



ТЕХНОСФЕРА

Особенности web-поиска. Спайдер.

Сергукова Юлия,
программист отдела инфраструктуры проекта
Поиск@Mail.Ru

Обо мне

Юлия Сергукова

МАИ, ф-т прикладной физики и
математики, 2012

С 2011г. - Mail.Ru, подразделение
Поиск



План лекции:

1. Web-поиск

- 1. Историческая справка**
- 2. Немного о рекламе**
- 3. Схемы**

2. Поисковый спайдер

- 1. Постановка задачи**
- 2. Выкачка**
- 3. Обновление**
- 4. Хранение**

Не забывайте о практике

Интернет и WWW

Интернет != WWW

Интернет и WWW

интернет != WWW

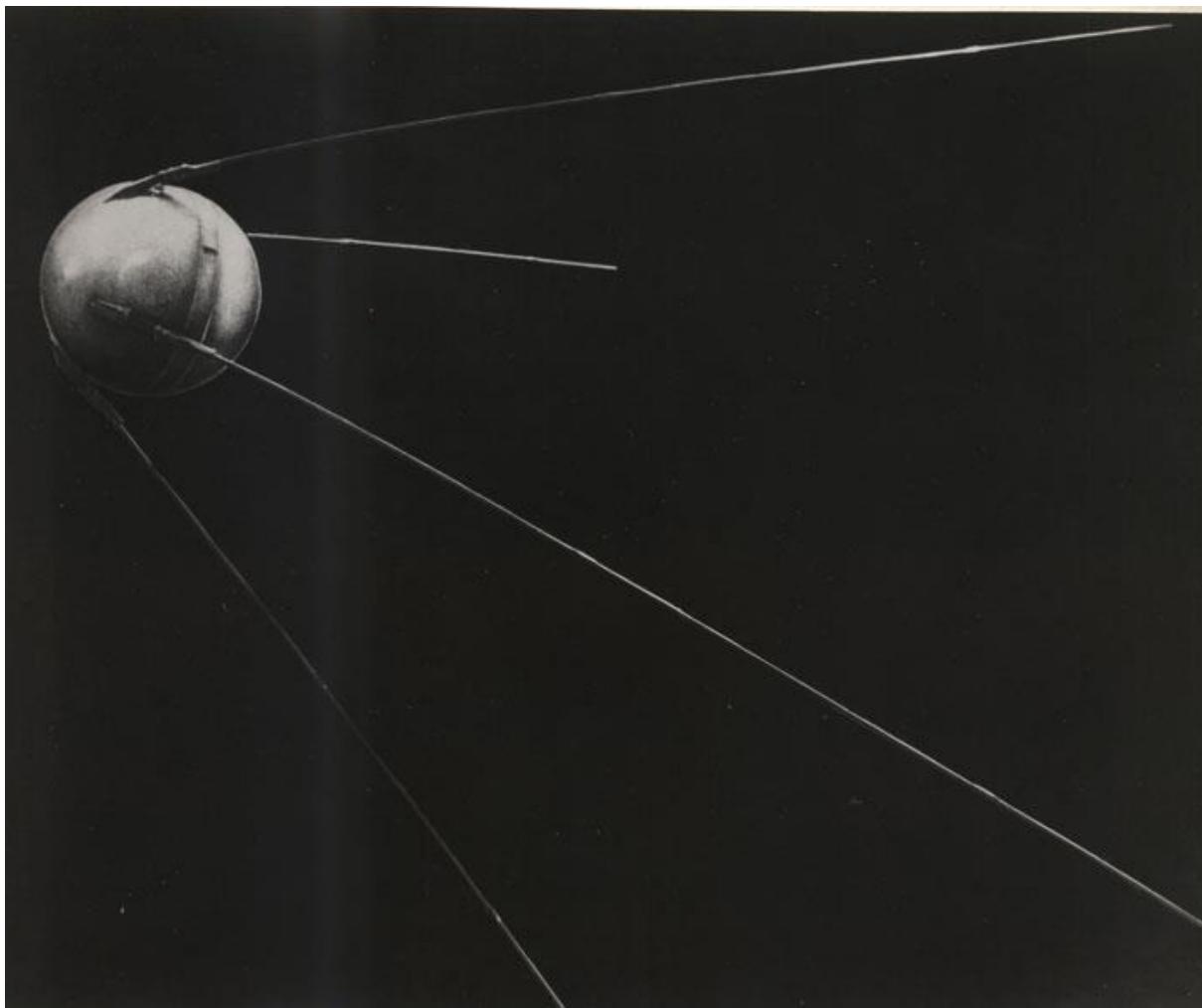
Разные уровни абстракции:

интернет - система объединенных компьютерных сетей

www - распределенная система, предоставляющая доступ к документам, расположенным на разных устройствах, подключенных к интернету

С чего всё началось?

С чего всё началось?



С чего всё началось?

29 октября 1969г. - рождение интернета (Калифорния - Массачусетс)

1971г. - первая почтовая программа (первые электронные адреса: user@machine)

1973г. - интернет становится международным (USA+GBR+NOR)

С чего всё началось?

1981г. - 200 машин. Интернет + www. IBM выпускает первый массовый (!) компьютер



С чего всё началось?

15 марта 1985г. - первый зарегистрированный домен: symbolics.com (в 2009г. перепродали - теперь рекламный сайт)

6 августа 1991г. - первый сайт. До сих пор функционирует:

<http://info.cern.ch/hypertext/WWW/TheProject.html>

Нам нужен Поиск!

1. *Пользователям* - находить нужную/актуальную информацию
2. *Владельцам сайтов* - об их сайтах узнают, их найдут
 - Возможность заработать на рекламе
 - Противостояние ПС и сайтов:
 - Сайты: как выдать свой контент за самый полезный
 - ПС: как отобразить **действительно** полезный контент
3. Кто-то должен поддерживать весь вэб:
 - Сервера, инфраструктура, создание контента
 - Поисковая реклама покрывает почти все расходы



Поисковые системы

1990г. - первая поисковая программа Archie. Поиск по заголовкам файлов на FTP-серверах.

1993г.

AliWeb (Archie-like indexing for the WEB) - готовые индексы от администрации сайтов. Первая ПС

W3Catalog - не обкачивает сайты, а использует чужие списки страниц. Первый агрегатор

WorldWideWebWanderer – первый поисковый робот. Цель - узнать все известные страницы.

Поисковые системы

1994 – keyword-based системы

AltaVista (до 8.07.2013) - первый короткий домен, первая "легкая" заглавная страница

Excite (<http://www.excite.com/>)

InfoSeek (в 1998 куплен The Walt Disney Company)

InkToMi (23.12.2002 поглощен Yahoo.com)

Keyword-based системы

1. "Найди мне то, что я сказал" (сейчас "найди мне то, что я хотел")
2. Не последняя роль - содержимое тега meta-keywords

```
<head>  
<meta charset="UTF-8">  
<meta name="description" content="Free Web tutorials">  
<meta name="keywords" content="HTML,CSS,XML,JavaScript">  
<meta name="author" content="Hege Refsnes">  
</head>
```

КТО-ТО ИХ ДО СИХ ПОР ИСПОЛЬЗУЕТ:

<http://www.ultersuite.ru/articles/meta/>

а Google нет:

<https://webmasters.googleblog.com/2009/09/google-does-not-use-keywords-meta-tag.html>

Поисковые системы

Люди начинают искать не конкретные файлы, а конкретную информацию.

Многообразие запросов.

<http://webmaster.mail.ru/querystat>

Запрос				вомбат							ПОКАЗАТЬ СТАТИСТИКУ						
Статистика запросов ?																	
Запрос	Хиты	Уники	Неизвестн.	М <13	М 13-18	М 19-25	М 26-34	М 35-49	М 50-65	М >65	Ж <13	Ж 13-18	Ж 19-25	Ж 26-34	Ж 35-49	Ж 50-65	Ж >65
вомбат	152	130	41	3	6	11	10	9	3	1	4	4	6	11	12	7	2
вомбат фото	43	42	10	0	1	1	3	2	2	0	1	3	0	8	11	0	0
вомбат фото животного	34	31	5	1	2	4	1	3	2	0	1	0	1	5	3	3	0
вомбат животное	11	10	4	0	0	0	1	1	0	0	0	0	1	0	0	2	1
кавашки вомбата	6	4	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0
опера хаус ехидна вомбат говядина что объединяет	5	5	1	0	0	0	0	0	1	0	0	0	0	0	1	2	0
бестолковый вомбат	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
детёныш вомбата фото	3	3	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
вомбат котяня	3	3	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
животное похожий на вомбата 4 буквы	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
что такое вомбат и опера хаус ехидна	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

Поисковые системы

1996 - “sponsored search”

место в выдаче по конкретному ключевому слову зависит от того, сколько вы за него заплатили

чем популярнее слово, тем дороже на нем продвигаться:
casino было очень дорогим

1998г - GoTo

8.11.2001 - переименован в Overture.com; поглотил
AltaVista

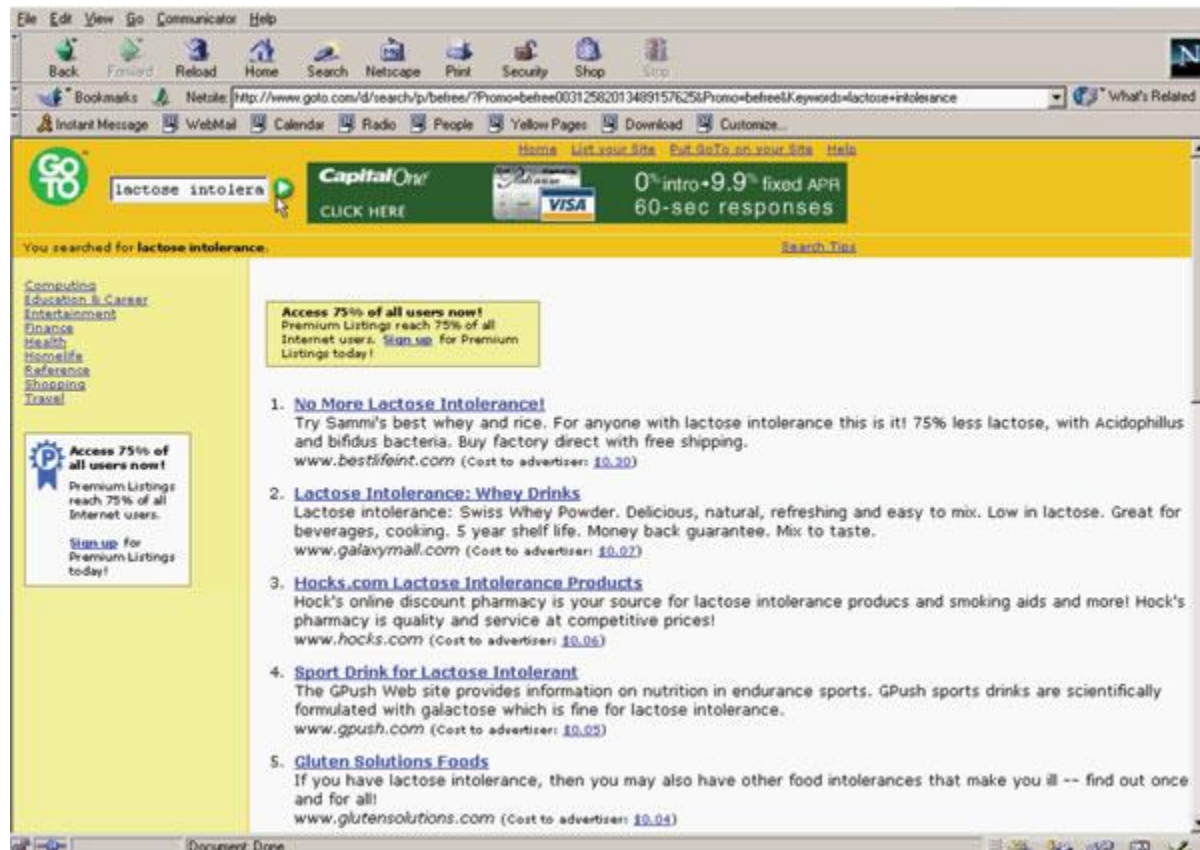
7.10.2013 - поглощен Yahoo.com

Sponsored search

Платят не только за то, **что** показывать, но и за то, **где** и **как**:

1. Платят только за переходы пользователя с ПС на сайт
2. Кто больше платит, тот выше в выдаче (выгода для ПС)
3. Чем выше в выдаче сайт, тем больше вероятность перехода (выгода для владельцев сайтов)

Начало эры поисковой рекламы



Google и PageRank

1998г.

Результаты выдачи ранжируются по своему "качеству" и релевантности запросу.

Релевантность - соответствие ожиданиям пользователя

Google и PageRank

1. Содержание страниц (не только keywords)
2. Популярность страниц (индекс цитирования)
 - Linkfarm - фермы по "разведению" ссылок

Google: PageRank + AdWords

PageRank не убил платную рекламу - он просто ее переместил

ПС нужен доход

Сайты продолжают платить, но уже за ранжирование в **рекламном блоке**

Баланс:

ПС получает прибыль за переходы, переходы происходят по релевантным объявлениям => удалять нерелевантные объявления и "плохой" контент

Реклама в ПС

Google [Advanced Search](#)

Web [Show options...](#) Results 1 - 50 of about 30,900,000 for florida vacation rentals on the beach. (0.94 second)

Google AdWords Ads

Florida Vacation Rentals
[CBFloridaVacations.com/Rentals](#) View Photos of Florida Beachfront Vacation Rentals with Details

Rent a Vacation Home
[www.HomeAway.com](#) Vacation Home Rentals from \$75 up. Book from the World's Leader today!

Relax at Westin® Resorts
[Westin.StarwoodHotels.com/Resorts](#) Exclusive Rates & Vacation Offers Westin® Hotels & Resorts

Vacation Rentals .com - Vacation Homes, Beach Houses, Vacation ...
 Find Vacation Rentals Deals and Discounts on Vacation Homes, Beach Houses, Villa Rentals, Condos, Cabins and Cottages in Florida, Colorado, California, ...
[Florida - California - Arizona - New York](#)
[www.vacationrentals.com/](#) - [Cached](#) - [Similar](#)

Florida Vacation Rentals, Florida Vacation Homes, Florida Beach ...
 But that's good news for anyone looking for a Florida family vacation rental near Disney romance-inspiring condo rental in Fort Myers Beach, ...
[www.vacationrentals.com > All Rentals > United States](#) - [Cached](#) - [Similar](#)

[Show more results from www.vacationrentals.com](#)

Gulf Coast Florida Rentals by Owner on the Gulf Coast of Florida ...
 Florida Gulf Coast Vacation Rentals lists property on Sanibel Island, Seaside, Bonita Beach, Captiva, Naples, Marco Island as well as Panama City Beach, ...
[Orlando / Disney - Destin - Alabama - Clearwater](#)
[www.gulfcoastrentals.com/](#) - [Cached](#) - [Similar](#)

Destin vacation rentals Destin Florida vacation condo beach house

Rent Florida Condo
 Save Over 50% On Florida Vacation Book Online, Low Rate Guarantee
[Florida.BookIt.com](#)

Key West - Tourism Site
 The Southern most city in the U.S. Palm trees, sunsets, fish & dive
[www.fla-keys.com/keywest](#) Florida

VacationRentals.com®
 Thousands of Vacation Rentals Worldwide-Book now for great deals
[VacationRentals.com](#)

Vacation Rentals by Owner
 Vacation Homes Around the World Rent Direct and Save!
[www.VRBO.com](#)

Florida Hotels On Beach
 Low Price Guaranteed on Hotels No Change/Cancel Fees. Book Now
[www.Travelocity.com/Hotels](#)

Luxury beach homes

Google [Advanced Search](#) [Preferences](#)

Web

Used Book - 95% off
[DealOz.com/Books](#) Searching 200+ Bookstores & 80,000 Sellers For Books- New, Used, Rare!

Cheap Textbooks 90% Off
[Textbooks.com](#) Immediate Free Shipping. Cheap Textbooks - Used & New. Save Today!




Textbooks For Sale
[www.Half.com/Textbooks](#) Buy and Sell New or Used Textbooks. Huge Discounts. Save Now!




Did you mean: [used books](#) Top 2 results shown

AbeBooks Official Site - New & Used Books, New & Used Textbooks ...
 AbeBooks is your source for Used, New, Rare and Out of print books. Find classic collectibles, rare signed editions, used textbooks, and inexpensive ...
[www.abebooks.com/](#) - 47k - [Cached](#) - [Similar pages](#)

Barnes & Noble - Books, Textbooks, Used Books, DVDs, Music, Toys ...
 Shop the Internet's Largest Bookstore for books, DVDs, music, textbooks, toys & games, and gift cards. Enjoy customer reviews, book clubs, and more.
[www.barnesandnoble.com/](#) - 77k - [Cached](#) - [Similar pages](#)

Results for: **used boks**

Used Boks | Used Books Blog  
 Jan 25, 2008 ... Welcome to the **Used Boks** Blog. **Used Boks** is a common misspelling for **Used Books**. Find **used** books reviews and commentary on fiction and ...
[usedbooksblog.com/blog/used-boks/](#) - 21k - [Cached](#) - [Similar pages](#) - 

Amazon.com: "The Best Childres's Boks"  
 The Best Childres's **Boks**. A Listmania! list by Nicole M. Riale (Gainesville, GA) ... A Light in the Attic by Shel Silverstein. \$12.91 **Used** & New from: \$1.93 ...
[www.amazon.com/The-Best-Childress-Boks/lm/32KBYL40JPDHX](#) - 249k - [Cached](#) - [Similar pages](#) - 

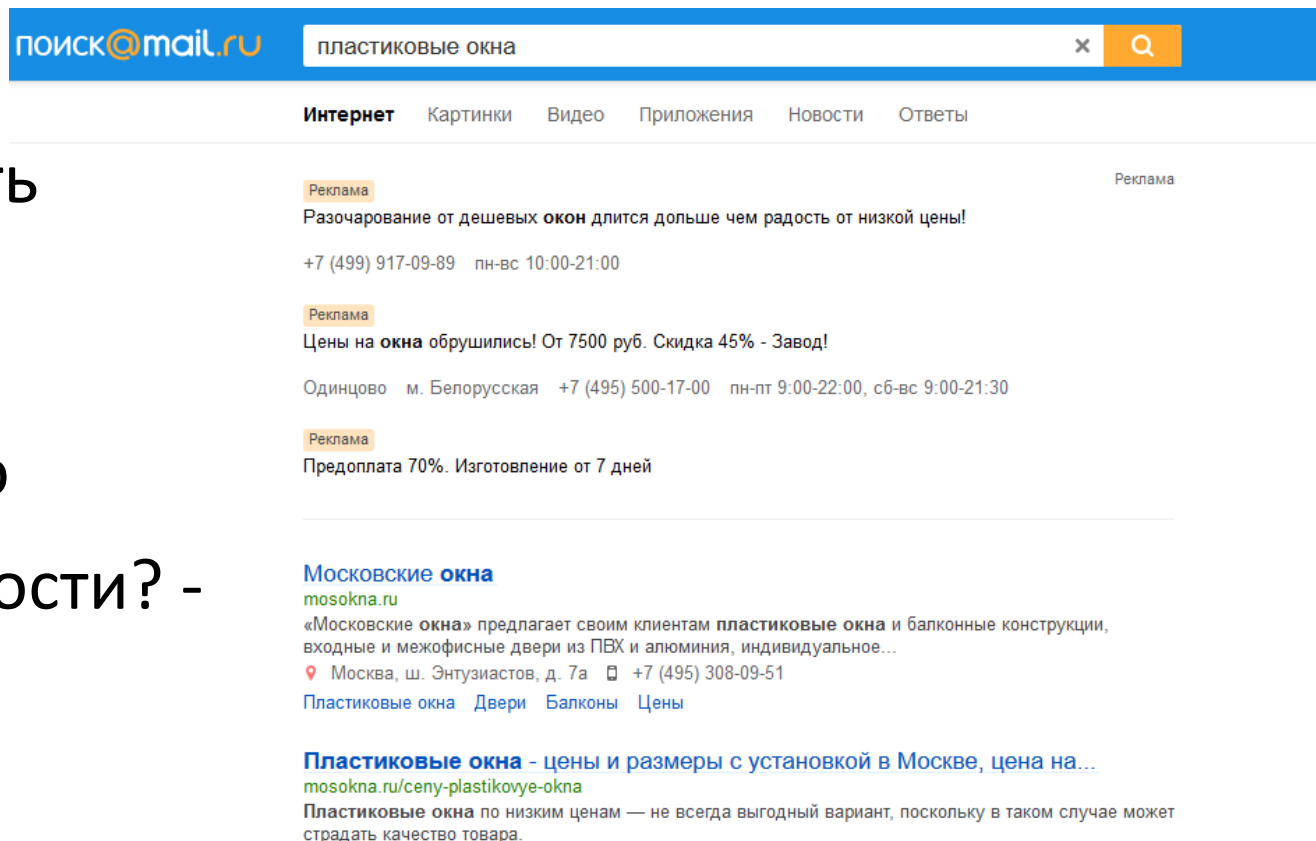
Реклама в ПС

Как ранжировать рекламу?

- по плате? - нерелевантно
- по релевантности? - невыгодно

Выход:

$$\text{rank} = f(\text{price}, \text{CTR})$$



поиск@mail.ru пластиковые окна x Q

Интернет Картинки Видео Приложения Новости Ответы

Реклама Реклама

Разочарование от дешевых окон длится дольше чем радость от низкой цены!
+7 (499) 917-09-89 пн-вс 10:00-21:00

Реклама

Цены на окна обрушились! От 7500 руб. Скидка 45% - Завод!
Одинцово м. Белорусская +7 (495) 500-17-00 пн-пт 9:00-22:00, сб-вс 9:00-21:30

Реклама

Предоплата 70%. Изготовление от 7 дней

Московские окна
mosokna.ru
«Московские окна» предлагает своим клиентам пластиковые окна и балконные конструкции, входные и межофисные двери из ПВХ и алюминия, индивидуальное...
📍 Москва, ш. Энтузиастов, д. 7а ☎ +7 (495) 308-09-51
[Пластиковые окна](#) [Двери](#) [Балконы](#) [Цены](#)

Пластиковые окна - цены и размеры с установкой в Москве, цена на...
mosokna.ru/ceny-plastikovye-okna
Пластиковые окна по низким ценам — не всегда выгодный вариант, поскольку в таком случае может страдать качество товара.

Second price auction

Аукцион Викри - однораундовый закрытый аукцион.
Победитель выплачивает вторую ставку

Закрытый => баланс между реальной оценкой и
максимально допустимыми затратами

Для рекламных систем - механизм Викри-Кларка-Гровса. N победителей, выплачивается сумма,
достаточная для удержания позиций

Second price auction

advertiser	bid	CTR			
A	4\$	0.01			
B	3\$	0.03			
C	2\$	0.06			
D	1\$	0.08			

bid - ставка за 1 переход с ПС

CTR - вероятность перехода:

$(\text{кол-во переходов}) / (\text{кол-во показов})$.

Это мера **релевантности**

Second price auction

advertiser	bid	CTR	ad rank		
A	4\$	0.01	0.04		
B	3\$	0.03	0.09		
C	2\$	0.06	0.12		
D	1\$	0.08	0.08		

ad rank - фактическая полезность объявления для ПС.

Самая простая формула:

$$V(\text{bid}, \text{CTR}) = \text{bid} * \text{CTR}$$

Second price auction

advertiser	bid	CTR	ad rank	rank	
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

rank - итоговое ранжирование. Определяет, какую в итоге позицию занимает каждое из объявлений.

Сравним с GoTo

Second price auction

advertiser	bid	CTR	ad rank	rank	GoTo
A	4\$	0.01	0.04	4	1
B	3\$	0.03	0.09	2	2
C	2\$	0.06	0.12	1	3
D	1\$	0.08	0.08	3	4

rank - итоговое ранжирование. Определяет, какую в итоге позицию занимает каждое из объявлений.

Сравним с GoTo

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

price - сколько в итоге заплатит рекламодатель за каждый переход.

$$V(\text{price}_1, \text{CTR}_1) = V(\text{bid}_2, \text{CTR}_2)$$

$$\text{price}_1 * \text{CTR}_1 = \text{bid}_2 * \text{CTR}_2$$

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	
B	3\$	0.03	0.09	2	
C	2\$	0.06	0.12	1	
D	1\$	0.08	0.08	3	

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

1: C, bid=2\$, CTR=0.06

2: B, bid=3\$, CTR=0.03

Second price auction

advertiser	bid	CTR	ad rank	rank	price
A	4\$	0.01	0.04	4	0.01\$
B	3\$	0.03	0.09	2	2.68\$
C	2\$	0.06	0.12	1	1.51\$
D	1\$	0.08	0.08	3	0.51\$

$$\text{price}_1 = \text{bid}_2 * \text{CTR}_2 / \text{CTR}_1 (+0.01\$)$$

Качество поиска

Качество поиска

- Релевантность
- Покрытие многозначности
- Простота UI
- Уменьшение ошибок пользователя
- Полнота

Качество поиска. Полнота

Большинству пользователей не нужна.

Проявляется на непопулярных запросах.

Найдите документы со словом
собакокрадство

Качество поиска

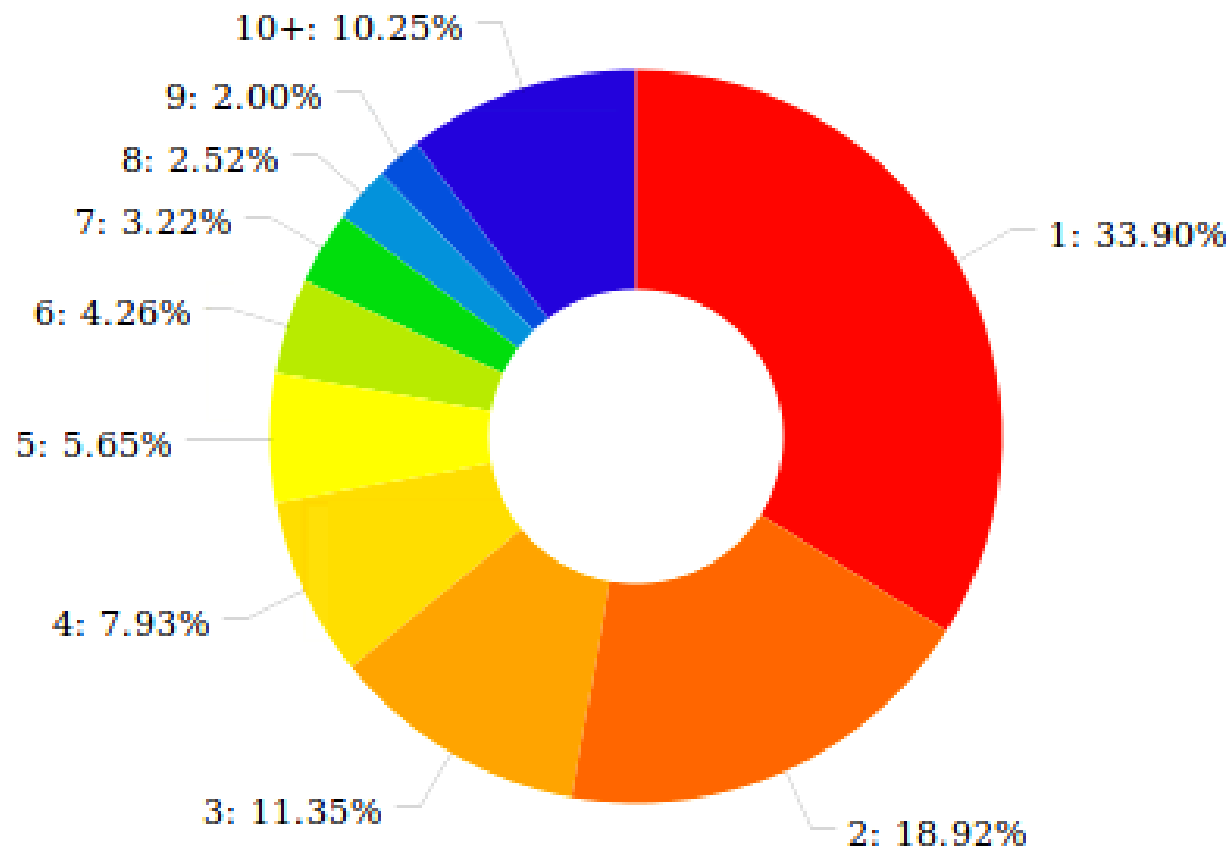
Полнота по непопулярным запросам - способ оценить размер полезного индекса.

<http://www.analyzethis.ru/?analyzer=rare&date=2016-09-01&lang=ru&location=ru>

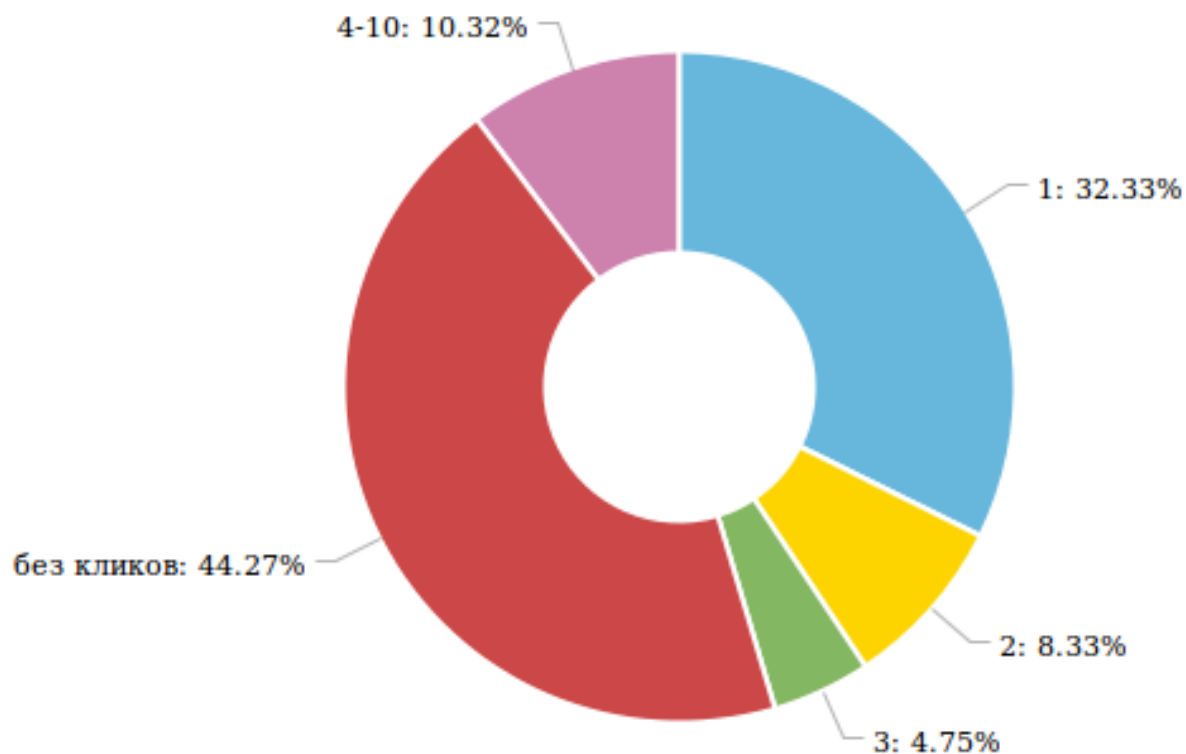
Независимые метрики различных аспектов ПС:

<http://www.analyzethis.ru/>

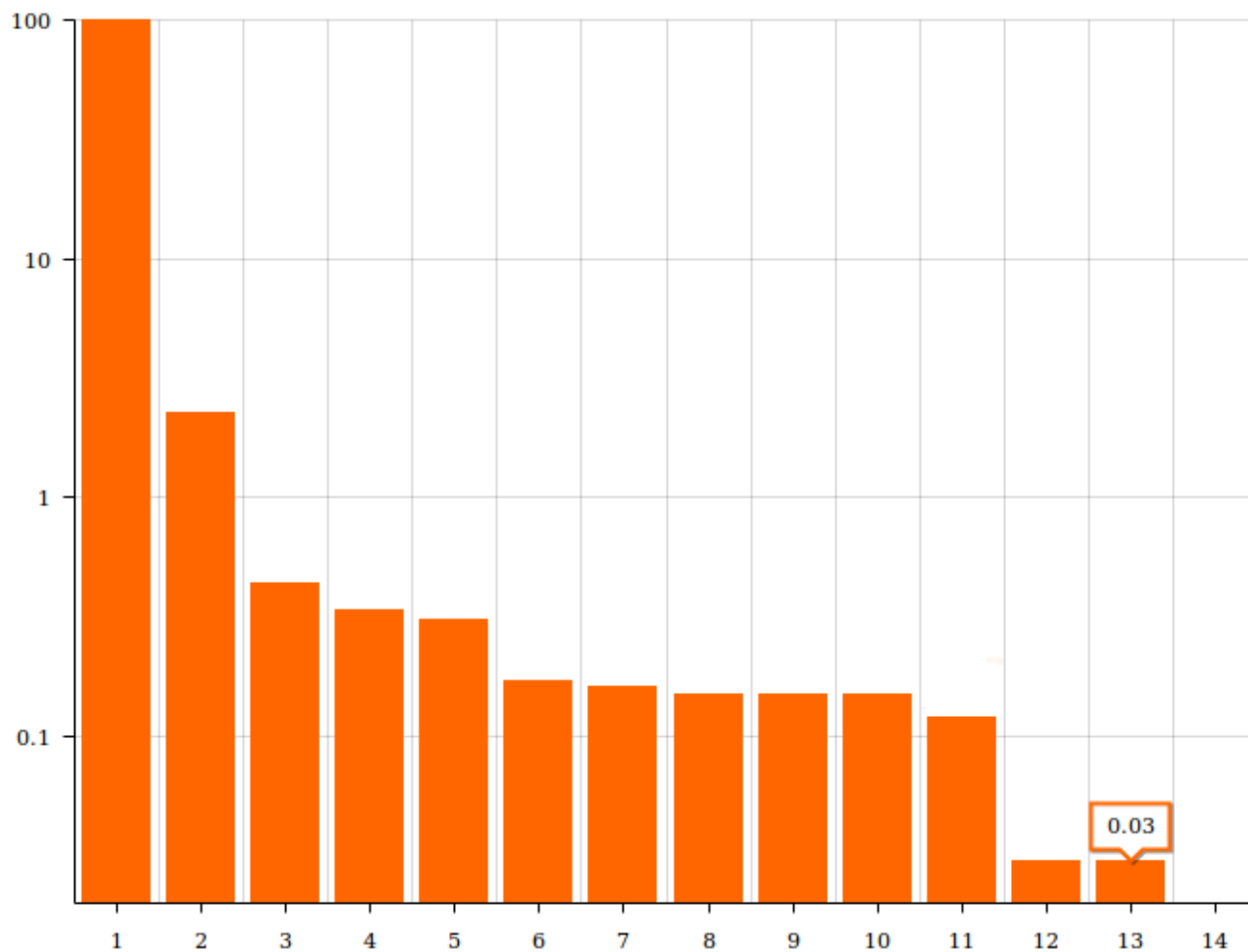
Пользователи и поиск. Запросы от 1 пользователя в день



Пользователи и поиск. Клики на первой странице



Пользователи и поиск. Как далеко заходят пользователи.



Перерыв

Поисковый спайдер



Задача:

Нужно скачать сайт.

Ваши предложения?

Проблема:

Сайтов много

Страниц еще больше

Времени мало

Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Требования к спайдеру

1. Politeness
2. Freshness
3. Actuality
4. Производительность
5. Масштабируемость

URL

RFC: <https://www.ietf.org/rfc/rfc1738.txt>

<http://site.ru/path?page=10>

http - схема

site.ru - хост

path - путь

page=10 - query

IP

Уникальный адрес сетевого узла

```
$ host go.mail.ru
```

```
$ host ru.wikipedia.org
```


DNS

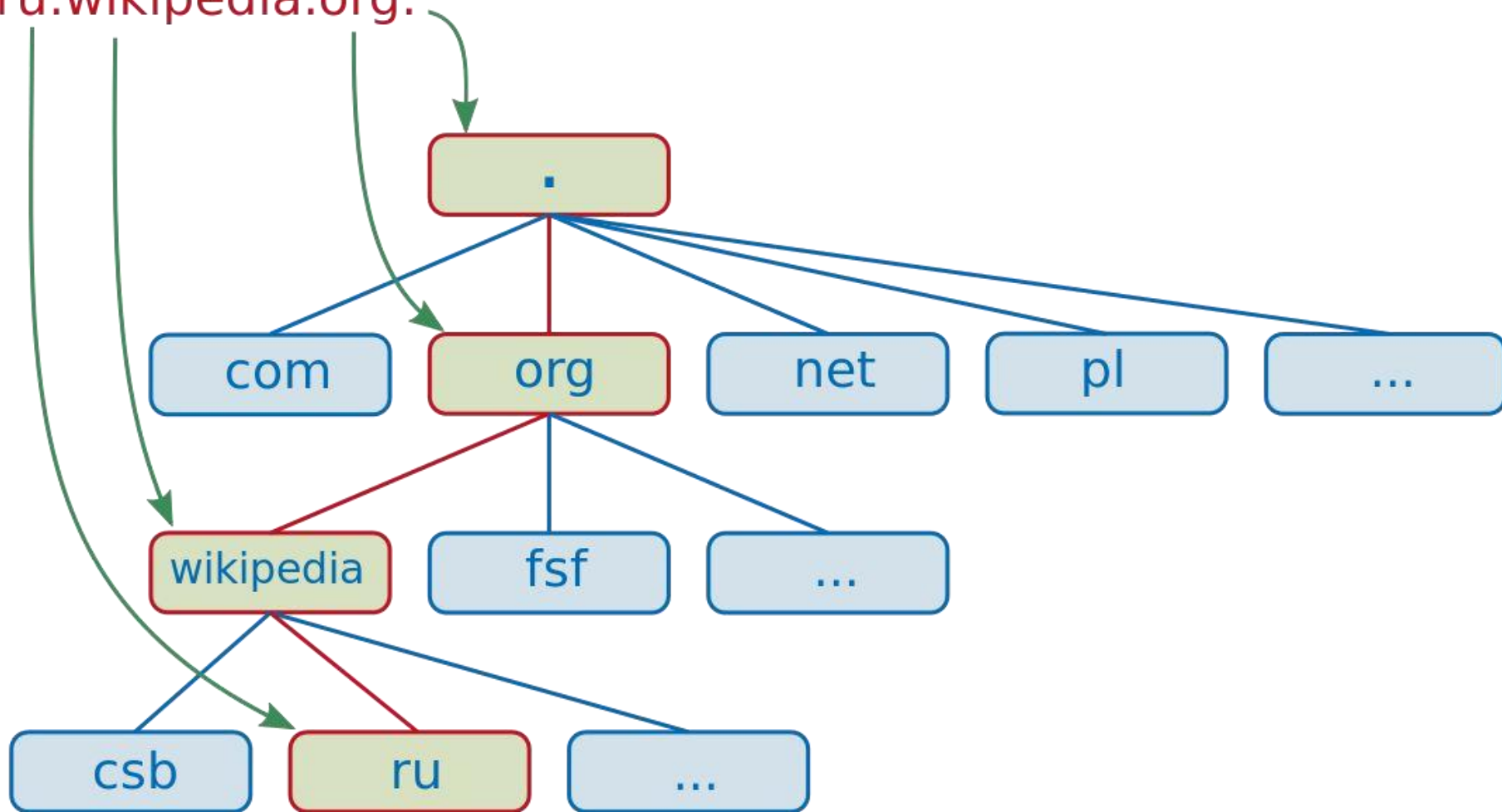
DNS - сервис для получения информации о доменах. Нам нужна информация об IP.

url -> ip

информация предоставляется иерархической системой серверов - может быть долго

DNS

ru.wikipedia.org.



DNS

```
$ dig -t NS .
```

```
$ dig -t NS @e.root-servers.net. ru
```

```
...
```

```
$ dig -t A @???? go.mail.ru
```

Сколько ір-адресов у сайта?

Сколько ip-адресов у сайта?

1. 1-1:

```
$ host -v -t A zonova.xyz
```

2. 1-n: снижение нагрузки (для высоконагруженных систем)

```
$ host -v -t A go.mail.ru
```

3. m-1: снижение стоимости

```
$ host -v -t A catalogr.ru
```

```
$ host -v -t A redbook73.ru
```

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Хорошие роботсы:

<http://lenta.ru/robots.txt>

Плохие роботсы:

<https://money.yandex.ru/robots.txt>

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Какие из этих документов
можно качать?

<http://site.ru/>

<http://site.ru/admin>

<http://site.ru/admin/article>

<http://site.ru/article/admin>

<http://site.ru/post>

Robots.txt

```
User-agent: *  
Crawl-delay: 50  
Disallow: /admin  
Allow: /article
```

Какие из этих документов
можно качать?

<http://site.ru/>

<http://site.ru/admin>

<http://site.ru/admin/article>

<http://site.ru/article/admin>

<http://site.ru/post>

Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Алгоритм

1. "Точка входа" - seed-урлы
2. Скачали
3. Распарсили, извлекли урлы, отправили урлы в очередь на обкачку
4. goto #2

Seed-урлы

КАТАЛОГ@mail.ru®

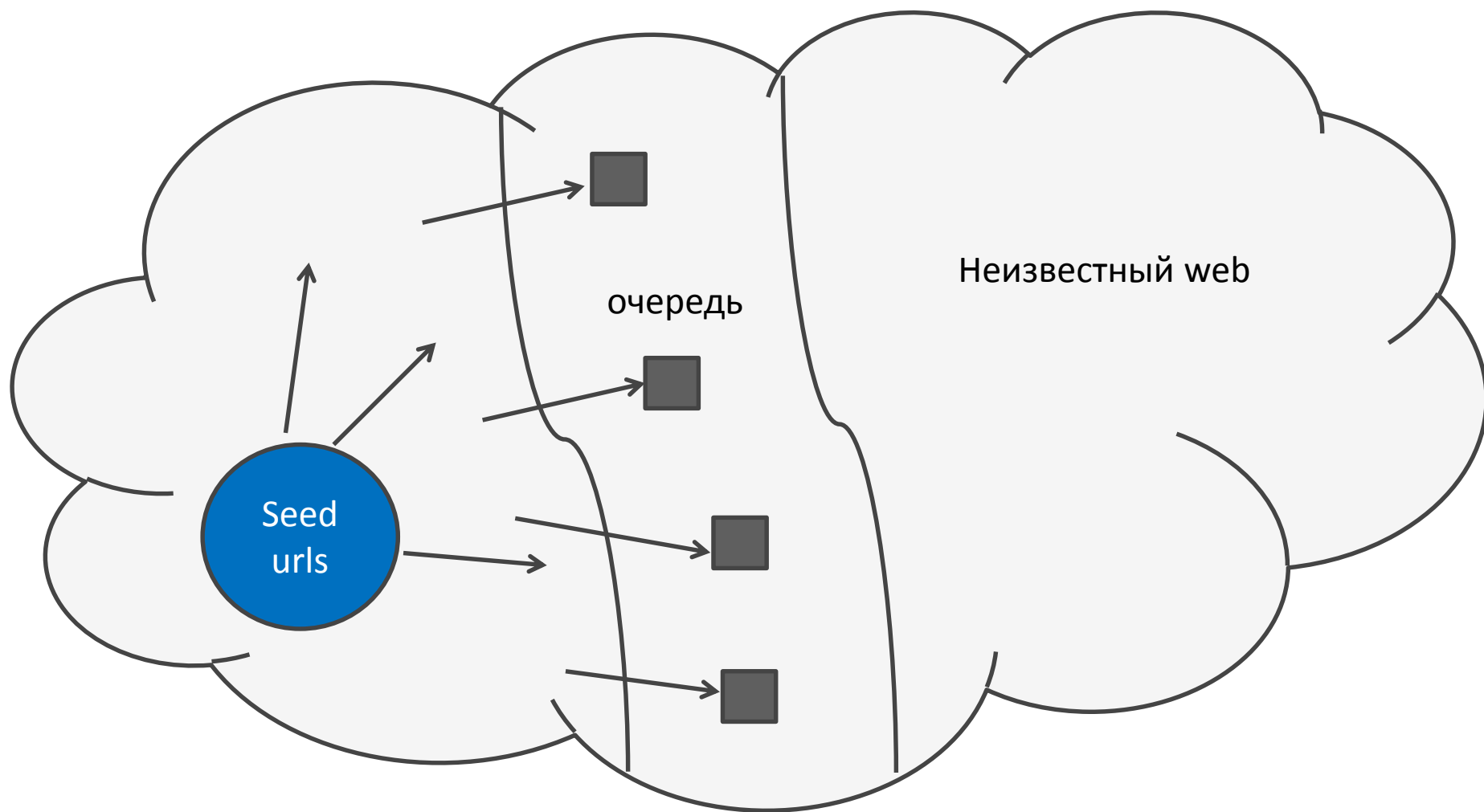
Яндекс
каталог

РЕЙТИНГ@mail.ru



ВИКИПЕДИЯ
Свободная энциклопедия

Выкачка



Выкачка



Ответы сервера

Какие бывают?

2xx - успешно

3xx - перенаправление

4xx - ошибка клиента

5xx - ошибка сервера

Особенности контента

1. Тип контента
2. Кодировка

Тип контента

html, jpeg, pdf, xml, mp3 и т.д.

Как определить:

1. Content-Type: text/html
2. По первым символам контента

```
1 <!DOCTYPE html>  
2 <html>  
3 <head>
```

Не всё так просто:

<http://kiev-ehudi.org.ua/>

БНОПНЯ

\$ echo БНОПНЯ | iconv -f CP1251 -t KOI8R



Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head

```
$ wget --spider -Sq https://en.wikipedia.org/wiki/Sicily  
2>&1 | grep charset
```

```
Content-Type: text/html; charset=UTF-8
```

Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head
2. Meta-charset

```
$ wget -SO ch1 http://solarboat.ru/catalog/lodki_solar/ 2>&1 | grep charset
```

```
Content-Type: text/html; charset=windows-1251
```

```
$ grep charset ./ch1
```

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1251" />
```

Какая кодировка?

Не надо быть умнее браузера.

1. Content-type: charset в http-head
2. Meta-charset

Определите кодировку:

<http://www.emalirovka-vann.ru/>

<http://ievpdgh.22web.org/?i=1>

Какая кодировка?

<http://www.emalirovka-vann.ru/>

Http-head: cp1251

Meta: utf8

Res: utf8

<http://ievpdgh.22web.org/?i=1>

Http-head: -

Meta: utf8;cp1251

Res: utf8

Извлечение ссылок (discovering)

```
<a href="...">
```

Помним о politeness:

```
<meta name="robots" content="nofollow" />
```

```
<a href="signin.php" rel="nofollow">Войти</a>
```

Извлечение ссылок (discovering)

Ссылки бывают:

1. Внутренние и внешние
2. Абсолютные и относительные
3. Валидные и невалидные

Минутка прекрасного:

<http://www.mongolianembassy.ru/>

Абсолютные и относительные ссылки

<http://site.ru/page/1>

`` --> <http://site.ru/page/2>

`` --> <http://site.ru/2>

`` --> <http://site.ru/d3>

`` --> <http://site.com/page>

`` --> <http://abc.org/g>

Нельзя брать все ссылки



Нельзя брать все ссылки

1. Robots.txt
2. Некоторые документы мы уже качали
3. Внутренний blacklist:
 1. Правильные ограничения: <http://go.mail.ru/robots.txt>
 2. <https://www.iconfinder.com/search/?q=search>

А еще сайты могут быть "бесконечными":

<http://www.calend.ru/day/1-2-2050/>

Что брать и сколько?

Решает внешняя задача - scheduler

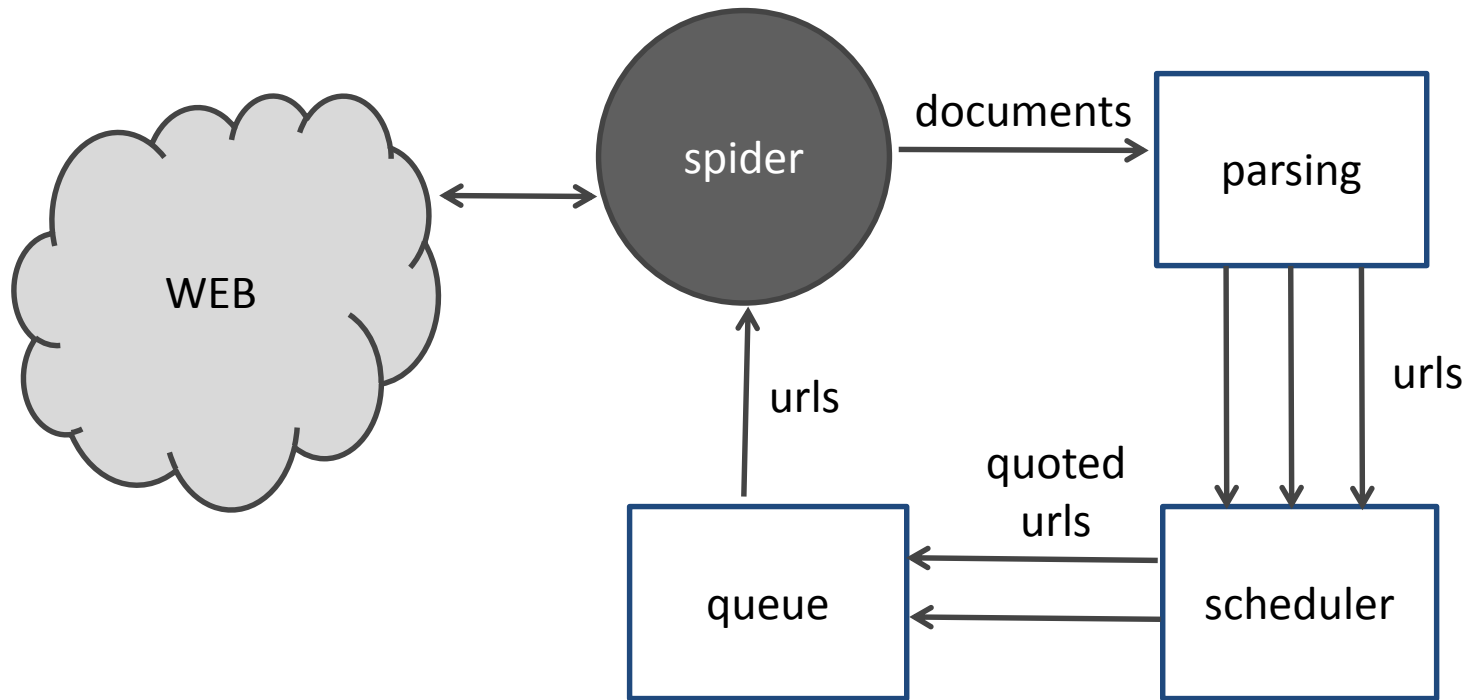
Учитывает:

1. Количество уже скачанных документов с сайта (успешно и нет)
2. Свойства скачанных документов (тип / язык)
3. Свойства самого сайта (посещаемость, CTR и т.д.)

Формируется квота.

Лекции про ранжирование

Spider & utils



Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Зачем перекачивать страницы?

1. Обновилось содержимое
2. Появились ссылки на новые страницы

Пример: главная страница сайта

Как часто перекачивать?

Простой подход:

если страница изменилась - $T = T/2$

если страница не изменилась - $T = T*2$

Усложнение:

- История выкачки
- Ранк сайта

Как понять, что страница изменилась?

Как понять, что страница изменилась?

<http://lenta.ru/>

<http://wellclix.net/>

<https://www.adme.ru/>

Как понять, что страница изменилась?

1. Брать только "чистый" контент
2. Удаление обвязки

Об этом - в другой лекции

Как понять, что страница изменилась?

Вэбмастера в одной лодке с нами

Http-response:

eTag

Last-Modified

В основном - для статического контента

Как понять, что страница изменилась?

```
$ HEAD http://s.imgur.com/images/loaders/ddddd1_181817/24.gif
200 OK
ETag: "57a25124-14f9"
Last-Modified: Wed, 03 Aug 2016 20:16:36 GMT
```

Как понять, что страница изменилась?

```
$ HEAD http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
200 OK
```

```
ETag: "57a25124-14f9"
```

```
Last-Modified: Wed, 03 Aug 2016 20:16:36 GMT
```

```
$ HEAD -H 'If-None-Match: "57a25124-14f9"'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
304 Not Modified
```

```
$ HEAD -H 'If-None-Match: "57a25124-14f8"'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
200 OK
```

```
$ HEAD -H 'If-Modified-Since: Wed, 03 Aug 2016 20:16:36 GMT'  
http://s.imgur.com/images/loaders/ddddd1_181817/24.gif  
304 Not Modified
```

Дополнительные источники информации

AliWeb - поисковик, который использовал заранее подготовленные "индексные файлы", содержащие список урлов и их описание (по усмотрению владельца ресурса)

А сейчас?

Дополнительные источники информации

<http://simonscat.tumblr.com/rss>

```
<?xml version="1.0" encoding="UTF-8"?>
<rss xmlns:dc="http://purl.org/dc/elements/1.1/" version="2.0">
<channel>
  <description>Channel description</description>
  <title>Simon's Cat</title>
  <item>
    <title>Simon's Cat refusing to face Monday! </title>
    <description>post description</description>
    <link>http://simonscat.tumblr.com/post/150306700829</link>
    <pubDate>Mon, 12 Sep 2016 12:33:35 +0100</pubDate>
  </item>
  ...
</channel>
```

Дополнительные источники информации

<http://all-t-shirts.ru/sitemap.xml?start=0>

```
<urlset>
```

```
  <url>
```

```
    <loc>http://all-t-shirts.ru/</loc>
```

```
    <lastmod>2016-03-28T00:03:15+03:00</lastmod>
```

```
    <changefreq>daily</changefreq>
```

```
  </url>
```

```
  ...
```

```
</urlset>
```


Спайдер

1. Постановка задачи
2. Выкачка
3. Обновление
4. Хранение

Хранение скачанных документов

Ваши варианты?

Хранение скачанных документов

Документ <--> урл

Ключ - f(url)

Практика. Сколько документов будет храниться в базе?

<https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BD%D0%B8>

<https://ru.wikipedia.org/wiki/Пони>

<https://ru.wikipedia.org/wiki/%CF%EE%ED%E8>

http://kikolani.com/blog-post-promotion-ultimate-guide?utm_source=kikolani&utm_medium=320banner&utm_campaign=bpp

<http://kikolani.com/blog-post-promotion-ultimate-guide>

<http://scifi.stackexchange.com/questions?page=4&sort=newest>

<http://scifi.stackexchange.com/questions?sort=newest&page=4>

<https://music.yandex.ru/album/3575649/track/29692077>

<http://music.yandex.ru/album/3575649/track/29692077/>

<https://www.music.yandex.ru/album/3575649/track/29692077>

http://opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

http://www.opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

<http://домены.рф/>

<http://xn--d1acufc5f.xn--p1ai/>

<http://domeny.rf/>

Хранение документов

Хэш - от **нормализованного** урла

RFC: <https://www.ietf.org/rfc/rfc1738.txt>

Хранение документов

RFC

<http://domeny.rf/> - .rf не существует

Хранение документов

Нормализованный URL - всегда в ASCII

Percent-encoding для query и пути

```
$ python -c "import urllib, sys; print urllib.quote(sys.argv[1])" Пони  
%D0%9F%D0%BE%D0%BD%D0%B8
```

Punycode для имени домена:

```
$ python -c "import urllib, sys; print sys.argv[1].decode('utf-8').encode('idna')"  
домены.рф  
xn--d1acufc5f.xn--p1ai  
$ python -c "import urllib, sys; print sys.argv[1].decode('idna')" xn--d1acufc5f.xn--  
p1ai  
домены.рф
```

Хранение документов

Нормализованный URL - всегда в ASCII

<https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BD%D0%B8>

<https://ru.wikipedia.org/wiki/Пони>

<https://ru.wikipedia.org/wiki/%CF%E8%ED%E8>

<http://домены.рф/>

<http://xn--d1acufc5f.xn--p1ai/>

Хранение документов

utm-метки для маркировки траффика

Параметры, которые игнорируются сервером, но учитываются в статистике

Позволяют оценить успешность рекламных кампаний (источники переходов)

Хранение документов

utm-метки для маркировки траффика

http://kikolani.com/blog-post-promotion-ultimate-guide?utm_source=kikolani&utm_medium=320banner&utm_campaign=bpp

<http://kikolani.com/blog-post-promotion-ultimate-guide>

Хранение документов

www. - наследие старого мира

Большинство - редиректят на нужную версию

Есть исключения:

www.music.yandex.ru - редиректит на корневик

[http://www.opennet.ru/](http://www.opennet.ru) и [http://opennet.ru/](http://opennet.ru) - обе отдают контент (одинаковый)

Хранение документов

www. - наследие старого мира

<https://music.yandex.ru/album/3575649/track/29692077>

<http://music.yandex.ru/album/3575649/track/29692077/>

<https://www.music.yandex.ru/album/3575649/track/29692077>

http://opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

http://www.opennet.ru/docs/RUS/inet_book/4/45/retr4514.html

Хранение документов

Зеркало - сайт (до 80%) дублирующий контент оригинала

1. Защита от падения
2. ... и от блокировок (lurkmore.to, lurklurk.com, lurkmirror.ml)
3. Дорогой внешний трафик - локальное зеркало

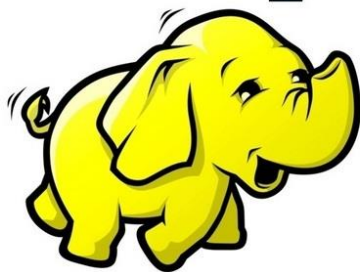
Как бороться? Искать дубликаты (другая лекция)

Хранение документов

> 15 Pb

> 50 млрд. документов

hadoop



Хранение документов



АEROSPIKE



Вопросы?