



**Темы дня в блогах:
Как это работает**

Андрей Мищенко
Антон Волнухин,

3.09.2008

Поиск по блогам: поиск по мнениям

Яндекс
Найдётся всё

поиск по блогам

[AntonMe:](#)

[Почта](#)

[Лента](#)

[Деньги](#)

[Помощь](#)

[Выход](#)

например, вакцина от птичьего гриппа [расширенный поиск](#)

[Везде](#)

[Новости](#)

[Маркет](#)

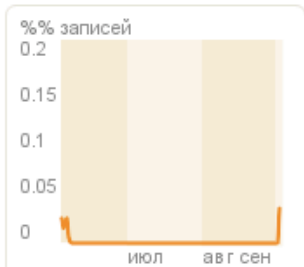
[Карты](#)

[Словари](#)

Блоги

[Картинки](#)

[Все службы...](#)



Пульс блогосферы

Главные темы дня

- ▶ [Санкции против России](#)
878 записей за три дня
- ▶ [Убит Глава Ингушетии.ру](#)
1 482 записи
- ▶ [День шахтера](#)
176 записей

Остальные темы

- [Ураган "Густав"](#)
- [Уссурийский тигр](#)
- [Победа Валуева в Берлине](#)
- [Заявление Гордона Брауна](#)
- [День кино](#)
- [Скоро день знаний](#)
- [Кинофильм "Гордость и предубеждение"](#)
- [Священный месяц Рамадан](#)
- [Изгнание Новодворской с "Эха Москвы"](#)

За последние три дня 2 536 записей посвящено трём самым популярным сегодня темам.

Из каталога: [Юмор](#) 70 блогов [Творчество](#) 297 [Развлечения](#) 353 [Дом](#) 186 [Технологии](#) 333 [Деловые](#) 200 [Ещё...](#)

Самое популярное и обсуждаемое в интернете

Сервисы

LiveJournal	49 751
LiveInternet	19 476

Блоги

drugoi	188 360
tema	166 691

Запросы

[http razvrat pomomolodye](#)
[griz_pskov](#)

Популярные записи

- [PR! до 15 сентября!](#)
Можешь принять участие ты, даже если твой журнал не на Livejournal.com. Что...
 [ar4i s](#) 23 отзыва

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.

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.

С чем мы имеем дело?

Новости

Пишут о событиях

Кодифицированный язык

События освещаются похоже

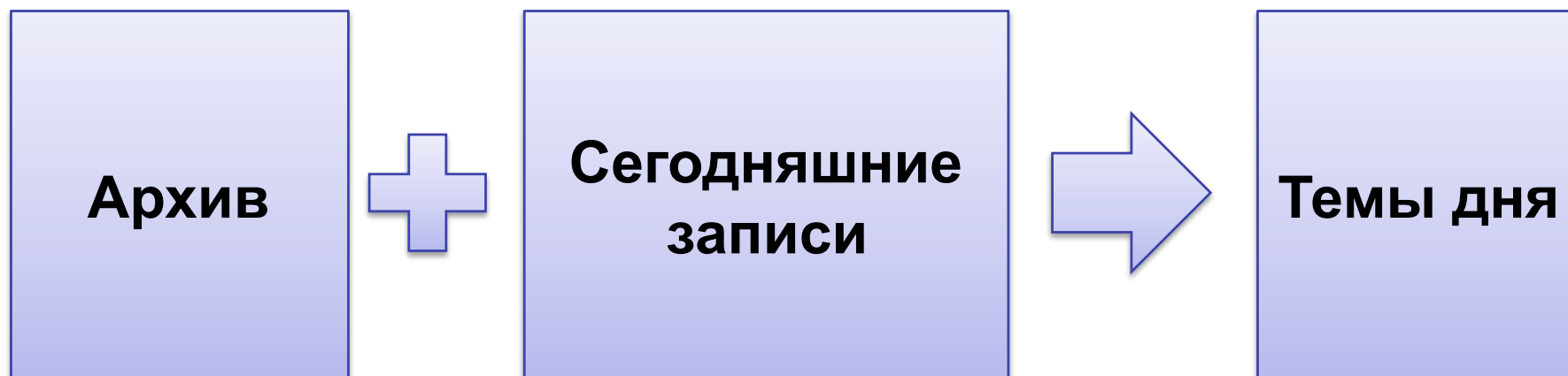
Блоги

Пишут и о событиях и о повседневном

Свободный, почти разговорный язык

Огромное количество разных способов назвать одно и то же

Кажется, что это работает так

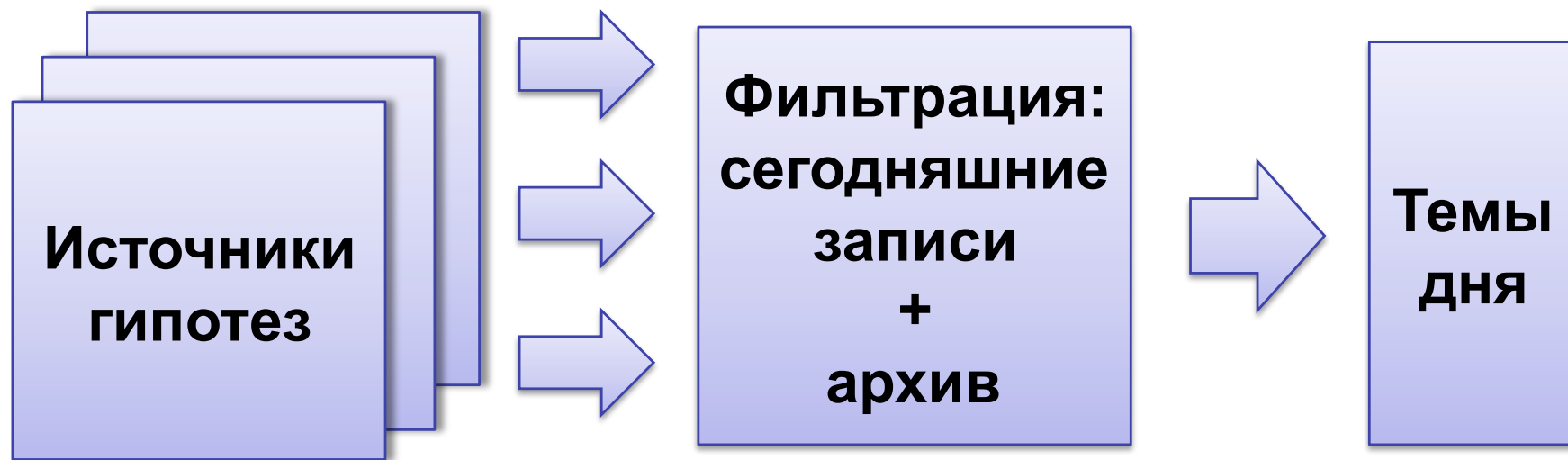


Нет, это работает не так. Почему?

- 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



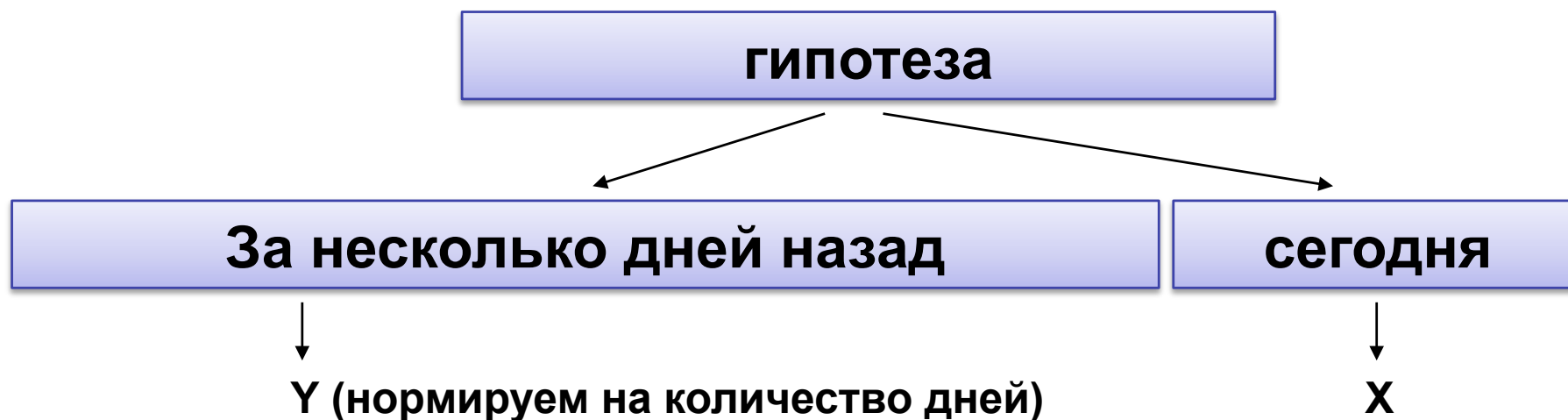
- Гипотезы тем дня – это согласованные, осмысленные, законченные словосочетания (обычно)
- Источники гипотез являются внешними по отношению к системе определения тем дня.
- Записи в блогах работают не как источник тем, а как фильтр гипотез.

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.

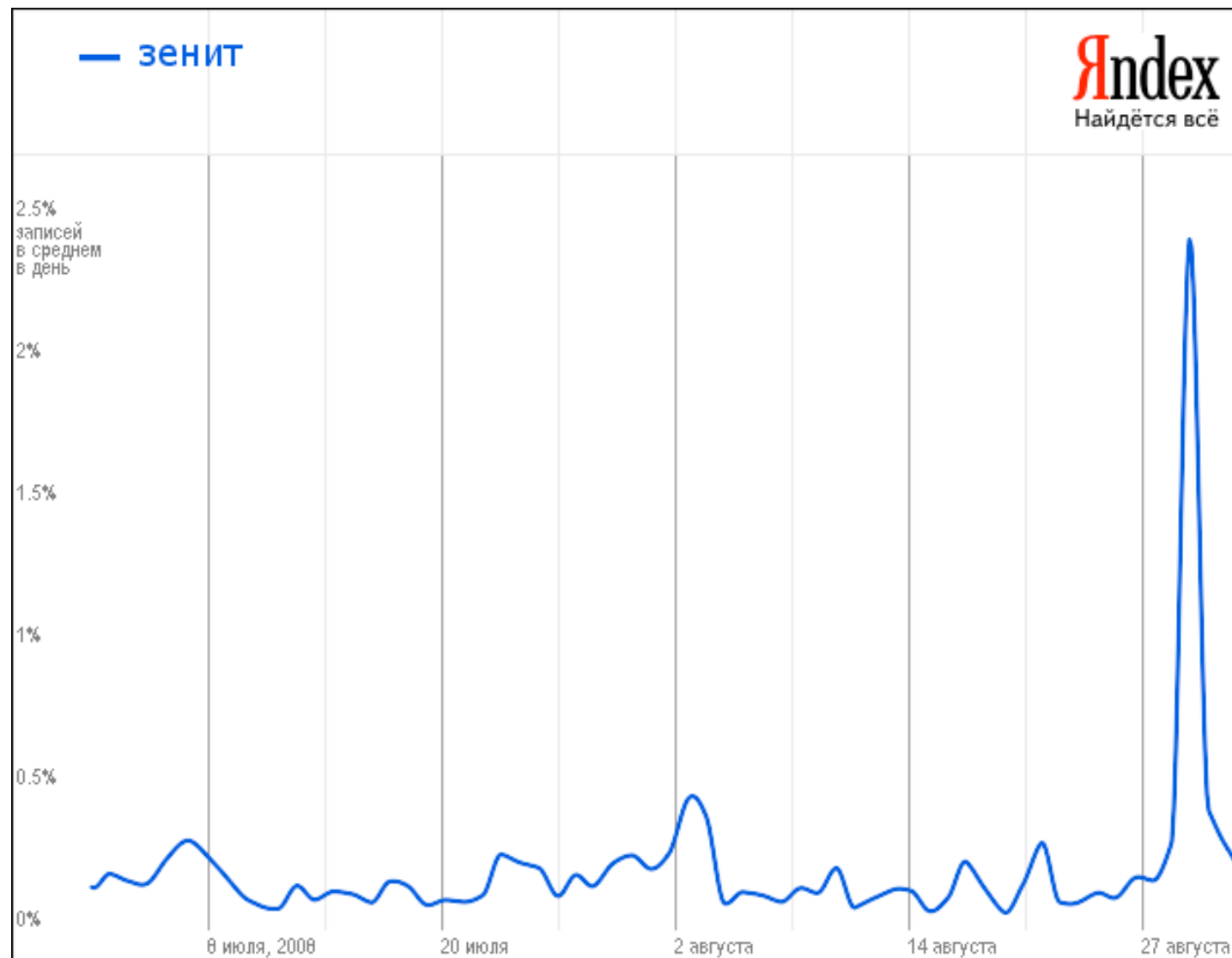
Принцип фильтрации

- 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.

Пример появления темы дня



Яндекс

Формула «темовитости»

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

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

Много гипотез одновременно

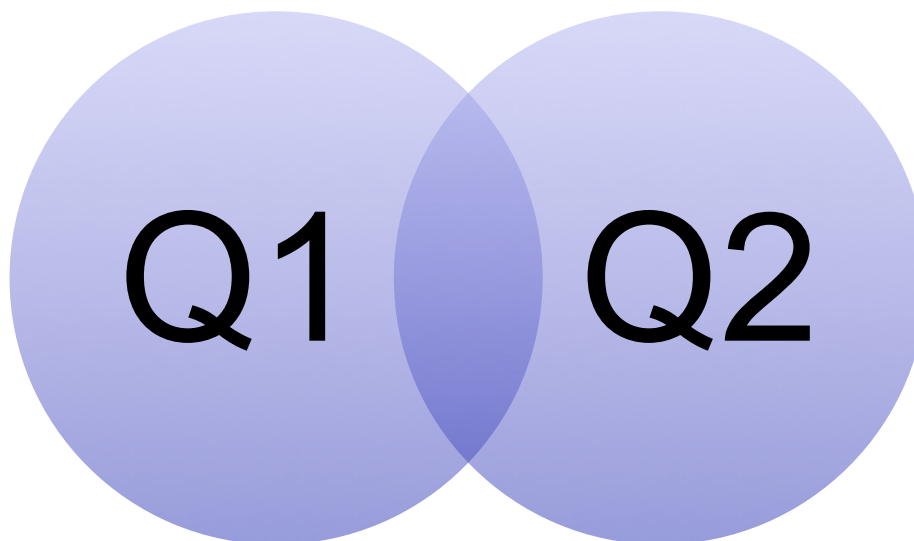


Много гипотез одновременно



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

ВОПРОСЫ?

<http://blogs.yandex.ru/>

Андрей Мищенко druха@yandex-team.ru

Антон Волнухин anton@yandex-team.ru

Яндекс