Новое ранжирование Рамблера

Протасов Сергей

Rambler Research

http://optimization.ru, ноябрь 2010

Обзор доклада

Rambler поиск

Обзор доклада

Rambler поиск

Проблемы контролируемого обучения

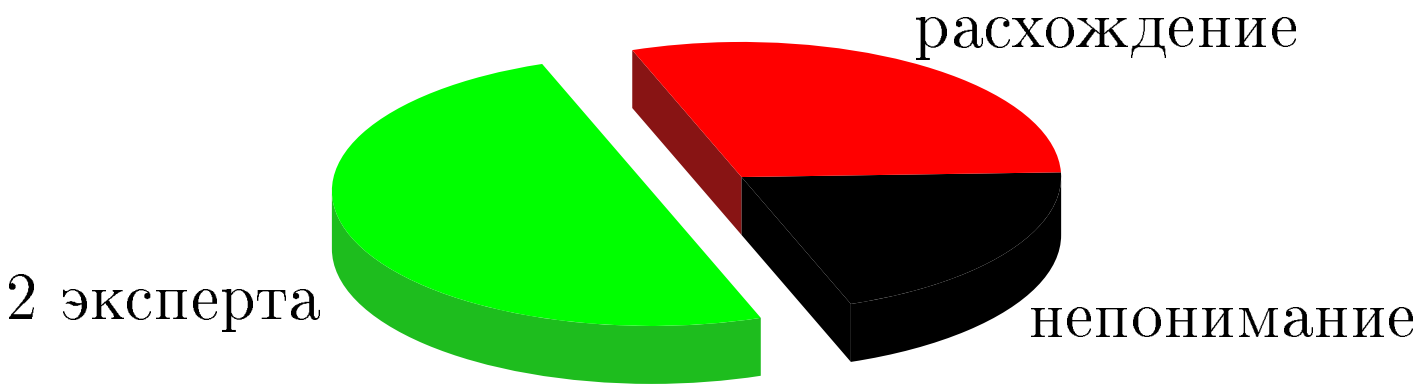
Проблемы любого контролируемого обучения

- Дефицит экспертов
- Шумящие факторы

Rambler
поиск

Проблемы контролируемого обучения

Эксперты: непонимание запросов, расхождение мнений.
Ни один человек не может понимать всего.



Проблемы контролируемого обучения - Дефицит экспертов

Дефицит экспертов ограничивает качество.

- Возьмем специалиста по взрослым сайтам и обучим на нем MatrixNet
- Возьмем врача и протестируем поиск по лекарствам на обученном алгоритме.

Качество для врача будет низким, так как для лекарств важны другие факторы.

Проблемы контролируемого обучения

Проблема расхождения мнений между экспертами.

	Эксперт 1	Эксперт 2	...	Эксперт 10
пластиковые окна	сайт1	сайт2	...	сайт10
натяжные потолки	сайт1	сайт2	...	сайт10
межкомнатные двери	сайт1	сайт2	...	сайт10

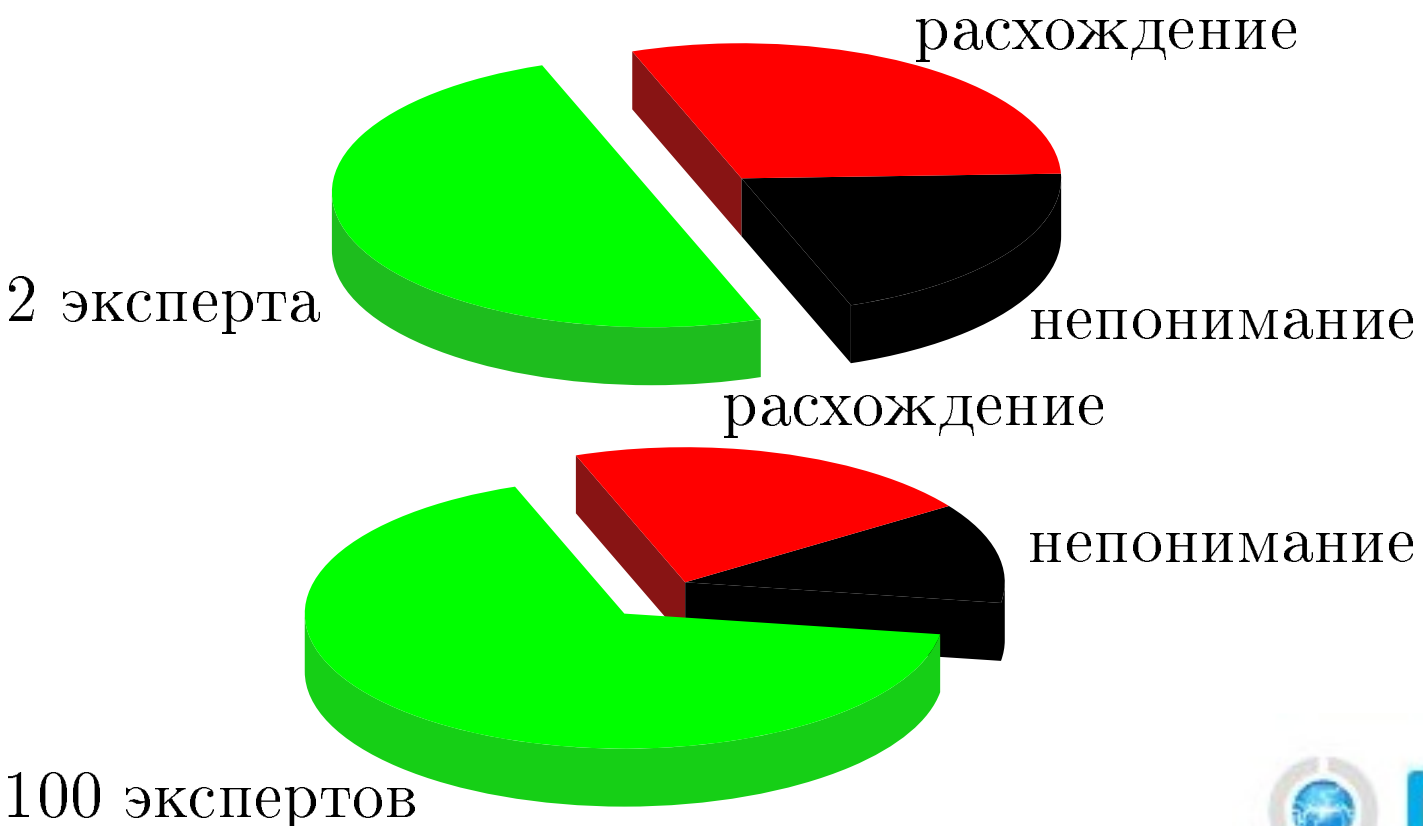
Таблица: *Каждый эксперт любит свои сайты*

MatrixNet не сможет обучится на противоречивой информации.
Расхождение 100%, качество близко к 0.

Rambler поиск

Проблемы контролируемого обучения - Дефицит экспертов

Что делать? Увеличивать число и разнообразие экспертов.



Rambler поиск

Проблемы контролируемого обучения - Дефицит экспертов

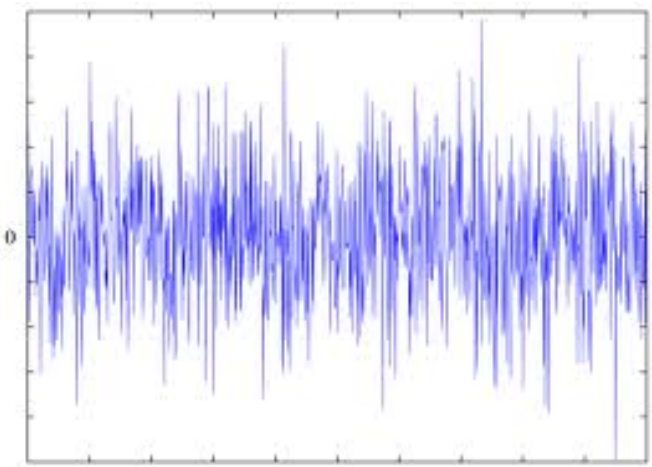
Увеличивать число экспертов до каких пор?

Проблемы контролируемого обучения

Факторы

- Сильные - определяют основное качество ранжирования.
- Слабые - все вместе могут создать силу.
- Шумящие - вредны.

При добавлении факторов качество сначала растет, потом падает из-за “шумящих” факторов.



Rambler поиск

Проблемы контролируемого обучения

Слабые или шумящие?

Проблема в том, что мы не можем точно отделить “шумящие” факторы от “слабых”. Алгоритм обучения с учителем не может справится с “шумящими” факторами, Сумма шумящих не создает сильный.

P.Long R.Servedio, Random Classification Noise Defeats All Convex Potential Boosters (Google Research, 2010)

Rambler поиск

Проблемы контролируемого обучения

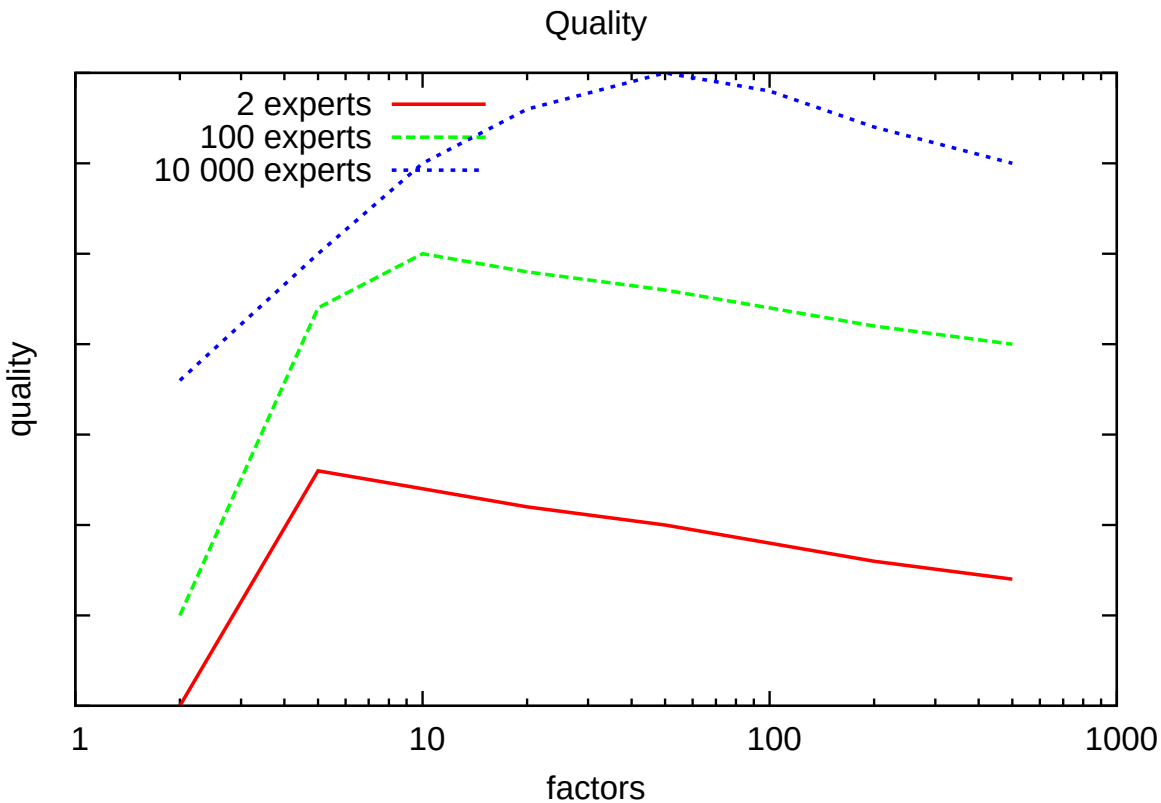
Надо больше экспертов

Для более точного выявления шумящих факторов надо увеличивать число экспертов.
Чем больше экспертов, тем больше слабых факторов мы можем добавить.

Rambler поиск

Проблемы контролируемого обучения

Больше экспертов → Больше понимания → Лучше качество



Качество надо измерять на отдельной группе экспертов.

Rambler поиск

Проблемы контролируемого обучения

Обучение экспертами.

- Эксперты смотрят на результаты друг-друга, пользуются поиском как эталоном.
- Быстрое устаревание актуальности оценок.
- Очень высокое значение цена/качество

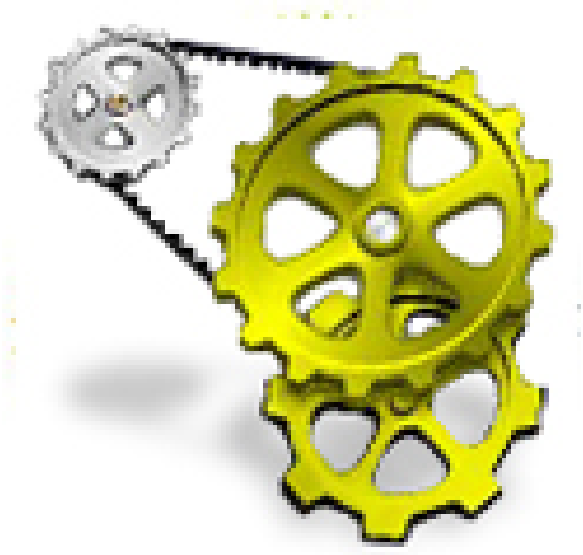
Проблемы приводят к частичному отказу от контролируемого обучения.

Rambler
поиск

Контролируемое обучение

Где можно использовать контролируемое обучение?

- Фильтрация документов (порноранк)
- Реинжениринг чужого ранжирования



Rambler
поиск

Реинжениринг чужого ранжирования

Реинжениринг чужого ранжирования

- (+) Легко (дешево) распарсить/разметить 10 млн запросов и обучить сотни факторов
- (-) Не можем подняться выше оригинала по качеству, если у нас нет нужных факторов.
- (-) Возникает похожесть результатов

Rambler поиск

Похожесть результатов между поисковиками

Похожесть результатов.

	yandex	google	mail	rambler	bing
yandex		46	38	39	37
google	46		40	41	39
mail	38	40		34	32
rambler	39	41	34		34
bing	37	39	32	34	

Таблица: *Доля совпадений по первому месту,%, top10k*

Rambler поиск

Обзор доклада

Rambler поиск

Новое ранжирование Рамблера

Новое ранжирование Рамблера - совокупность идей

- Неконтролируемое обучение без экспертов
- Вместо мнения экспертов - поведение пользователей
- Стараемся быть уникальными
- Заботимся о первой странице
- Рандомизация

Rambler поиск

Новое ранжирование Рамблера 2010

Рандомизация поисковой выдачи

- Борьба с обратной связью в поведенческих ранках
- Чем дольше работает, тем лучше качество
- Рандомизируем только плохое, хорошее не рандомизируем
- Затруднение реинжиниринга алгоритмов ранжирования поиска
- Размазывание трафика по первой сотне сайтов.

Каковы результаты?

Rambler
поиск

Обзор доклада

Rambler поиск

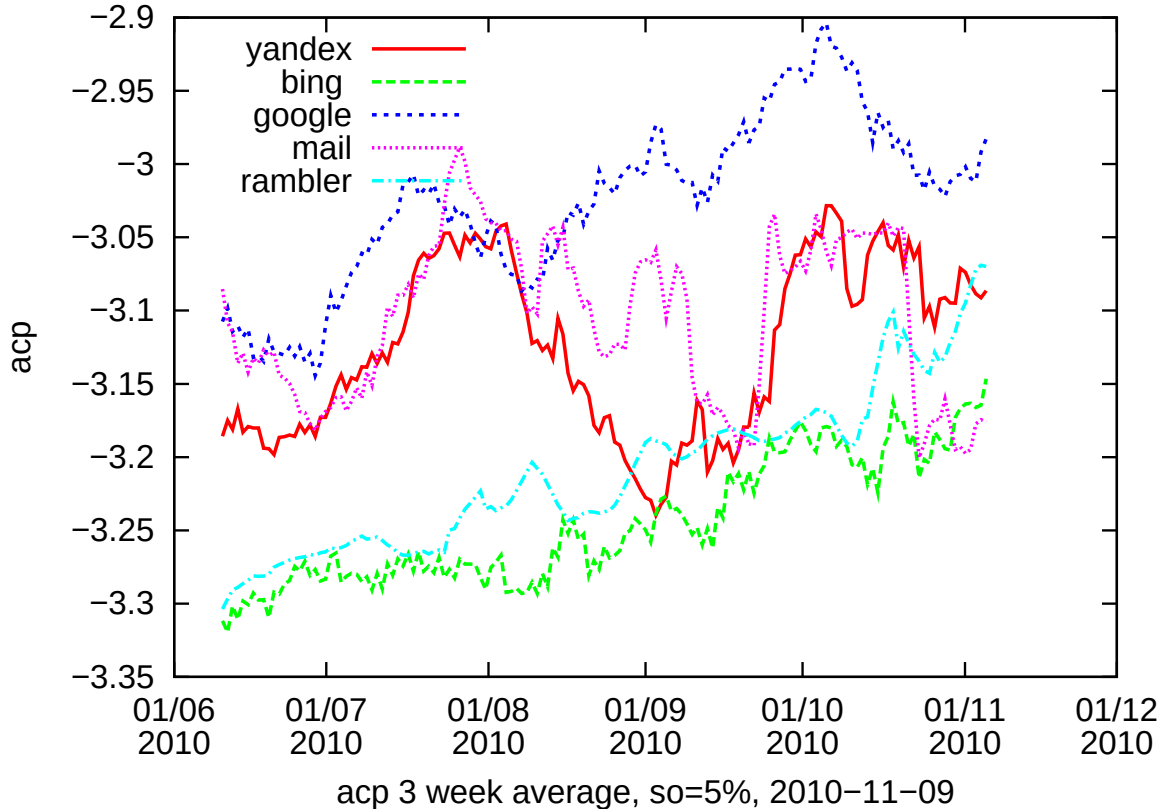
Графики: Качество поиска p1c

Своих метрик недостаточно.

Rambler поиск

Графики: Качество поиска acp конкурентов

Наилучшее качество не у лидера по доле рынка



Пользователи Рамблера - не гики.

Rambler поиск

Обзор доклада

Rambler поиск

Что влияет на долю рынка?

- Качество (ранжирование)
- Маркетинг (телевидение)
- Новые ноутбуки и плагины к браузерам.
- реклама в результатах поиска отталкивает
- Фишки: вертикали, сниппеты, опечаточник, подсказки.

Каково влияние основых двух факторов?

Rambler поиск

Оцениваем влияние качества

Падение метрик качества $\longrightarrow$ Падение доли

	дата	Качество %/макс.	Рост доли % в месяц
mail-gogo	2010/01	80	-5
google	2010/08	100	0
yandex	2010/08	90	0
rambler	2010/01	80	-5
rambler-new	2010/11	90	0
mail-gogo	2010/08	90	0
mail-google	2010/10	100	5

Таблица: *влияние качества*

На Google и Yandex работают дополнительные факторы.

Оцениваем влияние маркетинга

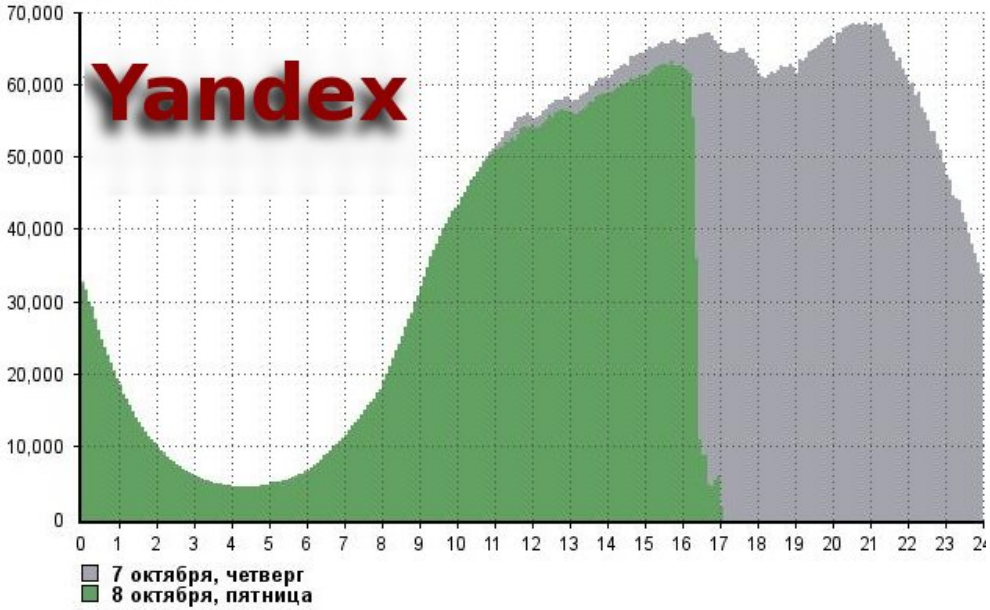
Маркетинг - сейчас главный фактор

- Аудитория поиска растет на 60% в год
- Основная часть аудитории сформирована в последние два года.
- Влияние качества усилится лишь после того как рост остановится.
- Новые ноутбуки, смартфоны.
- Плагины к браузерам, браузеры.

Rambler
поиск

Оцениваем влияние маркетинга (по данным LiveInternet)

Yandex: DDoS 2010/10/08 показал, что
2/3 пользователей Яндекса не воспользовались другими поисковиками.

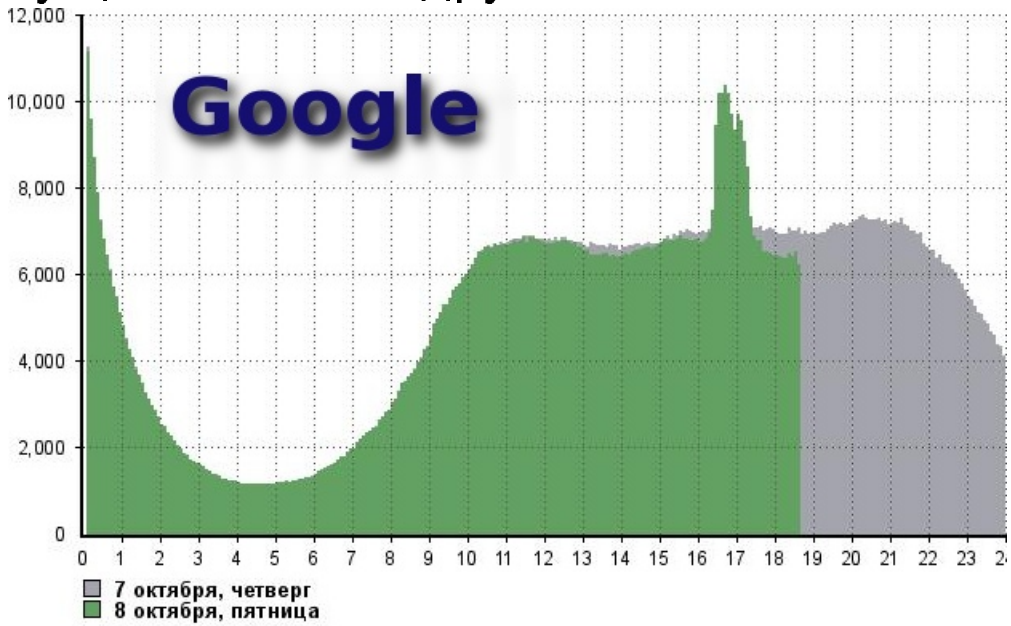


Rambler поиск

Оцениваем влияние маркетинга (по данным LiveInternet)

Yandex: DDoS 2010/10/08

Возможно, что 2/3 пользователей ничего не знают о существовании других поисковиков.

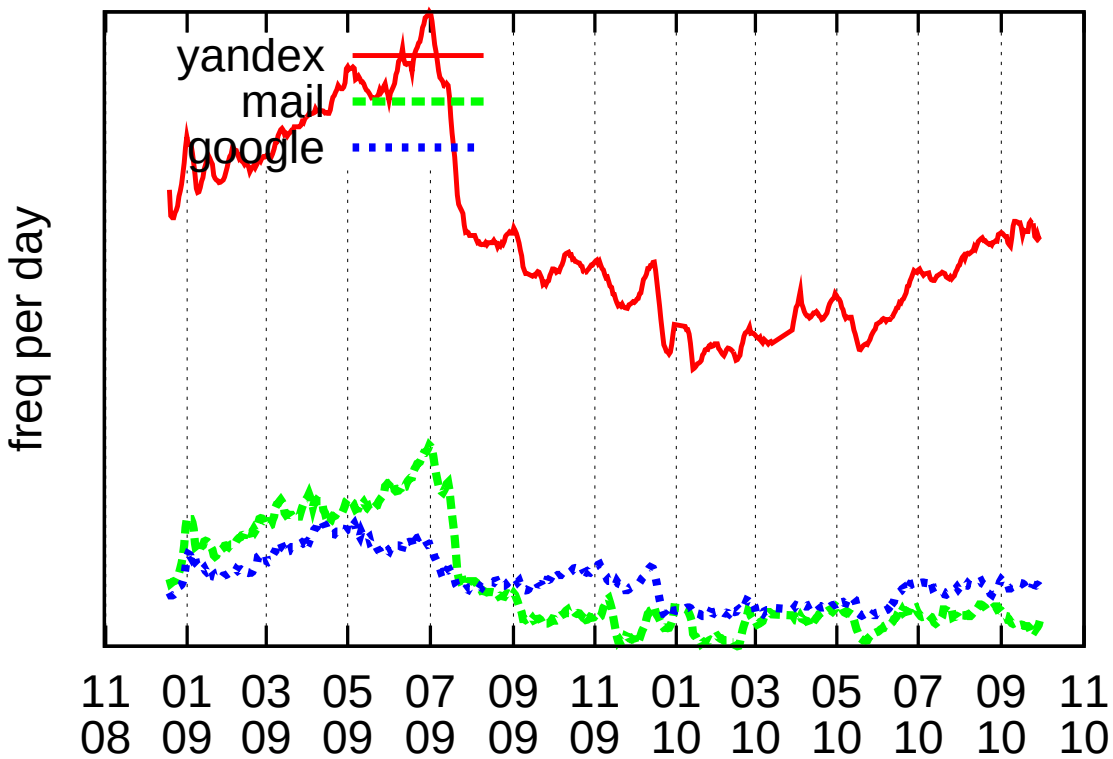


Rambler поиск

Маркетинг: сила бренда (по данным Рамблер)

Графики частотности запросов “яндекс”,“гугл”,“майл” в поиске Рамблера

Яндекс в 2 раза более на слуху.

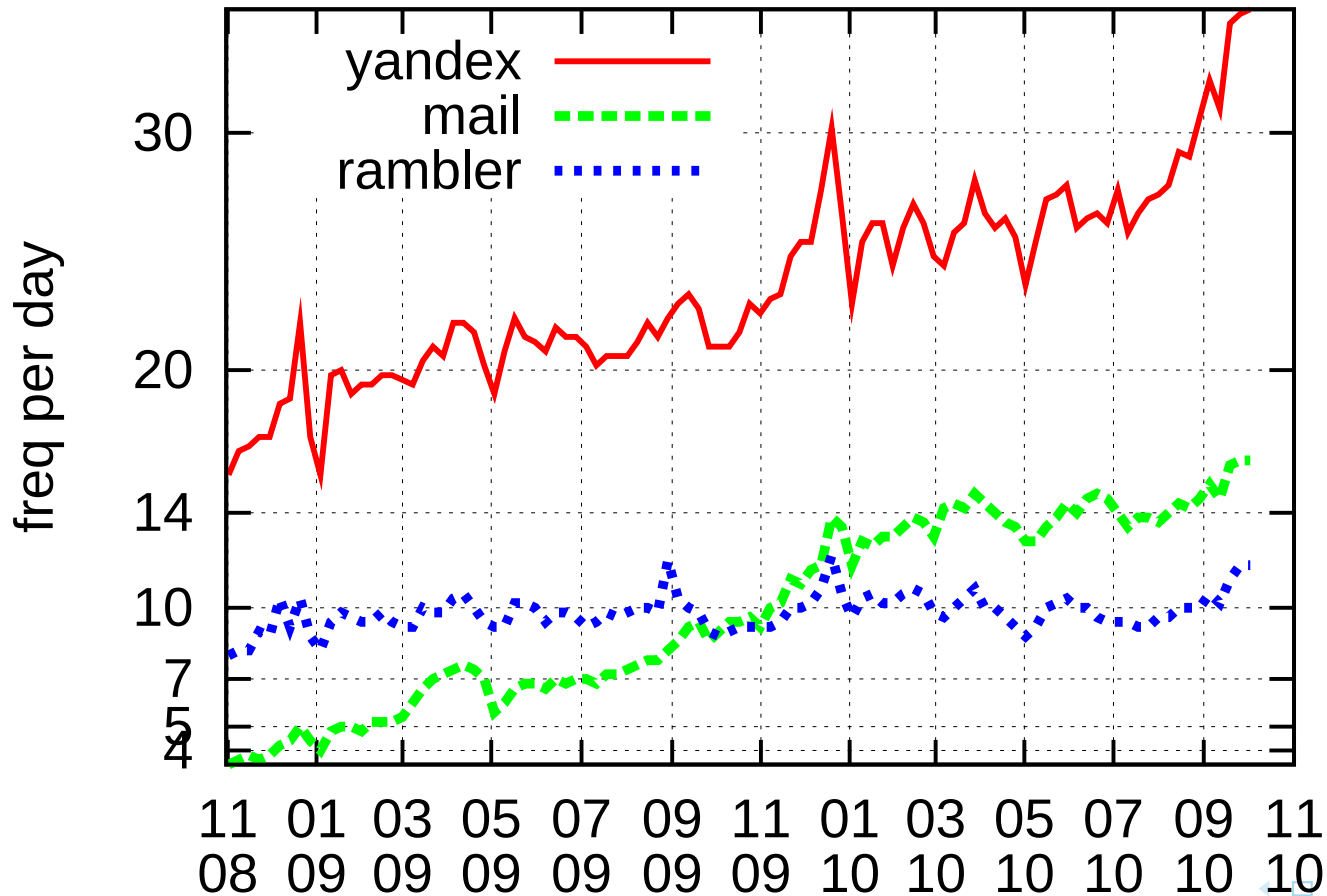


Rambler поиск

Маркетинг: сила бренда (Google Trends)

Графики частотности запросов “рамблер”,“яндекс”,“майл” в поиске Google

Яндекс в 3 раза более на слуху.



Google Trends LABS

Rambler поиск

Маркетинг: доля новых пользователей (Rambler Top 100)

Rambler Top 100 cookie creation age
Кука Рамблера (например '05F9DAFE4BED91CF000156714BA4DF01') внутри себя содержит дату создания.

	возраст куки, дней	доля новых, проценты
yandex	0.0051	11.1
google	7.85	4.0
rambler	34.9	2.61
bing	27.1	0.35

Таблица: *Средний возраст куки*

Яндекс иногда откуда-то берет толпы новых пользователей. Реклама на ТВ?

Качество и маркетинг

Качество и маркетинг

	качество место	маркетинг место	прогноз 1 год
yandex	II	I	рост
google	I	II	падение

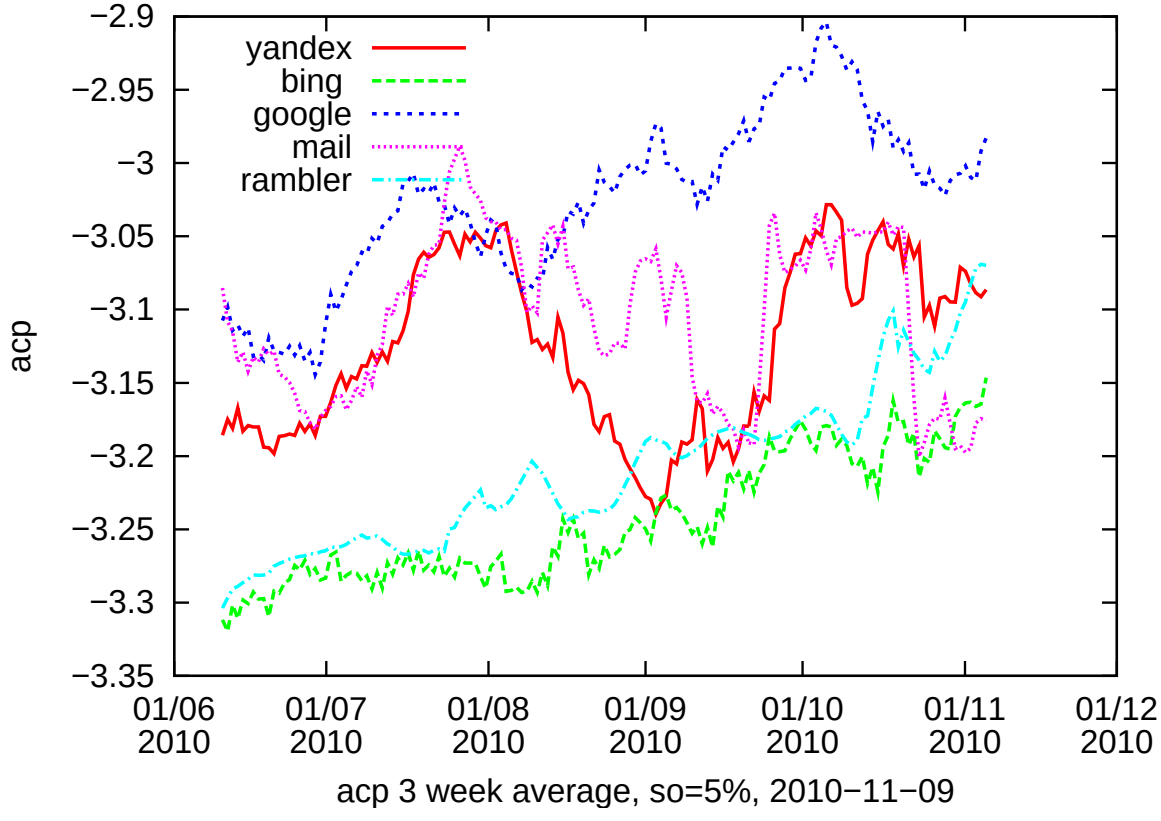
Таблица: *Качество и маркетинг*

Rambler поиск

Новое ранжирование Рамблера

- Выкатка: октябрь 2010

Будет ли эффект в росте доли рынка?



Rambler поиск

Новое ранжирование: доля рынка

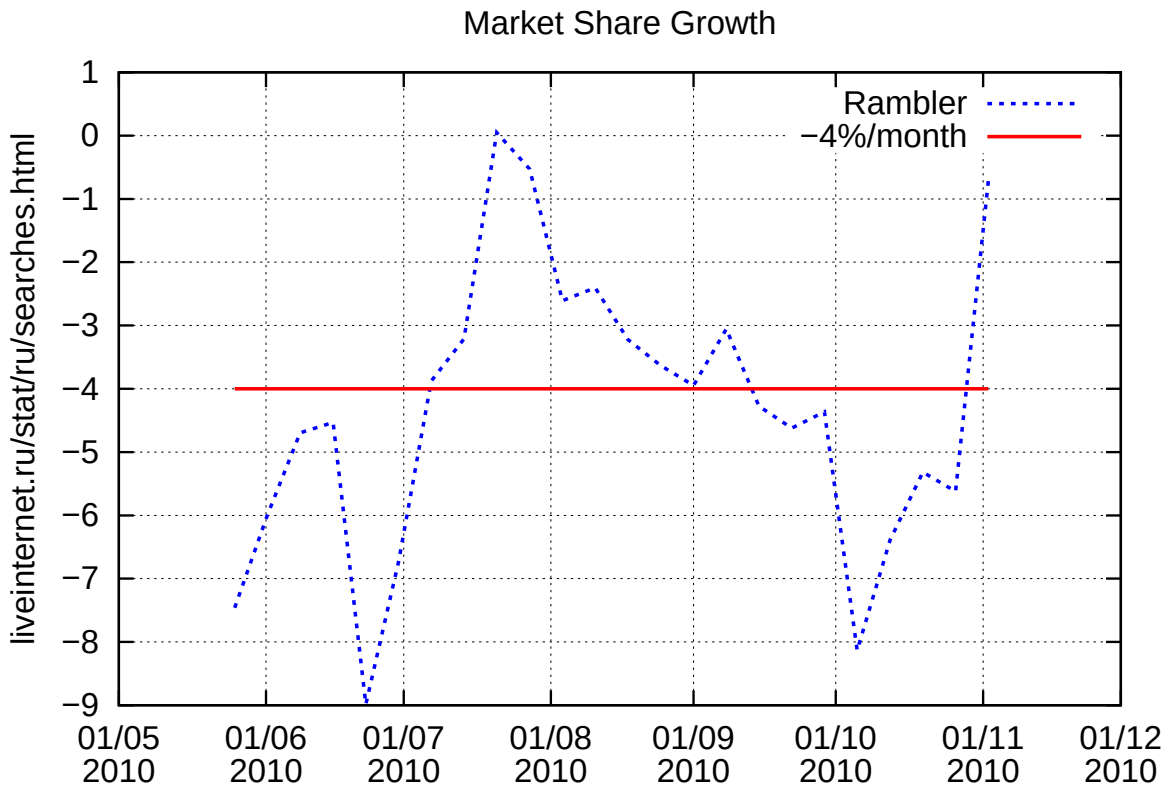
Доля рынка LiveInternet, как оценка поисковой системы.

- Зависит от чужого ранжирования (LiveInternet как фактор)
- Зависит от активности антивирусных роботов поисковиков.

Rambler
поиск

Новое ранжирование: доля рынка

Стабилизация по доле рынка означает рост на 60% в год.
Чем дольше работает - тем лучше качество.



Rambler
поиск