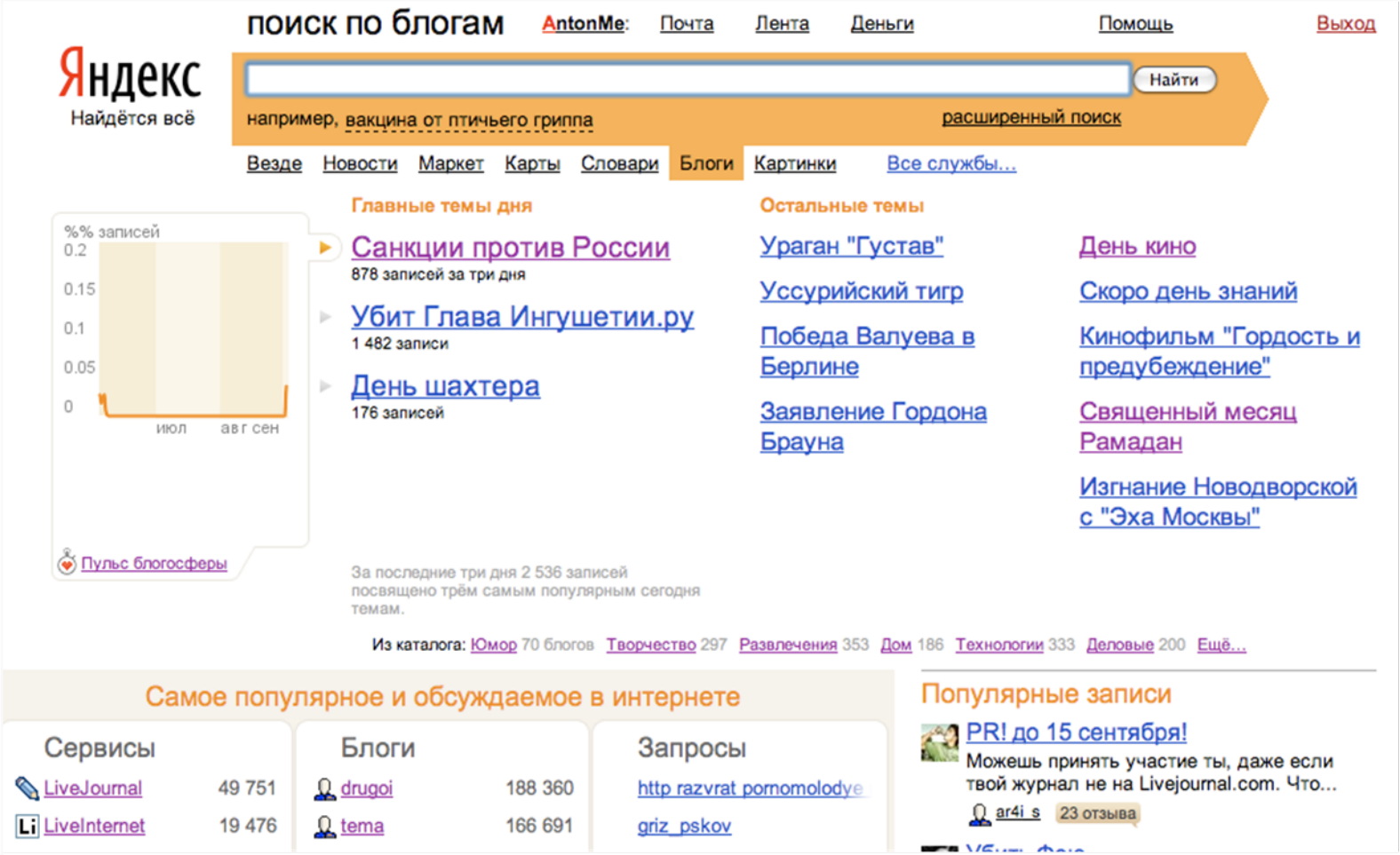
Темы дня в блогах:
Как это работает

Андрей Мищенко
Антон Волнухин,

3.09.2008

Поиск по блогам: поиск по мнениям



Яндекс

Что такое темы дня?

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

Т.е. это то, что больше всего обсуждают сегодня люди. В противоположность официальным новостям, где главными новостями считаются те, о которых больше пишут СМИ.

Яндекс

С чем мы имеем дело?

- Около **200 тысяч** записей блогах каждый день
- Около **400 тысяч** комментариев в день
- Более **380 миллионов** записей всего
- Более **400 миллионов** комментариев
- Более **5 миллионов** блогов

Для сравнения, в базе Яндекса по всему интернету, – около 4 миллиардов страниц.

Яндекс

С чем мы имеем дело?

Новости	Блоги
Пишут о событиях	Пишут и о событиях и о повседневном
Кодифицированный язык	Свободный, почти разговорный язык
События освещаются похоже	Огромное количество разных способов назвать одно и то же

Яндекс

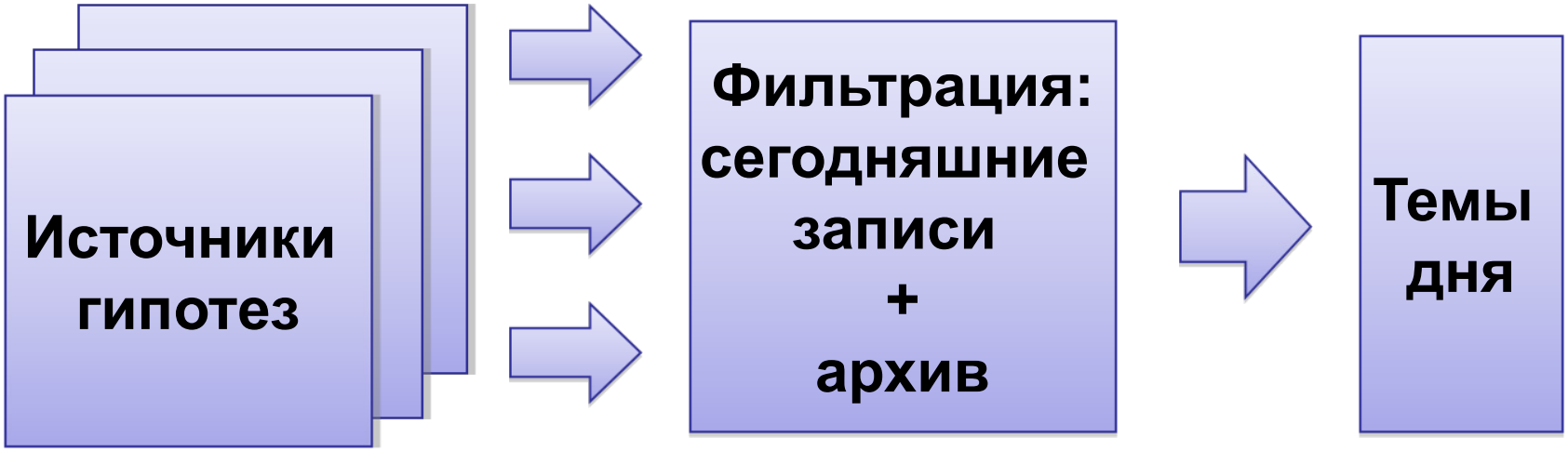
Кажется, что это работает так

Нет, это работает не так. Почему?

- Большие объемы данных, трудно обсчитать без привлечения какой-нибудь информации со стороны.
- Желательно, чтобы тема дня была не просто набором ключевых слов, а каким-то осмысленным словосочетанием.

Яндекс

Это работает так



- Гипотезы тем дня – это согласованные, осмысленные, законченные словосочетания (обычно)
- Источники гипотез являются внешними по отношению к системе определения тем дня.
- Записи в блогах работают не как источник тем, а как фильтр гипотез.

Яндекс

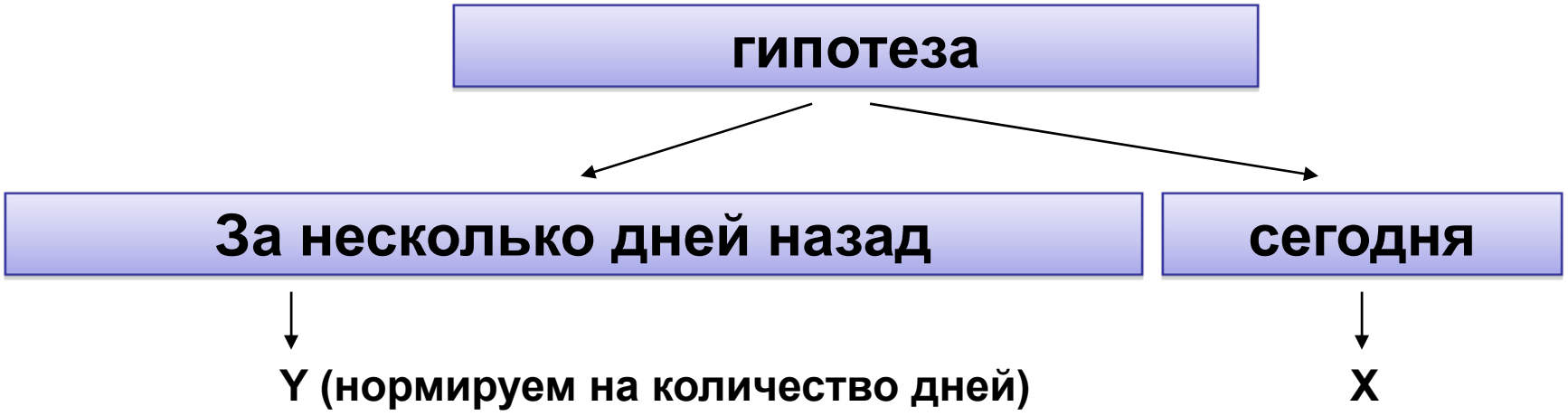
Источники гипотез тем дня

- 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.

Яндекс

Принцип фильтрации

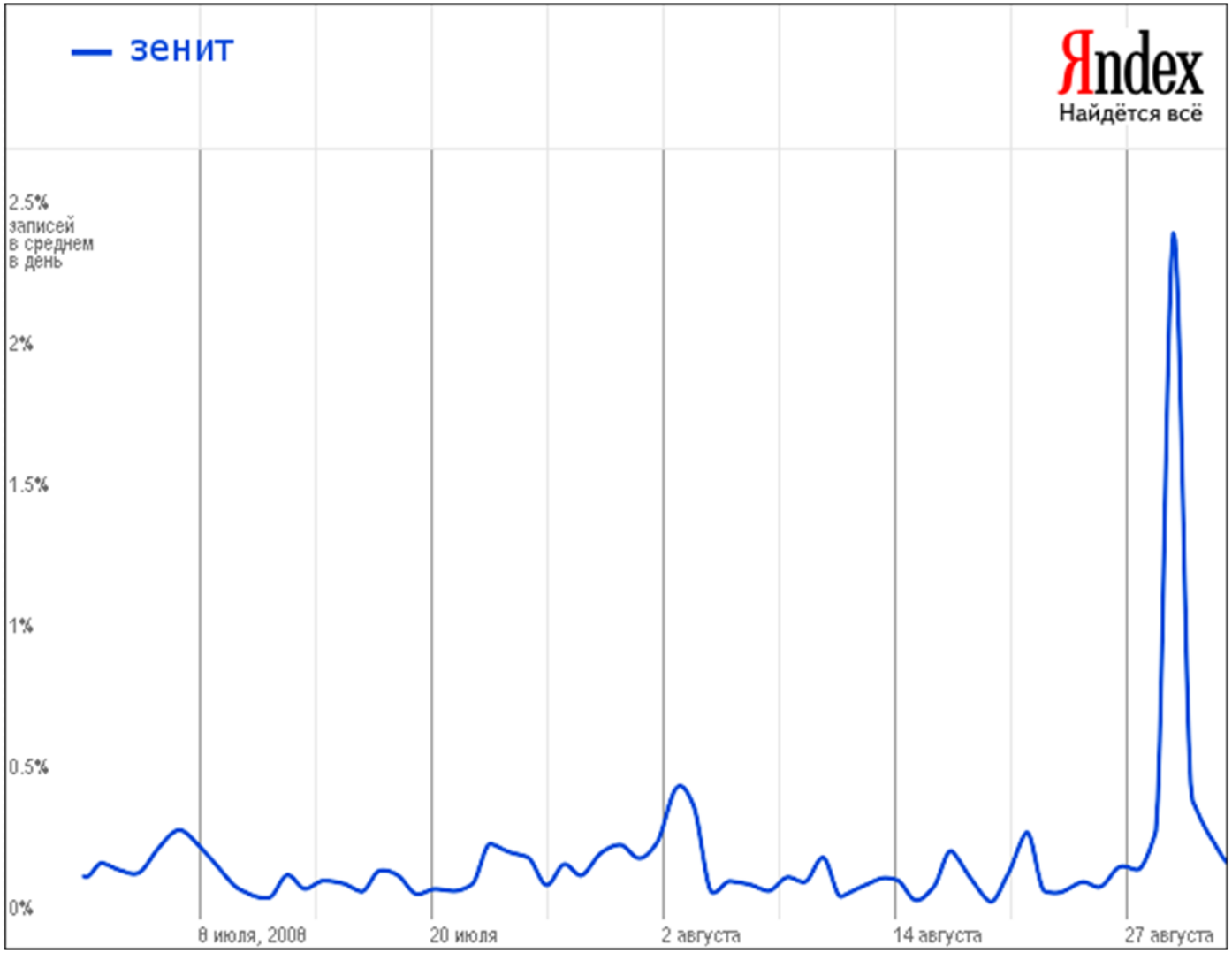
- Записи в блогах уже предобработаны, по ним построен поисковый индекс. Посмотрим на тему дня как на поисковый запрос.



- Вычисляем количество записей за временной интервал, подходящих под запрос-гипотезу.
- Сравнивая X и Y между собой, можно оценить, какая из гипотез больше подходит на роль темы.

Яндекс

Пример появления темы дня



Яндекс

Формула «темовитости»

- Вычитание? Плохо.
  Например, 100 → 200 и 10000 → 10500
- Деление? Тоже плохо.
  Например, 10 → 30 и 1000 → 2000
- Нужно подобрать «золотую середину».

$$\ln\left(\frac{x}{y}\right)(x-y)$$

Яндекс

Много гипотез одновременно

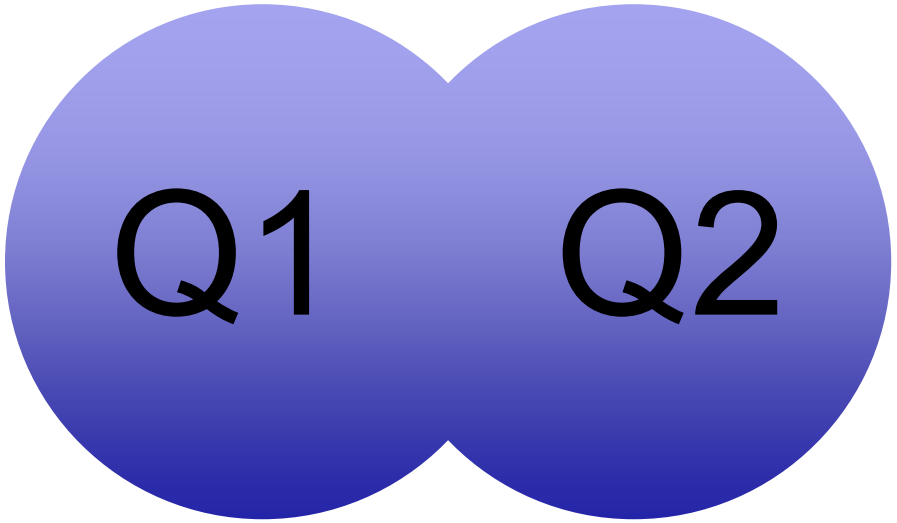
Яндекс

Много гипотез одновременно

Яндекс

Склеивание похожих тем

- Как установить связь между двумя гипотезами, не имеющими ничего общего в смысле текста? Снова с помощью поискового индекса.



- Если две гипотезы тем дня часто встречаются в одних и тех же записях, - это с большой вероятностью об одном и том же

Яндекс

ВОПРОСЫ?

http://blogs.yandex.ru/

Андрей Мищенко druxa@yandex-team.ru
Антон Волнухин anton@yandex-team.ru

Яндекс