Яндекс

Ранжирование: от строчки кода до Матрикснета

Федор Романенко
Менеджер отдела качества поиска

СтуДень, Москва, 1 апреля 2010

Ранжирование – что это?

— основной алгоритм в поиске

— с помощью **факторов** вычисляет **релевантность** документа в виде числа

— выбирает топ-10

— сложная вещь с простым результатом

— определяет долю на рынке

— самый большой секрет поисковых компаний

Я

Язык запросов
(Yahoo – 1994 *)

Бинарное ранжирование: 0 или 1?

- формируем правило отбора релевантных документов, используя «язык запросов»
- просматриваем все найденные документы

— *подходит только для специалистов*

— *нельзя использовать, когда найдено много*

* здесь и далее: первый пример удачного применения в интернет поиске

Я

Текстовое ранжирование: tf * idf
(Altavista – 1995)

«Близость» текста запроса к тексту документа учитывает:

- количество слов запроса в документе (**t**erm **f**req.)
- обратную частоту слова в языке (**i**nverted **d**ocument **f**req.)
- длину документа и запроса

Текстовая релевантность также учитывает:

- слова в заголовках
- близость слов запроса
- совпадение словоформ

Я

Лингвистика – понимание языка
(Яндекс – 1997)

- морфология: *ребенок шёл = дети идут*
- опечатки: *аднаклассники = одноклассники*
- расширения: *МГУ = Московский Государственный Университет*

Два пути компьютерной лингвистики:

— словарный: морфология, синтаксис

— статистический: языковые модели

Я

Рейтинги сайтов
(Рамблер Top100 – 1997, Яндекс тИЦ – 1999)

Использование внешней информации о странице:

- ссылки с тематических сайтов
- переходы из каталогов
- посещаемость по счетчикам

— *тИЦ Яндекса оброс мифами вебмастеров*

Я

PageRank
(Google - 1998)

PageRank – глобальная ссылочная авторитетность:

- на страницу много ссылок
- на страницу есть ссылки с авторитетных страниц

Модель случайного «блуждателя»:

1. выбираем случайную страницу
2. с вероятностью *0.85* переходим по выходящей ссылке
3. с вероятностью *0.15* устаем и переходим к п.1

Ссылочное ранжирование
(Google - 1998)

Текст ссылки описывает страницу, на которую ссылается

- голосование весами релевантных ссылок
- уточнение тематики страницы с помощью сопоставления текстов ссылок

Я

Метрика качества поиска

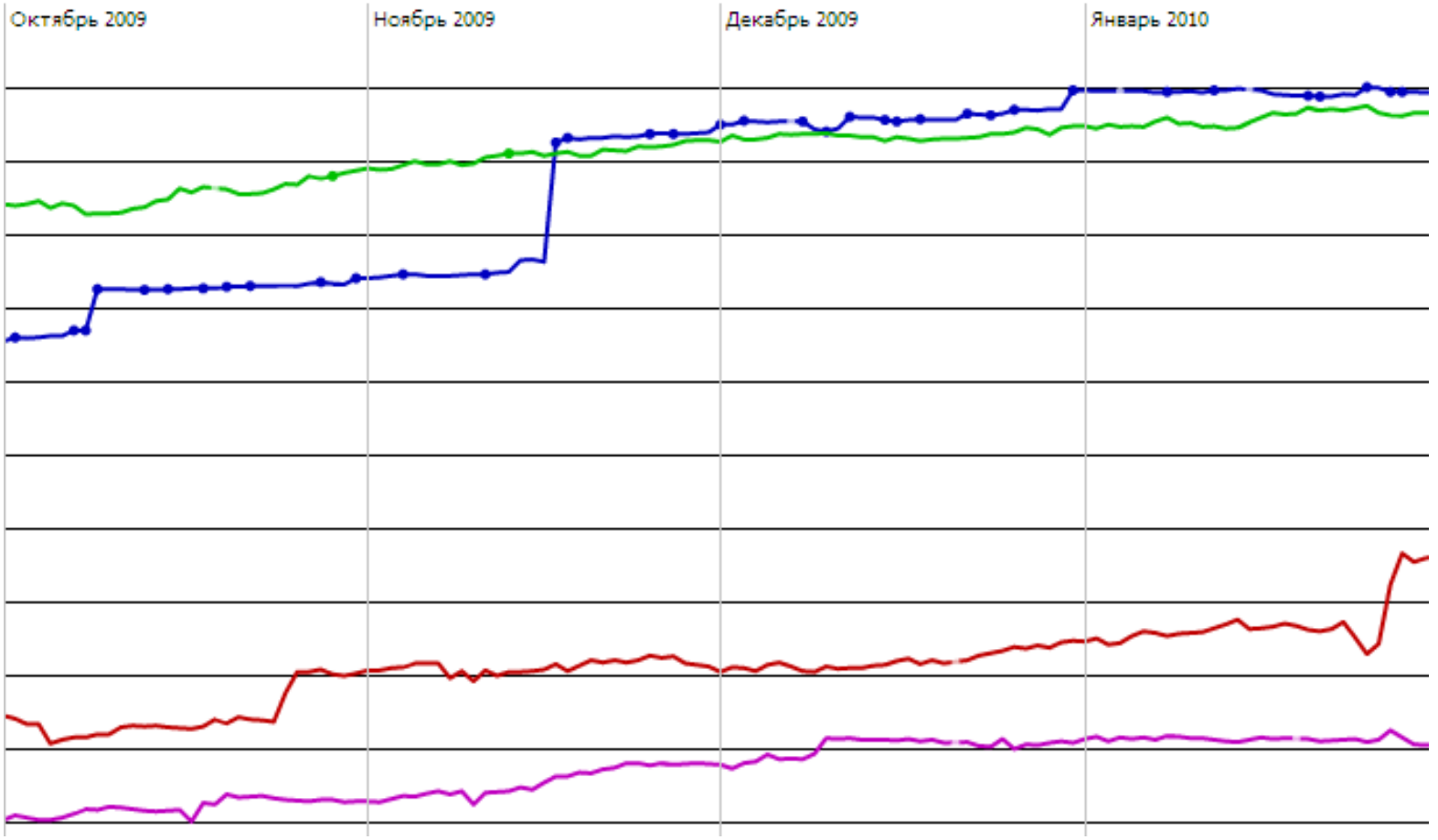
Как измерить удовлетворение пользователей поиском?

- из потока запросов пользователей строим репрезентативную **выборку**
- **асессоры** (люди) оценивают отдельные результаты поиска
- **оценки** используются в числовой **метрике** качества ранжирования

Яндекс: оценено 4 млн. документов по 100 тыс. запросов

Я

Метрика качества поиска



Я

Машинное обучение
(Яндекс, Yahoo, Bing – 200x)

Как учесть в ранжировании **400** параметров документов?

- **модель** (формула на **факторах**) умеет считать релевантность документа

- **машинное обучение** автоматически настраивает модель, максимизируя метрику качества на примерах

- машина учится ранжировать и показывает поведение, не заложенное в нее явно

Я

Проблемы машинного обучения

Метрика – *научить именно тому, что понадобится в бою*
нужно математически измерить «счастье» пользователей

Входные данные – *органы чувств*
позволяют отличить документы по качеству

Способность к обобщению – *аналитические способности*
учимся на примерах, ранжируем то, что никогда не видели

Переобучение – *паранойя*
мало оценок и умная модель – ложные закономерности

Кроссвалидация – *не дадим выучить контрольную*
нужна не память, а умение

Я

Матрикснет
(Яндекс – 2009)

Разработка **Яндекса** (релиз «Снежинск», ноябрь 2009)

- автоматически выбирает связанные факторы и диапазоны их значений
- генерирует тысячи комбинированных факторов
- очень сложная модель без склонности к переобучению

— *извлекает неиспользованный сигнал из старых данных*

— *позволяет добавлять много данных в ранжирование*

Я

Формула Матрикснет

Одна из формул

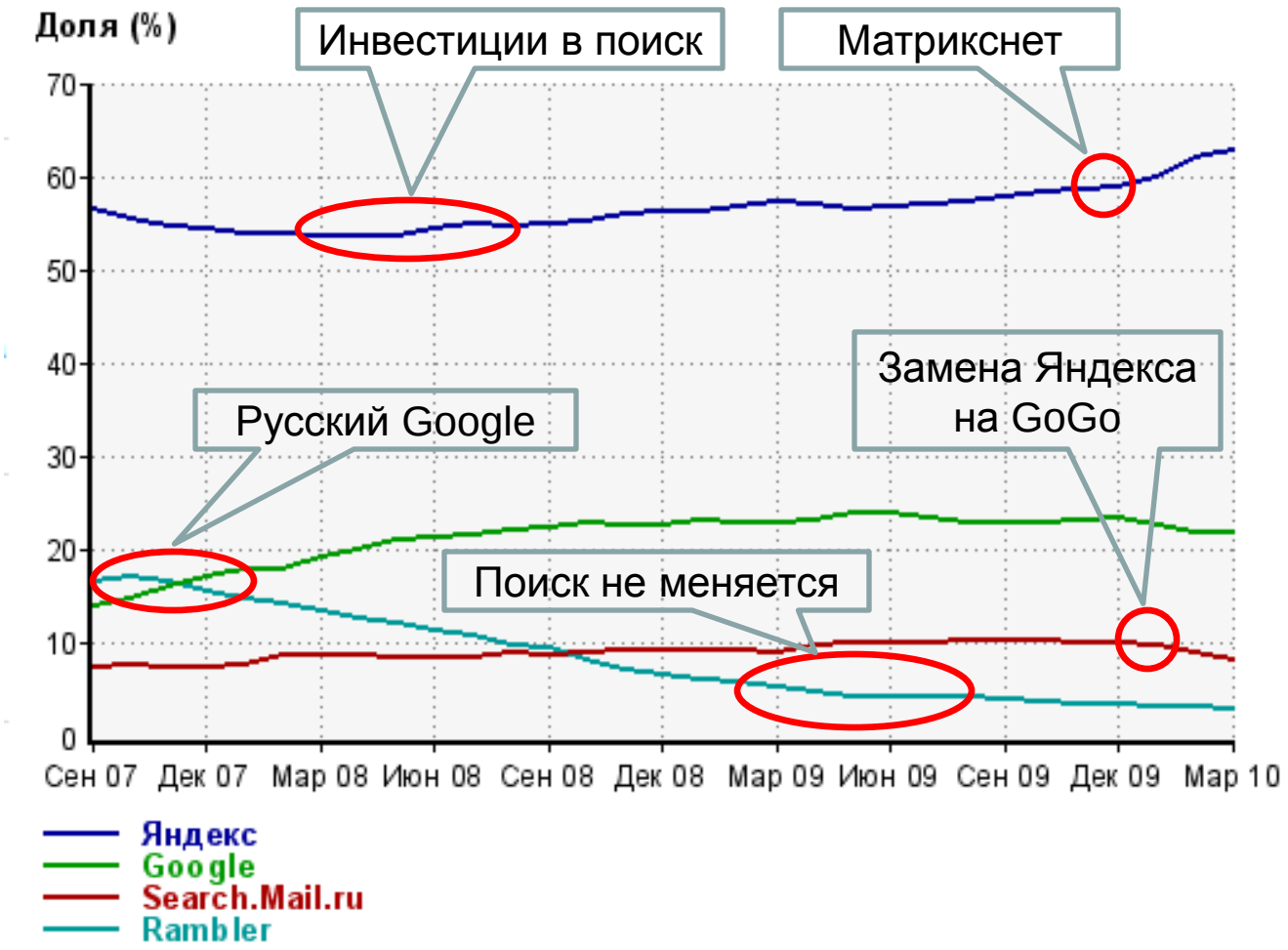
(вычисляет релевантность по факторам найденного документа)

```
...

-4025627,483990,-36437960,39979596,
-92842056,-50086892,-100233272,243162224,
-22282850,57163664,-24991620,-9889194,

...

vars[5699] = fFactorInt[376] > 1060970280 ? 1 : 0; // 0.738757
vars[5700] = fFactorInt[376] > 1061923687 ? 1 : 0; // 0.795584
vars[5701] = fFactorInt[376] > 1049721454 ? 1 : 0; // 0.284137
vars[5702] = fFactorInt[376] > 948291011 ? 1 : 0; // 6.37877e-05

...

,{376, .0f}
,{376, 0.6251018047f}
,{-1, .0}
,{376, 0.05682743713f}
,{376, 0.4546194971f}

...
```

- 20 тыс. строк на C++
- объем 3МБ
- 10 тыс. коэффициентов

Я

Liveinternet: доля рынка, Россия



Я

Независимые метрики

У поиска много «качеств»!

Навигационный поиск	Тематический поиск	Подсказки	Опечатки	Цитатный поиск	Оригиналы	Синонимы
98.0 ↑3.1	31.6 ↑1.1	96.1 ↑0.6	81.8 ↑1.1	66.0 ↑1.0	55.0	63.8 ↑0.4
97.0 ↑1.1	29.3 ↑3.6	95.5 ↓0.3	60.8 ↓0.1	60.0	42.0	62.4 ↓0.2
94.9 ↑2.0	27.2 ↑2.1	93.4 ↓0.9	55.2 ↑0.3	42.0 ↑2.0	42.0	48.3 ↓0.3
94.9	18.1 ↓1.6	91.4 ↑4.0	46.1 ↓3.1	36.0 ↑1.0	38.0 ↓2.0	47.7 ↓0.4
91.9 ↓2.0	17.5	91.0 ↓2.0	22.5 ↑0.3	35.0 ↓1.0	29.0	47.3 ↓0.1
90.9 ↑3.1	17.5 ↓0.6	85.5 ↑0.7	4.4 ↓1.2	35.0 ↓1.0	26.0 ↑2.0	47.3 ↓0.1
89.9 ↓4.0	15.6 ↑0.9	77.8 ↑0.6	1.4 ↓0.4	33.0	26.0 ↑1.0	37.4 ↑0.8
78.8 ↓3.9	2.7 ↓0.3	N/A	0.3	23.0 ↑4.0	8.0 ↓4.0	36.0 ↓0.4

Поисковый спам	SEO-прессинг	Порнография	Полнота индекса	Апдейты	Переходы
0.8 ↑0.3	27.7 ↑0.9	6.3 ↑0.1	75.0 ↑0.8	0.2 ↑0.1	63.2 ↑0.8
3.3 ↑0.4	34.2 ↑0.4	9.6 ↓0.8	64.5 ↑1.7	4.3 ↓0.6	22.2 ↓1.8
4.4 ↓0.2	34.7 ↑0.2	14.1 ↑0.2	59.2 ↓5.3	4.6 ↑1.8	8.9 ↑0.4
4.6	40.8 ↓0.7	15.1	48.5 ↑1.5	4.9 ↑0.3	3.2 ↑0.4
5.0 ↓0.4	42.4 ↑0.1	15.8 ↓1.0	42.7	8.1 ↓2.8	1.0 ↑0.2
5.0 ↓0.1	43.3 ↓0.5	16.8 ↑0.2	15.9 ↑3.3	15.5 ↓0.8	0.1
6.3 ↓1.3	45.5 ↓0.2	20.3 ↓0.7	7.8 ↓0.5	15.9 ↓1.6	0.0
6.5 ↓0.3	48.7 ↓0.8	20.4 ↓0.8	5.2 ↓0.1	36.0 ↓35.4	0.0

http://www.analyzethis.ru

Я

Поиск - вечная задача

— поиск нельзя «доделать»: тот, кто думает, что все сделал, проигрывает

— размер интернета и количество пользователей растут, люди решают в интернете новые задачи

— много направлений: локальность, глобальность, свежесть, разнообразие, скорость, представление, …

— сильные математики и программисты имеют дело с уникальными задачами на огромных объемах данных

— результаты их работы сразу видят миллионы людей

Я

Федор Романенко

Я