Особенности регионального ранжирования Яндекса. Украинская 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 $\mathbf{q}^{(i)}$

Rel($q^{(i)}$, $d_j^{(i)}$) - ручная оценка соответствия документа запросу

Конкурс «Интернет-математика – 2009»:

Rel(q, d) - значения из диапазона **[0, 4]**

(4 – «высокая релевантность»,…, 0 – «нерелевантно»)

ФАКТОРЫ РАНЖИРОВАНИЯ

Набор факторов ранжирования

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

Конкурс «Интернет-математика – 2009»:

N=245

«Яндекс на РОМИП’2009»:

N=163

(коллекция BY.WEB);

N=69

(коллекция KM.RU, без ссылочных факторов)

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_BM25.

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

$$\mathbf{Fr(q, d) = Fr(F(q,d)) = Fr(f_1(q,d), \ldots, f_N(q,d))}$$

Построение функции релевантности с помощью генетических алгоритмов:

1. Выбор метрики

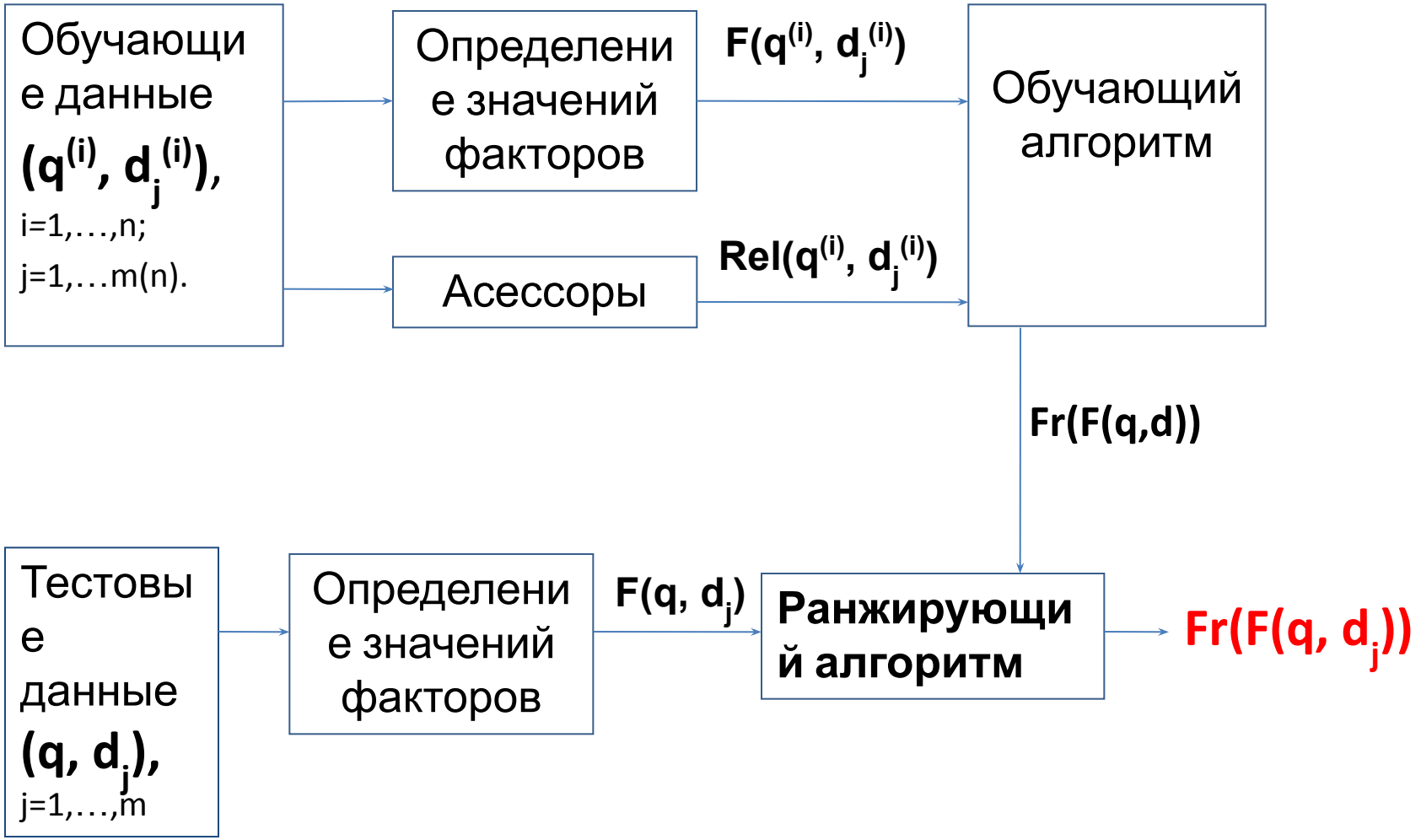
(«Яндекс на РОМИП'2009»: **pfound** – максимизация вероятности найти релевантный результат)

2. Подбор вида функции

(«Яндекс на РОМИП'2009»: полином $\sum a_l \mathbf{f}_1^{i_1} \mathbf{f}_2^{i_2} \ldots \mathbf{f}_N^{i_N}$)

3. 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

УКРАИНСКАЯ ФОРМУЛА

Фактор: Количество употреблений термина запроса (tf)

Характер зависимости: Прямая

Фактор: Длина документа в словах

Характер зависимости: Обратная

Фактор: Количество употреблений самого частотного термина

Характер зависимости: Обратная

Спасибо за внимание!

Пожалуйста, задавайте вопросы

Для продолжения темы посетите

bdbd.ru

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.ruhttp://rbsgroup.ru | http://bdbd.ruhttp://rbsgroup.ru | http://bdbd.ru | http://mediaguru.ruhttp://rbsgroup.ru | http://bdbd.ru | http://mediaguru.ru | http://webvisor.ru