

10 неправильных способов сравнивать качество ПОИСКОВИКОВ

Протасов Сергей

Rambler

<http://www.search-conf.ru/>, февраль 2010



Rambler

Интернет Новости

Например: качество поиска

Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Типичные ошибки при сравнении качества поиска



Основные ошибки при сравнении

- Нарушение репрезентативности выборки запросов
- Нарушение репрезентативности выборки аудитории
- Недостаточный объем выборки
- География, персонализация



Rambler

Интернет Новости

Например: качество поиска

Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Что такое маркерный метод?

Примеры маркеров для тестирования поиска

- Запрос “xxx”, сайт ууу на 1 стр., хорошо(+1).
- Запрос “xxx”, сайт ууу **на 1 месте**, хорошо(+10).
- Запрос “xxx”, в выдаче есть фраза “zzz”, хорошо(+3).

Качество по маркерному методу:

Мы вводим в поиск много запросов и считаем сумму правильных ответов

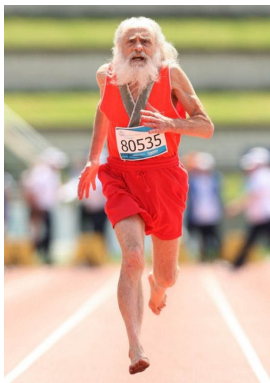


Rambler

Интернет Новости

Например: качество поиска

Маркерный метод



Бегун пробежал первые 50 метров за 5 сек..
За сколько времени он пробежал 100 метров?



Rambler

Интернет Новости

Например: качество поиска

Сложные запросы

Как найти людей

Как найти людей, которые способны оценивать сложные запросы?

- “snmp applet”
- “kafka svejk krusovice”
- “лучшее из оформления парилок”
- “java diagram drawing software”

и как пригласить их **поработать оценщиками** в поиске?



Особенности маркерного метода



Проблемы маркерного метода

- Не охватывает редкие запросы, не обеспечивает полноты.
- Не видит географии и **персонализации**
- **Разногласия** между экспертами
- Новые сайты - маркеры устаревают и “протухают”
- Плохая **точность** (ошибка в десятки процентов)

Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история**
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Персональная история запросов

Эксперты оценивают не чужие запросы - а свои собственные, они точно знают, какой ответ им понравился

- 21:04

Поиск ["Baker" "Trainable grammars * speech"](#) - [Количество просмотренных результатов: 2](#)

 - <http://aima.cs.berkeley.edu/newchapbib.pdf>
 - <http://www.cs.mu.oz.au/acl/C/C92/C92-2066.pdf>
- 15:29

Поиск [1558601244](#) - [Просмотрен 1 результат](#)

 - http://froogle.google.com/froogle_cluster?q=1558601244&pid=22...
- 13:10

Поиск [site:citeseer.ist.psu.edu Bahl Jelinek Mercer Maximum likelihood](#) - [Просмотрен 1 результат](#)

 - <http://citeseer.ist.psu.edu/context/56649/0>

(это был номер icq)



Персональная история запросов

Качество по персональной истории запросов двух экспертов

поисковик	Google	Yandex	Rambler	запросов/год
Эксперт 1	52	15	11	440
Эксперт 2	44	56	38	120

Оцениваем релевантность первого результата



Персональная история запросов: проблемы

Проблемы:

- где найти N тысяч людей?
- как их попросить оценить?
- будет ли это правильной выборкой?



Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики**
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Поведенческие метрики

Поведенческие метрики

их довольно много:

- Клики: c0, c1, ds, asr, p1c, click-entropy,
- Время: ltm, tmr, oldtime
- Глубина поиска: c1t1, kpg
- **Лояльность**, возвращаемость: kret

Метрики - статистические коэффициенты на основе логов действий посетителей



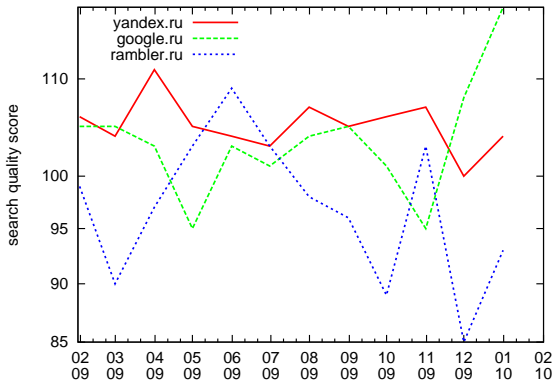
Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси**
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



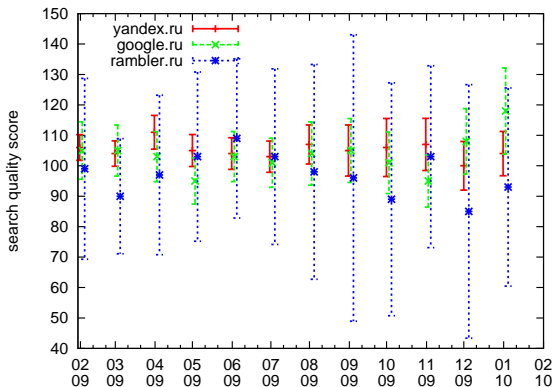
Анализ логов прокси

Берем логи любого доступного прокси сервера и анализируем время на результате, долю 1 клика и т.д.



Анализ логов прокси

А на самом деле нам нужно всегда считать погрешности.



Анализ логов прокси

Проблемы:

- мы не знаем, как вели бы себя юзеры на других поисковиках
- 200 человек - маленькая выборка, и графики “шумят”



Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики**
- 7 “прыгающие пользователи”
- 8 подмена результатов



Поведение по данным LiveInternet

Оказывается, Доля 1 клика для миллионов пользователей публично доступна!

показывается статистика, ограниченная критерием: переходы из поисковика **Search.Mail.ru** (отслеживать только сами переходы)

отчет:

просмотров за сессию ▾

[по дням](#) | [по неделям](#) | [по месяцам](#)

[<14 фев](#)

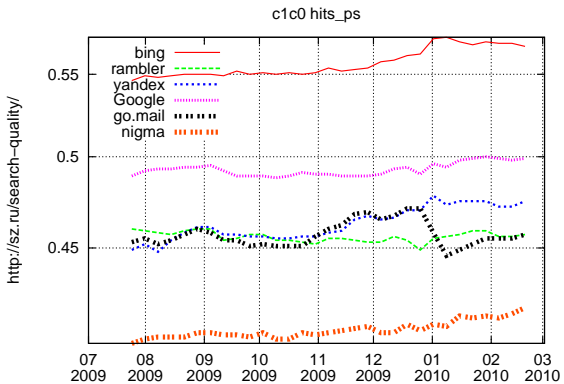
**с 15 по 21
февраля**

<input checked="" type="checkbox"/> 1 просмотр	1,205,597	45.8%
<input checked="" type="checkbox"/> 2-3 просмотра	794,543	30.2%



Поведение по данным LiveInternet

Теперь мы можем построить графики “качества” поисковиков



Rambler

Интернет Новости

Например: качество поиска

Разное число результатов в выдаче

10 результатов вместо 15 иногда лучше нового ранжирования
RN15

	R15	R10 split/64	RN15 split/64
аср	3.93	3.32	3.76
plc	3.49	2.76	3.28
tmr, сек	50.6	49.9	51.1
otk, %	25.8	26.4	25.7
c1t1, %	15.8	15.9	16.1



Rambler

Интернет Новости

Например: качество поиска

Поведение по данным LiveInternet

Проблемы:

- сильные искажения при маленькой доле рынка
- разное число результатов в выдаче
- мы не знаем, как вели бы себя юзеры на других поисковиках



Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Поведение прыгающих пользователей (данные Рамблер)

Анализируем индивидуальные сессии пользователей:

время	запрос	url	
15:28	“куздра”	поисковик1	неудовлетворен
15:35	“куздра”	поисковик2	неудовлетворен
15:40	“куздра”	поисковик3	удовлетворен

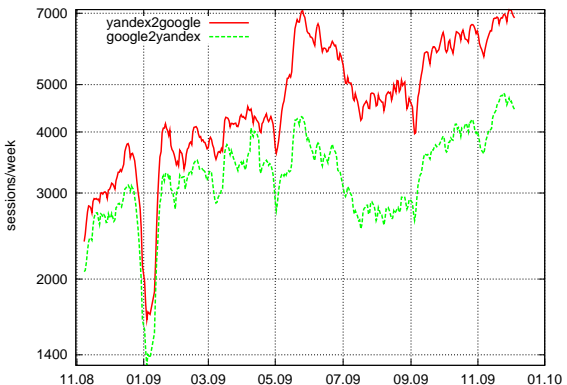


по логам плагинов к браузерам



Поведение прыгающих пользователей (данные Рамблер)

Одинаковые запросы в разных поисковиках.



Поведение прыгающих пользователей (данные Рамблер)

Относительная сила двух поисковиков.



Rambler

Интернет Новости

Например: качество поиска

Поведение прыгающих пользователей (данные Рамблер)

Влияние интерфейса (Яндекс vs Google)

Искомая комбинация слов нигде не встречается.

«ЛОМО 7К-44» · Запросов за месяц: ломо — 9 126, 7к — 904, 44 — 0.

в других поисковых системах: [Google](#) · [Bing](#) · [Yahoo!](#) · [Rambler](#) ·

[Яндекс.Каталог](#)



Поведение прыгающих пользователей

Проблемы:

- слишком умные пользователи (нерепрезентативная выборка)
- влияние интерфейса



Обзор доклада

- 1 типичные ошибки
- 2 маркерные базы
- 3 персональная история
- 4 поведенческие метрики
- 5 логи прокси
- 6 Web счетчики
- 7 “прыгающие пользователи”
- 8 подмена результатов



Подмена результатов с кэшированием

Собираем top 1 млн самых частотных запросов и подменяем выдачу. Как меняется поведение?



Подмена результатов с кэшированием

Проблемы:

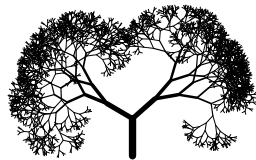
- плохая выборка - как у маркерного метода
- ранжирование оперативной информации
- география и персонализация?



Подмена результатов в реальном времени

Метапоиск чужих результатов. Варианты:

- берём свои сниппеты и заголовки
- берём подмеси, отключаем рекламу
- подмешивание или полная замена?



Подмена результатов в реальном времени

	R10 1/64	Y10 1/1024	G10 1/1024	B10 1/1024	M10 1/1024
аср	3.32	3.35	3.50	3.53	3.63
p1c	2.76	2.74	2.87	2.92	3.01
tmr, сек.	49.9	57.1	57.2	52.1	49.2
otk, %	26.4	23.9	23.5	25.2	27.8
c1t1, %	15.9	15.8	14.1	13.6	13.4

февраль 2010



Подмена результатов в реальном времени

Проблемы:

- метапоиск запрещен
- сильное влияние подмесов (новости, картинки, рубрики)
- результаты поиска используются как **закладки!**
- мы не знаем как вели бы себя **пользователи других ПОИСКОВИКОВ**



Осталось за кадром

Еще неправильные способы:

- подмена результатов на прокси (200 чел мало)
- оценка качества через долю рынка (причина и следствие)
- Rambler Top 100 (проблемы аналогичны LiveInternet)
- комбинация неправильных (ошибка слабого звена)



Итоги неправильных тестов

- Самый точный из неправильных - подмена результатов (метапоиск)
- На сплит тесте мы отстаем от лидера на 10 процентов по tnr,otk
- Выигрываем на 5 процентов по r1c,асr на сплите нового ранжирования

Спасибо. Вопросы?



Итоги неправильных тестов

- Самый точный из неправильных - подмена результатов (метапоиск)
- На сплит тесте мы отстаем от лидера на 10 процентов по tmr,otk
- Выигрываем на 5 процентов по p1c,аср на сплите нового ранжирования

Спасибо. Вопросы?



Интернет Новости

Интернет Новости

Например: качество поиска