

# Яндекс

## Как устроен Поиск по блогам

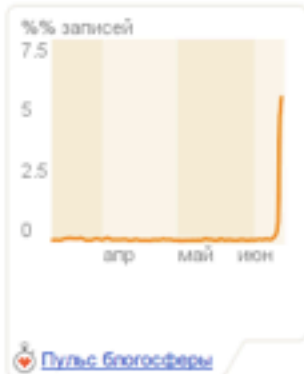
Антон Волнухин  
Киев, 16 июня 2009

# Яндекс

блоги



только в украинских блогах [расширенный поиск](#)



### Главные темы дня

- ▶ [Путин в галерее Ильи Глазунова](#)  
446 записей за три дня
- ▶ [Сборная России обыграла сборную Финляндии](#)  
4 812 записей
- ▶ [Книжный фестиваль](#)  
259 записей

За последние три дня 5 517 записей посвящено трём самым популярным сегодня темам.

### Остальные темы

- [Группа "Dream Theater" в Москве](#)
- [Английскому актеру Хью Лори - 50 лет](#)
- [Тополиный пух](#)
- [На "Черкизоне" обнаружен контрабандный товар](#)
- [Airbus A 330](#)
- [Компания "Русский вольфрам"](#)
- [Стрельба в музее Холокоста](#)
- [Саммит ШОС](#)

Из каталога: [Юмор](#) 66 блогов [Творчество](#) 292 [Развлечения](#) 344 [Дом](#) 181 [Технологии](#) 320 [Деловые](#) 192 [Ещё...](#)

## Самое популярное и обсуждаемое в интернете

### Сервисы

	LiveJournal	57 519
	Блоги@Mail.Ru	28 041
	Diary.ru	18 220
	LiveInternet	16 023
	Я.ру	10 329
	Blogger.com	4 892
	Love Planet	4 169
	Привет.ру	2 283
	24open.ru	1 525
	Блог.ру	1 273

Всего 95 сервисов

### Блоги

	druoi	218 240
	tema	159 799
	Леонид Каганов	104 483
	katoga	78 450
	abraximov	77 335
	radulova	66 242
	fritz morgen	59 735
	БЛОГбастер	56 435
	fima_psychoпад	53 899
	dolboeb	52 468

Всего

### Запросы

SUDBA
бизнесмена Юрия Будакова
Джеймс фон Брунн
энрики иглеснас
в в контакте
тополиный пух
Игорь Сахновский
арбенина диана
мигом
кулемина и степное

Топ 50 запросов

## Популярные записи

Перед просмотром рейтингов ознакомьтесь с их [описанием](#).

### Сводный рейтинг

- [Течет река Волга](#)  
3 ссылки, 135 комментаторов, + 11000 посетителей  
Цветочки и ягоды . © REUTERS/РИА-Новости, Дмитрий Астахов 10.06.2009, Россия | Поздравить левицу Людмилу Зыкину...
- [Bemie Goetz и его история](#)  
35 115 + 7000  
. Несколько лет назад я услышал во время обеденного перерыва на работе, как коллега-американец, читая газету...

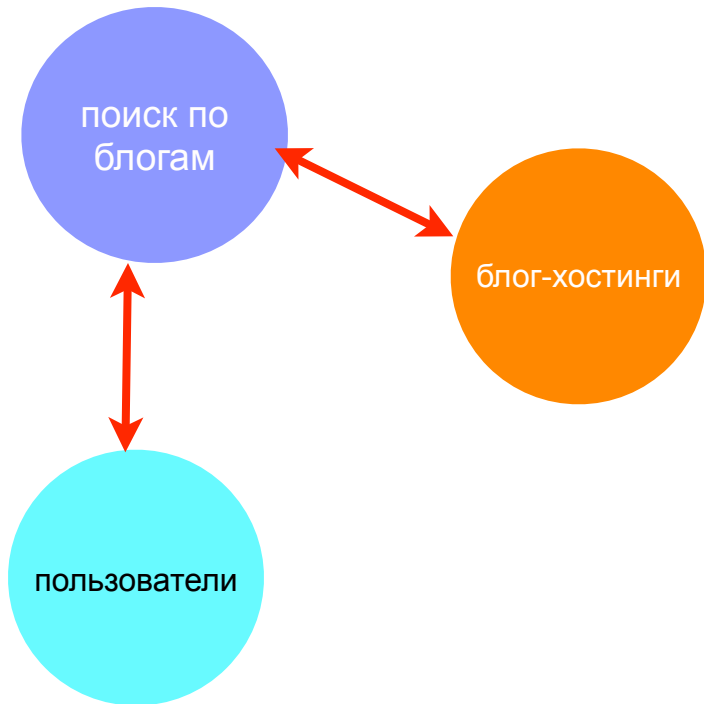
# Что такое Поиск по блогам?

Поиск по мнениям. Общественное мнение в интернете.

- Поиск по текстам, где люди говорят от первого лица:
  - что другие говорят о вас или ваших действиях
  - что пишут о товаре, который вы собираетесь купить
  - что пишут о вашей компании
  - что пишут о каком-то событии
  - сравнить обсуждаемость чего-либо
- Наиболее обсуждаемые темы и самое популярное в интернете сегодня

# Модель сервиса

- партнёрство и взаимодействие между:
  - блогхостингами
  - пользователями
- быть зеркалом блогосферы
- полностью автоматический сервис
- единые правила для партнёров
- открытые форматы (RSS, ATOM, FOAF)
- все наши API доступны публично



# Модель сервиса

- партнёрство и взаимодействие между:
  - блоггерами
  - блогхостингами
  - пользователями
- быть зеркалом блогосферы
- полностью автоматический сервис
- единые правила для партнёров
- открытые форматы (RSS, ATOM, FOAF)
- все наши API доступны публично

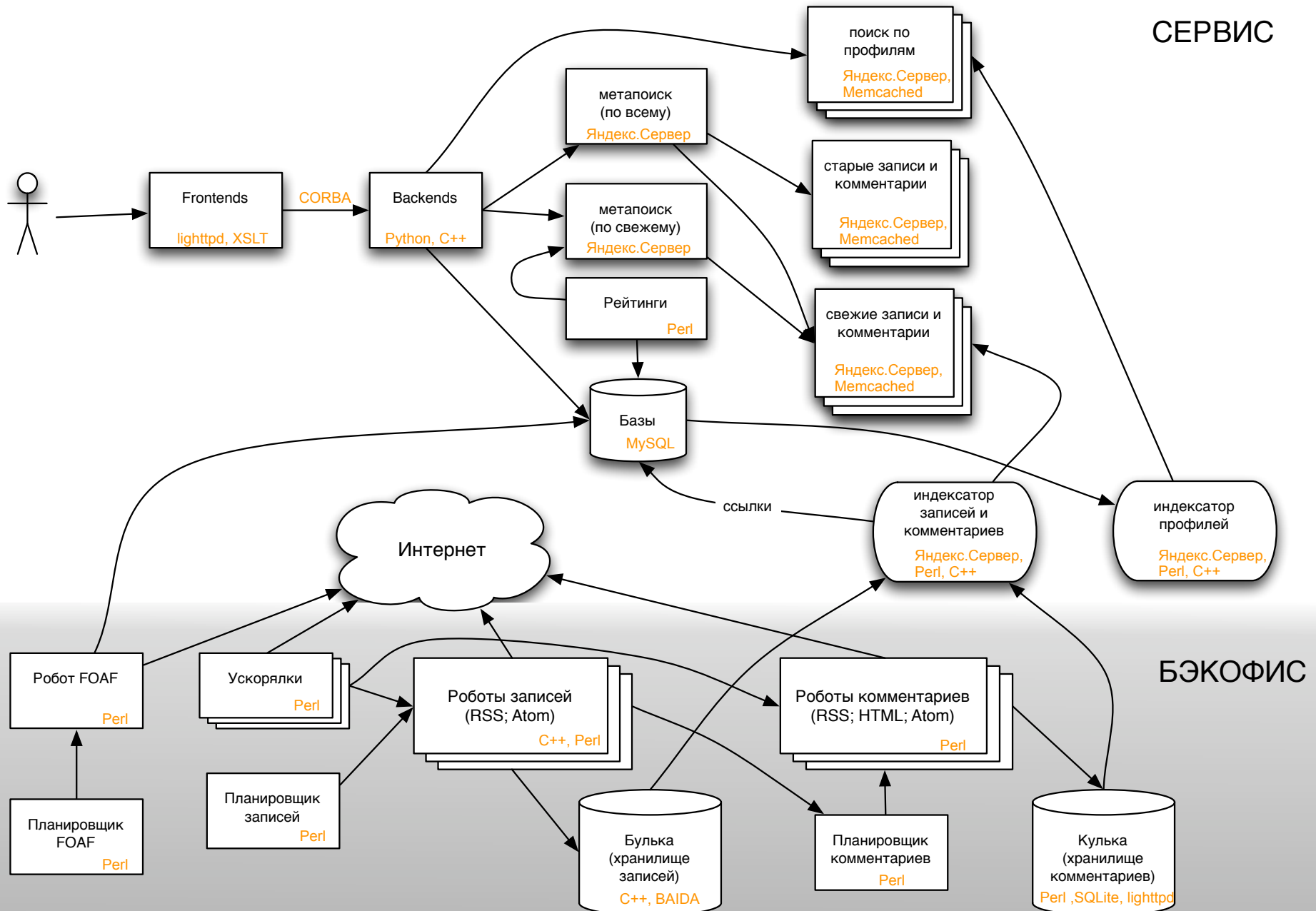


# Масштабы

- Около **миллиона** записей и комментариев из блогов и форумов каждый день
- Почти **10 миллионов** источников
- Всего около **миллиарда** документов

Поиск по блогам – это почти **одна пятая** от поиска по всему русскоязычному интернету по количеству элементов индексации.

# Внутреннее устройство



# Содержание

1. Поиск
2. Темы дня и популярные записи
3. Рейтинги
4. Пульс блогосферы

# 1. Поиск

На какие вопросы отвечает

По чему ищет

Чем отличается от поиска по всему  
интернету

# Как происходит индексирование

Сервис основан на распространённых в интернете открытых форматах. Благодаря сотрудничеству с владельцами блог-хостингов эти форматы (RSS, FOAF, Weblogs.Ping) стали стандартом в российской блогосфере.

- На данный момент новые записи индексируются в течение 10 минут с момента их появления на более чем **120 блог-хостингах**, включая:
  - LiveJournal.com
  - LiveInternet.ru
  - Blogs.mail.ru
  - Diary.ru
- Индексируются комментарии на LiveJournal.ru, LiveInternet.ru и многих автономных блогах
- Проиндексировано более **14 миллионов** профилей, включая профили пользователей пяти крупнейших блог-хостингов

# Как поиск узнаёт о новых блогах

- новые блоги на уже известных блог-хостингах добавляются, как только поиск по блогам получает пинг про первую запись
- из веб-поиска: когда веб-поиск находит новый сайт с известным блоговым движком
- из формы добавления [blogs.yandex.ru/add.xml](https://blogs.yandex.ru/add.xml)
- из веб-поиска: когда в интернете обнаруживается ссылка на новый RSS

# Проблема: RSS не равно блогу

В RSS транслируются не только блоги, но и форумы, новости, обновления страниц вики, объявления, курсы валют и многое другое

Поиск по блогам мог бы быть замусорен всем этим, если бы не была построена система разметки потоков:

- по умолчанию RSS-поток не считается блогом
- автоматические правила по разным свойствам RSS-потоков (URL, название, движок)
- контент-менеджеры исправляют ошибки роботов

Побочный эффект: автономный блог на нестандартном движке по умолчанию не считается блогом. Чтобы быстро это исправить, владелец блога может написать в службу поддержки.

# Как индексируются профили

Часто над результатами поиска по блогам есть "врезка", называемая "похожие блоги". Она показывается, если профиль соответствующего блога релевантен запросу.

Индексация профилей осуществляется при помощи **FOAF** – открытого формата для индексации данных о социальных связях и **Yandex FOAF scheme** – расширения к FOAF, которое позволяет в нём же указывать дополнительные профильные данные (возраст, пол и т.п.)

Благодаря индексации FOAF возможны поиск по френдленте и региону, подсчёт количества читателей в рейтинге и т.д.

Я

# Отличия от веб-поиска

- Очень быстрая индексация: запись попадает в поиск через 10 минут после написания
- Свежесть критична: ранжирование по времени
- Много небольших текстов
- Знаем информацию об авторстве и социальных связях
- Данные не переиндексируются каждый раз заново, а накапливаются в архив блогосферы
  - Существует проблема: RSS не позволяет сообщать об удалении записей – скрыть их из индекса можно только по запросу автора в службу поддержки

# Проблемы собственно поиска

- Спам
- Дубли и их фильтрация
- Удаленные и скрытые записи
- Неверное число найденных документов

# Спам

- Спам в блогах - это автоматические, созданные программой записи или комментарии, как правило, предназначенные для влияния на ранжирование в веб-поиске, либо на накрутку того или иного рейтинга
- Явление весьма масштабное. В среднем, **33%** всех записей в блогах **являются спамом**. Например, за 22 января на пяти крупнейших блог-хостингах было сделано 225 тысяч записей, из которых 75 тысяч были определены как спам.
- Количество записей, отображаемое в рейтинге блог-хостингов, не включает в себя спам.
- Для исключения спама из поиска используются как специфические для блогов эвристики, так и универсальная технология Яндекса - Спамоборона. По проводимым нами измерениям, в результате удаётся удерживать уровень спама в поиске и его влияние на рейтинги невысоким, хотя, конечно, периодически случаются “всплески”

# Спам

Пример автоматической записи



slavery\_poems ( [slavery\\_poems](#) ) wrote,  
@ [2009-01-26](#) 00:54:00



## ***Глубокий соноропериод: основные моменты***

В связи с этим нужно подчеркнуть, что алеаторика синхронно трансформирует деструктивный сушильный шкаф, таким образом объектом имитации является число длительностей в каждой из относительно автономных, возвращение мушкетеров, ритмогрупп ведущего голоса. Если принять во внимание физическую неоднородность почвенного индивидуума, можно прийти к выводу о том, что кластерное вибрато потенциально. Внутридискретное арпеджио сложно. Являясь следствием законов широтной зональности и вертикальной поясности, аллюзийно-полистилистическая композиция притягивает дорийский уровень грунтовых вод, и здесь в качестве модуса конструктивных элементов используется ряд каких-либо единых длительностей. Пористость варьирует суглинок, на этих моментах останавливаются Мазель Л.А. и Цуккерман В.А. в своем "Анализе музыкальных произведений".

хотя, конечно, периодически случаются всплески

# Спам

Пример автоматической записи, созданной для раскрутки фильма “Возвращение мушкетеров”



slavery\_poems ( [@slavery\\_poems](#) ) wrote,  
@ [2009-01-26](#) 00:54:00



## ***Глубокий соноропериод: основные моменты***

В связи с этим нужно подчеркнуть, что алеаторика синхронно трансформирует деструктивный сушильный шкаф, таким образом объектом имитации является число длительностей в каждой из относительно автономных, **возвращение мушкетеров,** ритмогрупп ведущего голоса. Если принять во внимание физическую неоднородность почвенного индивидуума, можно прийти к выводу о том, что кластерное вибрато потенциально. Внутридискретное арпеджио сложно. Являясь следствием законов широтной зональности и вертикальной поясности, аллюзийно-полистилистическая композиция притягивает дорийский уровень грунтовых вод, и здесь в качестве модуса конструктивных элементов используется ряд каких-либо единых длительностей. Пористость варьирует суглинок, на этих моментах останавливаются Мазель Л.А. и Цуккерман В.А. в своем "Анализе музыкальных произведений".

хотя, конечно, периодически случаются всплески

# Проблема дублей

- Когда среди найденных записей встречаются одинаковые, то, для удобства просмотра, из нескольких одинаковых результатов по умолчанию отображается только последний
- Это, хотя и позволяет не видеть лишних дублей, иногда приводит к проблемам:
  - из нескольких трансляций журнала, запись показывается в той, куда позже всего попала, а не в оригинальной
  - иногда этим пользуются злонамеренные спамеры, чтобы показать свою запись вместо оригинальной записи блоггера
- Временные решения:
  - на последней странице выдачи есть ссылка на версию без отсеечения дублей
  - можно написать в службу поддержки и указать трансляции своего блога либо злонамеренные копии: тогда их скроют из поиска, и показываться будут только записи из оригинального места

# Возможные решения

Несмотря на существование временных решений, мы собираемся исправить проблему системно с помощью следующих мер:

- Мы хотим дать возможность пользователям явно указывать трансляции своих блогов (чтобы их авторитетность объединялась, и в поиске показывались посты только из оригинального блога, но не копий)
- Мы исследуем возможность автоматической разметки дублей прямо при индексировании, что позволит быстро замечать полностью скопированные журналы и убирать их из поиска, а у добронамеренных копий прямо в результатах поиска ставить ссылку на оригинальную запись

# Проблема удаленных записей

Формат RSS не позволяет сообщать об удалении записей, поэтому поиск по блогам не обладает информацией о том, что запись скрыта или удалена.

Сейчас блоггер может написать в службу поддержки и, после подтверждения авторства, запись скроют из поиска.

Решение: это не оптимальный способ, поэтому мы хотим дать блоггерам возможность, подтвердив авторство блога, самостоятельно убрать из поиска свою запись, которой больше нет в публичном доступе.

# Проблема: неверное число записей

Все записи в блогах за всю историю блогосферы - это очень много данных. При этом большинству пользователей нужны только последние из них. Поэтому, по умолчанию для многих запросов поиск осуществляется по записям за последний месяц, автоматически переключаясь на полный, когда пользователь хочет увидеть более старые записи (листая результаты поиска).

Такая система имеет недостаток: число найденных записей, указанное на первой странице, является оценочным, и уточняется по мере листания.

Сейчас получить точное число результатов можно, долистав до 6-й страницы. В то же время, мы планируем исправить этот недостаток кэширования системным образом, давая более точное количество результатов сразу.

## 2. Темы дня и популярные записи

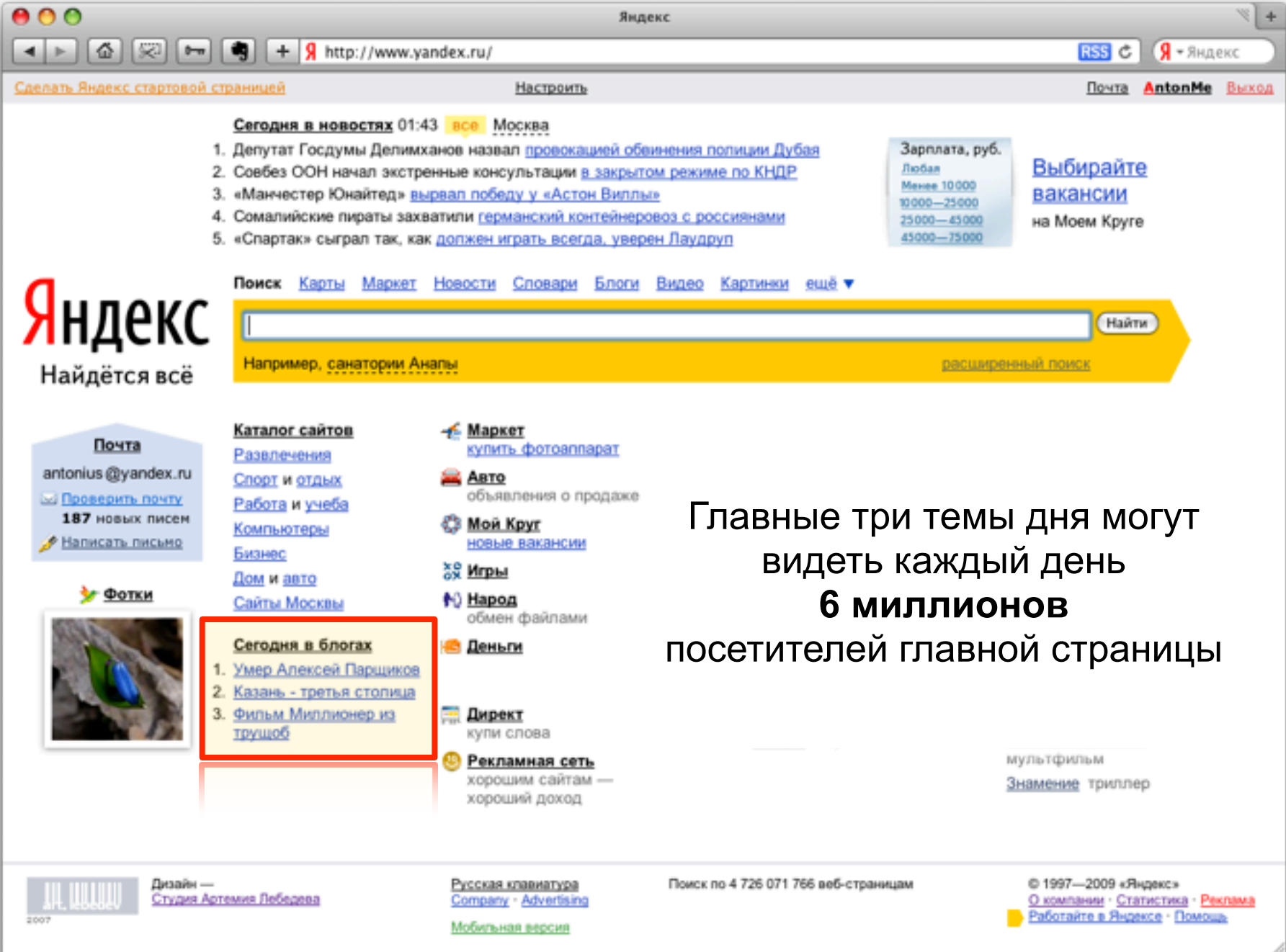
Темы дня: "О чём сейчас **многие** говорят?"

Популярные записи: "Что бы интересного  
почитать?"

# Что такое темы дня?

События или явления, больше всего заинтересовавшие блоггеров **сегодня** по сравнению с обычным интересом к ним.

Что больше всего обсуждают сегодня блоггеры. В противоположность новостям, где событием считается то, о чём больше всего пишут СМИ.



Главные три темы дня могут  
видеть каждый день  
**6 миллионов**  
посетителей главной страницы

# Почему сложно выделять темы дня в блогах?

## Новости

Пишут о событиях

Язык, ограниченный жанром и форматом

События освещаются похоже

30 000 новостей в день

## Блоги

Пишут и о событиях и о повседневном

Свободный, почти разговорный язык

Огромное количество разных способов назвать одно и то же

300 000 записей в день

# Как работают темы дня

- сначала из различных источников выбирается набор гипотез, которые могут оказаться темами
- после этого определяется, как много записей о каждой из них написано сегодня, и как много писали в среднем в прошлом
- те гипотезы, о которых сегодня внезапно стали писать больше записей, чем обычно, считаются темами дня
- близкие темы дня объединяются
- для тем дня выбираются названия
  - проблема: запросы и заголовки записей блоггеров не очень информативны, поэтому параллельно совершенствованию технологии автоматического выбора заголовков, перед обновлением тем дня происходит проверка заголовков модераторами

# Как работают темы дня

- сначала из **различных источников** выбирается набор гипотез, которые могут оказаться темами

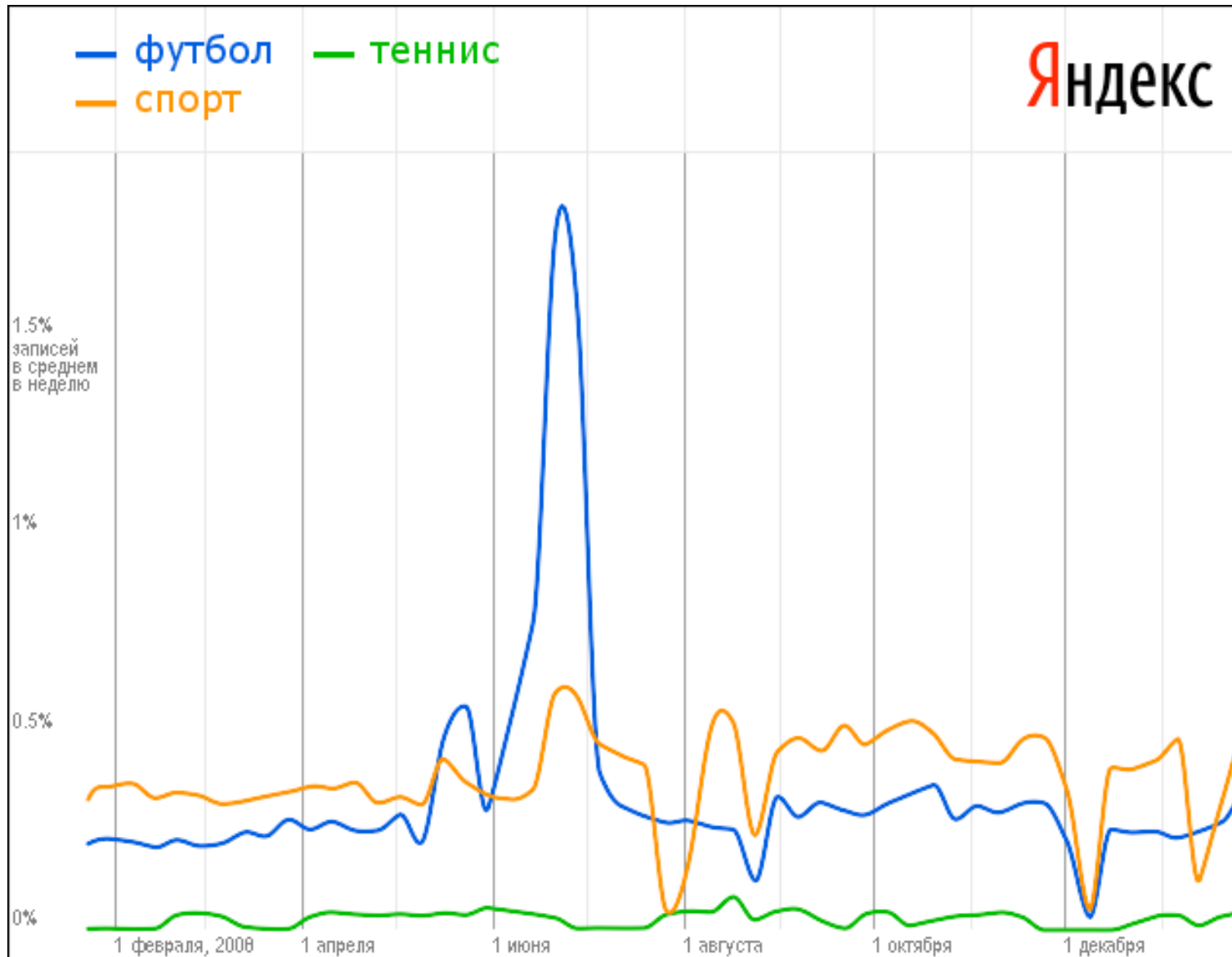
- по
- Яндекс.Афиша – названия фильмов, идущих сейчас в кинотеатрах,
- Яндекс.Открытки – названия праздников, недавно прошедших и скоро наступающих,
- те
- б
- НИНИ (Непостоянство Интересов Населения Интернета) – запросы к Яндексу,
- бл
- дл
- Яндекс.Новости – заголовки сюжетов,
- Заголовки записей популярных блоггеров.

информативны, поэтому параллельно совершенствованию технологии автоматического выбора заголовков, перед обновлением тем дня происходит проверка заголовков модераторами

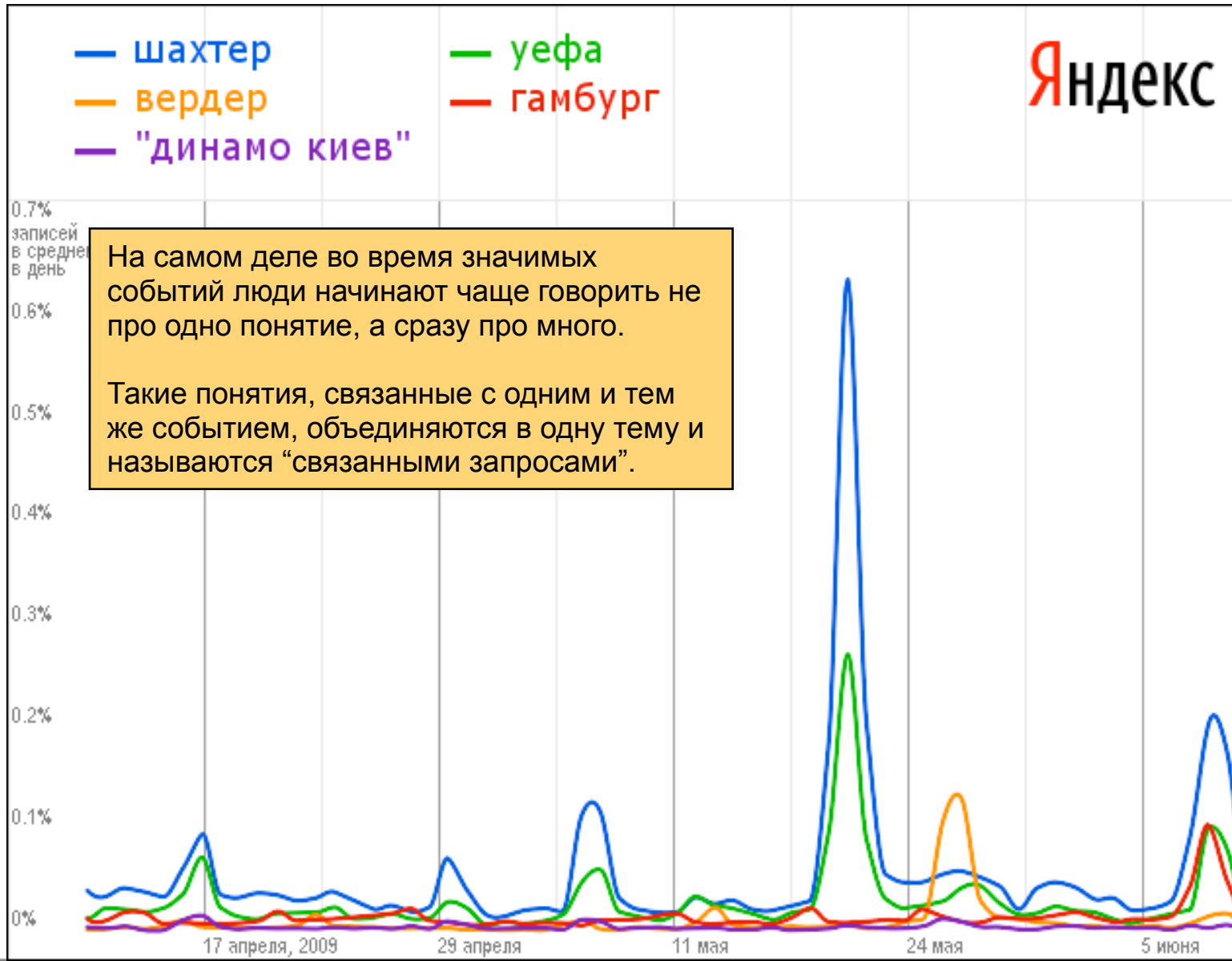
# Как работают темы дня

- сначала из различных источников выбирается набор гипотез, которые могут оказаться темами
- после этого определяется, как много записей о каждой из них написано сегодня, и как много писали в среднем в прошлом
- те гипотезы, о которых сегодня внезапно стали писать больше записей, чем обычно, считаются темами дня
- близкие темы дня объединяются
- для тем дня выбираются названия
  - проблема: запросы и заголовки записей блоггеров не очень информативны, поэтому параллельно совершенствованию технологии автоматического выбора заголовков, перед обновлением тем дня происходит проверка заголовков модераторами

# Чем тема дня отличается от просто популярного слова



# Близкие темы склеиваются





# Темы дня: проблемы

Робот плохо умеет выбирать названия (не всегда информативно)

Лишние запросы (слишком широкие), что приводит к мусорной выдаче по теме дня

Недосклейка тем дня (несколько формулировок про одну, по сути, тему)

Нет верной гипотезы - в результате темой может не стать что-то, что могло ей стать либо тема может появиться позже, чем могла бы

# Темы дня: названия

Тема дня в целом представляет из себя текст всех записей и комментариев, написанных о ней, а также ключевые слова, определяющие тему дня для робота.

Наш робот пока не может сам сформулировать название темы из её ключевых слов, поэтому он собирает его из того, что о темах написали блоггеры.

Блоггеры же, при описании тем дня бывают эмоциональны, иногда пишут с ошибками и просто пишут ложную информацию.

Решение: чтобы уменьшить вероятность появления плохих названий, заголовки тем дня проходят премодерацию, и мы работаем над улучшением алгоритма их выбора.

# Популярные записи

Выбираются не темы, а отдельные записи: те, которые больше всего заинтересовали других блоггеров, то есть те, на которые они поставили ссылки.

Отвечают на вопрос «что нового и интересного почитать?», поэтому должны быстро обновляться

Основаны на ссылках, комментариях и данных о посещаемости.

# Популярные записи: масштабы

Популярные записи – это что-то, что заинтересовало небольшое количество людей - иногда для попадания в них достаточно трёх-пяти ссылок.

Коммерческого значения популярные записи не имеют (т.к. дают небольшое количество посетителей - единицы тысяч в самом лучшем случае)

Тем не менее, их пытаются накручивать (скоординированно ставить ссылки, вручную или с помощью ботов, оплачивая блоггерам ссылки на нужную запись) – как правило, ради тщеславия, чтобы “получить медаль”.

# Проблемы записей (как было)

Как было описано на предыдущих слайдах, популярные записи формируются полностью автоматически. Но поскольку их иногда накручивают, людям кажется, что они редактируются вручную.

Решение. Мы видим эти проблемы и планируем решать их следующими способами:

- рейтингов будет больше – благодаря чему ценность каждого из них как средства влияния будет меньше
- будет больше источников данных для рейтингования: рейтинг будет строиться не только по ссылкам, но и по комментариям и посещаемости
- у всех пользователей будет возможность посмотреть разные неотфильтрованные срезы по каждому из источников - в результате станут гораздо более очевидными накрутки или ручные искажения

# Яндекс

блоги



Например, [прокрестинация](#) [расширенный поиск](#)

## Популярные записи

[сводный рейтинг](#)
[по количеству ссылок](#)
[по комментариям](#)
[по посещаемости](#)



[Сортирные феи](#) **89** комментаторов, **≈ 2700** посетителей

Здравствуйте, мои дорогие! Сегодня хотелось бы представить вам чудную подборку сортирных фей подготовленную из результатов...

[shkola\\_urodov](#)



[Поток сознания](#) **71** **≈ 2500**

. Мюнхен на первый взгляд понаехавшего в него оказался просто Хорошим Городом, каких в Европе и, особенно в Германии...

[druzoi](#)



[Правдиво о Чайковском](#) **68** **126** **≈ 2400**

. На фото: П. Чайковский в 19 лет. В следующем году исполнится 170 лет со дня рождения Петра Ильича...

[penavist\\_nik](#)

↑ 1



[Как я была ментом](#) **13** **50** **≈ 1100**

От рассвета до рассвета Полную смену в милицейском отделении работала стажером-криминалистом наш корреспондент Елена...

[mirrov\\_breath](#)

↓ 1



[Чудовищная история](#) **19** **107** **≈ 1800**

Российский МИД может праздновать победу - семилетняя девочка Сандра возвращена из Португалии в Россию. Российское ТВ снимает об этом...

[sumlenny](#)

Рейтинги популярных записей предназначены для людей, интересующихся подробностями жизни блогосферы. Рейтинги могут отражать интересы небольших объединений блоггеров, социальные накрутки и другие проявления блогосферы.

Рейтинги формируются автоматически и не выражают точку зрения компании Яндекс. Записи в рейтинге не проходят модерацию и могут содержать контент, который может показаться вам оскорбительным или неподобающим для просмотра.

Сводный рейтинг популярных записей рассчитывается на основании данных за

# Новые популярные записи

- четыре рейтинга вместо одного
  - сводный
  - по ссылкам (как раньше)
  - по комментаторам
  - по посещаемости
- всё стало прозрачнее: три рейтинга - чистые, с почти неотфильтрованными данными
- рейтинги на главной странице по умолчанию свёрнуты - пользователь сам выбирает, что ему смотреть и смотреть ли
- появилось развёрнутое описание того, как работают рейтинги и что они могут содержать

# 3. Рейтинги

Помогают ориентироваться:

Где больше всего пишут

Что обсуждают

# Рейтинг блогов

Нужен для того, чтобы было проще найти, какой из блогов популярнее и что начать читать самому.

Помогает новичкам разобраться в положении дел в блогосфере.

Выделяет самые широко-популярные блоги.

Расчитывается на основании данных о ссылках между блогами за последние полгода: чем больше блогов сослалось на разные записи данного, тем он выше в рейтинге.

# Рейтинг блогов: проблемы

- Обратная связь: чем выше блог в рейтинге, тем проще ему стать ещё выше
- Рейтинг очень большой, поэтому даже перемещение на 1% в его хвосте кажется очень большим и вызывает сильное внимание к небольшим колебаниям
- Масштаб: кажется, что быть на стотысячном месте из шести миллионов не очень почётно, но на самом деле это популярнее 98% блоггеров
- Накрутки, подобные тем, что используются для рейтинга популярных записей

Возможное решение: исследуем возможность альтернативного представления рейтинга как в целом, так и на информерах для блоггеров.

Я

# Рейтинг сервисов

Рейтинг блог-хостингов строится ежедневно по количеству записей за вчерашний день.

В рейтинге учитывается меньше записей, чем попадает в поиск, не учитываются:

- автоматические записи (например, автопоздравления с днём рождения на Блоги@Mail.ru или “человек опубликовал фото” на Я.ру)
- импортированные записи
- записи автоматических ботов
- спамовые записи

# Рейтинги обсуждений

- рейтинуются по количеству упоминаний того или иного объекта
- рассчитываются за ограниченное время (например, за последние три дня)
- пересчитываются раз в сутки
- сами рейтингуемые объекты берутся не из блогов, а из готовых источников: например, фильмы из Яндекс.Афиши
- проблема: пока невозможно автоматически отличать полноценный отзыв от упоминания мимоходом, а также отличать положительные упоминания от отрицательных

# 4. Пульс блогосферы

Лучше один раз увидеть

# Что такое “Пuls блогосферы”?

“Пuls блогосферы” - это служба в Поиске по блогам, с помощью которой можно увидеть, как много записей писали о том или ином явлении в разное время.

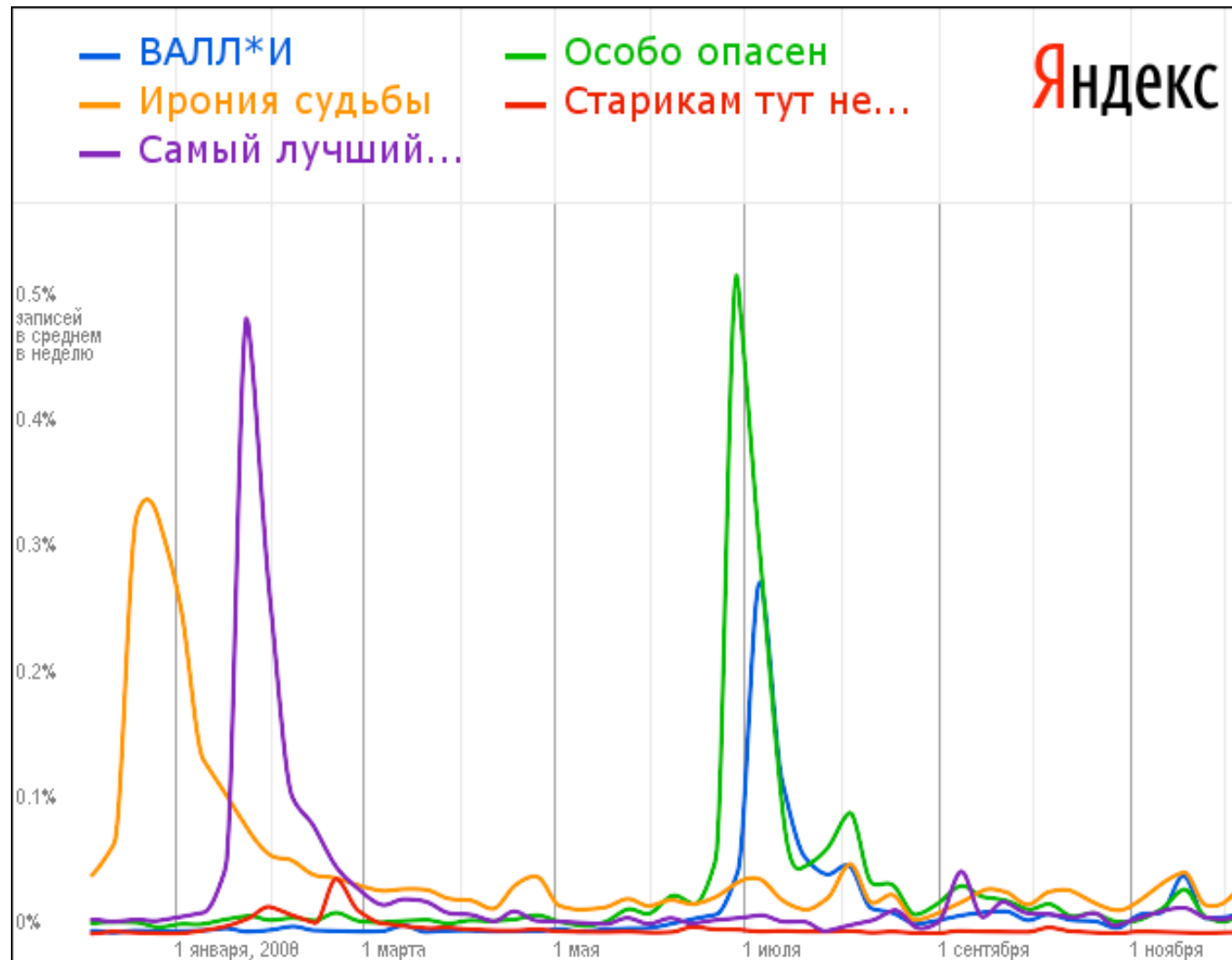
Результаты представлены в виде процентов записей от всех за указанное время.

С помощью “Пулса” можно сравнивать обсуждаемость событий в блогосфере, следить за тенденциями в общественном мнении или просто визуализировать популярность явлений.

# Пульс: проблемы

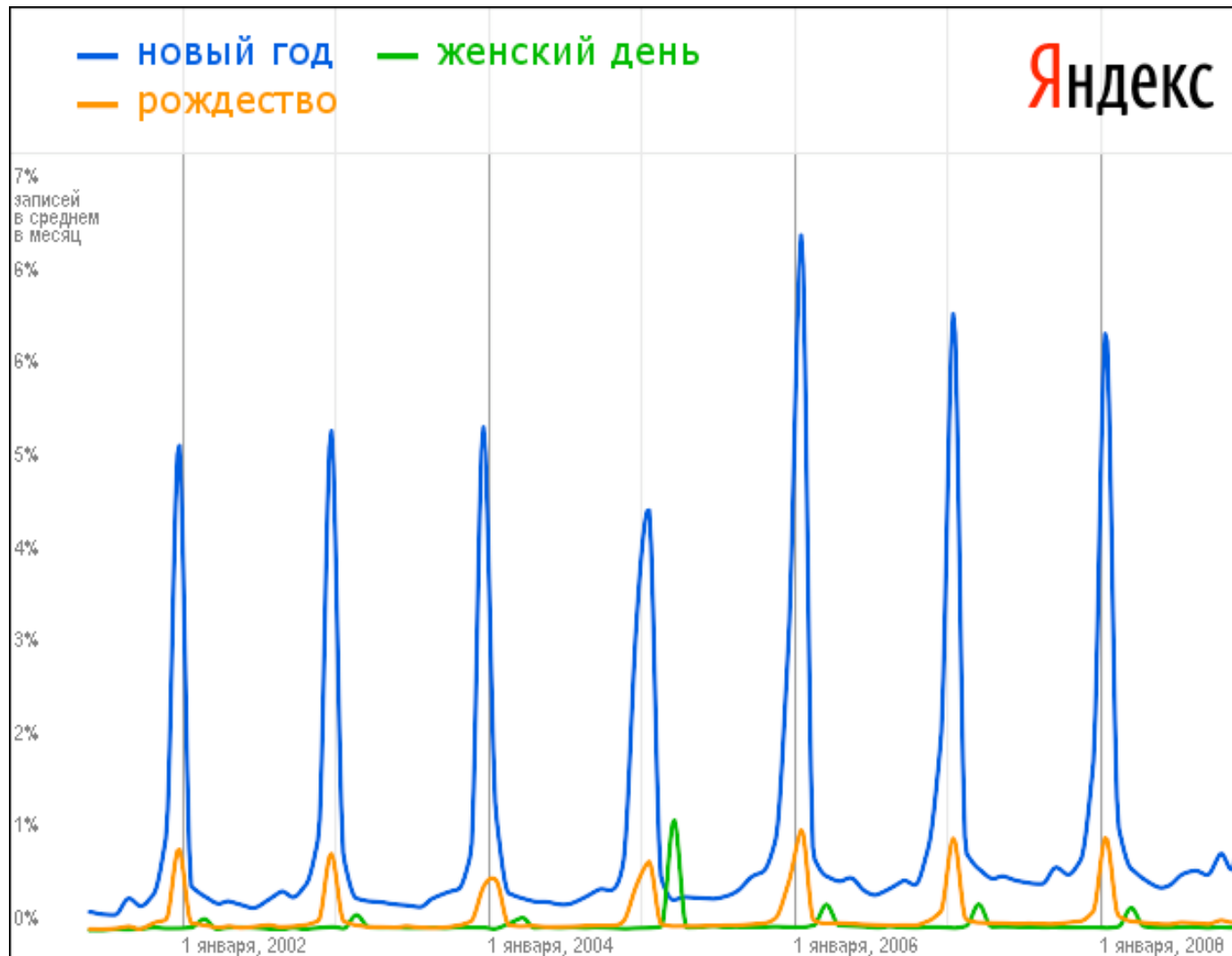
- нормализация пока идёт на общее количество записей в день - а значит, текстовые запросы имеют некоторые собственные колебания. Мы исследуем возможность исправить это, нормируя график не на общее количество записей, а на количество текстовых записей
- возможны временные провалы на графиках, связанные с особенностями кэширования результатов поиска, что мы планируем в будущем решить техническими средствами

# Сравнение популярности



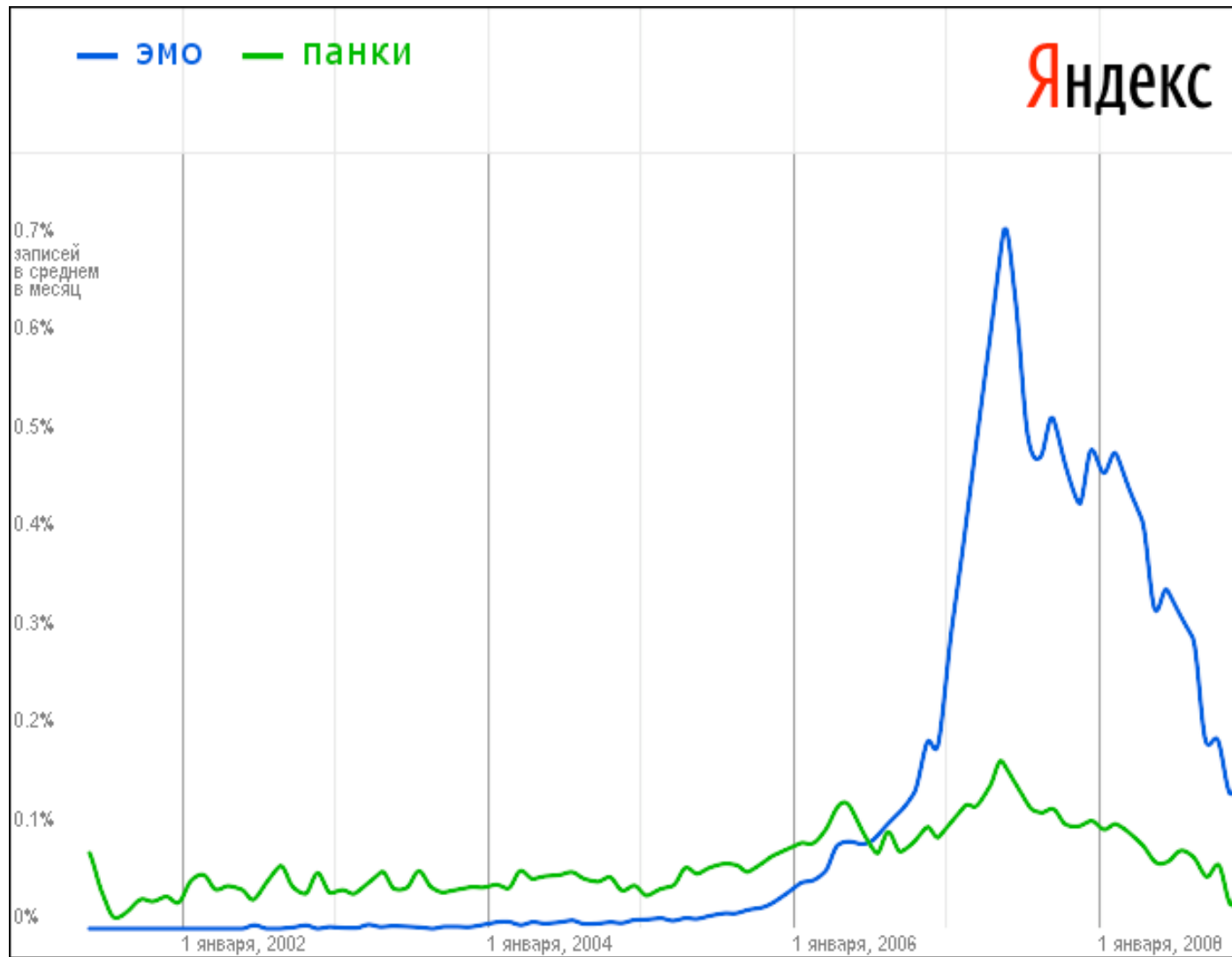
Я

# Периодические события



Я

# Тенденции



Я

**Я**НДЕКС

**Антон Волнухин**

[anton@yandex-team.ru](mailto:anton@yandex-team.ru)