Прошу! Пожалуйста, перебивайте!
Задавайте вопросы во время рассказа!

Как сделать свой Яndex/Google

(ну, почти)

Антон Алексеев
асп. ПОМИ РАН, SofIT Labs
(ex-Яndex.Вертикали)

Прошу! Пожалуйста, перебивайте!
Задавайте вопросы во время рассказа!

Почему поиск?

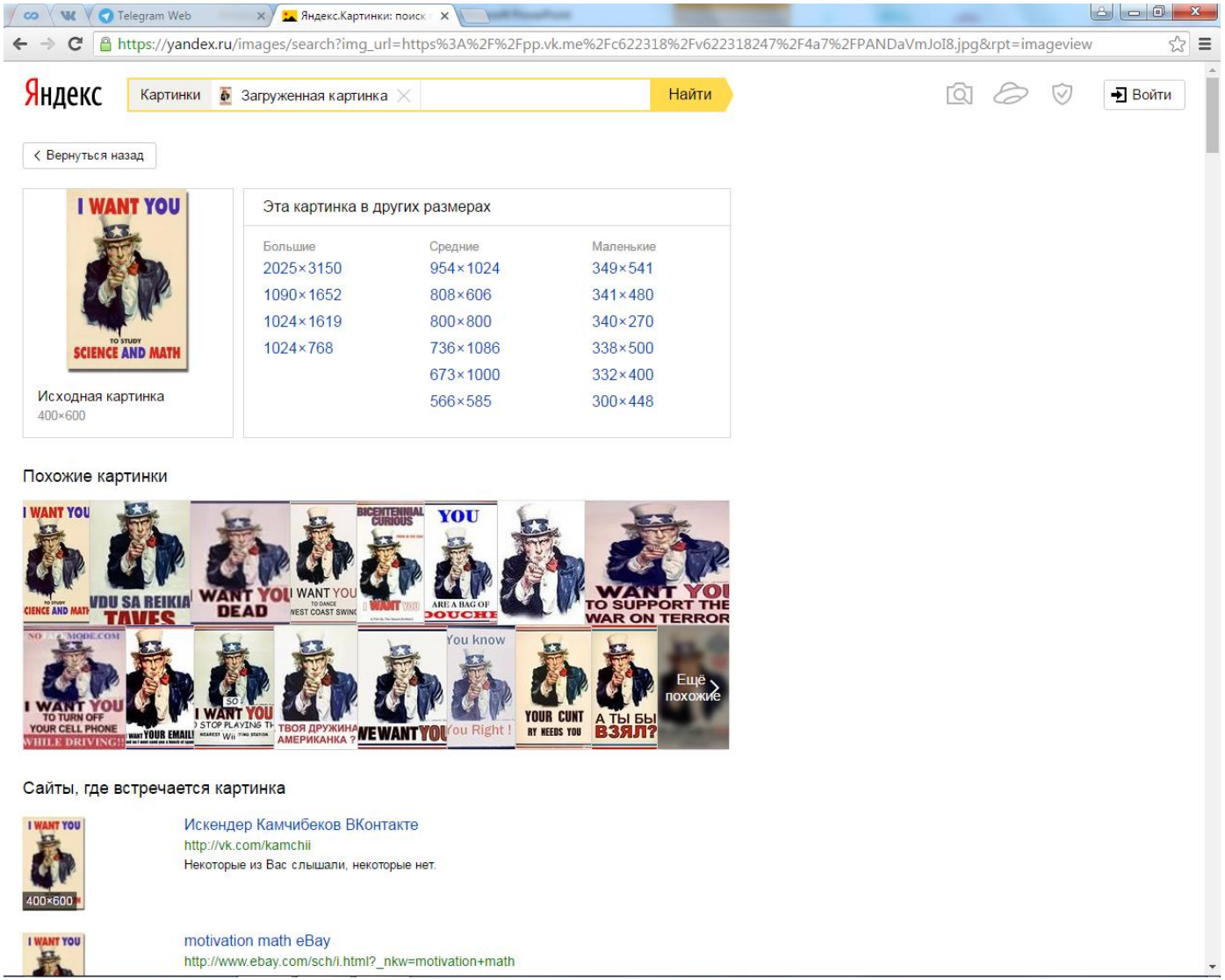
- Куча интересных алгоритмических, математических, лингвистических задач
- Вокруг компаний, индексирующих Интернет, много всего
- Поиск – «магия» и «фокусы», которые всегда приятно раскрывать!

История

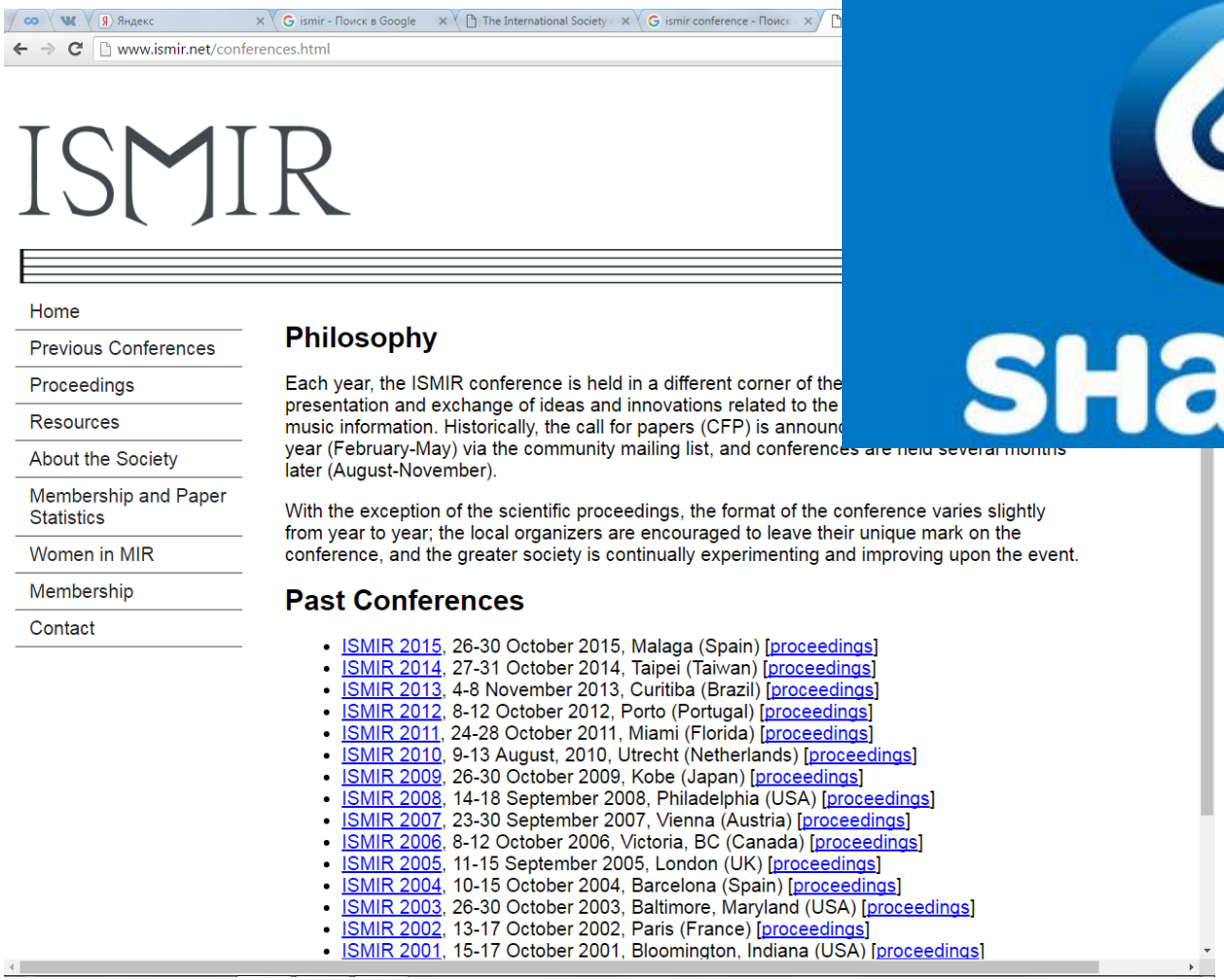
- Наука о поиске документов – с 1950-х
- Хитрые матмодели поиска – с 1970-80-х
- Интернет + дешёвые данные 1990-е -…

Timeline (full list)		
Year	Engine	Current status
1993	W3Catalog	Inactive
	Aliweb	Inactive
	JumpStation	Inactive
	WWW Worm	Inactive
1994	WebCrawler	Active, Aggregator
	Go.com	Inactive, redirects to Disney
	Lycos	Active
	Infoseek	Inactive
1995	AltaVista	Inactive, redirected to Yahoo!
	Daum	Active
	Magellan	Inactive
	Excite	Active
	SAPO	Active
	Yahoo!	Active, Launched as a directory
1996	Dogpile	Active, Aggregator
	Inktomi	Inactive, acquired by Yahoo!
	HotBot	Active (lycos.com)
	Ask Jeeves	Active (rebranded ask.com)
1997	Northern Light	Inactive
	Yandex	Active
1998	Google	Active
	Ixquick	Active also as Startpage
	MSN Search	Active as Bing
	empas	Inactive (merged with NATE)
1999	AlltheWeb	Inactive (URL redirected to Yahoo!)
	GenieKnows	Active, rebranded Yellowee.com
	Naver	Active
	Teoma	Inactive, redirects to Ask.com
	Vivisimo	Inactive
2000	Baidu	Active
	Exalead	Active
	Gigablast	Active

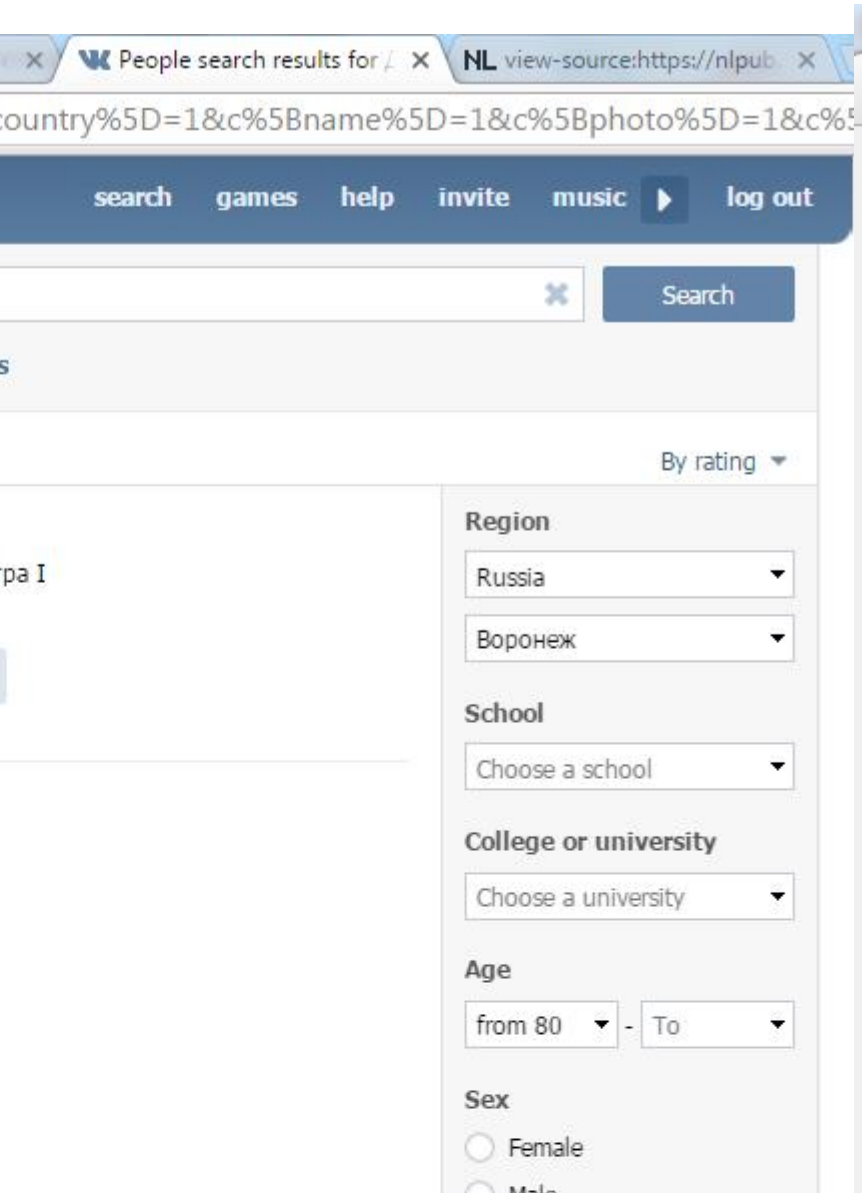
Поиск бывает разным: картинки

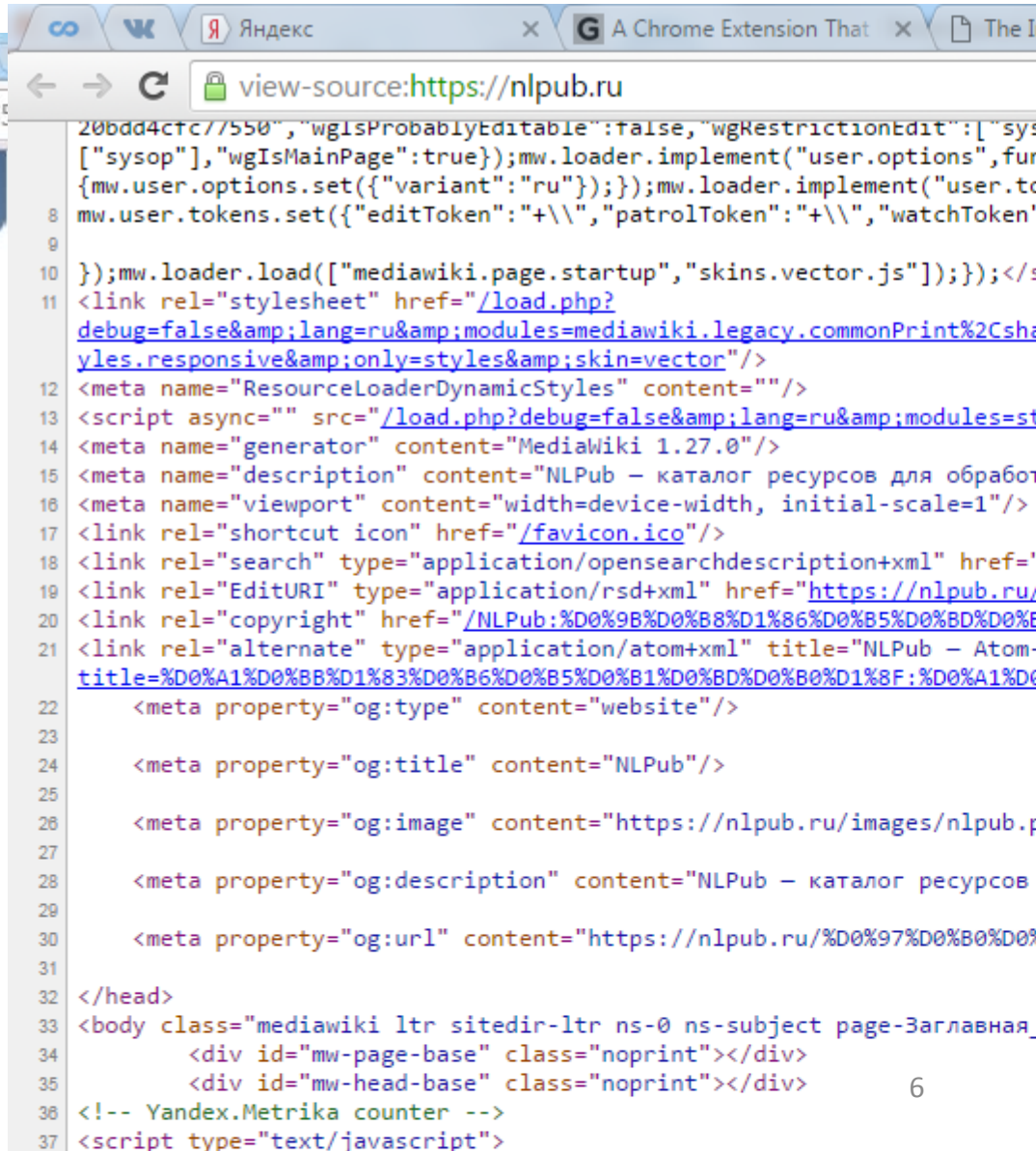


Поиск бывает разным: музыка

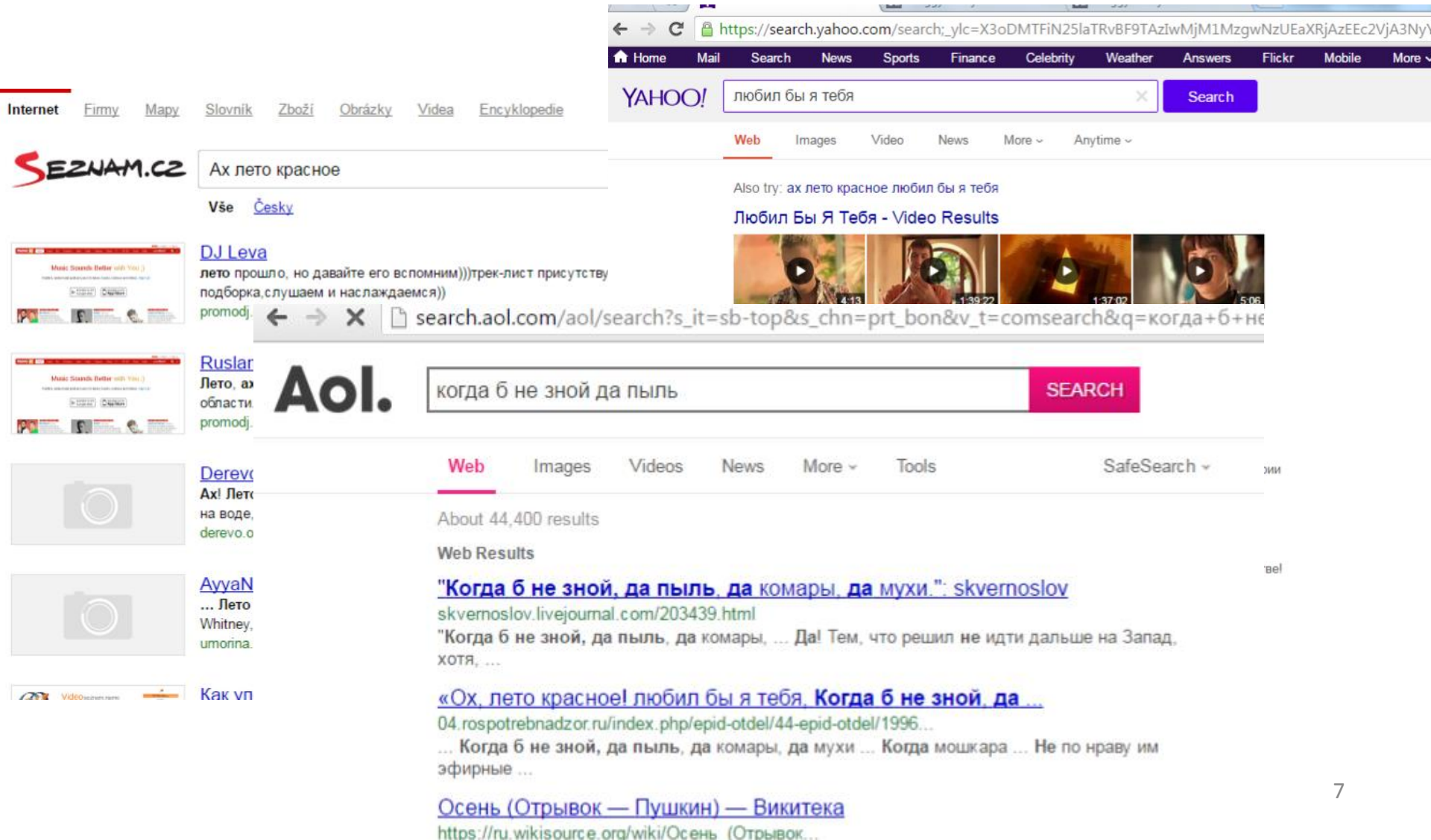


Поиск бывает разным: метаданные



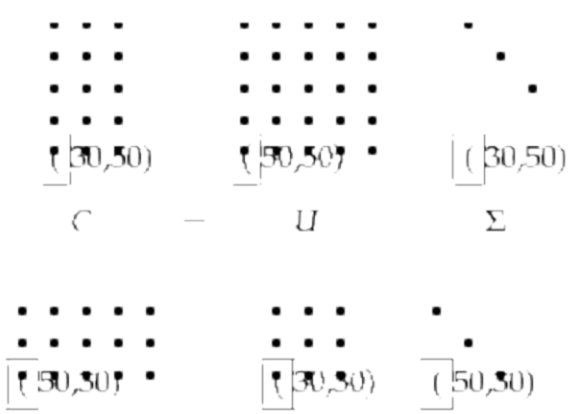


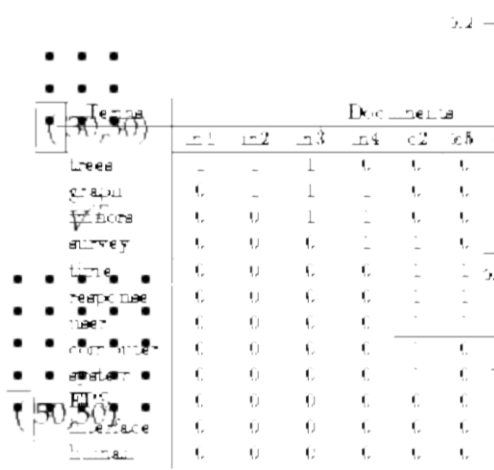
Поиск бывает разным: **полнотекстовый поиск**

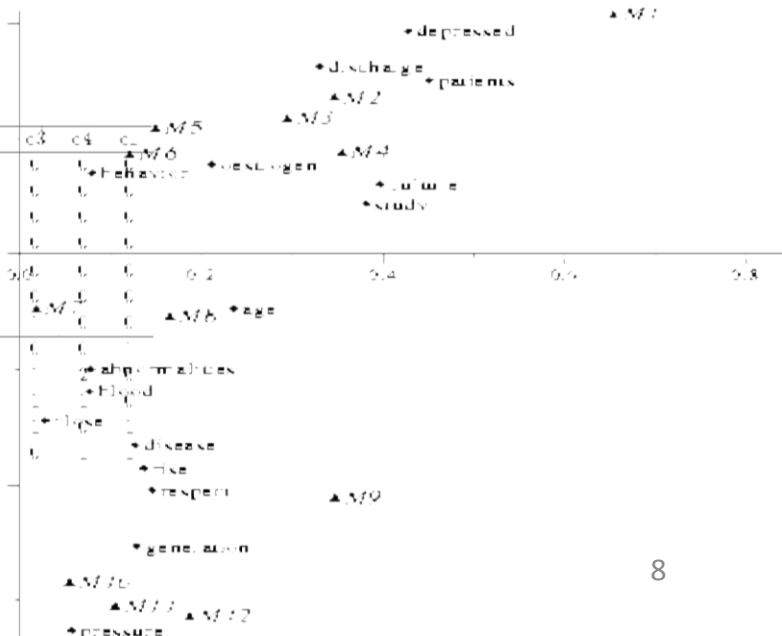


Предупреждение!

- Возможны оговорки
  - «документ» = «страница в вебе»
  - «терм» = «слово в документе»







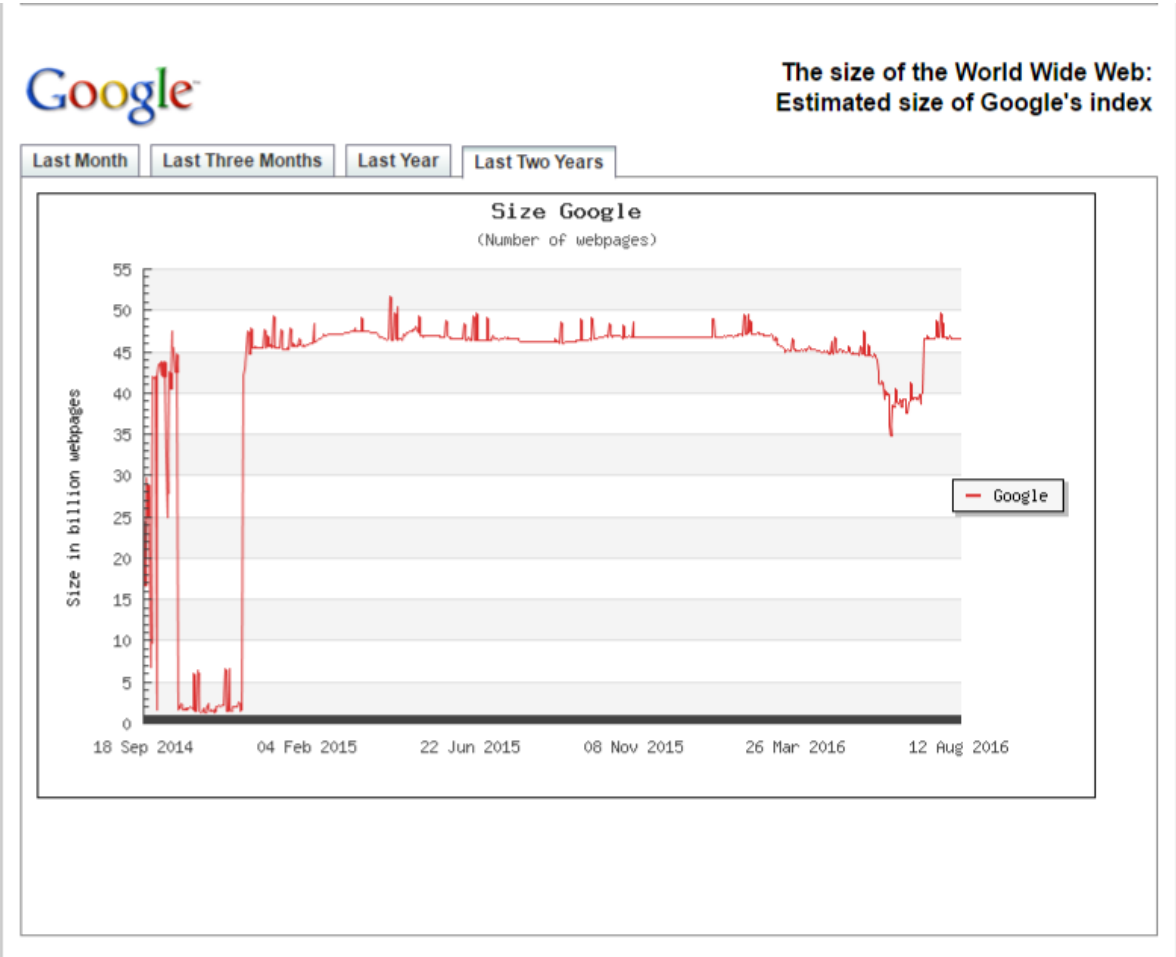
ure 18.1 Illustration of the singular-value decomposition. In this schematic illus- of (232), we see two cases illustrated. In the top half of the figure, we have a C for which $M > N$. The lower half illustrates the case $M < N$.

Поиск по страницам в Интернет – что сложного?

1. Собрать страницы
2. Искать по ним
3. Брать деньги за рекламу
4. ???????
5. PROFIT!


NO

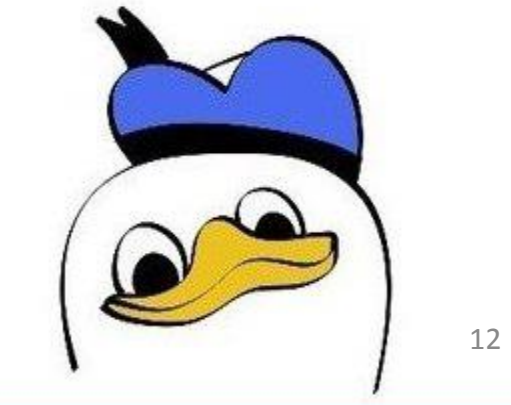
Поиск по страницам в Интернет – что сложного?



http://www.worldwidewebsize.com/

Поиск по страницам в Интернет – что сложного?

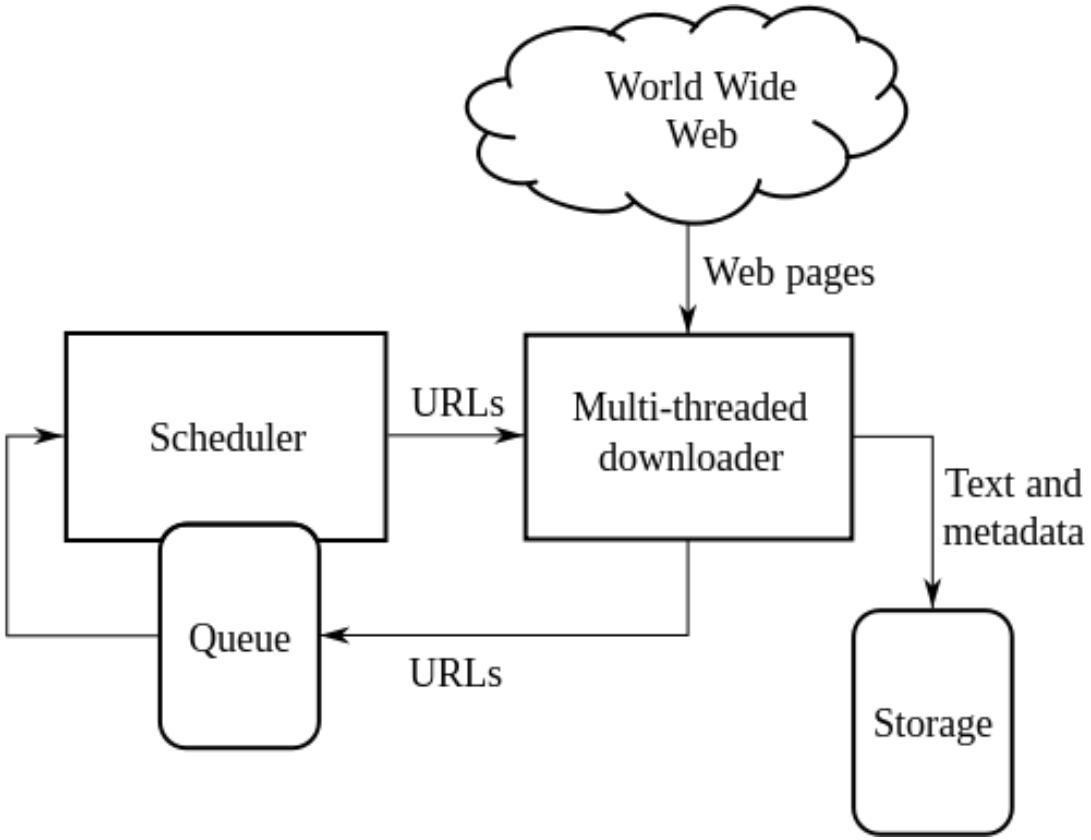
1. Документов *очень* много
2. Документы шумные
3. Страницы и ссылки умирают и меняются
4. Море разных языков
5. и т. д.



«Как сделать свой Яндекс»

1. Обход графа документов
2. Хитрое преобразование данных
3. Хитрая укладка данных (индексация)
4. Выборка по индексу по запросу
5. Задание порядка документов по ценности (ранжирование)

Обход и свежесть



«Как сделать свой Яндекс»

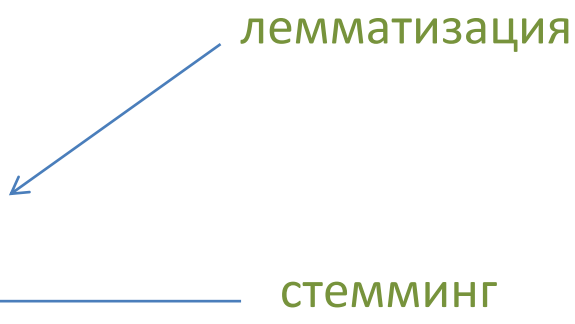
1. Обход графа документов
2. Хитрое преобразование данных
3. Хитрая укладка данных (индексация)
4. Выборка по индексу по запросу
5. Задание порядка документов по ценности (ранжирование)

Хитрое преобразование данных

«На берегу пустынных волн»</p>

на берегу пустынных волн

«русский язык!»



на берег пустын**ный** волн**а**
(на берег пустын волн)

берег пустынный волна

Нормализация слов: стемминг

- Классика: стемминг, Мартин Портер, 1980

(F)

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Нормализация слов: лемматизация

- **Лемматизация**: словарь и «умное угадывание»

./mystem.exe -e utf-8 -i --format json -d --weight

варькалось. хливкие шорьки


```
[
{"text":"варькалось","analysis":[
{
      "wt":0.6584257483,
      "lex":"варькаться",
      "gr":"V,несов,нп=прош,ед,изъяв,сред",
      "qual":"bastard"
}]},

{"text":"хливкие","analysis":[
{"wt":0.9958436489,"lex":"хливкий","gr":"A=вин,мн,полн,неод","qual":"bastard"},
{"wt":0.9958436489,"lex":"хливкий","gr":"A=им,мн,полн","qual":"bastard"}]},

{"text":"шорьки","analysis":[
{"wt":0.3092010319,"lex":"шорька","gr":"S,жен,неод=вин,мн","qual":"bastard"},
{"wt":0.3092010319,"lex":"шорька","gr":"S,жен,неод=род,ед","qual":"bastard"},
{"wt":0.3092010319,"lex":"шорька","gr":"S,жен,неод=им,мн","qual":"bastard"}]}
]
```

Определение языка

1. Статистика использованных символов

   א – «либо иврит, либо математика» ☺
   ѯ – «церковнославянский»

2. Марковские модели (*что это?*)

Марковские модели

- Вероятность языка при условии, что встретилась та или иная последовательность

  p(**русский**|”_ноч_”) = p(“_”|”оч”) p(“ч”|”но”) p(“о”|”_н”) …p(“_”)
  [п.ч. на русских текстах]

  p(**белорусский**|”_ноч_”) = p(“_”|”оч”) p(“ч”|”но”) p(“о”|”_н”) …p(“_”)
  [п.ч. на белорусских текстах]

  Самый простой способ оценить вероятность множителей:

  p(x|yz) = $\frac{\text{сколько\_раз\_встретилось}(\text{"}\boldsymbol{yzx}\text{"})}{\text{сколько\_раз\_встретилось}(\text{"}\boldsymbol{yz}\text{"})}$

- **p(белорусский|”_ноч_”) > p(русский|”_ноч_”) –**

  значит, белорусский текст

Какие ещё хитрости?

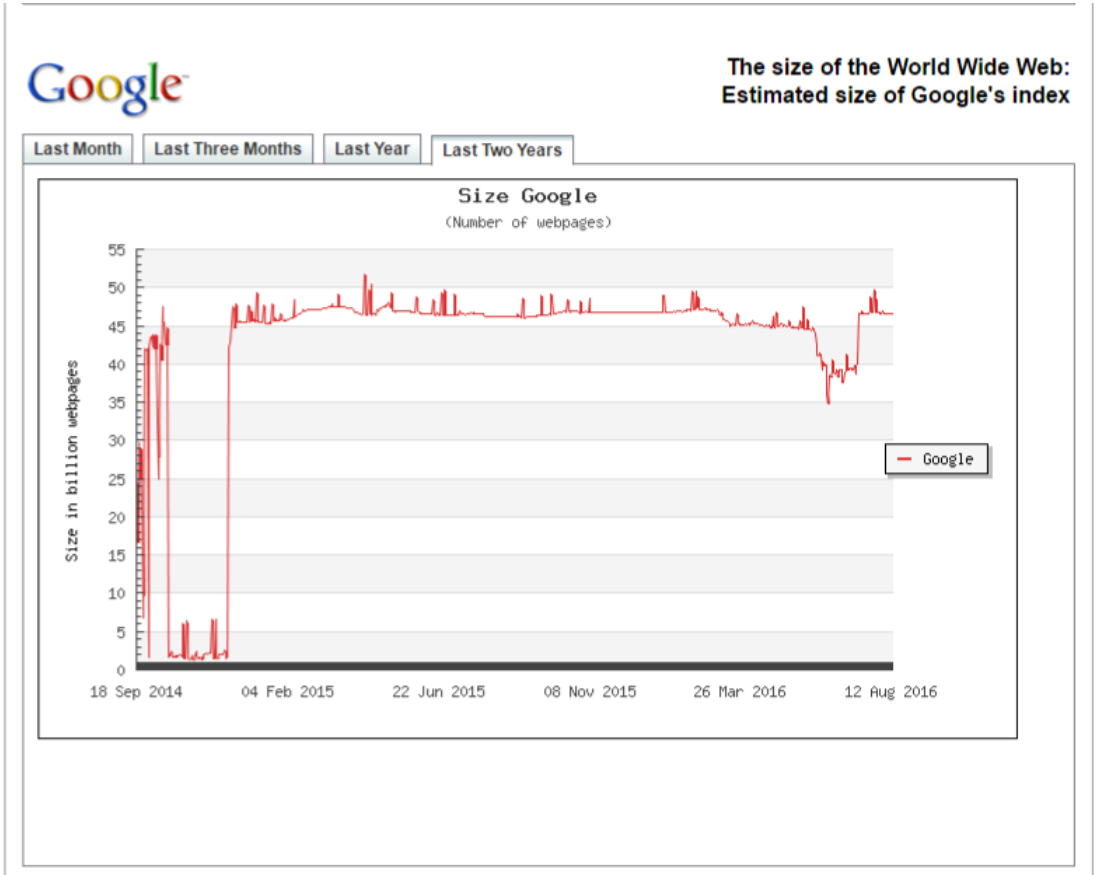
- Тысячи их
  - исправление опечаток и орфографических ошибок
    «На рублиштейна»!
  - Неизвестные слова люди пишут «как слышат»
    prittany spirse = britney spears

«Как сделать свой Яндекс»

1. Обход графа документов
2. Хитрое преобразование данных
3. Хитрая укладка данных (индексация)
4. Выборка по индексу по запросу
5. Задание порядка документов по ценности (ранжирование)

Как хранить и искать

- Напоминание: на каждый запрос просто пройти по документам мы не можем



Как хранить и искать

(1)
На берегу
пустынных волн

(2)
Море ты море,
Ты родина волн

(3)
Меня, колеблемого
на морской волне

Частоты, позиции
в документе и т. д.

Запрос = булевская формула

Примеры запросов
«на волне»
«на волне» с расстоянием
«берег –(волны)»

Как хранить и искать

- Проблема 1: слово может и не встречаться в нужном документе!
  «Элитные машины» и страница о Mazerati

- Проблема 2: жулики, вплетающие в текст не соответсвтующие теме слова
  купить скачать бесплатно без смс без регистрации 100% не развод

- Об этом дальше

«Как сделать свой Яндекс»

1. Обход графа документов
2. Хитрое преобразование данных
3. Хитрая укладка данных (индексация)
4. Выборка по индексу по запросу
5. Задание порядка документов по ценности (ранжирование)

Ранжирование

Ранжирование

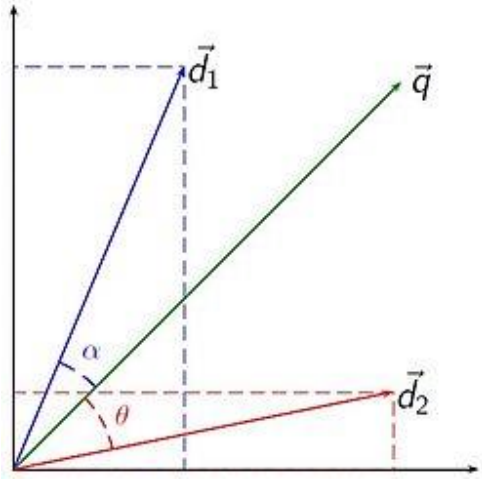
- Самый страшный «секрет фирмы»
- Миллиарды документов, надо top 10
- Учитывать
  1. Соответствие дока запросу
  2. Качество документа
  3. Соответствие интересам пользователя*
  4. Другое

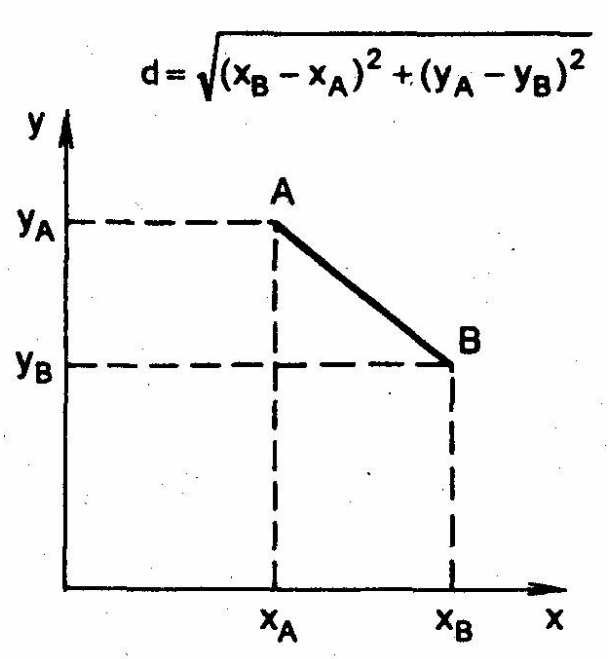
Классика: Vector Space Model

Документ огромный вектор **d** длиной в количество слов в словаре
Запрос огромный вектор **q** длиной в количество слов в словаре

абакан	абырвалг	азкабан	аксакал	аксолоть	аксон	берег	волна	заноза	…
0	0	0	0	0	0	1	1	0	…

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$





Есть много других моделей поиска…

Современные подходы, пример

- Посчитаем статистики и другие показатели
  - по документу
    - сколько по нему кликнули
    - сколько на него ссылок
    - …
  - по документу и запросу
    - доля слов из запроса в документе
    - …

(для разработки показателей есть специальные люди)

Современные подходы, пример

vec(d, q) =

Фича0	Фича1	Фича2	…
123.7	12.6	1	…

Rank(vec(d1, q)) > **Rank**(vec(d2, q))

<=>

d1 более релевантен, чем d2

Если в выдаче так же, как в размеченной аннотаторами выборке, то алгоритм подбора параметров «хвалят», иначе – «ругают»

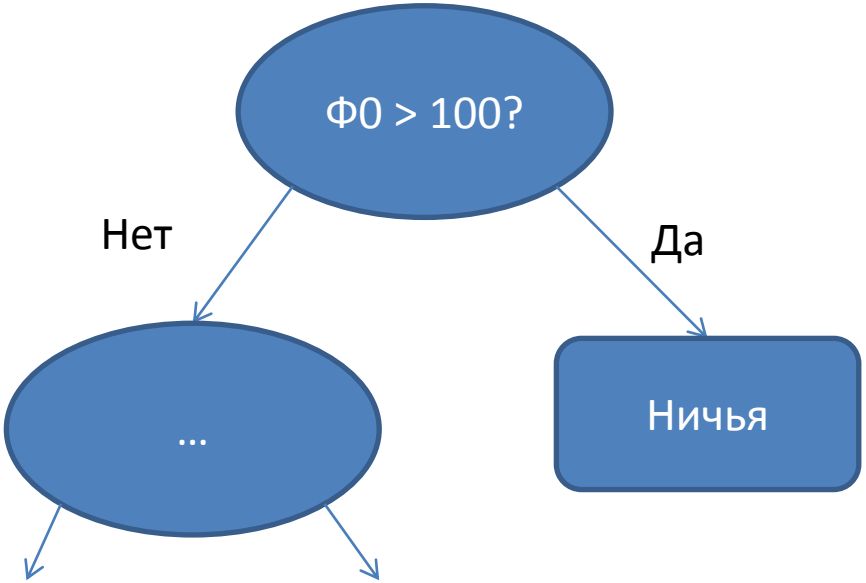
Что такое MatrixNet?

Gradient
Boosted
(Oblivious)
Regression
Trees

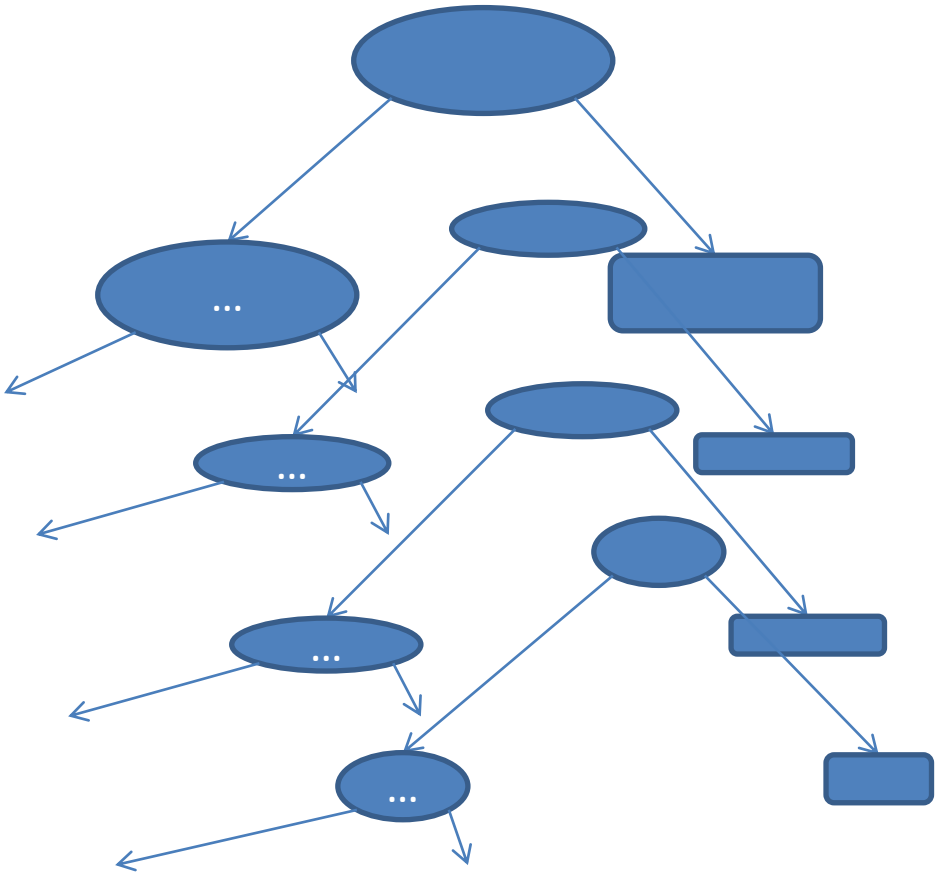
Решающие деревья

- Проще на примере о классификации

Ф0	Ф1	Ф2	Ф3	класс
12	0	56	474	Победа
3	0	4	46	Поражение
11	0	4	46	Победа
132	0	4	768	Ничья
155	0	4	53	Ничья
67	0	6	37	Поражение



Ансамбль решающих деревьев



Вы великолепны!
Базовый поиск готов.

Не пора ли закругляться?

- Много, много других и более «математичных» задач
  - исправление запросов
  - «расширение» запросов: морфология, семантика
  - персонализация поиска и подсказок
  - разрешение семантической неоднозначности запросов
  - извлечение фактов (hot!)
  - масса задач для **компьютерной лингвистики**
  - «умные» подсказки
  - классификация запросов
  - кластеризация документов
  - обнаружение дубликатов
  - распознавание вирусов на страницах
  - распознавание событий в мире

    …

Будет интересно

Research at Google

Home Publications People Teams Outreach Blog Work at Google

YAHOO! RESEARCH

research papers each year. Publishing is important to us; it enables us to collaborate and share ideas with, as well as learn from, the broader scientific community. Submissions are often made stronger by the fact that ideas have been tested through real product implementation by the time

tructures of publishing computer science ltiple ways of ion. We encourage ventional scientific her venues such as rds bodies, and open

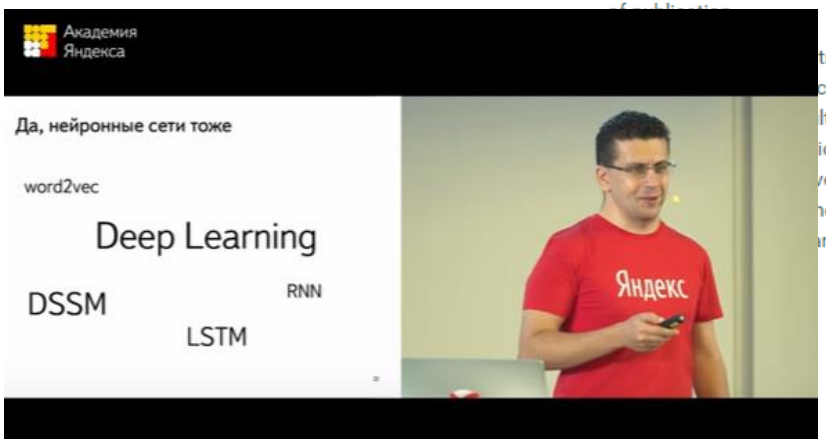
21 Research Areas

Algorithms and Theory

Data Management

Data Mining and Modeling

Distributed Systems and Parallel Computing



Спасибо :)

Прошу! Пожалуйста, перебивайте!
Задавайте вопросы во время рассказа!

Как сделать свой Яndex/Google

(ну, почти)

Антон Алексеев
асп. ПОМИ РАН, SofIT Labs
(ex-Яndex.Вертикали)

Прошу! Пожалуйста, перебивайте!
Задавайте вопросы во время рассказа!