



# **Ранжирование: от строчки кода до Матрикснета**

**И**лья Сегалович, **Ф**едор Романенко  
Технологии и разработка

Студень, Новосибирск, 24 сентября 2010 г.

# Ранжирование – что это?

- Основной алгоритм в поиске.
- С помощью **факторов** вычисляет **релевантность** документа в виде числа.
- ыбирает топ-10.
- сложная вещь с простым результатом
- определяет долю на рынке
- самый большой секрет поисковых компаний

# Ранжирование – что это?

- Основной алгоритм в поиске.
- С помощью **факторов** вычисляет **релевантность** документа в виде числа.
- Выбирает топ-10.
- Сложная вещь с простым результатом.
- Определяет долю на рынке.
- Самый большой секрет поисковых компаний.

# Язык запросов

(Yahoo – 1994 \*)

**Бинарное ранжирование: 0 или 1?**

- формируем правило отбора релевантных документов, используя «язык запросов»
- просматриваем все найденные документы

— *подходит только для специалистов*

— *нельзя использовать, когда найдено много*

\* здесь и далее: первый пример удачного применения в интернет поиске

Я

# Текстовое ранжирование: $tf * idf$

(Altavista – 1995)

**«Близость» текста запроса к тексту документа учитывает:**

- количество слов запроса в документе (term freq.)
- обратную частоту слова в языке (inverted document freq.)
- длину документа и запроса

**Текстовая релевантность также учитывает:**

- слова в заголовках
- близость слов запроса
- совпадение словоформ

# Лингвистика – понимание языка

(Яндекс – 1997)

- морфология: *ребенок шёл = дети идут*
- опечатки: *аднакласники = одноклассники*
- расширения: *МГУ = Московский Государственный Университет*

## Два пути компьютерной лингвистики:

- словарный: морфология, синтаксис
- статистический: языковые модели

# Рейтинги сайтов

(Рамблер Top100 – 1997, Яндекс тИЦ – 1999)

## Использование внешней информации о странице:

- ссылки с тематических сайтов
- переходы из каталогов
- посещаемость по счетчикам

— *тИЦ Яндекса оброс мифами вебмастеров*

# PageRank

(Google – 1998)

**PageRank – глобальная ссылочная авторитетность:**

- на страницу много ссылок
- на страницу есть ссылки с авторитетных страниц

**Модель случайного «блуждателя»:**

1. выбираем случайную страницу
2. с вероятностью  $0.85$  переходим по выходящей ссылке
3. с вероятностью  $0.15$  устаем и переходим к п.1



# Ссылочное ранжирование

(Google – 1998)

**Текст ссылки описывает страницу, на которую ссылается**

- голосование весами релевантных ссылок
- уточнение тематики страницы с помощью сопоставления текстов ссылок

# Ссылочное ранжирование (DirectHit)

**Релевантность** выше у тех страниц, на которые чаще переходят пользователи.

# Метрика качества поиска

Как измерить удовлетворение пользователей поиском?

- Из потока запросов пользователей строим репрезентативную **выборку**.
- **Асессоры** (люди) оценивают отдельные результаты поиска.
- **Оценки** используются в числовой **метрике** качества ранжирования.

**Яндекс**: оценено 4 млн. документов по 100 тыс. запросов

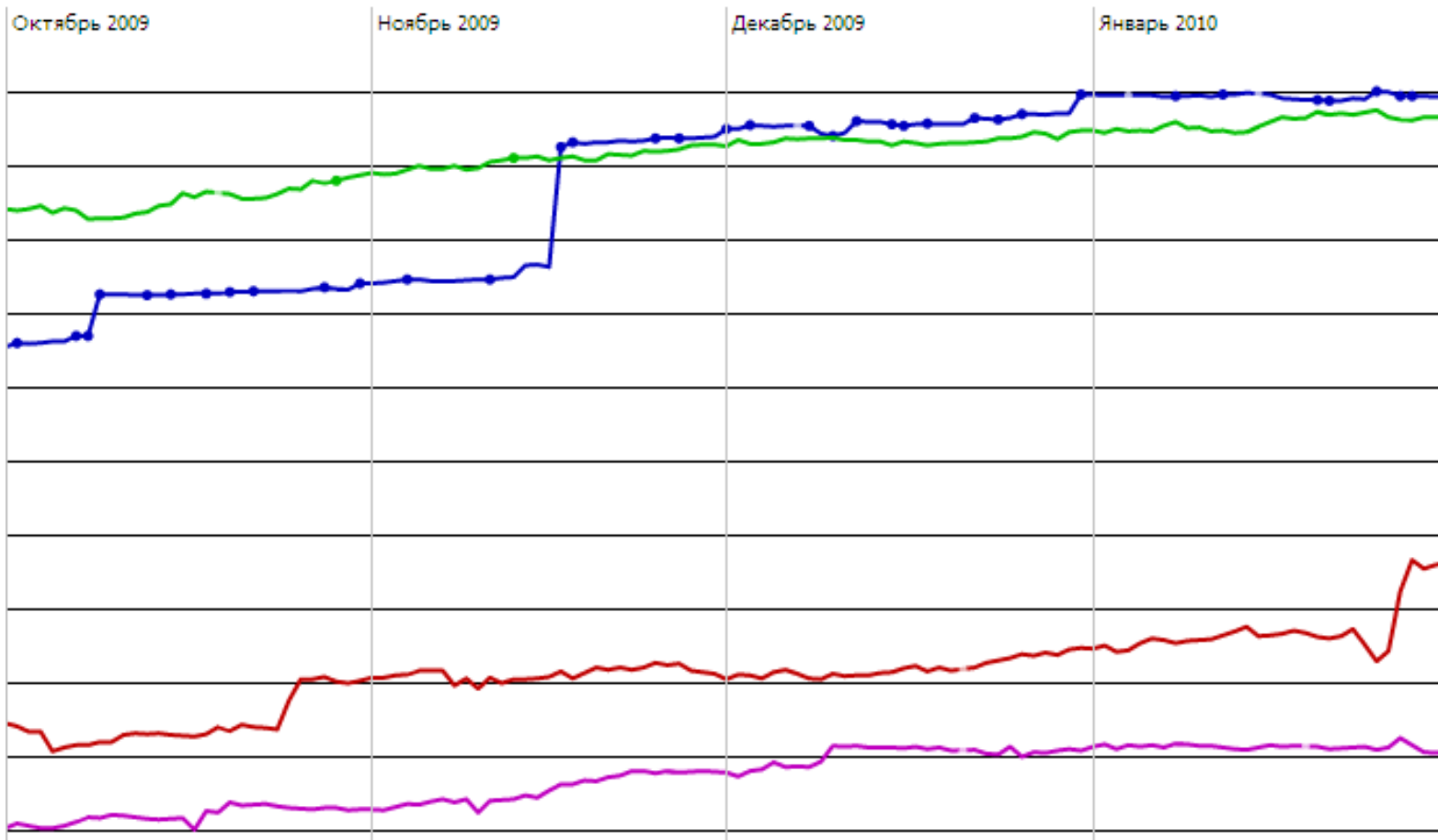
# Probability of User Satisfaction

## Optimization goal at Yandex since 2007

- >  $pFound$  – **P**robability of an answer to be **FOUND**
- >  $pBreak$  – **P**robability of abandonment at each position (**BREAK**)
- >  $pRel$  – **P**robability of user satisfaction at a given position (**RE**Levance)

$$pFound = \sum_{r=1}^n (1 - pBreak)^{r-1} pRel_r \prod_{i=1}^{r-1} (1 - pRel_i)$$

# Метрика качества поиска



# Машинное обучение

(Яндекс, Yahoo, Bing: 2006-2008)

Как учесть в ранжировании **400** параметров документов?

- **Модель** (формула на **факторах**) умеет считать релевантность документа.
- **Машинное обучение** автоматически настраивает модель, максимизируя метрику качества на примерах.
- Машина учится ранжировать и показывает поведение, не заложенное в нее явно.

# Проблемы машинного обучения

**Метрика** – *научить именно тому, что понадобится в бою*  
нужно математически измерить «счастье» пользователей.

**Входные данные** – *органы чувств*  
позволяют отличить документы по качеству.

**Способность к обобщению** – *аналитические способности*  
учимся на примерах, ранжируем то, что никогда не видели.

**Переобучение** – *паранойя*  
мало оценок и умная модель – ложные закономерности.

**Кроссвалидация** – *не дадим выучить контрольную*  
нужна не память, а умение.

# Матрикснет

(Яндекс – 2009)

Разработка **Яндекса** (релиз «Снежинск», ноябрь 2009)

- автоматически выбирает связанные факторы и диапазоны их значений;
- генерирует тысячи комбинированных факторов;
- очень сложная модель без склонности к переобучению.

— *извлекает неиспользованный сигнал из старых данных*

— *позволяет добавлять много данных в ранжирование*



# Формула Матрикснет

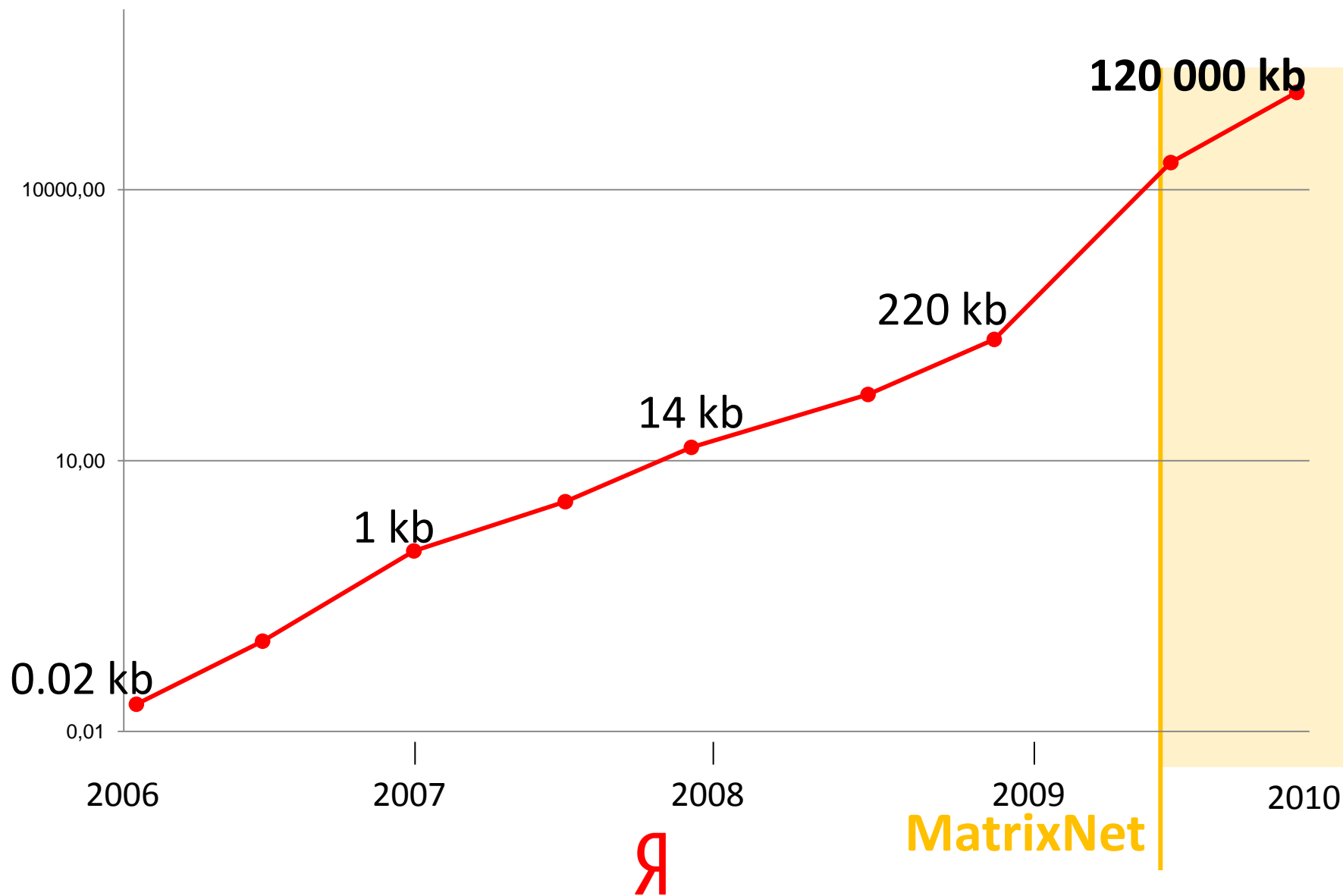
## Одна из формул

(вычисляет релевантность по факторам найденного документа)

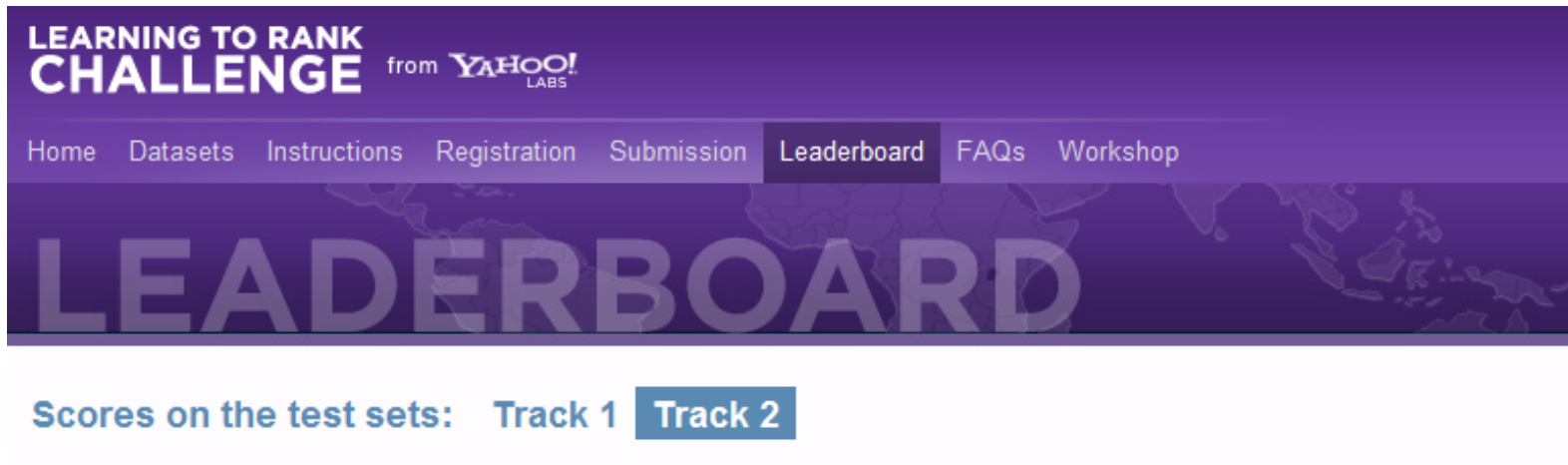
```
...  
  
-4025627,483990,-36437960,39979596,  
-92842056,-50086892,-100233272,243162224,  
-22282850,57163664,-24991620,-9889194,  
  
...  
  
vars[5699] = fFactorInt[376] > 1060970280 ? 1 : 0; // 0.738757  
vars[5700] = fFactorInt[376] > 1061923687 ? 1 : 0; // 0.795584  
vars[5701] = fFactorInt[376] > 1049721454 ? 1 : 0; // 0.284137  
vars[5702] = fFactorInt[376] > 948291011 ? 1 : 0; // 6.37877e-05  
  
...  
  
, {376, .0f}  
, {376, 0.6251018047f}  
, {-1, .0}  
, {376, 0.05682743713f}  
, {376, 0.4546194971f}  
  
...
```

- 20 тыс. строк на C++
- объем 3МБ
- 10 тыс. коэффициентов

# MLR: сложность формулы ранжирования



# Yandex MLR на соревнованиях по IR



LEARNING TO RANK CHALLENGE from YAHOO! LABS

Home Datasets Instructions Registration Submission **Leaderboard** FAQs Workshop

## LEADERBOARD

Scores on the test sets: [Track 1](#) **[Track 2](#)**

Rank	Team Name	ERR Score	NDCG Score
1	MN-U	0.463476	0.7863
2	arizona	0.463169	0.7876
3	Joker	0.463113	0.7887
4	ULG-PG	0.461686	0.7819
5	VeryGoodSignal	0.461632	0.7849
6	ya	0.461492	0.7828
7	WashU in Saint Louis	0.461184	0.7838
8	catonakeyboardinspace	0.461146	0.7833
9	CLTeam	0.460897	0.7815
10	yareg	0.460519	0.7782

# №1

MatrixNet at Yahoo Challenge: #1, 3, 10  
(Track 2), also BagBoo, AG



# Независимые метрики

У поиска много «качеств»!

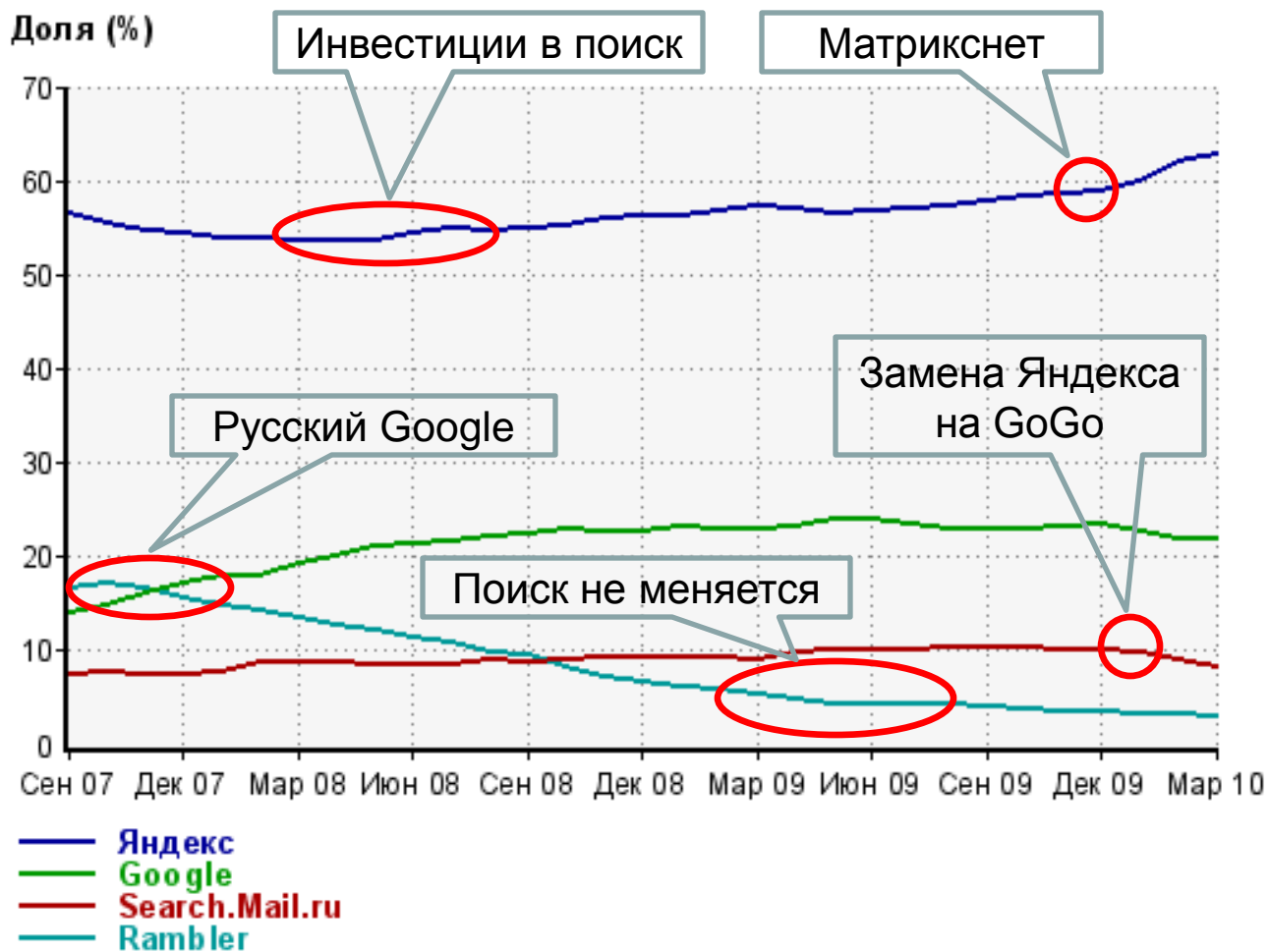
Навигационный поиск	Тематический поиск	Подсказки	Опечатки	Цитатный поиск	Оригиналы	Синонимы
Я 98.0 ↑3.1	Я 31.6 ↑1.1	@ 96.1 ↑0.6	b 81.8 ↑1.1	Я 66.0 ↑1.0	Я 55.0	Я 63.8 ↑0.4
Я 97.0 ↑1.1	Я 29.3 ↑3.6	Я 95.5 ↓0.3	Я 60.8 ↓0.1	Я 60.0	Я 42.0	Я 62.4 ↓0.2
Я 94.9 ↑2.0	Я 27.2 ↑2.1	Я 93.4 ↓0.9	@ 55.2 ↑0.3	Я 42.0 ↑2.0	Я 42.0	Я 48.3 ↓0.3
Я 94.9	b 18.1 ↓1.6	Я 91.4 ↑4.0	Я 46.1 ↓3.1	b 36.0 ↑1.0	b 38.0 ↓2.0	@ 47.7 ↓0.4
Я 91.9 ↓2.0	@ 17.5	Я 91.0 ↓2.0	Я 22.5 ↑0.3	Я 35.0 ↓1.0	Я 29.0	Я 47.3 ↓0.1
Я 90.9 ↑3.1	Я 17.5 ↓0.6	b 85.5 ↑0.7	Я 4.4 ↓1.2	@ 35.0 ↓1.0	Я 26.0 ↑2.0	Я 47.3 ↓0.1
@ 89.9 ↓4.0	Я 15.6 ↑0.9	Я 77.8 ↑0.6	Я 1.4 ↓0.4	Я 33.0	@ 26.0 ↑1.0	Я 37.4 ↑0.8
Я 78.8 ↓3.9	Я 2.7 ↓0.3	Я N/A	Я 0.3	Я 23.0 ↑4.0	Я 8.0 ↓4.0	Я 36.0 ↓0.4

Поисковый спам	SEO-прессинг	Порнография	Полнота индекса	Апдейты	Переходы
Я 0.8 ↑0.3	Я 27.7 ↑0.9	Я 6.3 ↑0.1	Я 75.0 ↑0.8	Я 0.2 ↑0.1	Я 63.2 ↑0.8
Я 3.3 ↑0.4	Я 34.2 ↑0.4	Я 9.6 ↓0.8	Я 64.5 ↑1.7	b 4.3 ↓0.6	Я 22.2 ↓1.8
Я 4.4 ↓0.2	@ 34.7 ↑0.2	b 14.1 ↑0.2	Я 59.2 ↓5.3	Я 4.6 ↑1.8	@ 8.9 ↑0.4
b 4.6	Я 40.8 ↓0.7	Я 15.1	Я 48.5 ↑1.5	Я 4.9 ↑0.3	Я 3.2 ↑0.4
@ 5.0 ↓0.4	b 42.4 ↑0.1	Я 15.8 ↓1.0	@ 42.7	Я 8.1 ↓2.8	b 1.0 ↑0.2
Я 5.0 ↓0.1	Я 43.3 ↓0.5	Я 16.8 ↑0.2	b 15.9 ↑3.3	Я 15.5 ↓0.8	Я 0.1
Я 6.3 ↓1.3	Я 45.5 ↓0.2	Я 20.3 ↓0.7	Я 7.8 ↓0.5	@ 15.9 ↓1.6	Я 0.0
Я 6.5 ↓0.3	Я 48.7 ↓0.8	@ 20.4 ↓0.8	Я 5.2 ↓0.1	Я 36.0 ↓35.4	Я 0.0

<http://www.analyzethis.ru>



# Liveinternet: доля рынка, Россия



# Поиск — вечная задача

- Поиск нельзя «доделать»: тот, кто думает, что все сделал, проигрывает.
- Размер интернета и количество пользователей растут, люди решают в интернете новые задачи.
- Много направлений: локальность, глобальность, свежесть, разнообразие, скорость, представление, ...
- Сильные математики и программисты имеют дело с уникальными задачами на огромных объемах данных.
- Результаты их работы сразу видят миллионы людей.



Илья Сегалович, Федор Романенко

Я