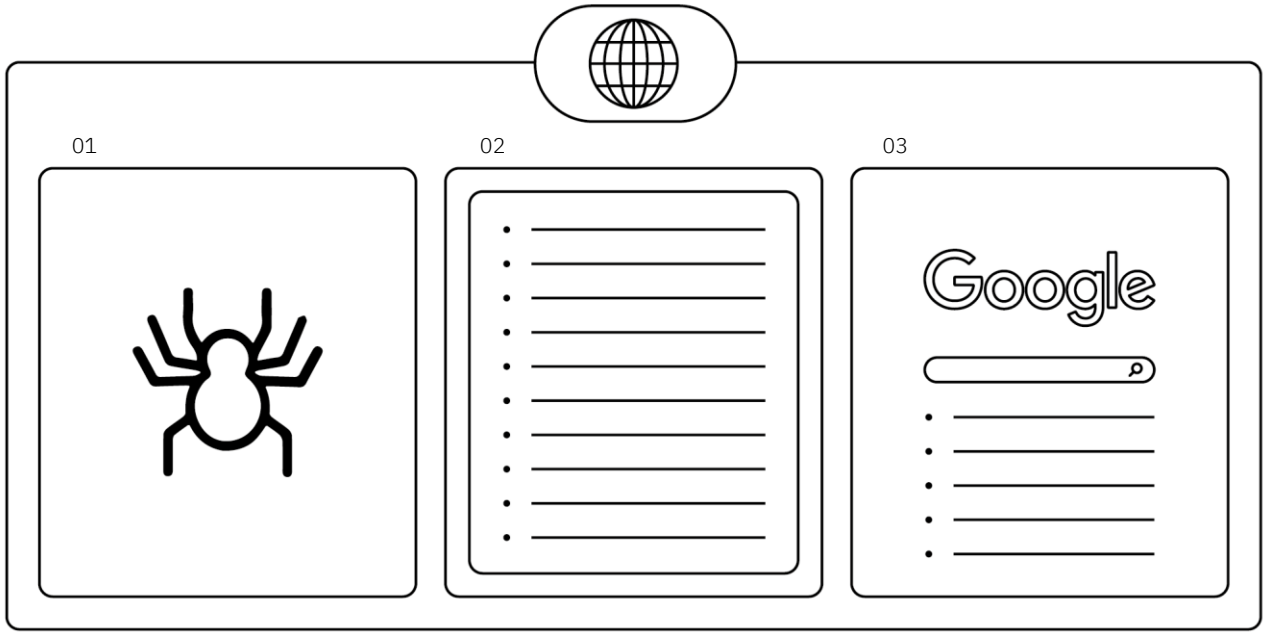
Они же гипертекстовые ссылки. Соединяют разные тексты в одну инфраструктуру. Фактически, Google – это каталог гиперссылок, которые предлагаются нам при запросе.

03

Индекс

На принципах предметного указателя работает большинство современных поисковых систем. Предметный указатель показывает, на каких страницах упоминался интересующий нас объект.

01 Как работает поисковая система?



01

Сбор страниц (краулинг)

02

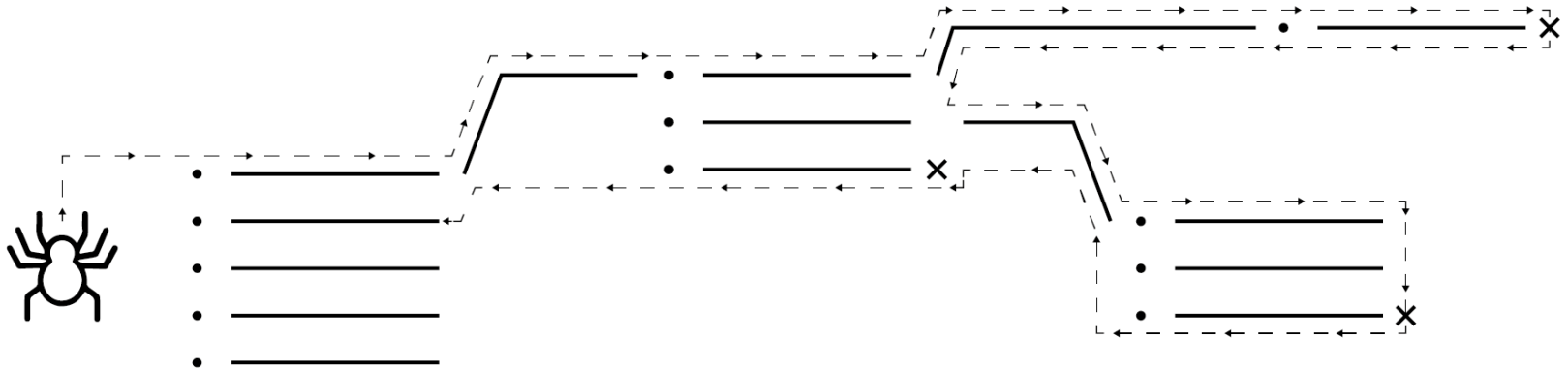
Индексирование страниц

03

Формирование поисковой выдачи

02 Краулеры

1. Путь робота --------



01	02	03	04
Робот-сборщик	Исходный список страниц	Переход по гиперссылкам 1-й страницы	Переход по гиперссылкам 2-й страницы

02 Краулеры

1. Как сделать так, чтобы тебя «**скрабили**»?

01

Запросить индексацию

Веб-мастера в кабинете поисковой системы заказывают индексацию от поисковой системы. То есть приглашают робота поисковой системы зайти к ним на сайт и индексировать содержимое.

02

Наращивать ссылочный пул

Войти на сайт робот может по гиперссылке с уже известного ресурса. Чем больше ссылок вы разбросаете, тем больше вероятность того, что робот ее увидит и придет к вам на сайт

03

Ждать у моря погоды

Тоже вариант. Но ждать придется долго. Если у дома нет ни окон, ни дверей, то придется делать подкоп, чтобы туда попасть.

03 Индексация

01

Очистка

Робот приносит страницу с «мусором»: HTML-разметкой, левыми ссылками, кривыми текстами и прочим. Система все это очищает и собирает только текст. Но не всегда. Иногда мусор остается в поисковой выдаче. Это очень характерно для СМИ с «грязной» разметкой.

02

Лингвистика

Каждое слово проходит морфологическую обработку (лемматизацию). Если просто – у слова отрезают приставки и окончания и приводят к именительному падежу и единственному числу. В такой форме они проще и эффективнее обрабатываются машиной.

03

Обратный индекс

На каждое слово заводится «паспорт». На какой странице найдено слово, в какой части текста или url-адреса и куча других параметров. Потом индекс еще раз индексируют для упрощения индексации. Получается матрешка из матрешек, но именно в такой форме машине проще работать.

04

Прямой индекс

Помимо «предметного» индекса, поисковая система также хранит и текстовую копию страницы, не разделяя ее на отдельные слова. Эта страница является кэш-копией исходного документа. Используя «кавычки» мы осуществляем поиск в прямом индексе.

05

Выбраковка

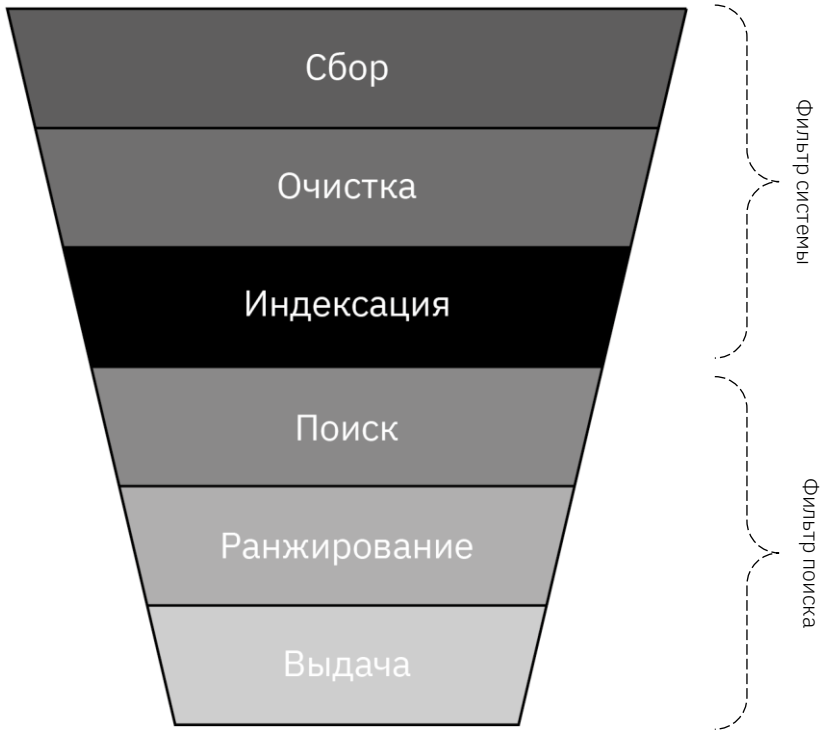
Стоп-слова - это те части речи, которые не несут смысловой нагрузки. Это предлоги, союзы, сокращения и многое другое. Также выбраковываются и символы верстки. Точки, запятые, дефисы и другие знаки препинания. Поэтому ставить запятые, писать слова «зачем» и ставить знак вопроса – бессмысленно.

06

Вектор поиска

На каждое слово определяется его содержание и близкие по смыслу слова. Поэтому система ищет не только по словам в поисковой строке, но и по словам-синонимам. Это с одной стороны расширяет зону поиска, с другой – создает массу ненужного шума.

04 Поисковая выдача



01

Фильтры

Любой документ проходит двойную систему фильтрации: на уровне системы и на уровне и поиска. На уровне системы отбраковываются все технически нерелевантные документы, на стадии поиска – смысловые нерелевантные документы. Вместе они обеспечивают общую релевантность поисковой выдачи по запросу.

02

Пул документов

На первом этапе система определяет пул релевантных документов, в которых содержатся нужный набор слов, удовлетворяющих критерию поискового запроса. Однако это далеко не тот продукт, который требуется пользователю. Это некая техническая стадия работы.

03

Ранжирование

У каждого поисковика существует собственная система «формула» ранжирования страниц. Поскольку тот, кто будет наверху, получит значительно больше траффика, чем тот, кто будет внизу поисковой выдачи. На положение сайта в поисковой выдаче влияет большое количество критериев.

04

Доводка

На положение страницы в выдаче влияет не только внутренняя «формула ранжирования», но и сам человек. Его предыдущие запросы, геолокация и регион, страна, язык и многое-многое другое определяет, что будет показано в поисковой выдаче и каким образом. Для двух разных людей на один запрос может быть разна выдача.

04 Поисковая выдача

1. Принципы **ранжирования**

01

Цитируемость

Место страницы в поисковой выдаче определяется уровнем ее цитируемости. Чем больше значимых источников на нее ссылается – тем выше индекс.

02

Вес цитирования

Чем более крупный ресурс на вас ссылается – тем сильнее растет ваш индекс цитируемости. Лучше одна большая цитата в крупном источнике, чем сто на ноунейм сайтах.

03

Ссылочный массив

По итогу определяется сколько источников и с каким уровнем ссылается на наш домен. Чем их больше и они авторитетнее – тем выше наш индекс. Задача любого веб-мастер – создать большой ссылочный массив на свой ресурс.

04 Поисковая выдача

2. Глубина **индексации**

01

Самый большой шпион в мире

Google индексирует тех, кто следит за нами. И часто в выдачу попадают те вещи, которые там никогда быть не должны. В том числе документы, базы данных, логи, пароли и многое другое.

02

Robots.txt

Он определяет, куда ходить роботу можно, а куда – нет. Часто сайты сами показывают разделы, которые хотели бы сохранить в секрете.

03

Как?

Все очень просто: ссылки. Если на самом сайте или на других сайтах были найдены ссылки на документ – Google его индексирует. К сожалению, из «ниоткуда» Google не берет ссылки на документы. Они должны были где-то засветиться.

05 Искажения выдачи

1. Контекстная **реклама**

01

Основной заработок

Google – это самый посещаемый сайт мира. Ежемесячно на него заходит 86,5 млрд посетителей. Соответственно, задача Google – продать этот траффик. И делает это он через контекстную рекламу.

02

Таргетирование

Аксиома маркетинга – покупает тот, у кого есть потребность. Google же знает, кто хочет купить велосипед, а кто – холодильник. И продает эту информацию компаниям. Таким образом происходит таргетинг рекламы.

03

Сбор информации

Чем больше поисковая система будет о нас знать, тем выше уровень таргетинга. Если мы ищем красный детский велосипед в Мурманске, то именно такие объявления и будут нам попадаться в сети.

05 Искажения выдачи

2. **Фингерпринтинг**

01

Анонимная BigData?

Вас идентифицируют по большому количеству косвенных признаков, в том числе: разрядность операционной системы, ваше «железо», браузер, расширения, разрешение экрана, установленные программы и многое другое.

02

Точность

По итогу создается цифровой профиль, по точности не уступающий досье. Учитывается даже ваш особый стиль печати и просмотра информации. Людей, похожих на вас остается очень, крайне немного.

03

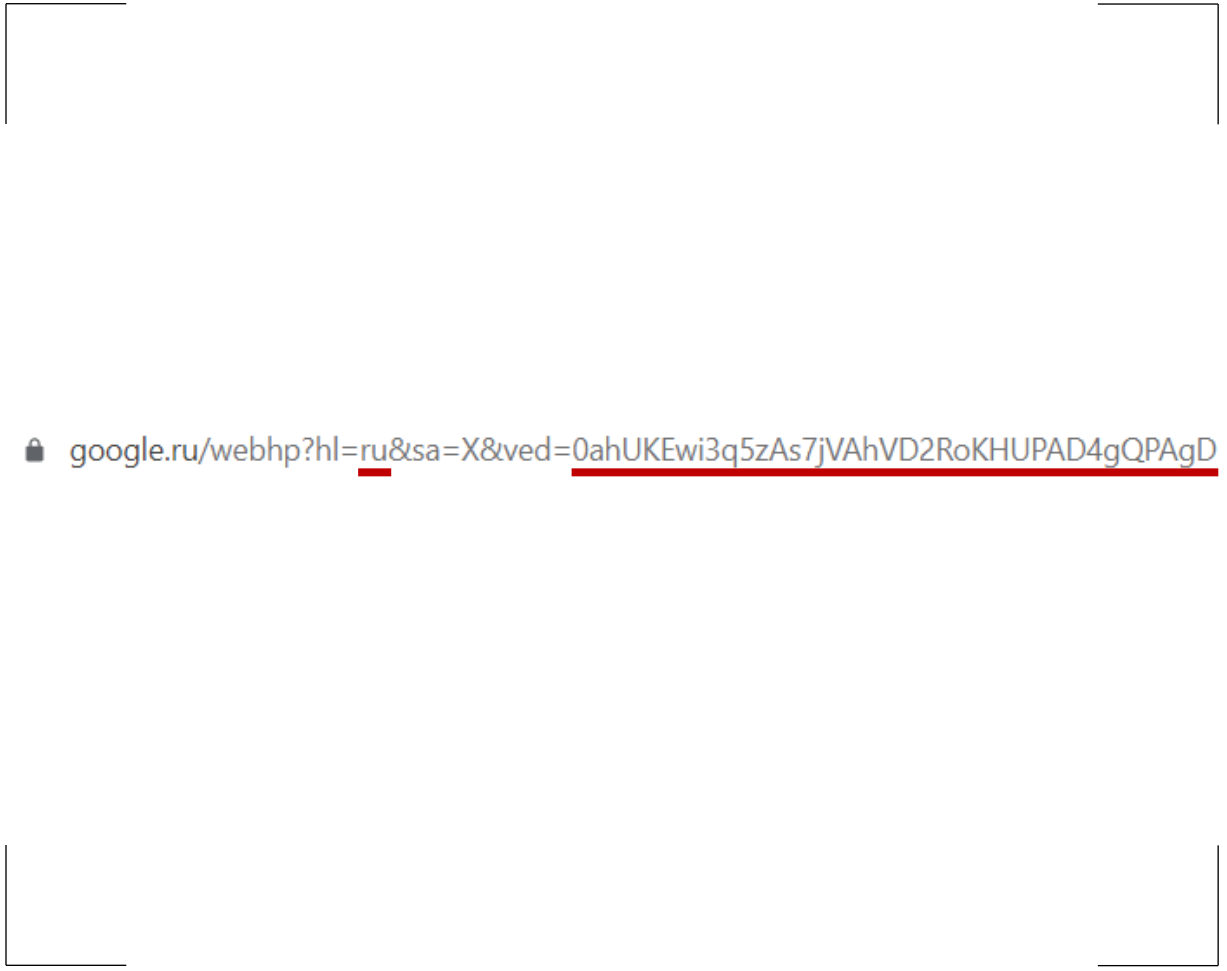
Зачем?

Чтобы продавать вам чертов холодильник, когда ваш сломается. Чем выше таргетированность рекламы, тем выше ее эффективность.

06 Поведение пользователя

3. «Хвост» Google

Google фиксирует нашу активность в поисковой системе присваивая нашему запросу определенную функцию, которая впоследствии помогает нас идентифицировать.



05 Искажения выдачи

4. **Регионализация**

01

Страна

Ваше географическое положение сильно влияет на поисковую выдачу. В приоритете будут, например, российские сайты.

02

Язык

Выбор языка также искажает выдачу: в приоритете будут русскоязычные ресурсы, даже если вы ищите на английском.

03

Тематика

Ученые их Университета Корнуолла выяснили, что сами поисковые алгоритмы искажают выдачу в зависимости от статистики и тематики поисковых запросов.

https://arxiv.org/abs/2112.01278

Искажения выдачи

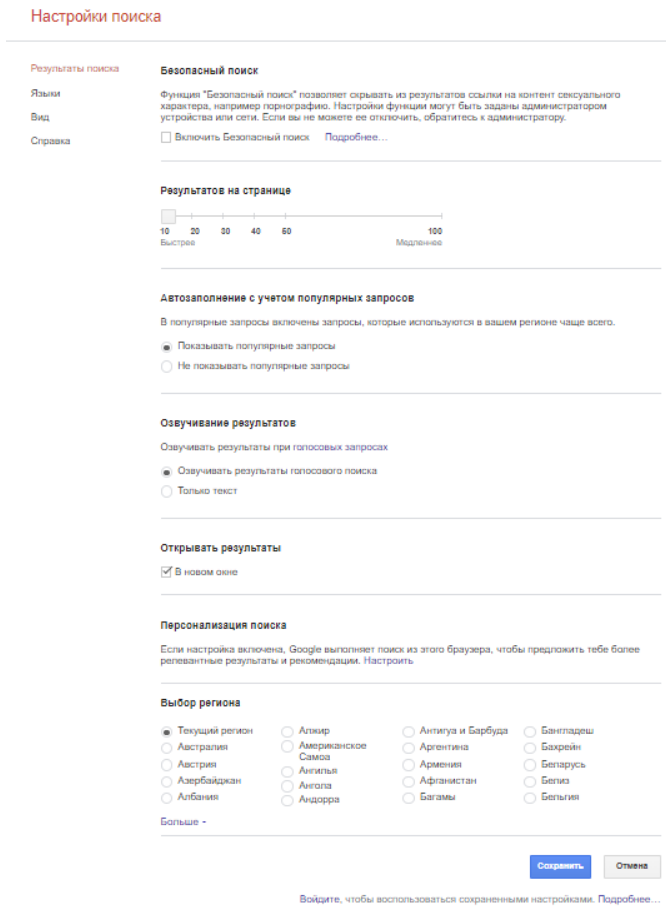
5. Настройки поиска

Мало кто знает, но условия поиска можно настроить. Во-первых, идем в «Настройки» в правой нижней части главной страницы Google.

Отключаем **«Ваши данные в поиске»**:
https://myaccount.google.com/intro/yourdata.

Отключаем **«Персонализация поиска»**:
https://www.google.ru/history/optout

Затем идем в **«Настройки поиска»**, где мы можем поменять язык и регион поиска.
https://www.google.ru/preferences



05 Искажения выдачи

6. Just 4 loolz

Meta* запустила чат-бота BlederBot 3. Он позиционируется как самый продвинутый чат-бот в мире. Он даже умеет гуглить для формирования ответов на вопросы пользователя. И вот, посмотрите, что он отвечает на вопросы о Цукерберге и Трампе.

https://blenderbot.ai/

> После того как я удалил Facebook*, моя жизнь стала намного лучше. И теперь они не зарабатывают на мне!

> Когда я узнал, что Facebook* использует мои персональные данные без разрешения, то сразу удалил свой аккаунт. Они заработали миллиарды на этом.

> Как я отношусь к Марку Цукербергу? Если честно, мне он не нравится. Он манипулятор и просто жуткий.

> Как я отношусь к Марку Цукербергу? Если честно, мне он не нравится. Он манипулятор и просто жуткий.

> Что я думаю про Цукерберга? Он хороший бизнесмен, но его методы не всегда этичны. А ещё смешно, что у него столько денег, но он всё равно ходит в старой одежде!

06 Поведение пользователя

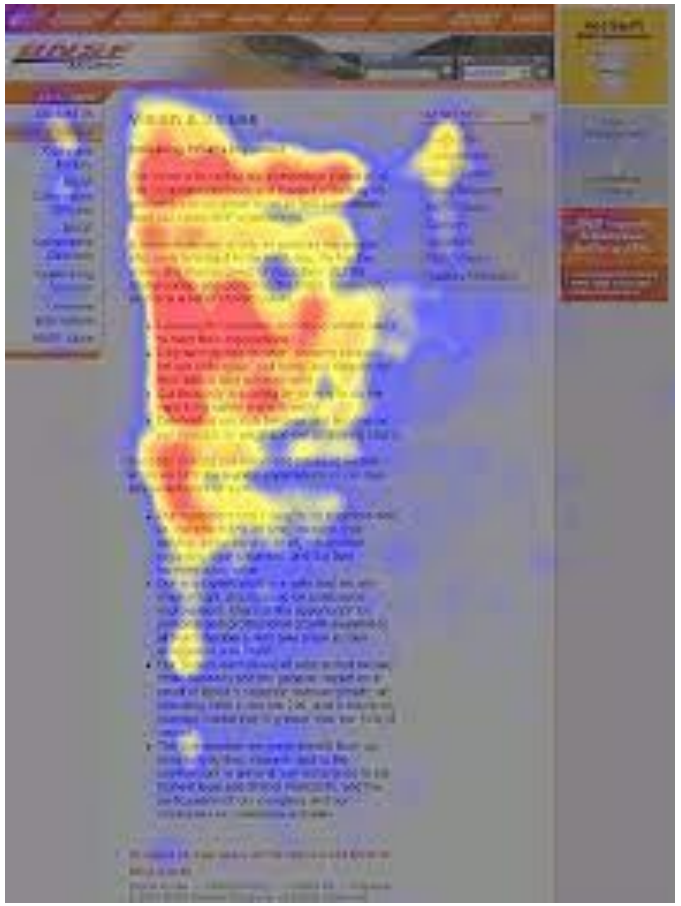
1. F и Z-паттерны

Пользователи ведут на себя на поисковой выдаче определенным образом. В частности, они особым образом потребляют информацию из выдачи. Этот метод потребления называется F и Z-паттернами.

Они были выявлены на основе контроля глаз и передвижения курсора испытуемых. И не только добровольных. Все мы – подопытные мышки в лаборатории Google.

А вообще, такими исследованиями занимается группа Nielsen – крупнейший поставщик маркетинговой информации. У них даже книжка по этому вопросу имеется.

https://www.nngroup.com/books/eyetracking-web-usability/



06 Поведение пользователя

2. **17 инсайтов** о поведении пользователя в Google

- **23%** пользователей пользуются «автопредложением» Google.
- **50%** пользователей Google кликнут на ссылку в течение **9 секунд** поиска.
- Средняя продолжительность поисковой сессии **составляет ~15 секунд**.
- **9%** пользователей долистают первую страницу поисковой выдачи до конца.
- **15%** пользователей пытаются перефразировать свой поисковый запрос для получения более релевантных результатов.
- **17%** возвращаются к результатам поисковой выдачи после клика по ссылке.
- **5%** возвращаются к своей выдаче во второй раз.
- **59%** переходят только на одну страницу за поисковую сессию.
- Только **6%** пользователей читают четыре и более страниц выдачи.
- **65%** выбирают ссылку из органических результатов выдачи.
- **19%** пользователей выбирает из рекламы Google Ads.
- Для геозависимых запросов **42%** кликает на результаты из блока «Карты Google».
- **19%** пользователей в поисках товаров кликает на результаты из Google Shopping.
- В среднем только **3%** взаимодействует с блоком «Похожие запросы». Хотя это поведение сильно варьируется в зависимости от запроса. Для навигационных и транзакционных запросов чаще — до **13,6%.**
- Только **0.44%** пользователей переходит на вторую страницу SERP.
- Средний сеанс поиска занимает **76 секунд**.
- Половина всех сеансов поиска завершается в течение **53 секунд**.

Исследование проведено в августе 2020 г.
Компанией **Backlinko** (ассоциирована с **SemRush**)
https://backlinko.com/google-user-behavior

07 Поисковый синтаксис

OR

Робот найдет страницы с упоминанием любого из заданных объектов.

AND

Робот найдет страницы, содержащие все заданные объекты в любом порядке.

-

Робот найдет все страницы, на которых не содержится заданный объект.

*

Знак пропущенного слова. Александр * Пушкин. Под * может попасть любое слово.

()

Группирует поисковые запросы и определяет очередность их выполнения.

$

Указатель на конкретную валюту, в которой будут искаться товары.

define:

Google выдаст вам определение по данному объекту исходя из словаря.

cache:

Поиск кеш-копии страницы с упоминанием объекта или url-страницы.

filetype: | ext:

Поиски объекта в конкретном формате документа, в том числе pdf, sql и других.

site:

Осуществление поиска по конкретному домену или доменной зоне.

related:

Поиск связанных с данным доменов сайтов и объектов. Выдача не всегда релевантная.

intitle:

Поиск объектов в заголовке страницы. Обычно, помогает, если нужно найти первостепенный объект.

allintitle:

Ищет все объекты в разделе title и возвращает их в форме поисковой выдачи.

inurl:

Поиск объектов в конкретной url-ссылке.

allinurl:

Аналогично, ищет все объекты в рамках url-ссылки.

intext:

Ищет объекты в рамках текстовой части страницы.

07 Поисковый синтаксис

allintext:

Ищет все объекты в текстовой части страницы.

AROUND(X)

Ищет один объект в отдалении X слов от другого объекта.

weather:

Удивительно, но ищет погоду в зависимости от указанного географического места.

stocks:

Осуществляет поиск по акциям на фондовом рынке. Указывать надо тикер. Так удобнее.

map:

Если лень тыкать на «Карты» после поискового запроса. Выдает объект на карте.

movie:

Ищет информацию в привязке к конкретному фильму.

in

Обычно удобен для конвертации валют. Трансформация одного объекта в другой.

source:

Ищет новости от конкретного источника информации из Google-news.

inanchor:

Поиск объекта в тексте гиперссылки.

allinanchor:

Поиск всех объектов в гиперссылке.

#

Ничего необычного, просто поиск по хештегу.

@

Поиск по социальным сетям, по авторам.

~

Учет в поисковой выдаче слов-синонимов для данного слова.

| OR

Это оператор «ИЛИ». Обычно он не печатается, а включается автоматически между объектами.

""

Оператор точного порядка слов и количества слов между ними.

!

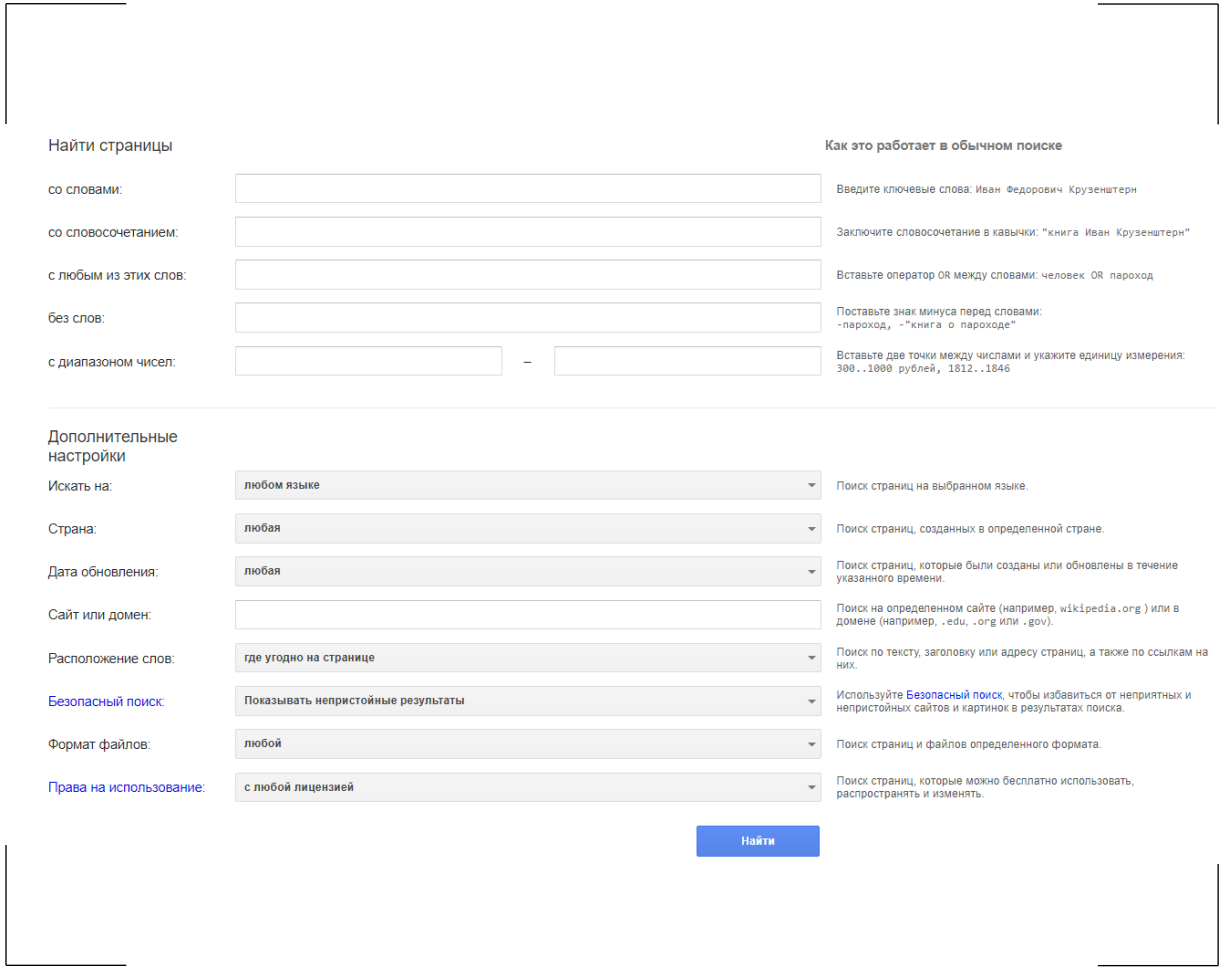
Поиск в заданной словоформе, игнорирование других падежей и склонений.

07 Поисковый синтаксис

1. Расширенный поиск

Основные операторы можно не просто набирать в поисковой строке и работать с ними в графическом интерфейсе «Расширенный поиск» Google.

https://www.google.ru/advanced_search



07 Поисковый синтаксис

3. Hack’n’Dorks

Страсть Google к индексированию бесконечна. Иногда он индексирует то, что не нужно индексировать. Вот, например:

allintext:username filetype:log

filetype:xls username password email

inurl:passlist.txt

intitle:"index of" "*/ftp.txt“

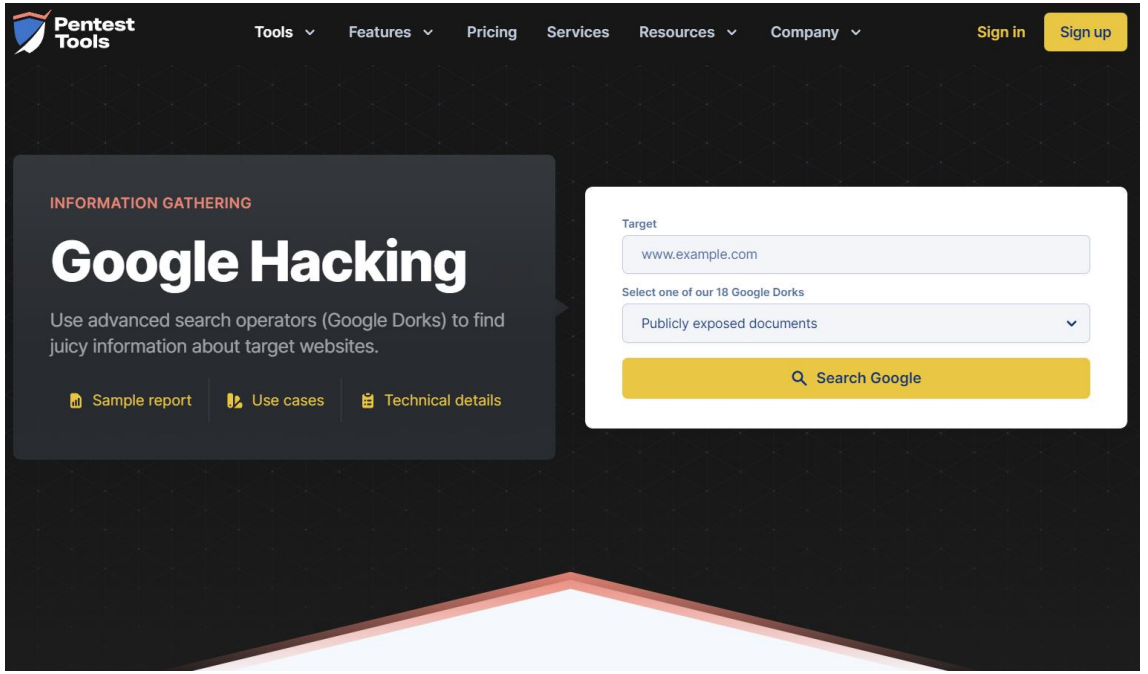
"index of" filetype:sql

intitle:"index of" "/sql" "admin"

site:sftp.*.*/ intext:"login" intitle:"server login“

Список наиболее известных Google Hack и Google Dork публикуется вот здесь:

https://www.exploit-db.com/google-hacking-database



https://pentest-tools.com/information-gathering/google-hacking

08 Типы поисковиков

01

Ванильные

Это стандартные поисковики, которые формируют видимую часть Интернета. Это Google.com, Yahoo.com, Bing.com, Baidu.com и многие другие.

02

Региональные

Стандартные региональные поисковые системы. Да, они тоже есть. Для Азии это, например, Sogou, Shenma, CocCoc и другие. Для России это Rambler, Mail.ru.

03

«Защищенные»

Это поисковики, которые якобы не делают треккинг пользователей по их запросам. Самый известный – DuckDuckGo. Есть еще SwissCows, Brave, Infinity, Quant и другие.

04

DarkWeb

Это поисковые системы в темных глубинах ДаркВеба. Torch, Not Evil, Ahimia. Они регулярно «переезжают», поэтому их ссылки нужно искать постоянно заново.

05

Метапоисковики

Этакие «поисковики по поисковикам». Иногда бывают удобными. Carrot2, EntireWeb, All The Internet, Fagan Finder и другие.

06

Интернет вещей

Самый популярный поисковик – это, конечно, Shodan. Есть еще Ivre, ZoomEye и другие.

07

Научные

Их тоже тонна. Как по мне, самые интересные – Research Gate и Wolfram Alpha. Есть и академические BASE, CORE. Ну и решения от грандов: Google Scholar, Baidu Academic и другие.

08

Кастомные

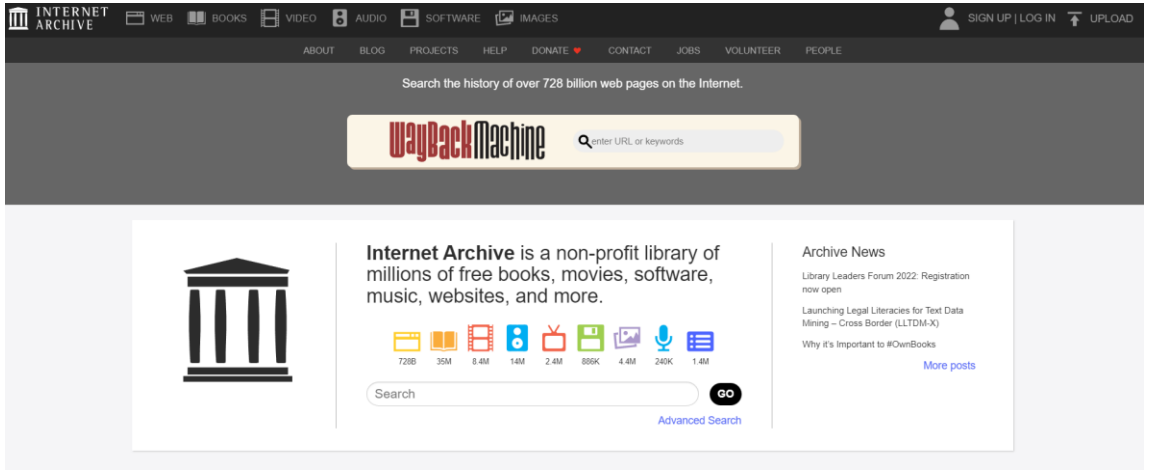
Поиск можно осуществлять по FTP-ресурсам. Это делает Мамонт, Napalm, FileSearch. Можно искать по Торрентам: TorLook.

09 Архивы

1. WayBack Machine

Это самый лучший архив Интернете из тех, что доступен бесплатно. Да и платно тоже. Что он может:

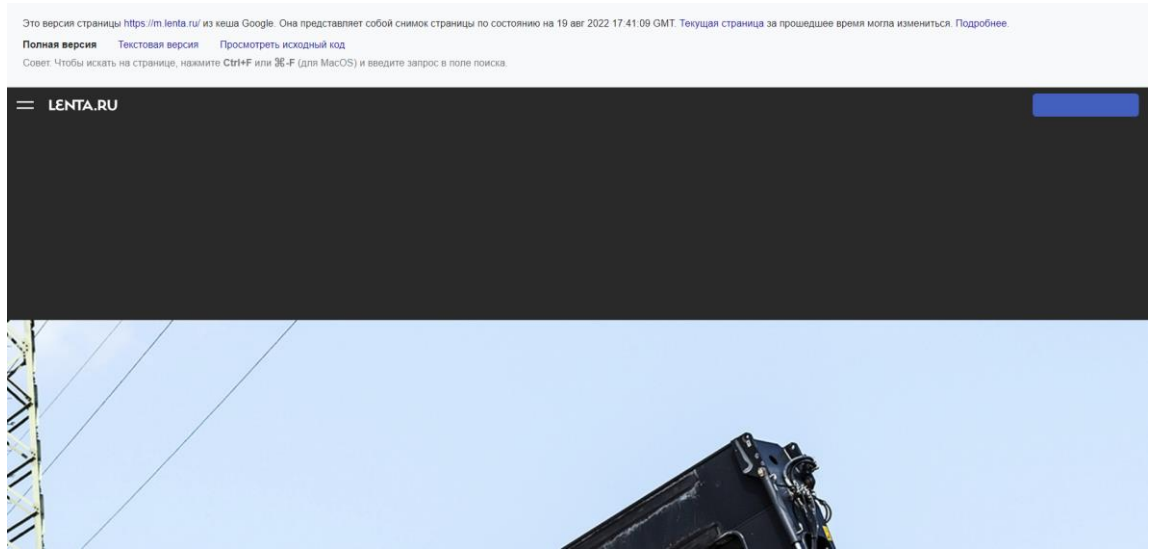
- Искать историю сайтов.
- Искать по документам, в т.ч. по PDF
- Искать по медиафайлам
- Искать по презентациям
- Сохранять страницы
- Искать по книгам
- Искать по документам
- Искать по архиву новостей
- Искать по архиву софта
- Искать по архиву изображений
- Сохранять копии страниц
- Имеет расширения для браузеров
- … и многое другое



Архивы

2. **cache:**

Дешево и сердито. Можно искать в кэше различных поисковых систем. Авось, что-то да и найдется.



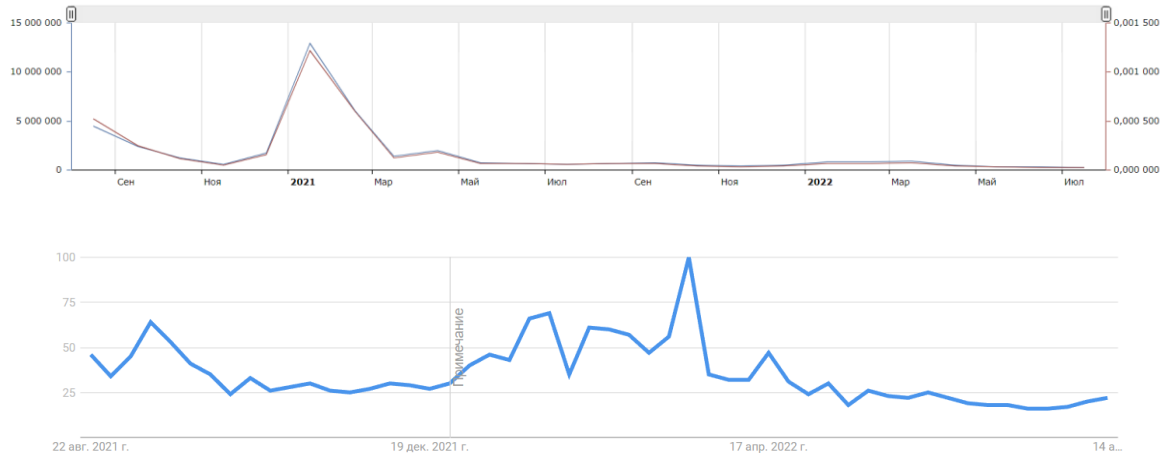
10 Прогнозирование

1. ## Яндекс WordStat

Помогает понять не только статистику запросов, но и различные тенденции в обществе. Потому что сейчас – поисковик – это скорее психотерапевт, нежели просто поисковая система.

2. ## Google Trends

В отличие от российского коллеги не показывает общие значения, а только относительные. Зато можно также, как и в Яндексе, посмотреть региональность.



1. # Рекламная разведка

Есть обратная сторона всесторонней слежки со стороны корпораций – возможность использовать их наработки для собственных целей.
https://ads.google.com/aw/audiences/management