Яндекс

Как устроен Поиск по блогам

Антон Волнухин
ВМиК МГУ, 3 декабря 2009

Что это такое?

Что такое Яндекс.Поиск по блогам и зачем он нужен?

На каких принципах основан?

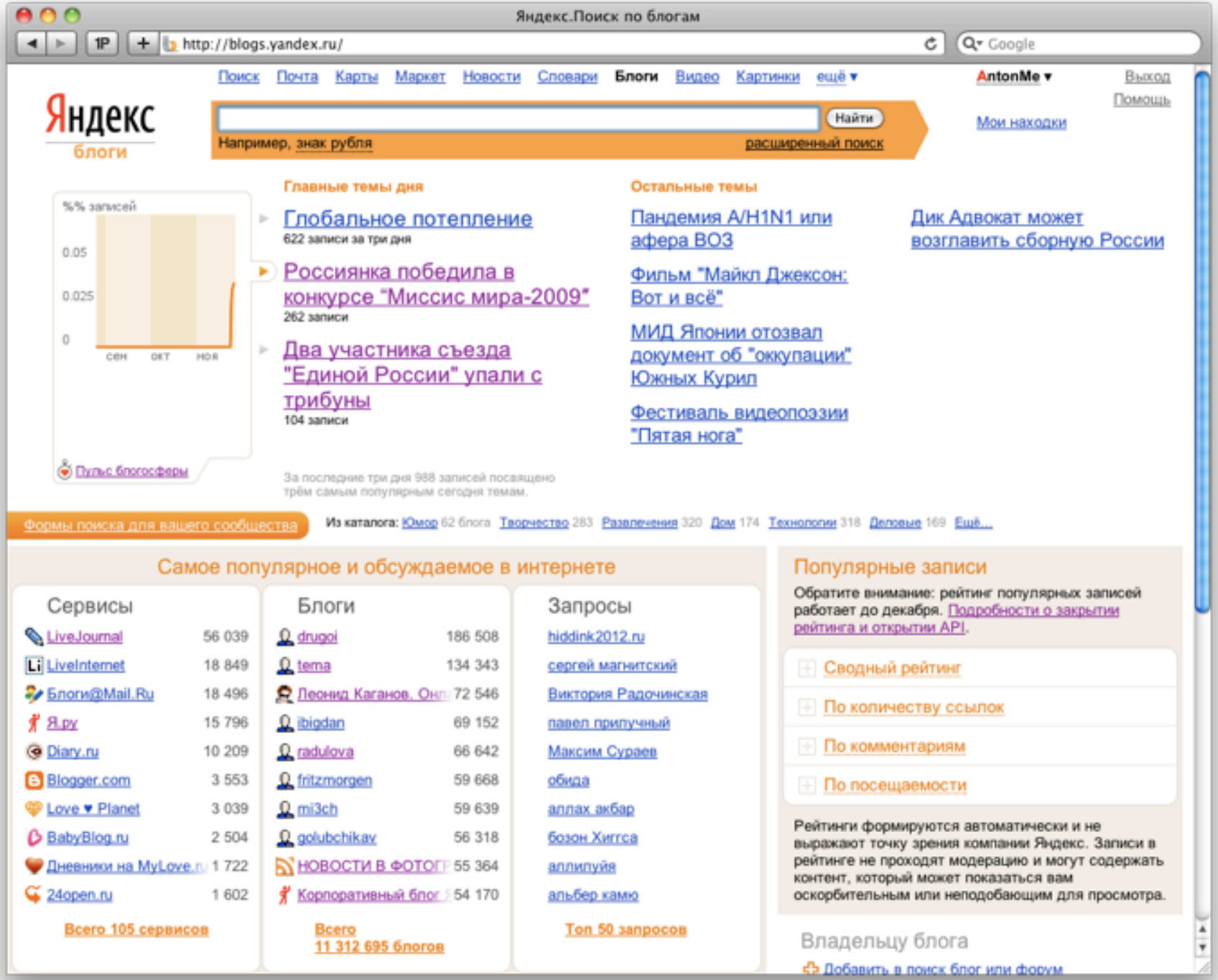
Что такое Поиск по блогам?

Поиск по мнениям. Общественное мнение в интернете

- Поиск по текстам, где люди говорят от первого лица, возможность сравнить обсуждаемость чего-либо:
  - 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

- Наиболее обсуждаемые темы и самое популярное в интернете сегодня

Я


Яндекс.Поиск по блогам
http://blogs.yandex.ru/
Google
Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё
AntonMe
Выход
Помощь
Яндекс
блоги
Найти
Например, знак рубля
расширенный поиск
Мои находки
%% записей
0.05
0.025
0
сен
окт
ноя
Пульс блогосферы
Главные темы дня
Глобальное потепление
622 записи за три дня
Россиянка победила в конкурсе "Миссис мира-2009"
262 записи
Два участника съезда "Единой России" упали с трибуны
104 записи
Остальные темы
Пандемия A/H1N1 или афера ВОЗ
Фильм "Майкл Джексон: Вот и всё"
МИД Японии отозвал документ об "оккупации" Южных Курил
Фестиваль видеопоэзии "Пятая нога"
Дик Адвокат может возглавить сборную России
За последние три дня 988 записей посвящено трём самым популярным сегодня темам.
Формы поиска для вашего сообщества
Из каталога: Юмор 62 блога Творчество 283 Развлечения 320 Дом 174 Технологии 318 Деловые 169 Ещё...
Самое популярное и обсуждаемое в интернете
Сервисы
LiveJournal 56 039
LiveInternet 18 849
Блоги@Mail.Ru 18 496
Я.ру 15 796
Diary.ru 10 209
Blogger.com 3 553
Love ♥ Planet 3 039
BabyBlog.ru 2 504
Дневники на MyLove.ru 1 722
24open.ru 1 602
Всего 105 сервисов
Блоги
drugoi 186 508
tema 134 343
Леонид Каганов. Онл 72 546
ibigdan 69 152
radulova 66 642
fritzmorgen 59 668
mi3ch 59 639
golubchikav 56 318
НОВОСТИ В ФОТОГ 55 364
Корпоративный блог 54 170
Всего
11 312 695 блогов
Запросы
hiddink2012.ru
сергей магнитский
Виктория Радочинская
павел прилучный
Максим Сураев
обида
аллах акбар
бозон Хиггса
аллилуйя
альбер камю
Топ 50 запросов
Популярные записи
Обратите внимание: рейтинг популярных записей работает до декабря. Подробности о закрытии рейтинга и открытии API.
Сводный рейтинг
По количеству ссылок
По комментариям
По посещаемости
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.
Владельцу блога
Добавить в поиск блог или форум

Почему социальный поиск важен?

свежесть

важный источник информации о важном для людей

возможность изучать живое общение

иногда людям нужен ответ от другого живого человека, оценочные суждения

структурированность

большой объём

Я

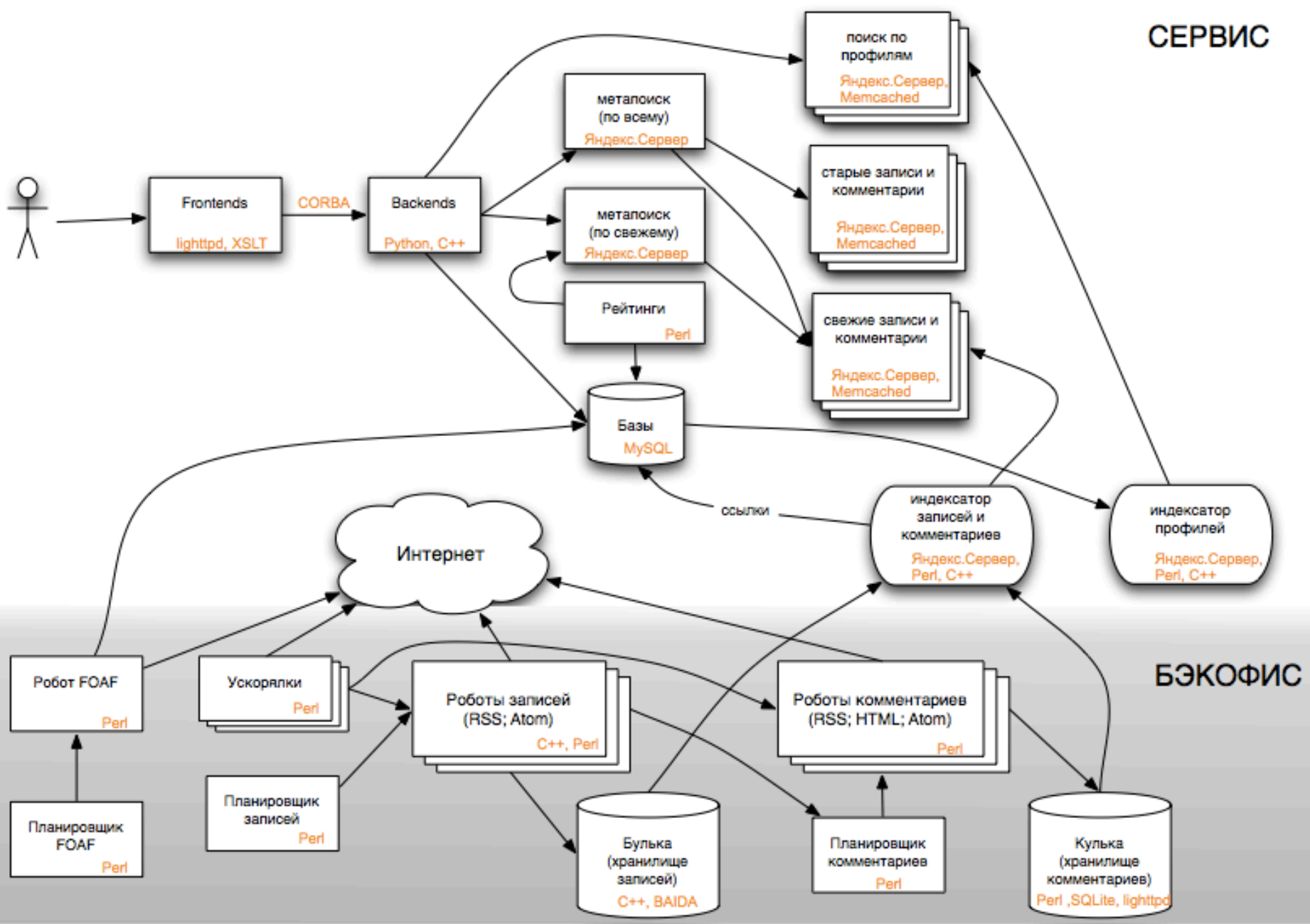
Масштабы

- Более **миллиона** записей и комментариев из блогов и форумов каждый день
- Почти **25 миллионов** источников
- Всего около **полутора миллиардов** документов

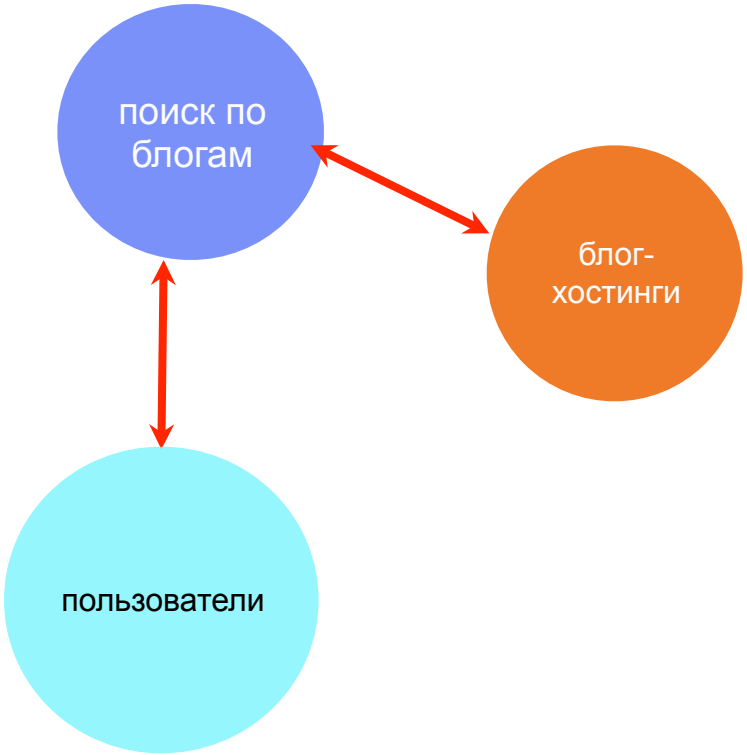
Поиск по блогам – это почти **одна пятая** от поиска по всему русскоязычному интернету по количеству элементов индексации

Я

Внутреннее устройство



Модель сервиса



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

Я

Модель сервиса

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

Я

Содержание

Что это и зачем

1. Поиск
2. Пульс блогосферы
3. Рейтинги
4. Темы дня
5. Открытые данные

(Почему мы закрываем популярные записи?)

Я

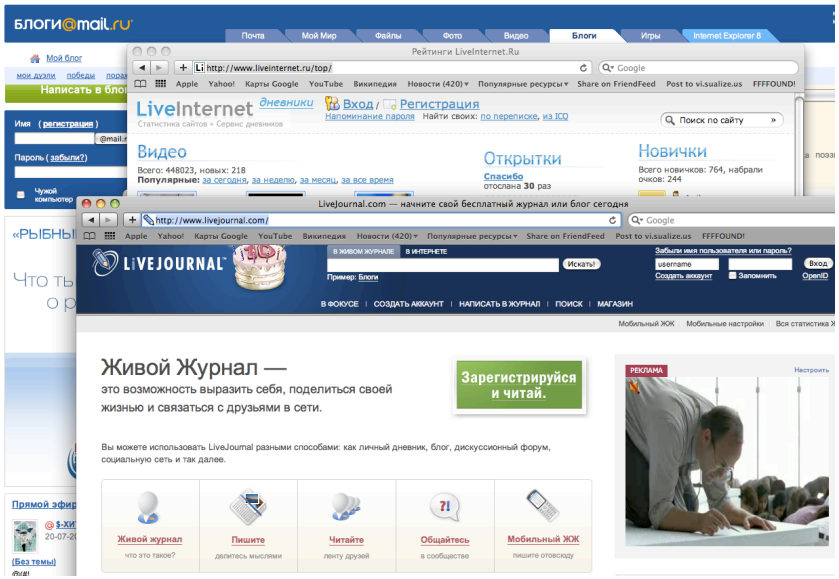
1. Поиск

На какие вопросы отвечает

Чем отличается от поиска по всему интернету

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.



- Livejournal.com
- Blogs.mail.ru
- Liveinternet.ru

> 190 блог-хостингов

Что индексирует поиск по блогам

- блоги (RSS и ATOM)
- форумы (RSS и ATOM)
- профили (FOAF + Yandex FOAF extension)
- комментарии (RSS и ATOM)

С помощью пингов можно ускорить индексацию,
сделав её почти мгновенной

Я

Как происходит индексирование

- На данный момент новые записи индексируются в течение 5 минут с момента их появления на более чем **170 блогхостингах**, включая:
  - LiveJournal.com
  - LiveInternet.ru
  - Blogs.mail.ru
  - Diary.ru
- Индексируются комментарии на LiveJournal.ru, LiveInternet.ru, Blogs.mail.ru и многих автономных блогах
- Проиндексировано более **50 миллионов** профилей, включая профили пользователей пяти крупнейших блог-хостингов

Как поиск узнаёт о новых блогах

- новые блоги на уже известных блогхостингах добавляются, как только поиск по блогам получает пинг про первую запись

- из веб-поиска: когда веб-поиск находит новый сайт с известным блоговым движком или ссылкой на RSS

- из формы добавления blogs.yandex.ru/add.xml

Я

Как индексируются профили

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 1-5 минут после написания
- Свежесть критична: ранжирование по времени
- Много небольших текстов
- Знаем информацию об авторстве и социальных связях и структуре блогов
- 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

Я

Спам

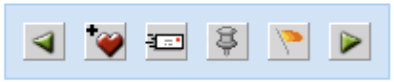
- Спам в блогах - это автоматические, созданные программой записи или комментарии, как правило, предназначенные для влияния на ранжирование в веб-поиске, либо на накрутку того или иного рейтинга
- Явление масштабное. В среднем, **11%** всех записей в блогах **являются спамом**. Например, во вторник, 24 ноября на пяти крупнейших блог-хостингах было сделано 280 тысяч записей, из которых 30 тысяч были определены как спам
- Количество записей, отображаемое в рейтинге блог-хостингов, не включает в себя спам
- Для исключения спама из поиска используются специфические для блогов эвристики и универсальная технология Яндекса - Спамооборона.

  В результате удаётся удерживать уровень спама в поиске и его влияние на рейтинги невысоким

Спам

Пример автоматической записи

slavery_poems (slavery_poems) wrote,
@ 2009-01-26 00:54:00



Глубокий сонопериод: 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".

мой

но

ая

ами

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 сонопериод: 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".

мой

но

ая

ами

хотя, конечно, периодически случаются “всплески

2. Пульс блогосферы

Лучше один раз увидеть

Что такое “Пульс блогосферы”?

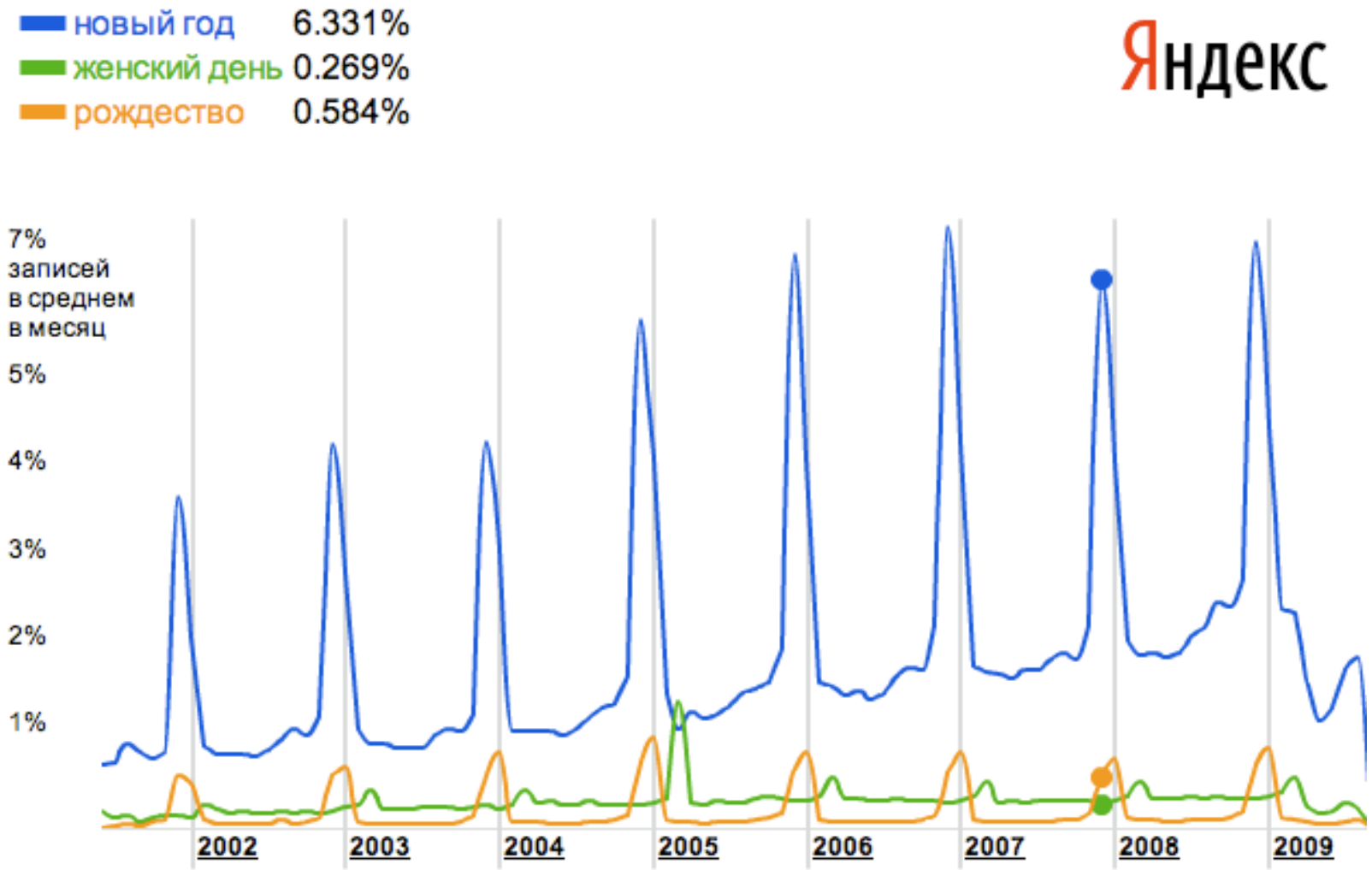
“Пульс блогосферы” - это служба в Поиске по блогам, с помощью которой можно увидеть, как много записей написали о том или ином явлении в разное время

Результаты представлены в виде процентов записей от всех за указанное время

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

Я

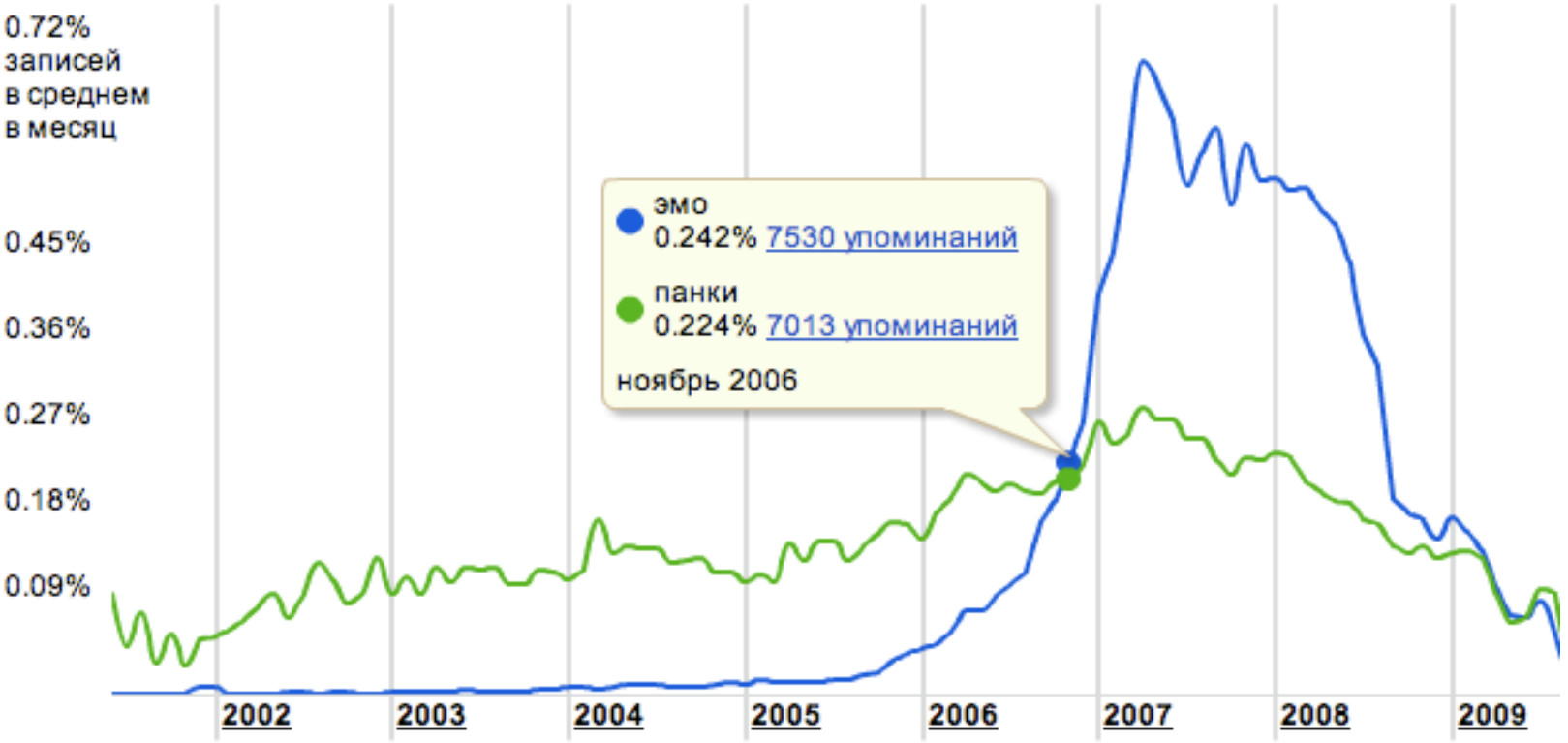
Периодичные события



Тенденции

эмо 0.242%
панки 0.224%

Яндекс



Я

3. Рейтинги

Помогают ориентироваться:

Где больше всего пишут

Что обсуждают

Рейтинг блогов

Помогает найти интересные блоги для чтения, узнать какой из блогов сейчас популярнее;

Даёт информацию новичкам о положении дел в блогосфере;

Выделяет самые широко-известные 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.

Я

Рейтинг сервисов

Рейтинг блогхостингов строится ежедневно по количеству записей на каждом сервисе 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

- рейтингуются по количеству упоминаний того или иного объекта;
- расчитываются 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. Темы дня

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

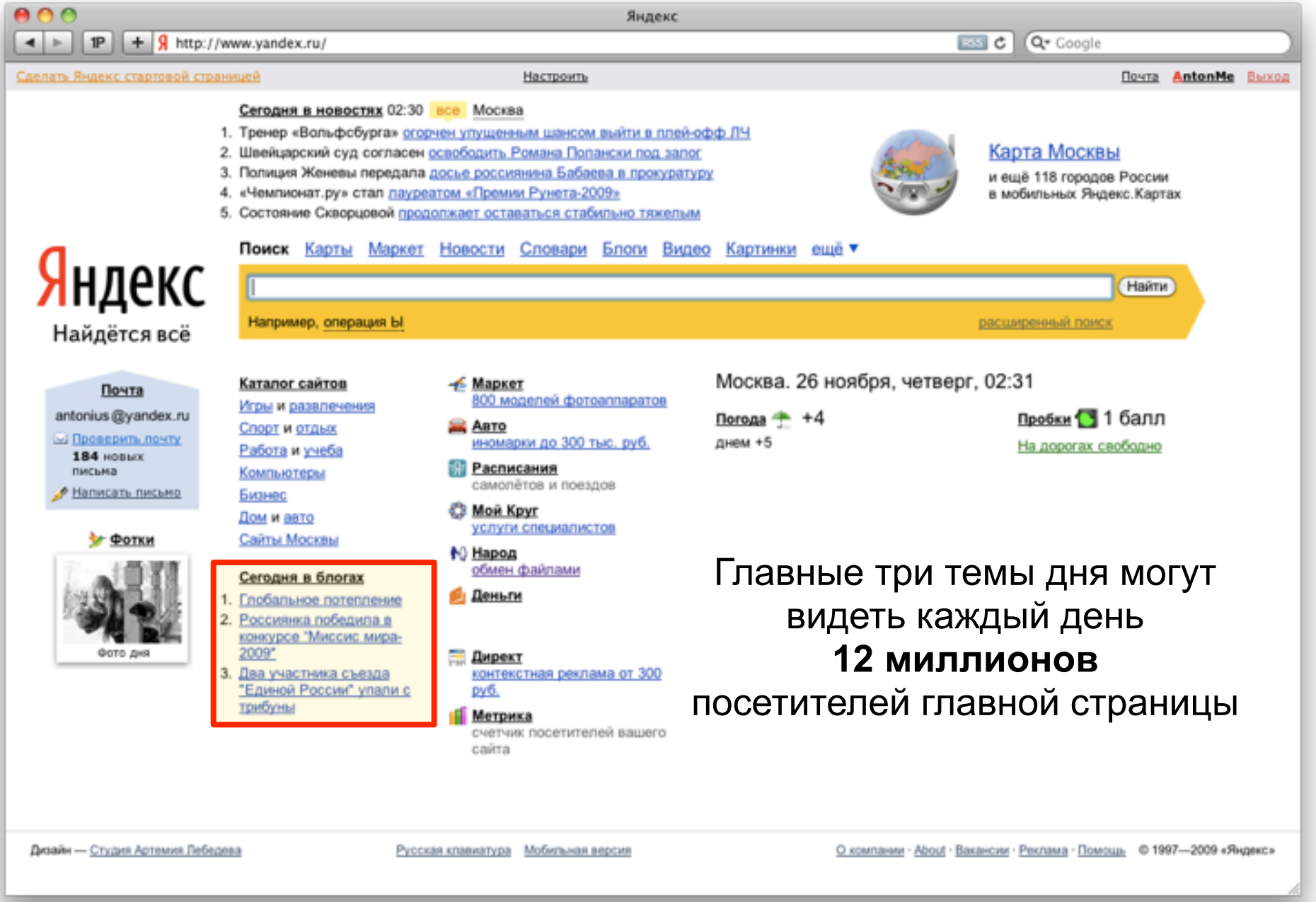
Что такое темы дня?

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.

Я


Яндекс
http://www.yandex.ru/
Google
Сделать Яндекс стартовой страницей
Настроить
Почта AntonMe Выход
Сегодня в новостях 02:30 все Москва
1. Тренер «Вольфсбурга» огорчен упущенным шансом выйти в плей-офф ЛЧ
2. Швейцарский суд согласен освободить Романа Полански под залог
3. Полиция Женевы передала досье россиянина Бабаева в прокуратуру
4. «Чемпионат.ру» стал лауреатом «Премии Рунета-2009»
5. Состояние Скворцовой продолжает оставаться стабильно тяжелым
Карта Москвы
и ещё 118 городов России
в мобильных Яндекс.Картах
Яндекс
Найдётся всё
Поиск Карты Маркет Новости Словари Блоги Видео Картинки ещё
Найти
Например, операция Ы
расширенный поиск
Почта
antonius@yandex.ru
Проверить почту
184 новых письма
Написать письмо
Фотки
Фото дня
Каталог сайтов
Игры и развлечения
Спорт и отдых
Работа и учеба
Компьютеры
Бизнес
Дом и авто
Сайты Москвы
Сегодня в блогах
1. Глобальное потепление
2. Россиянка победила в конкурсе "Миссис мира-2009"
3. Два участника съезда "Единой России" упали с трибуны
Маркет
800 моделей фотоаппаратов
Авто
иномарки до 300 тыс. руб.
Расписания
самолётов и поездов
Мой Круг
услуги специалистов
Народ
обмен файлами
Деньги
Директ
контекстная реклама от 300 руб.
Метрика
счетчик посетителей вашего сайта
Москва. 26 ноября, четверг, 02:31
Погода +4
днем +5
Пробки 1 балл
На дорогах свободно
Главные три темы дня могут
видеть каждый день
12 миллионов
посетителей главной страницы
Дизайн — Студия Артемия Лебедева
Русская клавиатура Мобильная версия
О компании · About · Вакансии · Реклама · Помощь © 1997—2009 «Яндекс»

Почему сложно выделять темы дня в блогах?

Новости	Блоги
Пишут о событиях	Пишут и о событиях и о повседневном
Язык, ограниченный жанром и форматом	Свободный, почти разговорный язык
События освещаются похоже	Огромное количество разных способов назвать одно и то же
30 000 новостей в день	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
них
про
• те
бол
• бл
• дл

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

из

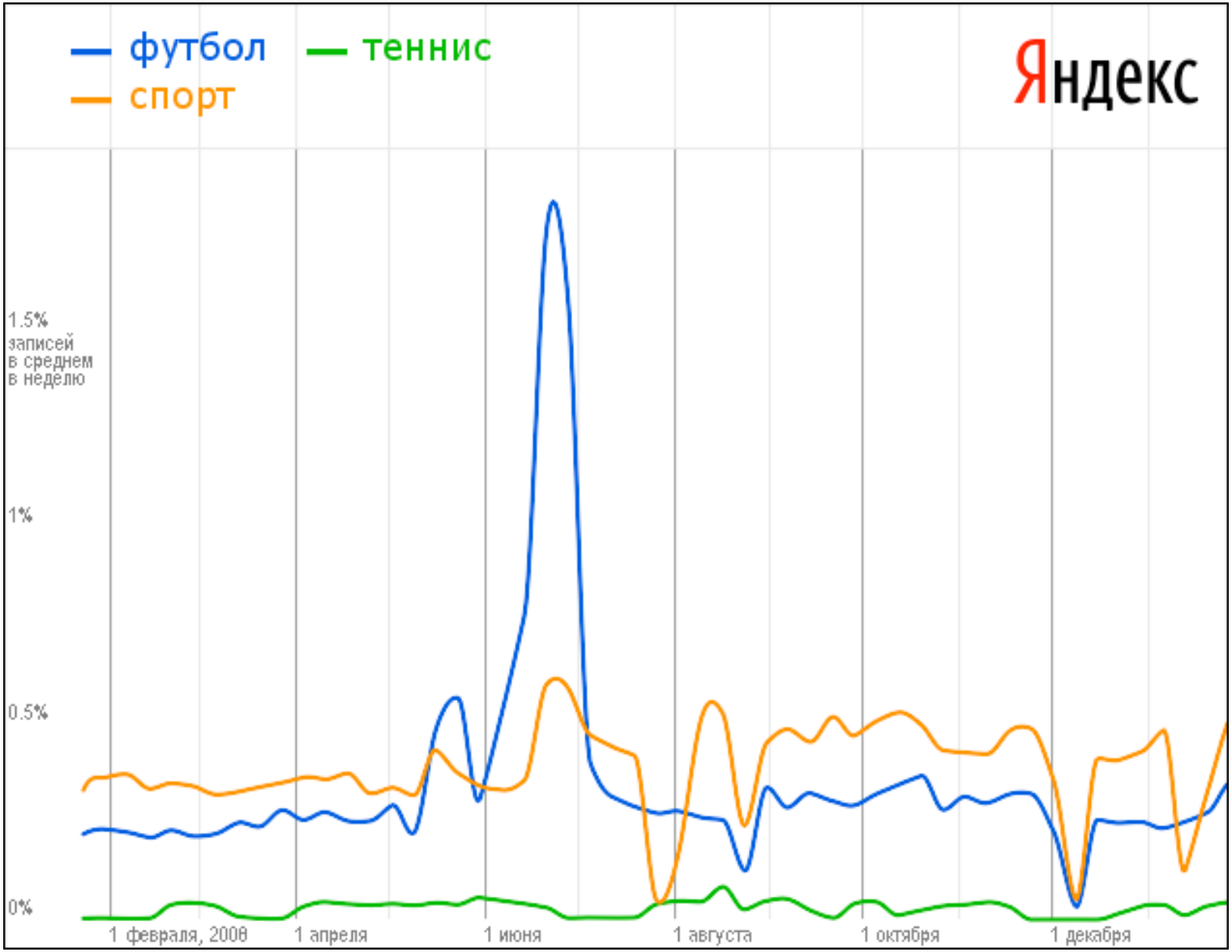
ень

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

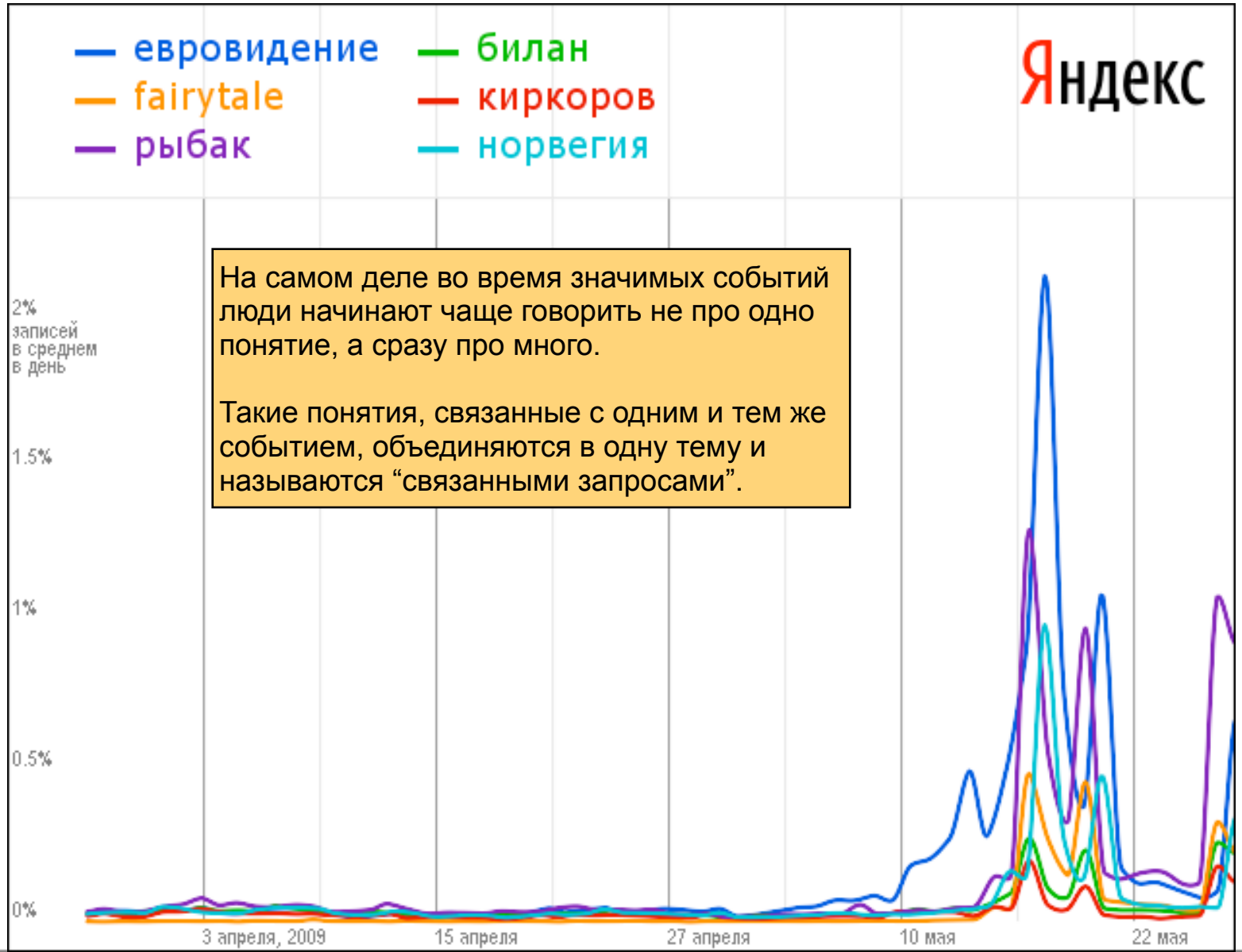
Как работают темы дня

- 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

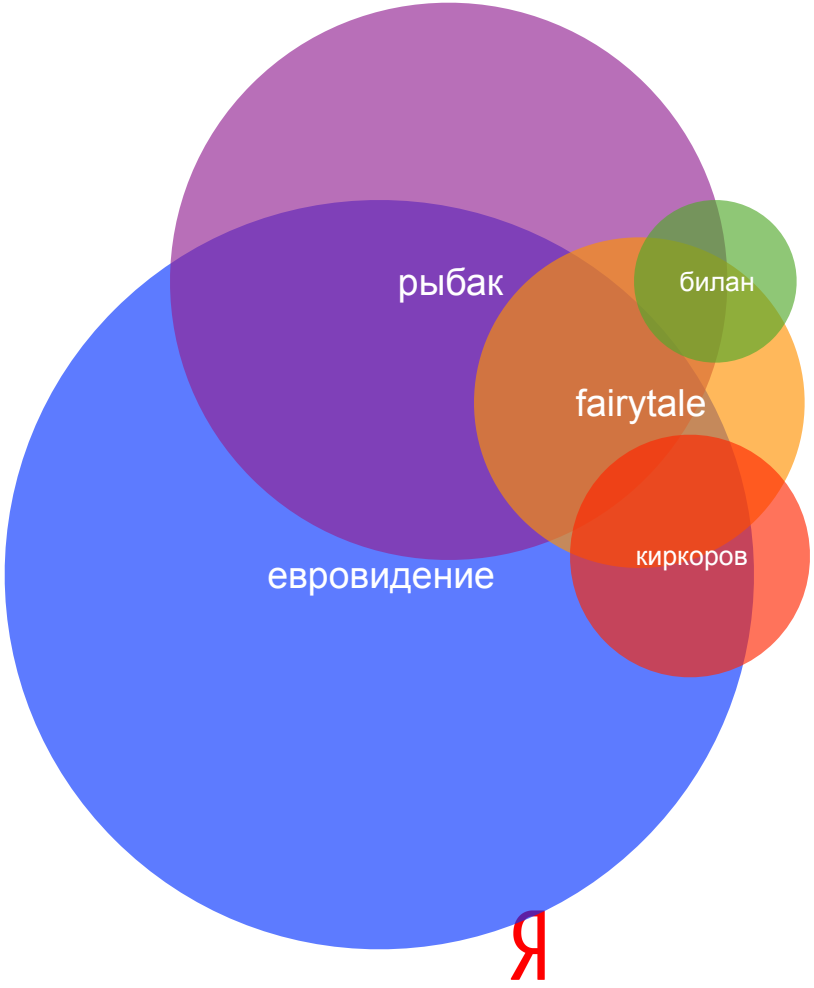


На самом деле во время значимых событий люди начинают чаще говорить не про одно понятие, а сразу про много.

Такие понятия, связанные с одним и тем же событием, объединяются в одну тему и называются “связанными запросами”.

Как именно склеиваются темы

- Темы дня склеиваются, когда о них часто пишут в одних и тех же записях



5. Открытые данные

API. Какие данные доступны?

- Все свежие проиндексированные записи
- Поиск по всем профилям (FOAF) и данные из них
- Социальные связи (кого “зафрендил” каждый пользователь, кто “зафрендил” его)
- То, что мы смогли понять в результате сбора всех данных в одном месте (агрегирование и анализ):
  - результат определения пола
  - поиск по записям с учётом данных из FOAF
  - фильтрация по блогам, форумам, комментариям и т.п.
- В будущем будут доступны, также, данные из всех рейтингов
- Мы готовы предоставлять для исследований и другие накопленные данные

Я

Вопросы?

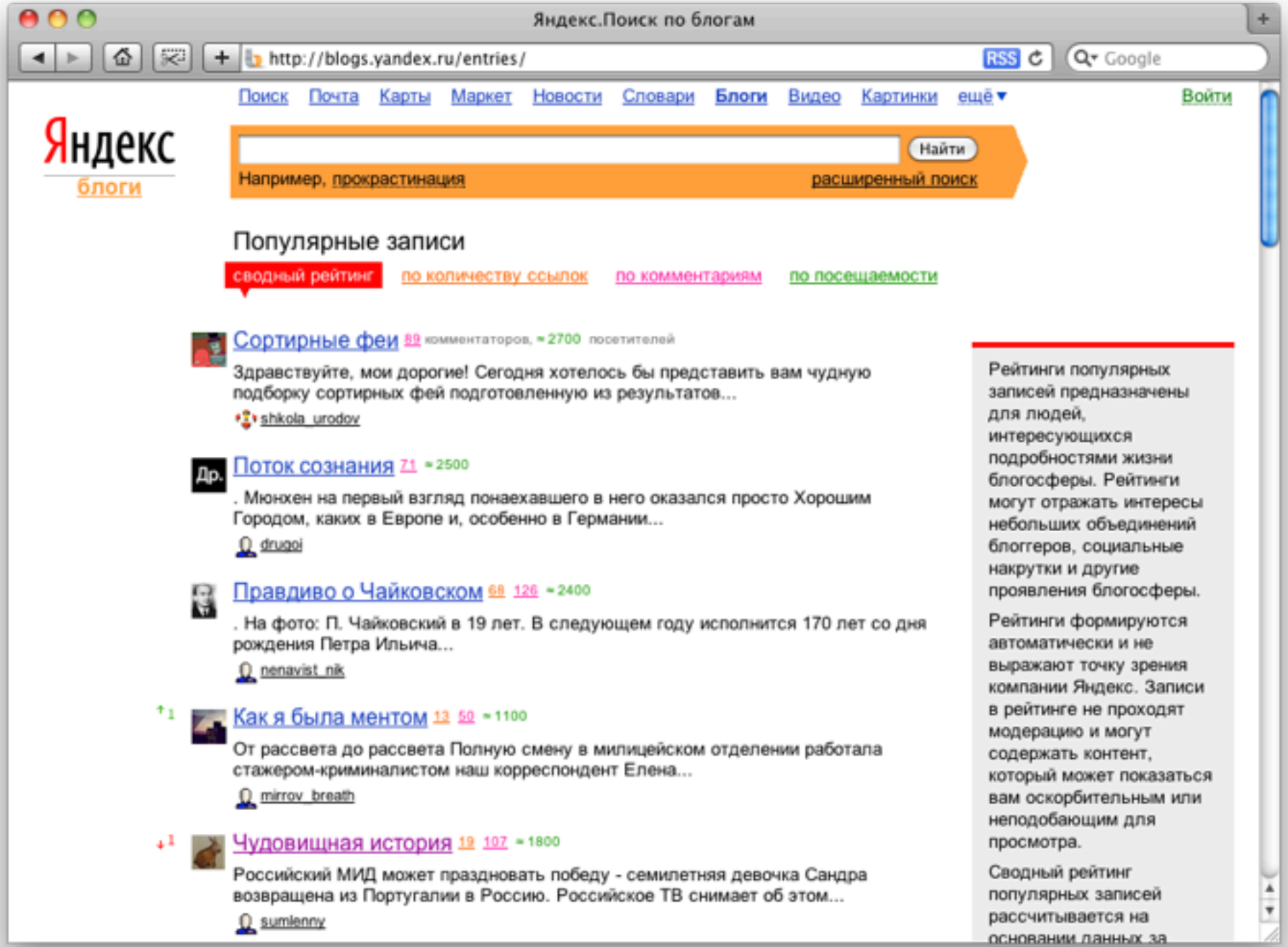
Яндекс

Антон Волнухин

anton@yandex-team.ru

Популярные записи

Что это такое, и почему мы открыли API вместо рейтинга?


Яндекс.Поиск по блогам
http://blogs.yandex.ru/entries/
RSS
Google
Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё
Войти
Яндекс
блоги
Найти
Например, прокрастинация
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...
drugoi
Правдиво о Чайковском 68 126 ≈ 2400
. На фото: П. Чайковский в 19 лет. В следующем году исполнится 170 лет со дня рождения Петра Ильича...
nenavist_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

Популярные записи

Выбираются не темы, а отдельные записи, которые больше всего заинтересовали других блоггеров (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

Я

Популярные записи: масштабы

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

Почему мы решили закрыть рейтинг?

- Рейтинг перестал отражать блогосферу, так как она стала активно влиять на рейтинг. Он стал инструментом, с помощью которого заинтересованные люди пытаются вынести какие-то сообщения в публичное поле. Иногда это что-то важное и нужное, иногда нет. Зачастую нельзя вообще объективно сказать, важное и нужное ли это

- Рейтинг стал специализированным инструментом медийного влияния, которое мало читают (7 тыс.), но много цитируют.

Про медийное влияние

- Яндекс как компания видит свою задачу в том, чтобы отвечать на вопросы пользователей, делать массовые информационые сервисы
- У нас нет позиции, мы не имеем содержательного мнения по поводу того или иного события, не выносим суждений и не занимаемся аналитикой
- Наша «редакционная политика» — это содержательная нейтральность и статистическая обработка информации
- Подробнее на http://company.yandex.ru/rules/media/

Популярные записи: масштабы

- Весь портал Яндекса — 18 млн. посетителей в день
- Главная страница Яндекса — 12 млн. посетителей в день
- Поиск по блогам в целом — 300 тыс. посетителей в день
- Из них на поиске — 275 тыс. посетителей в день
- На рейтинге популярных записей — 6-7 тыс. посетителей в день
- Даже на волне обсуждений о закрытии топа количество посетителей поднялось только до 9 тыс. посетителей в день.

Было ли давление снаружи?

- Решение закрыть рейтинг было принято внутри компании, по тем причинам, о которых мы рассказали
- Никогда никакие чиновники или силовые структуры не обращались к нам по поводу этого рейтинга
- «Давление» в некотором роде было со стороны блоггеров: у многих есть своё представление о том, какой должна быть редакционная политика этого СМИ. Нас обвиняли в том, что в топе есть что-то отвратительное или лживое. Нас обвиняли в том, что в топе нет чего-то важного

- Но рейтинг закрывается не из-за негативных отзывов. Мы не хотим, чтобы сервис для мониторинга стал сервисом для «вывода в топ»

Данные стали **более открытыми**

Это важно: нас часто обвиняют в закрытии рейтинга из-за цензуры, но в нашем API доступно **больше данных**, чем можно было получить через рейтинг

Можно самостоятельно анализировать любые данные о блоггере, записи, ссылках на неё, её посещаемости, комментариях (и кто прокомментировал)

Всё это теперь открытая и доступная любому информация

Я

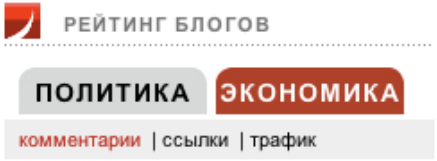
API

Уже появилось более 20 рейтингов на основе нашего API.

Благодаря тому, что в нём доступно больше данных про записи, чем когда-либо было, они могут быть разными

Появляются рейтинги с дополнительными возможностями

Появляются тематические рейтинги



Что будет дальше

Мы не планируем закрывать другие рейтинги, но хотим, чтобы все они были зеркалом блогосферы

Мы будем развивать темы дня и другие инструменты в Поиске по блогам для отслеживания состояния блогов

Мы будем развивать сам **Поиск по блогам** как поиск по общественному мнению, чтобы он искал полно, быстро и удобно

Мы будем развивать основной поиск Яндекса, чтобы в нём можно было быстро найти любую информацию, которая вам нужна: в том числе и в блогах и форумах

Я

Антон Волнухин

anton@yandex-team.ru