



## Особенности регионального ранжирования Яндекса. Белорусская формула

Сергей ЛЮДКЕВИЧ, начальник отдела исследований и аналитики



## ТЕКУЩИЙ АЛГОРИТМ. МАШИННОЕ ОБУЧЕНИЕ

Обучающие данные

Набор запросов  $\mathbf{q}^{(i)}$

Набор документов  $\mathbf{d}_j^{(i)}$  для каждого запроса  $\mathbf{q}^{(i)}$

$\text{Rel}(\mathbf{q}^{(i)}, \mathbf{d}_j^{(i)})$  - ручная оценка соответствия документа запросу

Конкурс «Интернет-математика — 2009»:

$\text{Rel}(\mathbf{q}, \mathbf{d})$  - значения из диапазона  $[0, 4]$

(4 — «высокая релевантность», ..., 0 — «нерелевантно»)



## ФАКТОРЫ РАНЖИРОВАНИЯ

Набор факторов ранжирования

$$F = (f_1(q,d) , \dots, f_N(q,d))$$

Конкурс «Интернет-математика — 2009»:

**N=245**

«Яндекс на РОМИП'2009»:

**N=163** (коллекция VU.WEB);

**N=69** (коллекция KM.RU, без ссылочных факторов)

Алгоритм «Снежинск»:

**N** — несколько тысяч



## ПРИМЕРЫ ФАКТОРОВ РАНЖИРОВАНИЯ

### Запросные

- длина документа в словах;
- язык запроса.

### Текстовые

- наличие точного вхождения запроса в тексте документа;
- наличие точного вхождения запроса в заголовке документа;
- $tf*idf$ ;
- различные модификации формулы Okapi<sub>BM25</sub>.



## ПРИМЕРЫ ФАКТОРОВ РАНЖИРОВАНИЯ

### Ссылочные

- PageRank;
- логарифм количества ссылок на документ;
- процент ссылок на документ, содержащих точное вхождение запроса.

### Географические

- регион сайта;
- язык документа.



## ФУНКЦИЯ РЕЛЕВАНТНОСТИ

Числовое соответствие документа запросу

$$\text{Fr}(q, d) = \text{Fr}(F(q,d)) = \text{Fr}(f_1(q,d), \dots, f_N(q,d))$$

Методы построения функции релевантности:

«Яндекс на РОМИП'2009»: генетический алгоритм

«Снежинск»: жадный (greedy) алгоритм



# ПОСТРОЕНИЕ ФУНКЦИИ РЕЛЕВАНТНОСТИ

## 1. Выбор метрики

(«Яндекс на РОМИП'2009»: **pfound** – максимизация вероятности найти релевантный результат;

«Снежинск»: минимизация невязок между измеренными и вычисленными значениями релевантности)

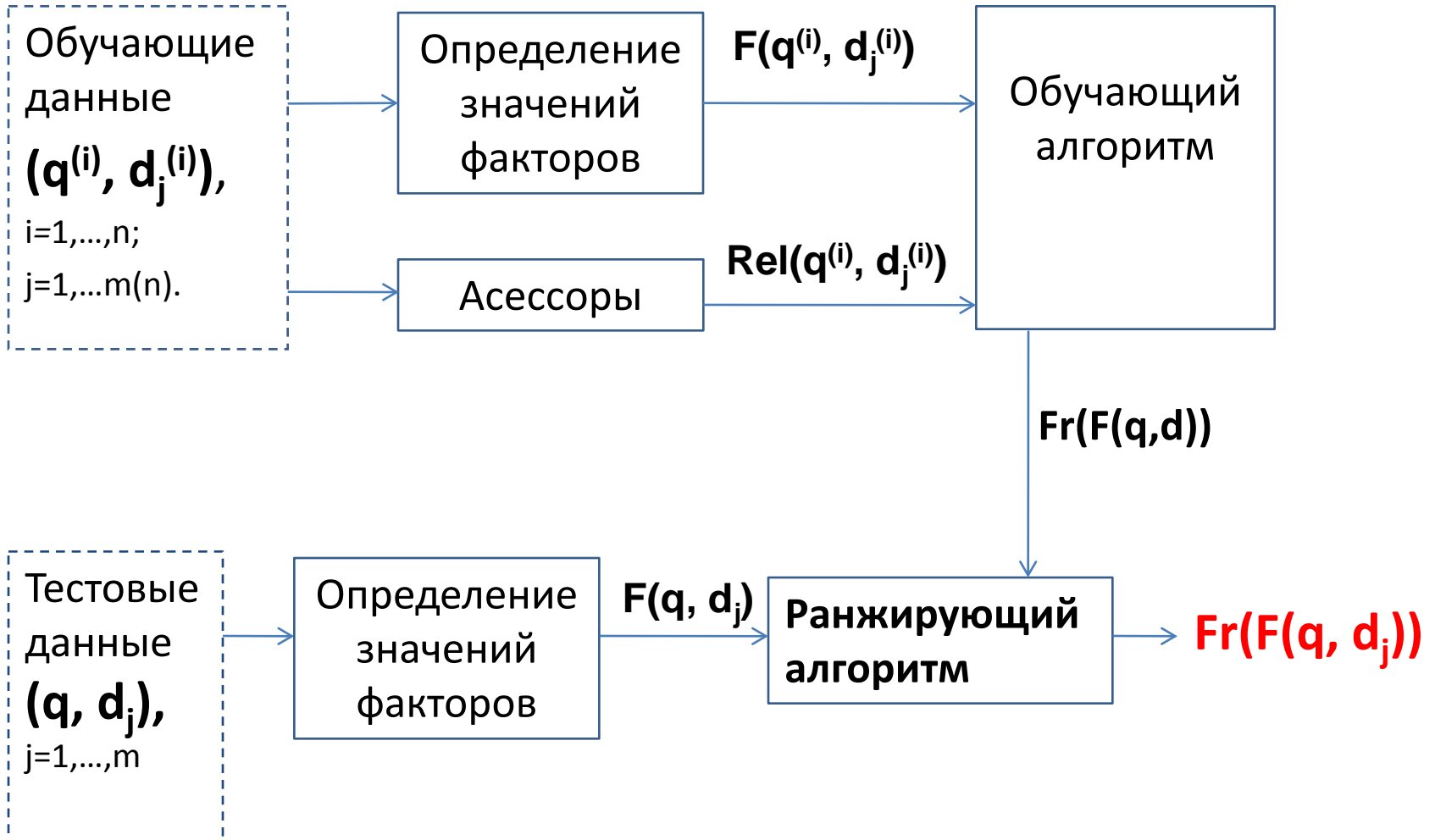
## 3. Подбор вида функции

(«Яндекс на РОМИП'2009»: полином)

## 4. Подбор коэффициентов



# СХЕМА ОБУЧАЮЩЕГО АЛГОРИТМА





## РЕГИОНАЛЬНЫЕ ФОРМУЛЫ

Отдельные функции релевантности:

- 19 городов России: Москва, Санкт-Петербург, Екатеринбург, Новосибирск и др.
- Общероссийская
- Украина
- Белоруссия
- Казахстан

Отличаться могут не только коэффициенты, но и сам вид функций!



# ИССЛЕДОВАНИЕ ФУНКЦИИ РЕЛЕВАНТНОСТИ

## Постановка эксперимента

Выбор исследуемого фактора

Генерация тестовых коллекций

- Варьирование исследуемого фактора
- Фиксация остальных факторов

Индексация тестовых коллекций

Анализ результатов

Принятие решения о характере влияния исследуемого фактора на функцию релевантности



## БЕЛОРУССКАЯ ФОРМУЛА

Фактор: Количество употреблений термина запроса (tf)

Характер зависимости: Прямая

Фактор: Длина документа в словах

Характер зависимости:

Однословные запросы — прямая

Двухсловные запросы — обратная

Трехсловные запросы — не установлена



## БЕЛОРУССКАЯ ФОРМУЛА

Фактор: Количество употреблений самого частотного термина

Характер зависимости:

Однословные запросы — обратная

Двухсловные и трехсловные запросы — прямая



Спасибо за внимание!

Пожалуйста, задавайте вопросы

Для продолжения темы посетите



Корпорация РБС  
115191, Россия, Москва,  
ул. Б. Тульская, д. 13, 4-й этаж ТЦ «Ереван Плаза»  
Телефон: (495) 772-97-91 (многоканальный)  
ICQ-консультант: 377-169-437

<http://rbsgroup.ru> | <http://bdbd.ru> | <http://mediaguru.ru> | <http://webvisor.ru>