

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

Как сделать свой Яндекс/Google (ну, почти)

Антон Алексеев
асп. ПОМИ РАН, SofIT Labs
(ex-Яндекс.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	AltheWeb	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

Telegram Web | Яндекс.Картинки: поиск | https://yandex.ru/images/search?img_url=https%3A%2F%2Fpp.vk.me%2F622318%2F622318247%2F4a7%2FPANDaVmJoi8.jpg&rpt=imageview

Яндекс | Картинки | Загруженная картинка | Найти | Войти

[← Вернуться назад](#)

Исходная картинка
400×600

Эта картинка в других размерах		
Большие	Средние	Маленькие
2025×3150	954×1024	349×541
1090×1652	808×606	341×480
1024×1619	800×800	340×270
1024×768	736×1086	338×500
	673×1000	332×400
	566×585	300×448

Похожие картинки

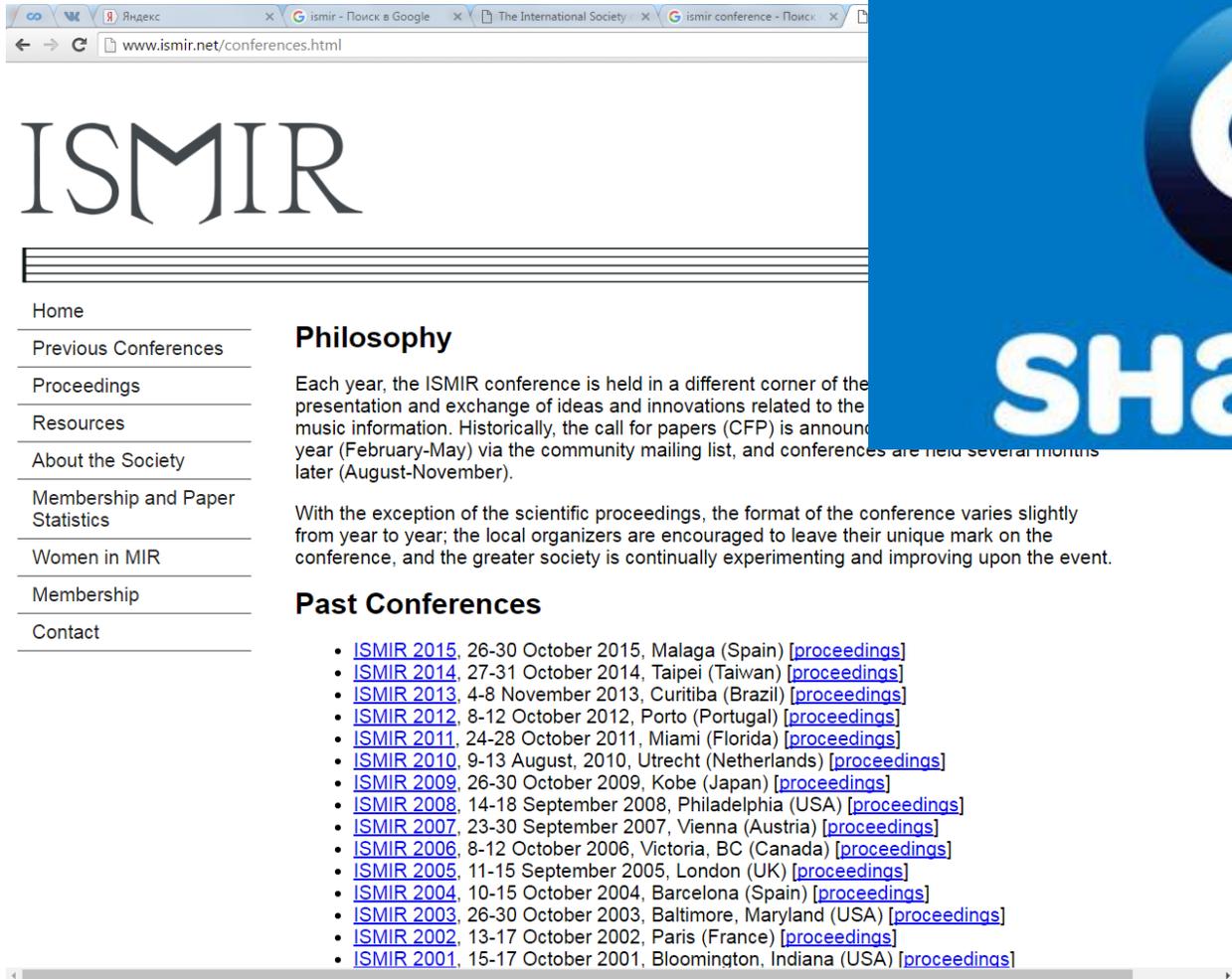
Еще [похожие](#)

Сайты, где встречается картинка

Искендер Камчибеков ВКонтакте
<http://vk.com/kamchi>
Некоторые из Вас слышали, некоторые нет.

motivation math eBay
http://www.ebay.com/sch/i.html?_nkw=motivation+math

Поиск бывает разным: музыка



The screenshot shows the ISMIR website with a navigation menu on the left and a main content area. The navigation menu includes: Home, Previous Conferences, Proceedings, Resources, About the Society, Membership and Paper Statistics, Women in MIR, Membership, and Contact. The main content area features a 'Philosophy' section with text about the conference's purpose and a 'Past Conferences' section with a list of events from 2001 to 2015, each with a link to its proceedings.

Home

Previous Conferences

Proceedings

Resources

About the Society

Membership and Paper Statistics

Women in MIR

Membership

Contact

Philosophy

Each year, the ISMIR conference is held in a different corner of the world for the presentation and exchange of ideas and innovations related to the music information. Historically, the call for papers (CFP) is announced one year (February-May) via the community mailing list, and conferences are held several months later (August-November).

With the exception of the scientific proceedings, the format of the conference varies slightly from year to year; the local organizers are encouraged to leave their unique mark on the conference, and the greater society is continually experimenting and improving upon the event.

Past Conferences

- [ISMIR 2015](#), 26-30 October 2015, Malaga (Spain) [[proceedings](#)]
- [ISMIR 2014](#), 27-31 October 2014, Taipei (Taiwan) [[proceedings](#)]
- [ISMIR 2013](#), 4-8 November 2013, Curitiba (Brazil) [[proceedings](#)]
- [ISMIR 2012](#), 8-12 October 2012, Porto (Portugal) [[proceedings](#)]
- [ISMIR 2011](#), 24-28 October 2011, Miami (Florida) [[proceedings](#)]
- [ISMIR 2010](#), 9-13 August, 2010, Utrecht (Netherlands) [[proceedings](#)]
- [ISMIR 2009](#), 26-30 October 2009, Kobe (Japan) [[proceedings](#)]
- [ISMIR 2008](#), 14-18 September 2008, Philadelphia (USA) [[proceedings](#)]
- [ISMIR 2007](#), 23-30 September 2007, Vienna (Austria) [[proceedings](#)]
- [ISMIR 2006](#), 8-12 October 2006, Victoria, BC (Canada) [[proceedings](#)]
- [ISMIR 2005](#), 11-15 September 2005, London (UK) [[proceedings](#)]
- [ISMIR 2004](#), 10-15 October 2004, Barcelona (Spain) [[proceedings](#)]
- [ISMIR 2003](#), 26-30 October 2003, Baltimore, Maryland (USA) [[proceedings](#)]
- [ISMIR 2002](#), 13-17 October 2002, Paris (France) [[proceedings](#)]
- [ISMIR 2001](#), 15-17 October 2001, Bloomington, Indiana (USA) [[proceedings](#)]



Поиск бывает разным: ПОЛНОТЕКСТОВЫЙ ПОИСК

The screenshot displays a search engine interface with the following elements:

- Search Bar:** Contains the text "когда б не зной да пыль" and a pink "SEARCH" button.
- Navigation:** Tabs for "Web", "Images", "Videos", "News", "More", and "Tools". A "SafeSearch" dropdown is visible on the right.
- Results:**
 - Summary: "About 44,400 results"
 - Section: "Web Results"
 - Result 1: ["Когда б не зной, да пыль, да комары, да мухи.": skvernoslov](http://skvernoslov.livejournal.com/203439.html)
 - Result 2: [«Ох, лето красное! любил бы я тебя, Когда б не зной, да ...»](http://04.rosпотреbnadzor.ru/index.php/epid-otdel/44-epid-otdel/1996...)
 - Result 3: [Осень \(Отрывок — Пушкин\) — Викитека](https://ru.wikisource.org/wiki/Осень_(Отрывок...)

Предупреждение!

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

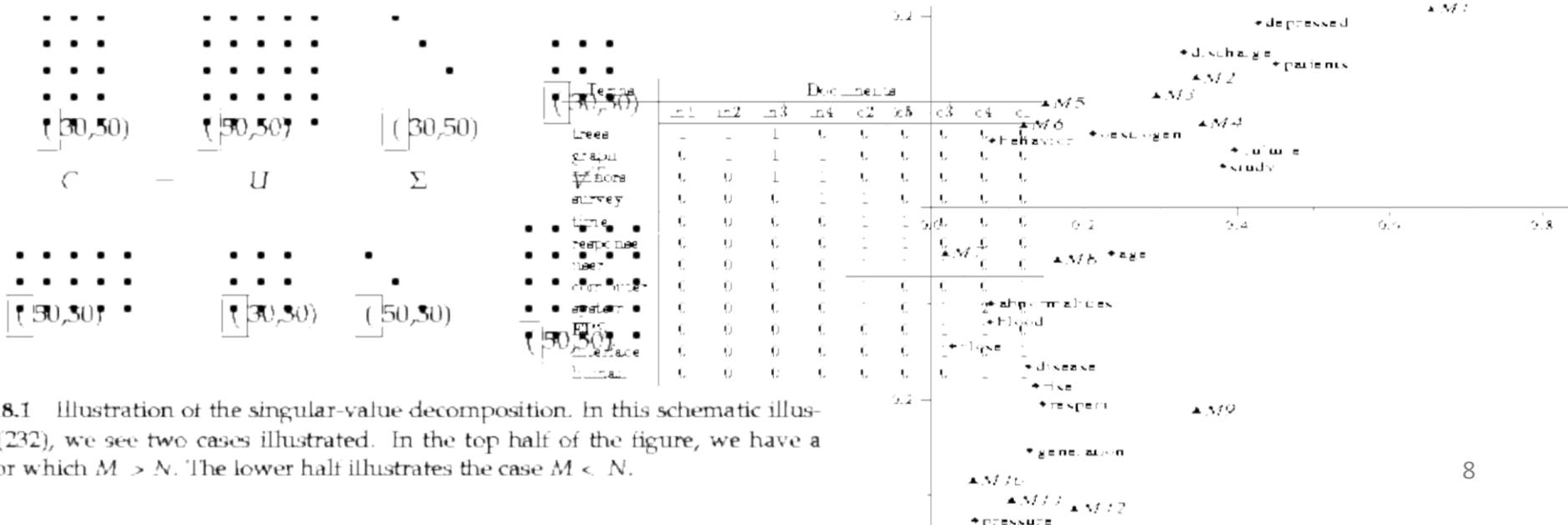


Figure 18.1 Illustration of the singular-value decomposition. In this schematic illustration of (232), we see two cases illustrated. In the top half of the figure, we have a C for which $M > N$. The lower half illustrates the case $M < N$.

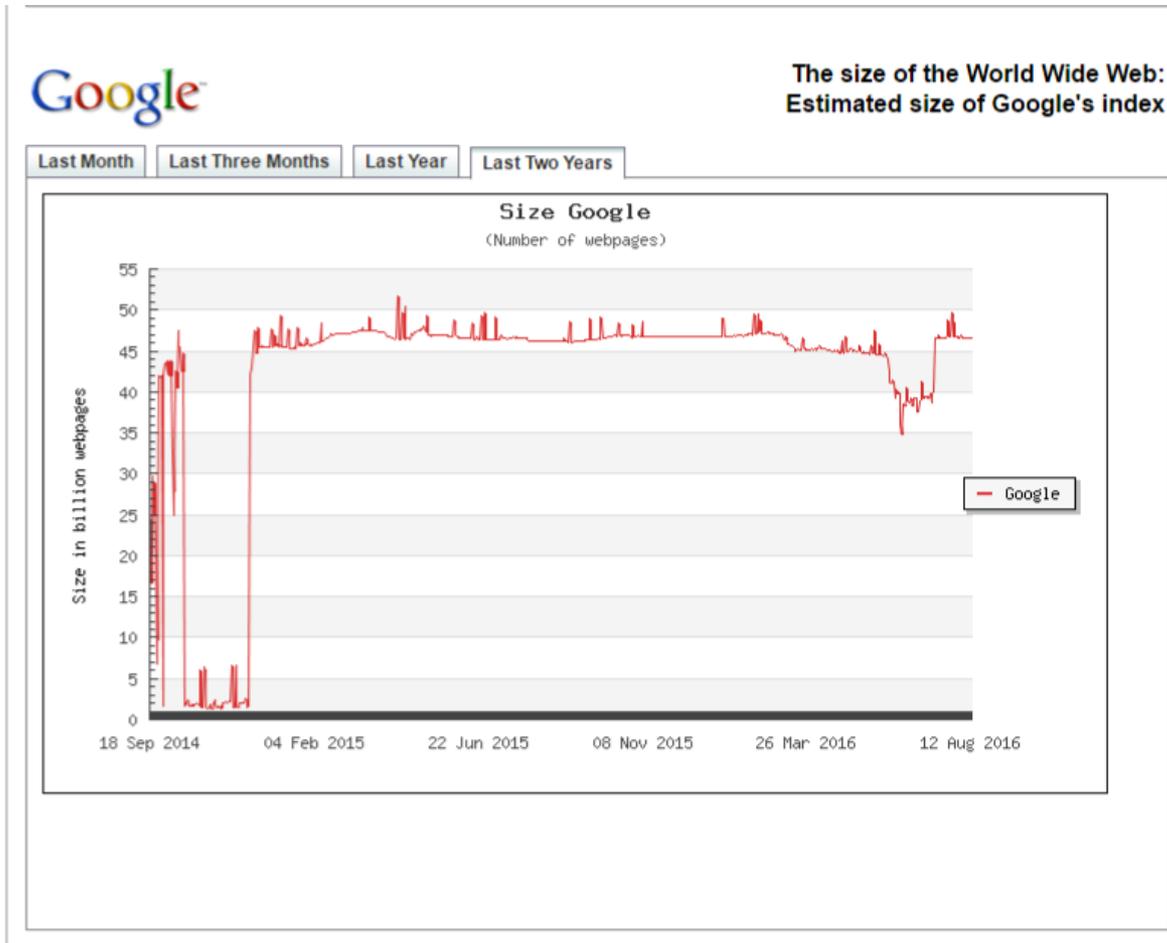
Поиск по страницам в Интернет – что сложного?

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





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.

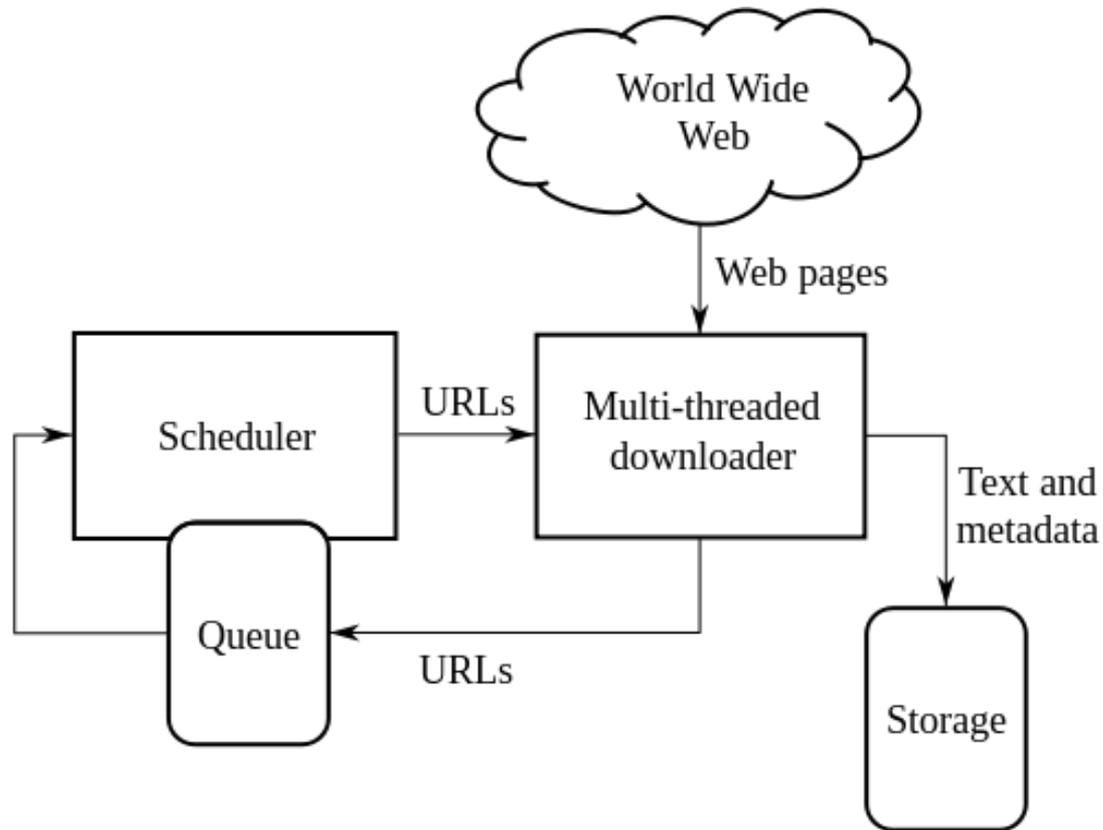
ВСЕ ОЧЕНЬ СЛОЖНО



«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»

на берегу пустынных волн

«русский язык!»

на берег пустынный волна

(на берег пустын волн)

берег пустынный волна

лемматизация

стемминг

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

\mathcal{X} – «либо иврит, либо математика» 😊

\mathcal{Y} – «церковнославянский»

2. Марковские модели (*что это?*)

Марковские модели

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

$$p(\text{русский} | \text{"_ноч_"}) = p(\text{"_" | "оч"}) p(\text{"ч" | "но"}) p(\text{"о" | "_н"}) \dots p(\text{"_"})$$

[п.ч. на русских текстах]

$$p(\text{белорусский} | \text{"_ноч_"}) = p(\text{"_" | "оч"}) p(\text{"ч" | "но"}) p(\text{"о" | "_н"}) \dots p(\text{"_"})$$

[п.ч. на белорусских текстах]

Самый простой способ оценить вероятность множителей:

$$p(x | yz) = \frac{\text{сколько_раз_встретилось}("yzx")}{\text{сколько_раз_встретилось}("yz")}$$

- $p(\text{белорусский} | \text{"_ноч_"}) > p(\text{русский} | \text{"_ноч_"})$ –
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

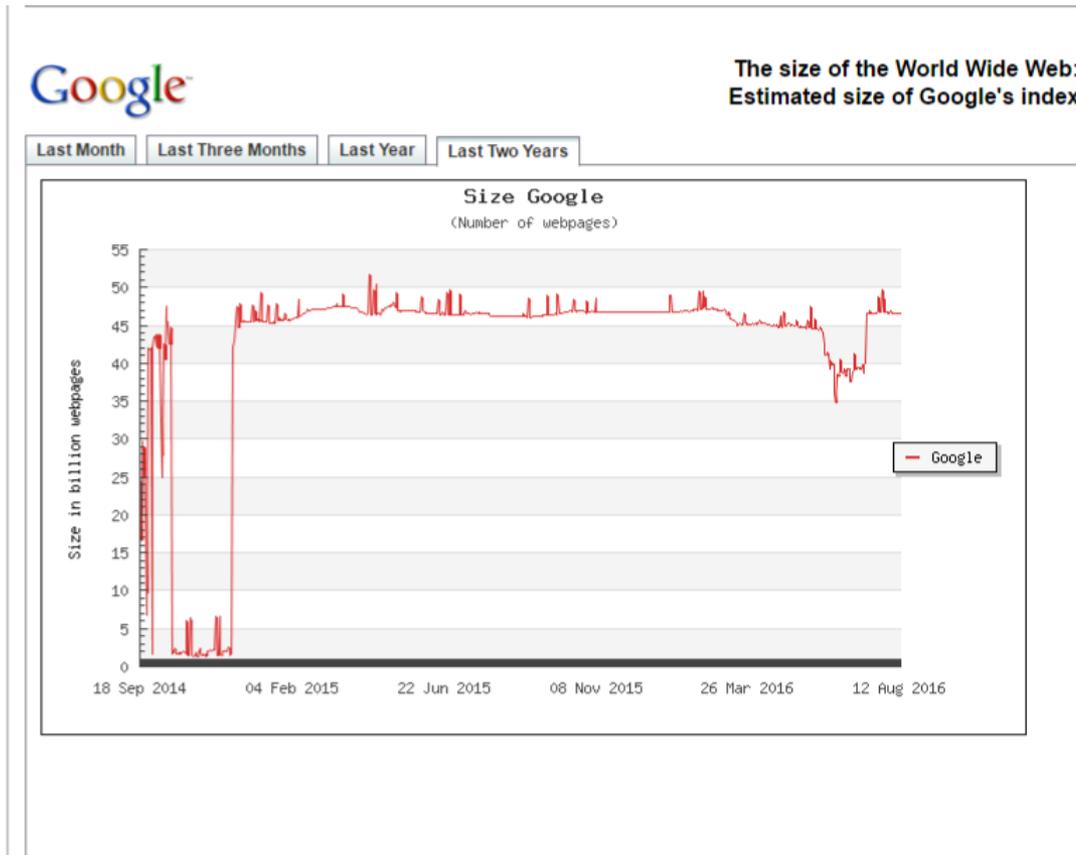


«Как сделать свой Яндекс»

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

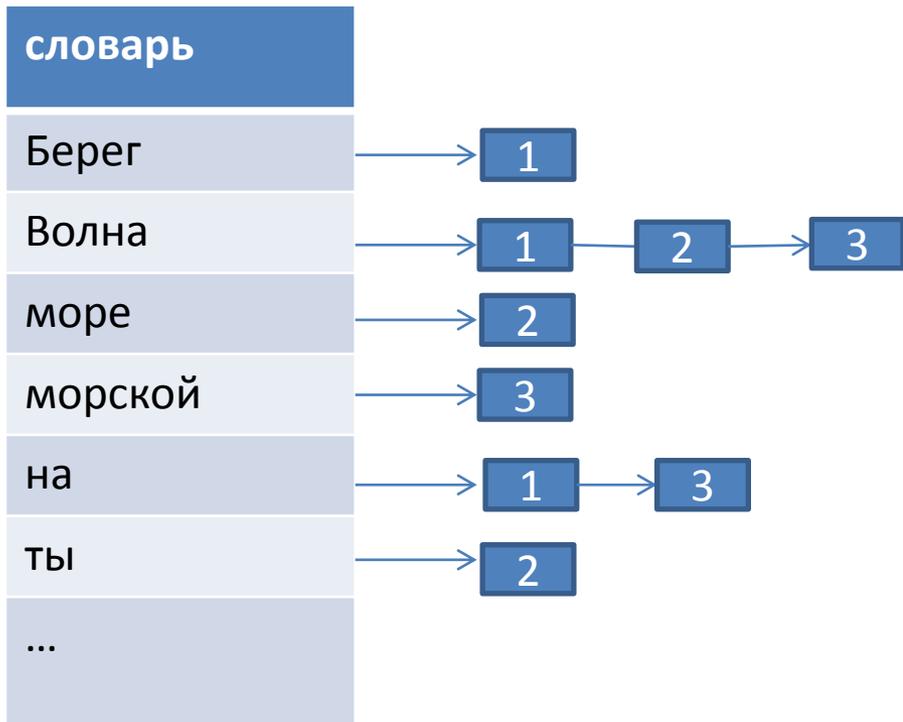


Как хранить и искать

(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 Maserati
- Проблема 2: жулики, вплетающие в текст не соответствующие теме слова
купить скачать бесплатно без смс без регистрации 100% не развод
- Об этом дальше

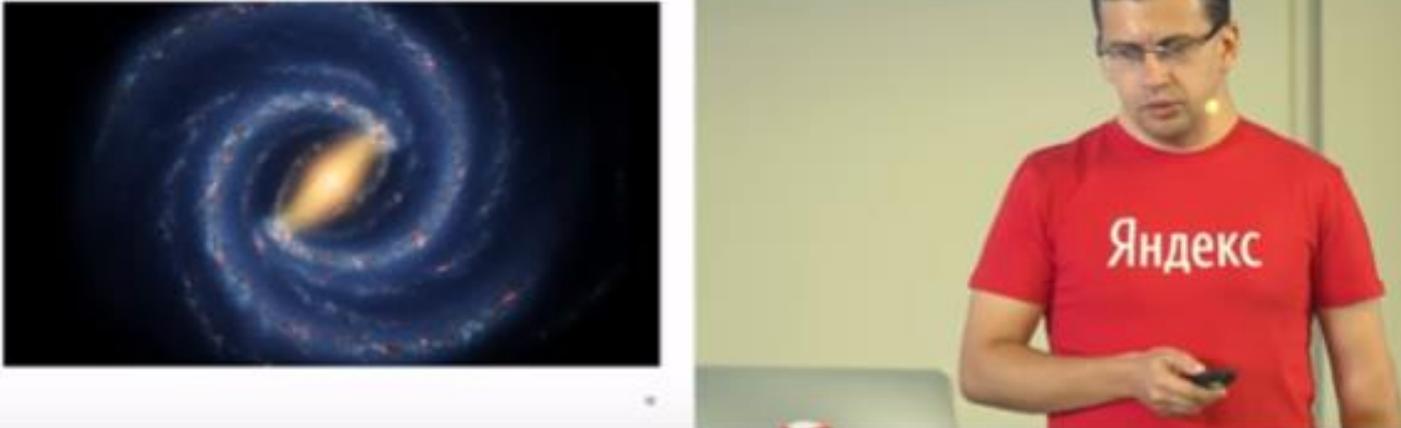
«Как сделать свой Яндекс»

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

Академия
Яндекса

Ранжирование — это сложно



2:28:42 / 3:25:10

Settings and Full Screen icons

The image is a video player interface. At the top left, there is a logo for 'Академия Яндекса' (Yandex Academy) consisting of a 3x3 grid of colored squares. Below the logo, the text 'Академия Яндекса' is displayed. The main content area is split into two panels. The left panel shows a title 'Ранжирование — это сложно' (Ranking is difficult) above a high-resolution image of a spiral galaxy with a bright yellow core and blue-tinted arms. The right panel shows a man with glasses and a red t-shirt with the word 'Яндекс' (Yandex) written on it in white. He is looking down at a small device in his hands. At the bottom of the video player, there is a black control bar with a red progress line. On the left side of the bar are icons for play, next, and volume. In the center, the time '2:28:42 / 3:25:10' is shown. On the right side, there are icons for settings (a gear) and full screen (a square).

Ранжирование

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

TOP 10

The text 'TOP 10' is written in a bold, red, hand-drawn style. Below it is a solid black horizontal line. At the bottom right, a hand is visible holding a black marker, positioned as if it just finished drawing the line.

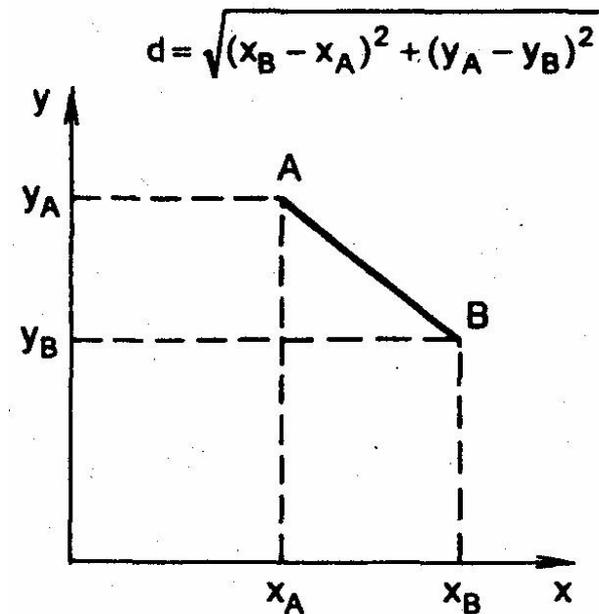
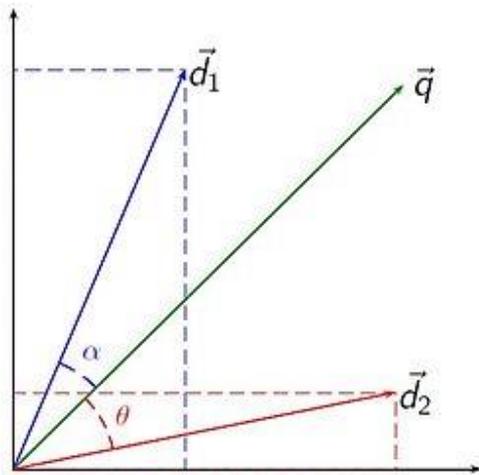
Классика: Vector Space Model

Документ огромный вектор \mathbf{d} длиной в количество слов в словаре

Запрос огромный вектор \mathbf{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

$\text{vec}(d, q) =$

Фича0	Фича1	Фича2	...
123.7	12.6	1	...

$$\text{Rank}(\text{vec}(d1, q)) > \text{Rank}(\text{vec}(d2, q))$$

\Leftrightarrow

$d1$ более релевантен, чем $d2$

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

Что такое MatrixNet?

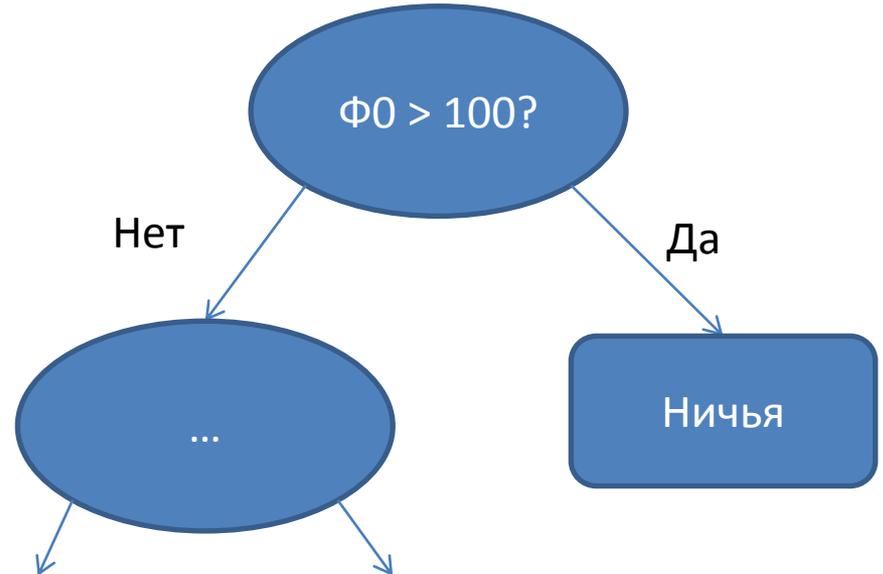
Gradient
Boosted
(Oblivious)
Regression
Trees



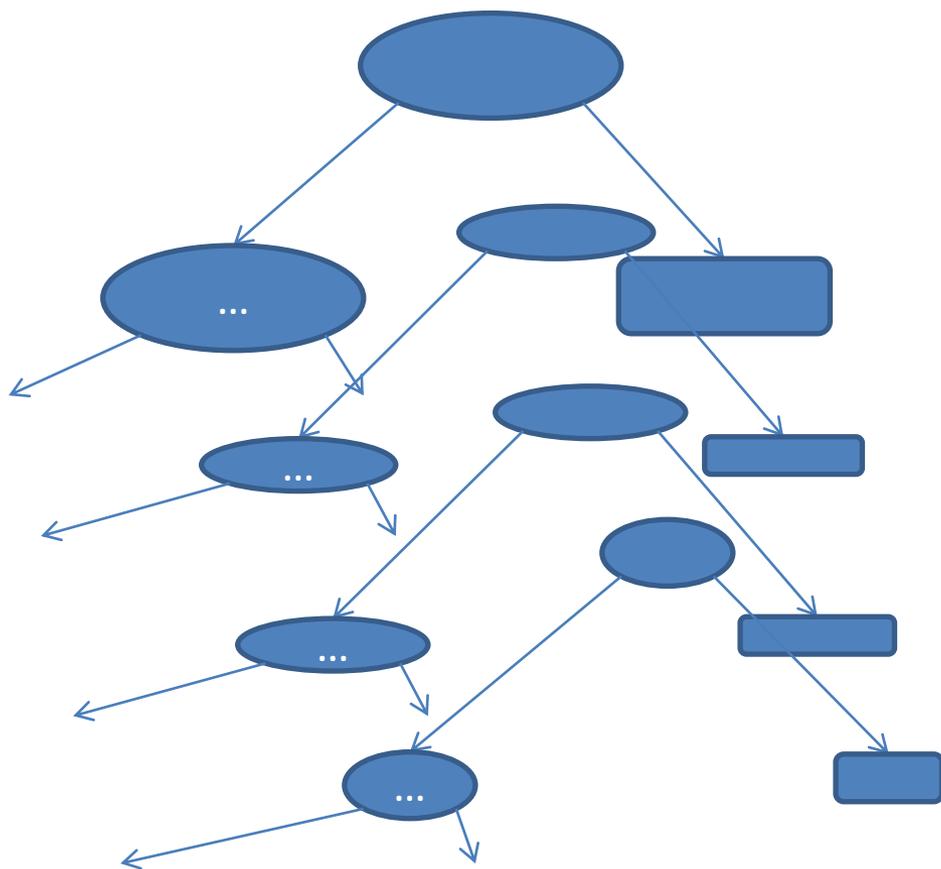
Решающие деревья

- 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



[Home](#) [Publications](#) [People](#) [Teams](#) [Outreach](#) [Blog](#) [Work at Google](#)

YANHOO! 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 of publication.

structures of publishing in computer science. Multiple ways of publication. We encourage traditional scientific venues such as journals, books, 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 Яндекс/Google (ну, почти)

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

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