Алгоритмы ранжирования, основанные на идее суперпозиции, и их применение к задаче информационного поиска

Алескеров Ф. Т. (ИПУ РАН, НИУ ВШЭ)
Митичкин Е. О. (Университет Мангейма)
Швыдун С. В. (ИПУ РАН, НИУ ВШЭ)

Задача поиска

Рассмотрим конечное множество A альтернатив, оцениваемых по n критериям, т.е. каждой альтернативе x из A ставится в соответствие вектор $(x_1, \dots, x_n)$.

$$x \in A \rightarrow (x_1, \dots, x_n)$$

Задача заключается в построении преобразования φ такого, что

$$\varphi : A \rightarrow R^1$$

Надпороговый подход, основанный на идее суперпозиции

Суперпозиция заключается в последовательной композиции функций $f_1(x),\ldots,f_n(x)$ по каждому критерию, где n – число критериев.

$$f_1(f_2(x)) \neq f_2(f_1(x))$$

Надпороговый подход, основанный на идее суперпозиции, заключается в последовательной композиции надпороговых функций $f_1(x),\ldots,f_n(x)$, т.е. тех функций, значения альтернатив по которым должны быть больше или меньше назначенных порогов.

Область исследования

Обучение ранжированию – класс задач машинного обучения, суть которых состоит в автоматизированном построении ограничений для ранжирующей модели по обучающей выборке, для их последующего применения к неизвестным объектам со сходной структурой.

Возможные области применения:

- Информационный поиск;
- Рекомендательные системы и системы поддержки принятия решений
- Задачи машинного перевода
- Системы защиты от сетевых атак

Цель и задачи работы

Цель работы: разработка ранжирующих алгоритмов на основе суперпозиционного подхода для последующего применения в поисковых системах

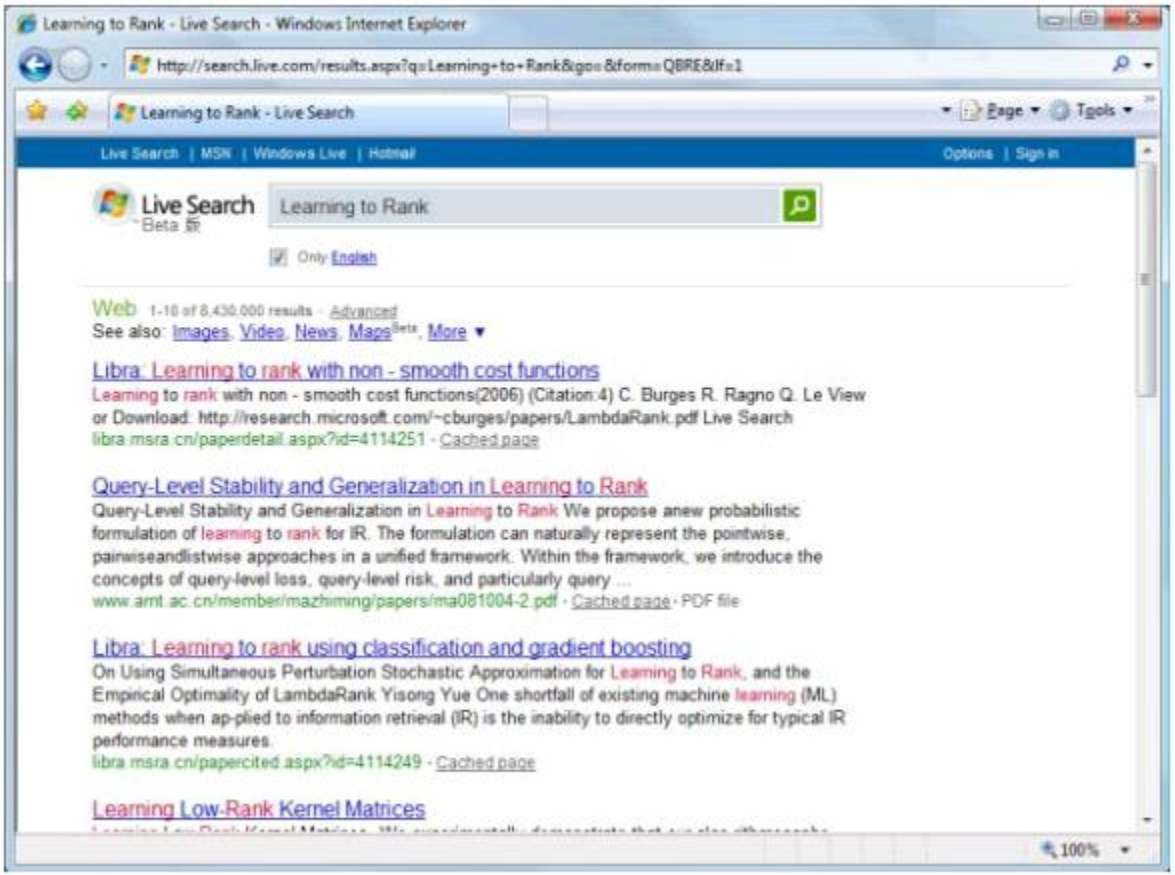
Задачи работы:

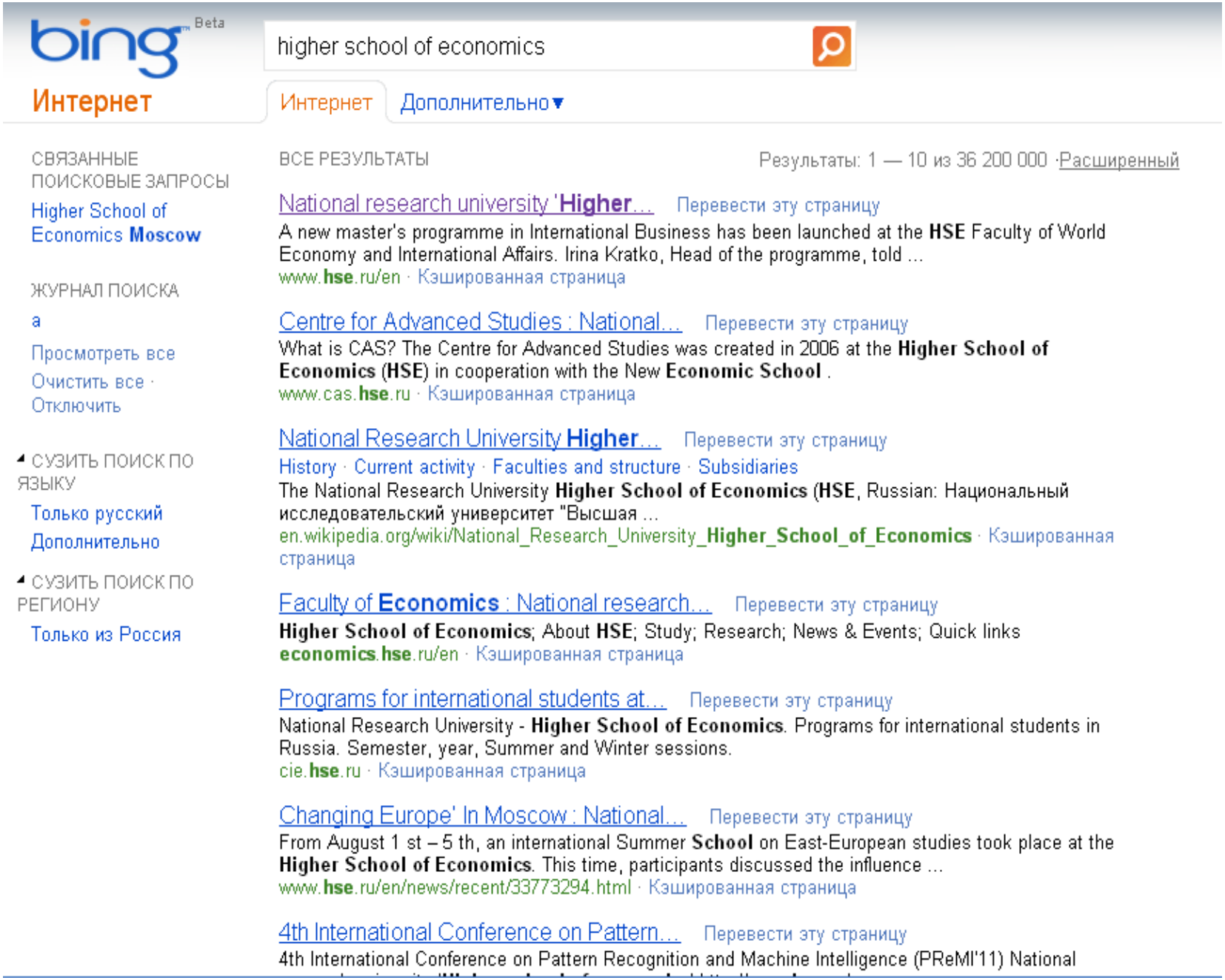
1) Выявление существенных факторов, влияющих на ранжирование
2) Кластеризация схожих запросов
3) Разработка алгоритмов суперпозиции
4) Тестирование алгоритма надпороговой суперпозиции на данных Microsoft LETOR 4.0

Актуальность работы

- Современные поисковые системы не всегда удовлетворяют информационные потребности пользователей
- Экспоненциальный рост количества информации и числа узлов в Сети, появление новых областей применения ранжирующих алгоритмов;
- Большинство существующих алгоритмов ранжирования являются коммерческой тайной

Составные элементы поисковых систем




bing Beta
higher school of economics
Интернет
Интернет
Дополнительно
СВЯЗАННЫЕ ПОИСКОВЫЕ ЗАПРОСЫ
Higher School of Economics Moscow
ЖУРНАЛ ПОИСКА
а
Просмотреть все
Очистить все · Отключить
СУЗИТЬ ПОИСК ПО ЯЗЫКУ
Только русский
Дополнительно
СУЗИТЬ ПОИСК ПО РЕГИОНУ
Только из Россия
ВСЕ РЕЗУЛЬТАТЫ
Результаты: 1 — 10 из 36 200 000 · Расширенный
National research university 'Higher... Перевести эту страницу
A new master's programme in International Business has been launched at the HSE Faculty of World Economy and International Affairs. Irina Kratko, Head of the programme, told ...
www.hse.ru/en · Кэшированная страница
Centre for Advanced Studies : National... Перевести эту страницу
What is CAS? The Centre for Advanced Studies was created in 2006 at the Higher School of Economics (HSE) in cooperation with the New Economic School .
www.cas.hse.ru · Кэшированная страница
National Research University Higher... Перевести эту страницу
History · Current activity · Faculties and structure · Subsidiaries
The National Research University Higher School of Economics (HSE, Russian: Национальный исследовательский университет "Высшая ...
en.wikipedia.org/wiki/National_Research_University_Higher_School_of_Economics · Кэшированная страница
Faculty of Economics : National research... Перевести эту страницу
Higher School of Economics; About HSE; Study; Research; News & Events; Quick links
economics.hse.ru/en · Кэшированная страница
Programs for international students at... Перевести эту страницу
National Research University - Higher School of Economics. Programs for international students in Russia. Semester, year, Summer and Winter sessions.
cie.hse.ru · Кэшированная страница
Changing Europe' In Moscow : National... Перевести эту страницу
From August 1 st – 5 th, an international Summer School on East-European studies took place at the Higher School of Economics. This time, participants discussed the influence ...
www.hse.ru/en/news/recent/33773294.html · Кэшированная страница
4th International Conference on Pattern... Перевести эту страницу
4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11) National

Классификация факторов

1. Факторы, *не зависящие от запроса.* Непосредственно описывают статические свойства (характеристики) страницы.
2. Факторы, *зависящие от только запроса.*
3. Факторы, *зависящие как от запроса, так и от самого документа*. Представляют собой динамические характеристики страницы.

Частота слова (Term Frequency)

Фактор, показывающий, сколько раз *i-ое* слово запроса встречаются в некоторой части документа.

$$TF_i = \frac{n_i}{\sum_{k=1}^{n} n_k},$$

Каждый документ представляется в виде вектора ($TF_1, \ldots, TF_n$). Таким образом,

$$TF = cos(q, d_i)$$

Недостаток

Значение *TF* будет высоким для наиболее употребительных слов.

Обратная частота документа (IDF)

Обратная частота документа (IDF) инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$IDF(t_i, d) = log \frac{|D|}{|\{d \in D : t_i \in d\}|}$$, где

|D| — общее число документов в коллекции, $|\{d \in D : t_i \in d\}|$ — число документов, в которых есть слово t_i введенного запроса.

$$IDF = \sum_{i=1}^{n} IDF(t_i, D)$$

TF-IDF и BM25

TF-IDF статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов. Величина TF-IDF представляет собой произведение величин TF и IDF.

$$TF{-}IDF_i = TF_i * IDF_i$$

Фактор **BM25** является аналогом TF-IDF

$$BM25(P,Q) = \sum_{i=1}^{n} IDF(t_i,d) * \frac{TF(q_i,P) * (k_1+1)}{TF(q_i,P) + k_1 * (1 - b + b * \frac{|P|}{avgdl})}$$

Где *avgdl* — средняя длина документа в коллекции.k_1 и b — свободные коэффициенты, обычно их выбирают как k_1 = 2.0 и b = 0.75.

Булева модель

Может быть поделена на 2 этапа:

1. Представление запроса в виде логического выражения
2. Вычисление соответствия некоторого документа D_i составленному логическому выражению.

Пример:

Пользователь хочет получить некоторую информацию о информационном или текстовом поиске, однако эта информацию не должна включать слово “теория”.

В результате запрос Q представляется в виде выражения:

((текстовый или информационный) и поиск и (не)теория)

Булева модель

В коллекции существуют документы со следующей информацией:

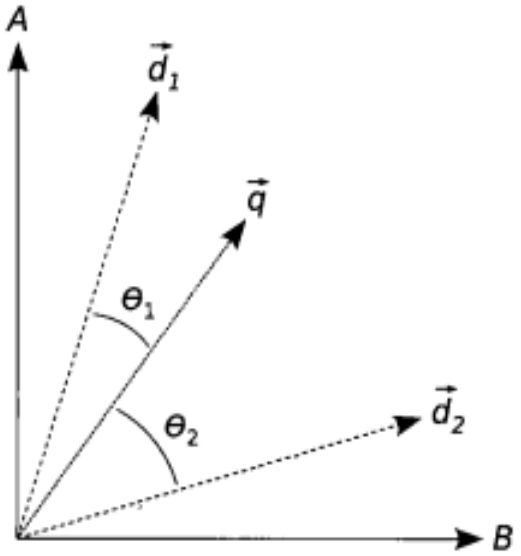
- *D1 = “Информационный поиск”*
- *D2 = “Теория информации”*
- *D3 = “Современный информационный поиск: Теория и практика”*
- *D4 = “Сжатие текста”*

Очевидно, что документ D1 единственный удовлетворяет введенному запросу. Стоит отметить, что, если бы пользователь не указал “*(не)теория*”, то релевантным был бы также документ D3.

Векторная модель

Векторная модель (vector space model) — представление коллекции документов векторами из одного общего для всей коллекции векторного пространства. Задействует веса TF, TF-IDF, может использовать Boolean model.

Значение векторной модели – есть Значение близости векторов запроса и документа ($cos(q, d_i)$).



Языковая модель информационного поиска

Языковая модель информационного представляет собой модель, показывающая вероятность нахождения некоторой последовательности слов в документе.

1. Запрос Q состоит из m слов $q_1,\ldots,q_m$.
2. Происходит вычисление вероятности нахождения слова q_i в документе D

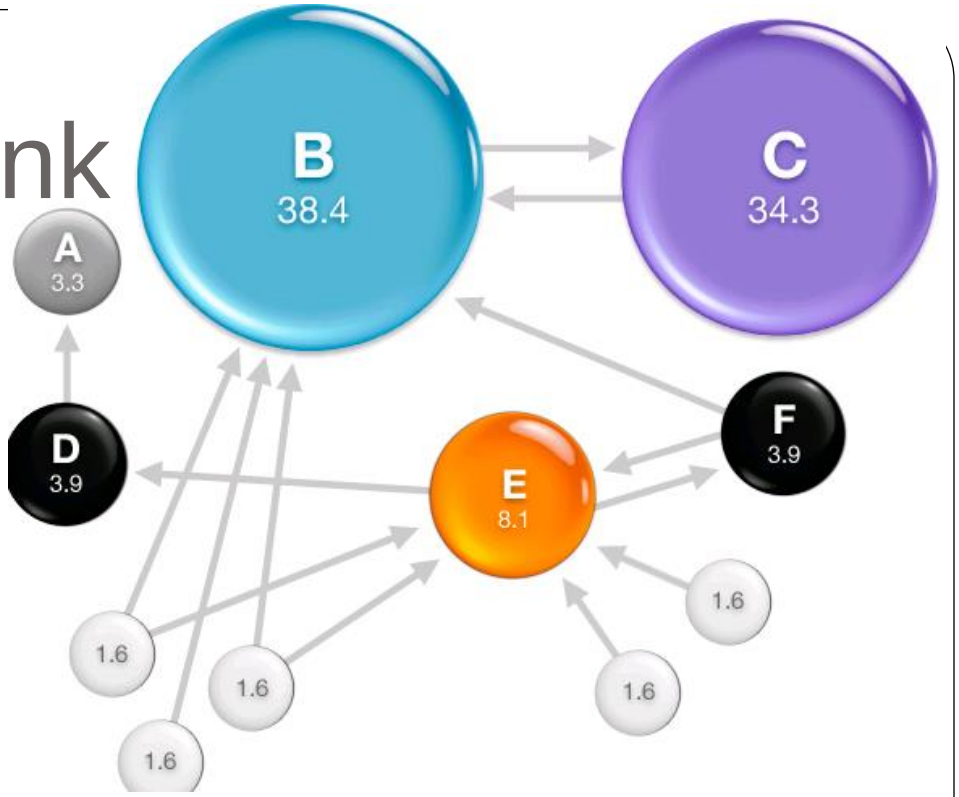
- *Абсолютная модель*

$$P(q_1, \ldots, q_m) = \prod_{i=1}^{m} P(q_i).$$

- *Байесовская модель.*

$$P(q_1, \ldots, q_m) = P(q_1) * P(q_2|q_1) * \cdots * P(q_m|q_{m-1}).$$

PageRank, SiteRank

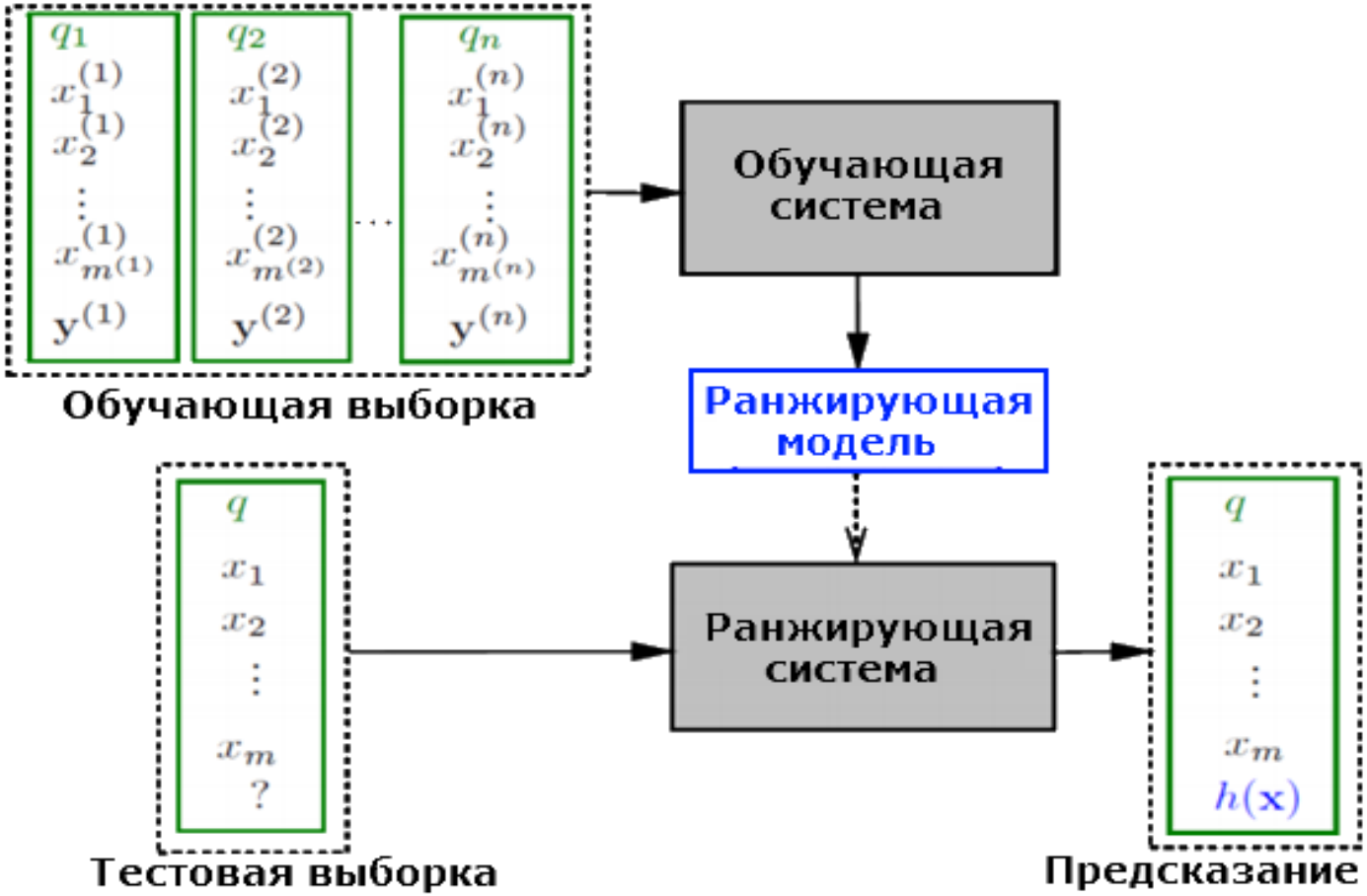


- **PageRank** – один из алгоритмов ссылочного ранжирования, отображающий вес страницы. Численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов.
- Фактор **SiteRank** является аналогом PageRank, но показывает вес не отдельной страницы, а всего сайта, на котором она находится.

Факторы пользовательского поведения

- **Query-URL click count** - сколько раз пользователи открывали (“кликали”) документ в запросе.
- **URL click count -** сколько раз пользователи вводили URL-адрес документа.
- **URL dwell time -** сколько времени прошло с того момента, как пользователь открыл документ, до того момента, как пользователь его покинул.

Принцип обучения ранжированию



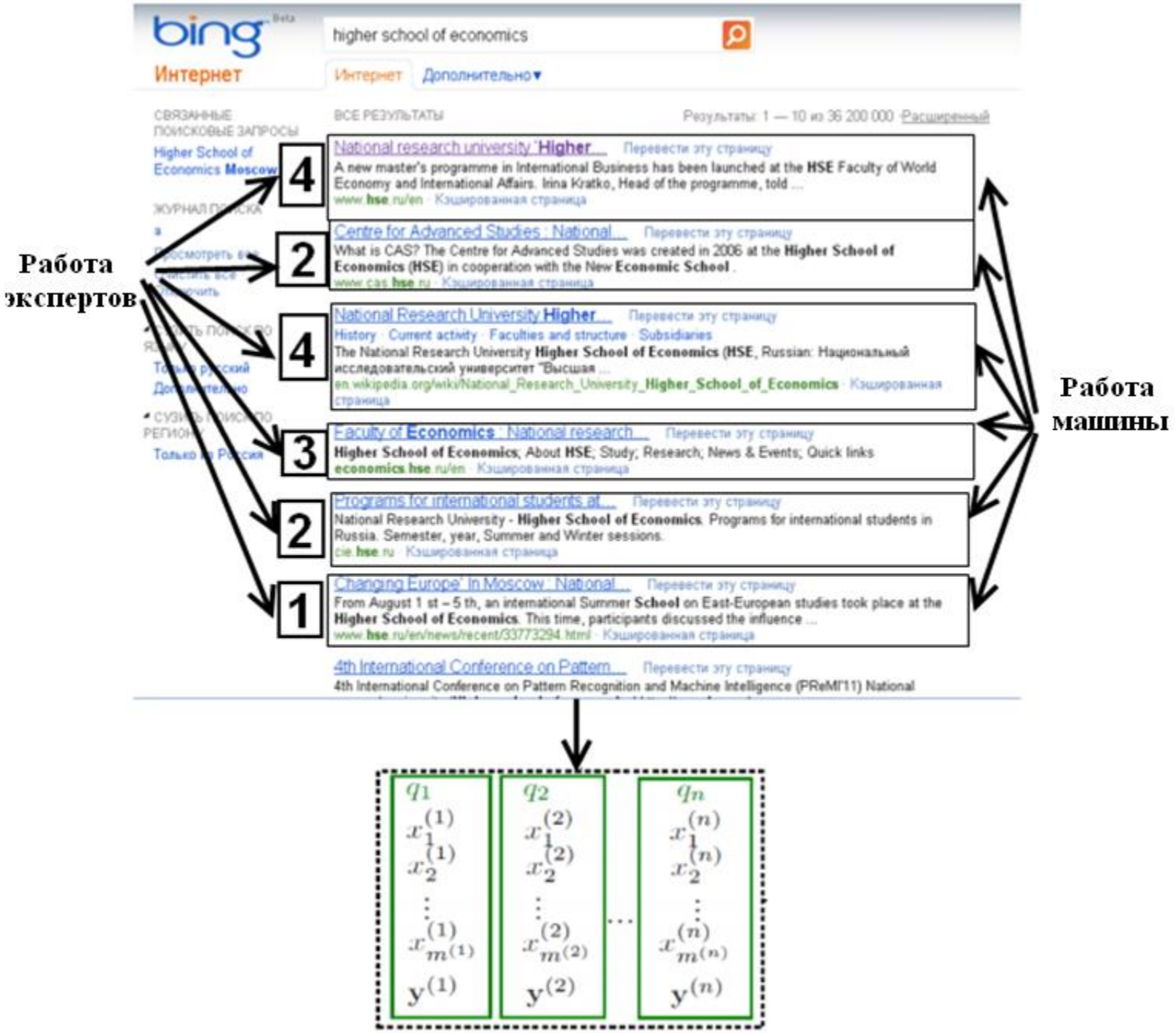
Значение релевантности документа

– Бинарная: релевантный vs. нереллевантный

– По категориям:

Идеально > Отлично > Хорошо >

> Удовлетворительно > Плохо



Обучающая выборка

Структура обучающей выборки

Документ D	Оценка релевантности R	Факторы, характеризующие документ F
D_1	$R(D_1)$	$\{f_1(D_1),\ldots, f_n(D_1)\}$
D_2	$R(D_2)$	$\{f_1(D_2),\ldots, f_n(D_2)\}$
…		
D_N	$R(D_N)$	$\{f_1(D_N),\ldots, f_N(D_N)\}$

Наборы данных:

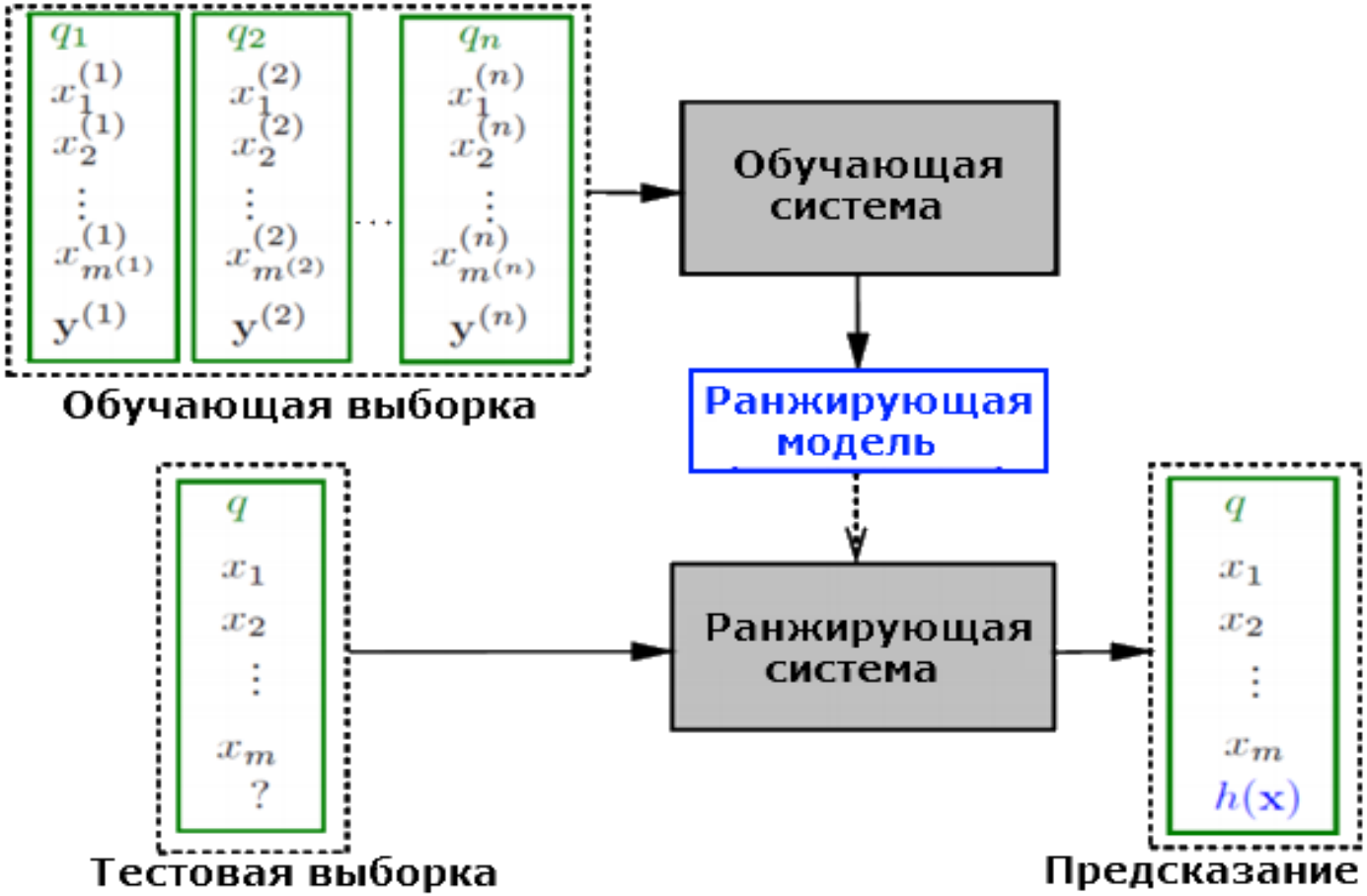
- LETOR (Microsoft Research Department)
- Яндекс Интернет-Математика
- Yahoo Learning to rank challenge

LETOR 4.0 dataset

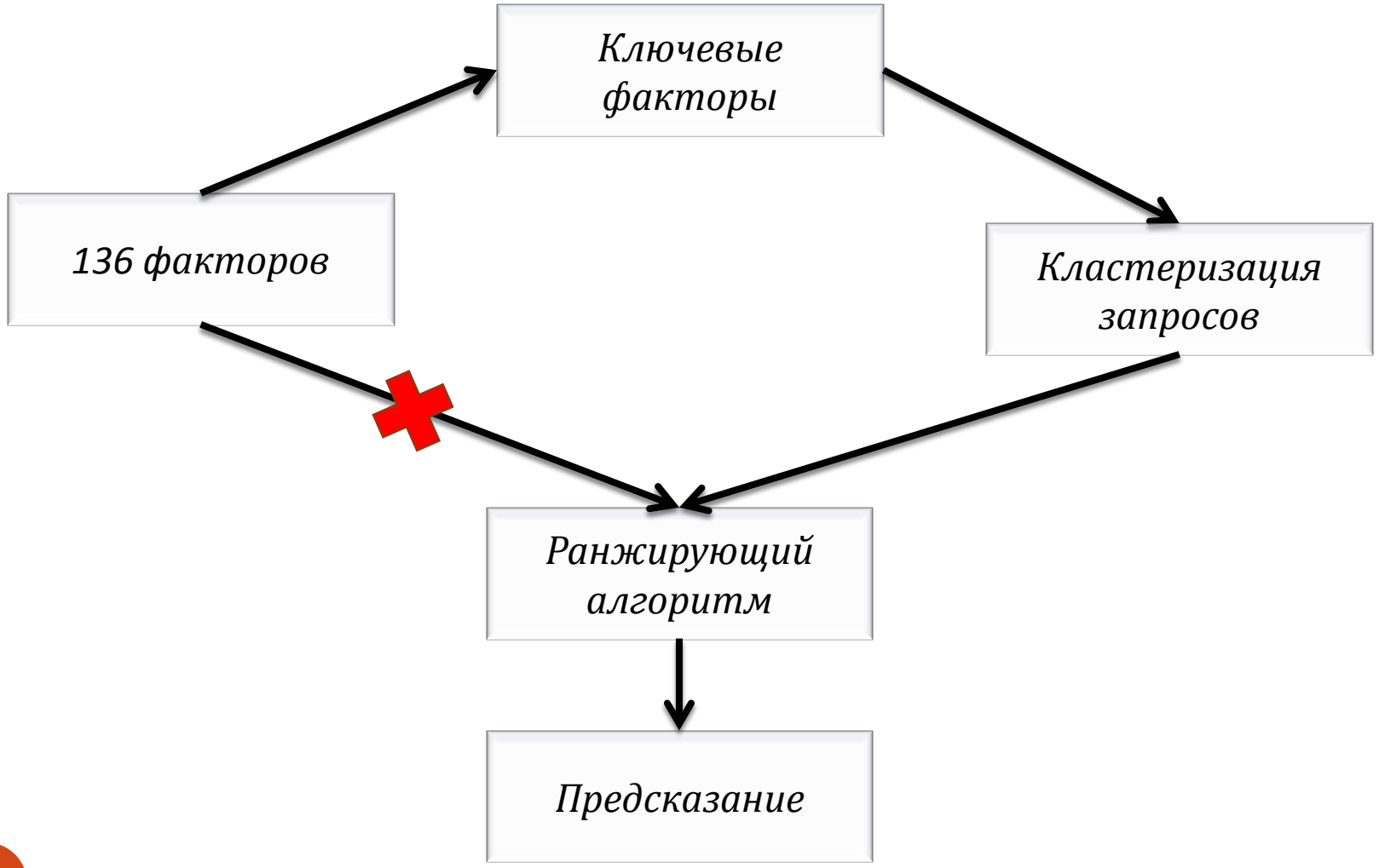
- 10000 запросов, более 1175000 документов;
- Обучающая, тестирующая и проверочная выборки (3:1:1);
- Каждый документ имеет целочисленную оценку релевантности от 0 до 4;
- 136 факторов для оценки документов и веб-страниц;

Обучающая выборка	Проверочная выборка	Тестовая выборка
{S1,S2,S3}	S4	S5
{S2,S3,S4}	S5	S1
{S3,S4,S5}	S1	S2
{S4,S5,S1}	S2	S3
{S5,S1,S2}	S3	S4

Принцип обучения ранжированию



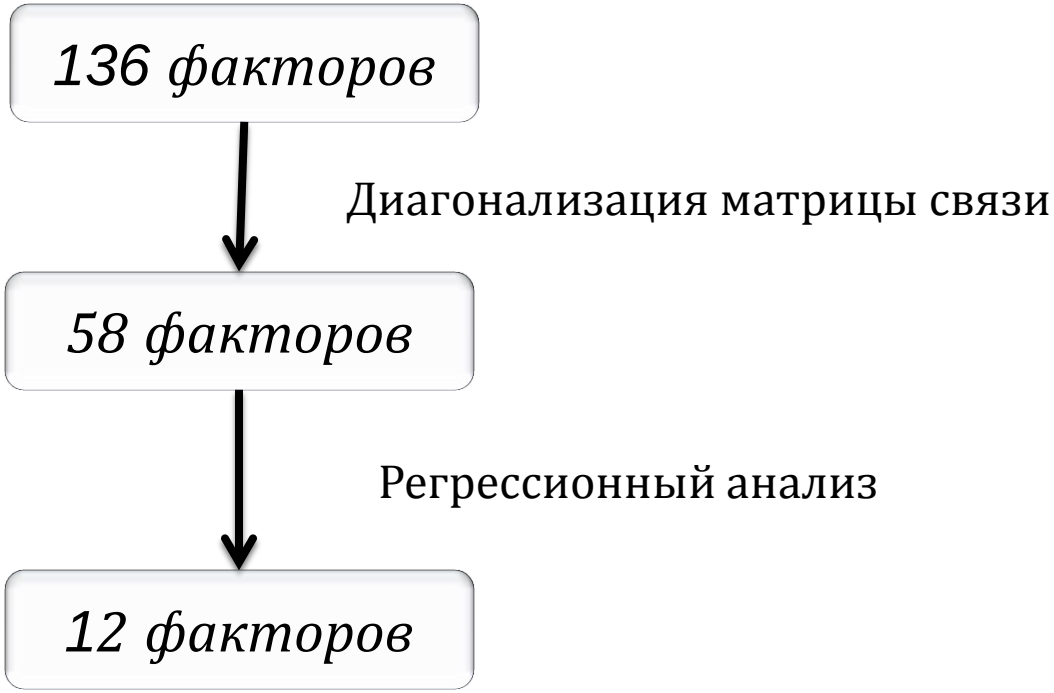
Ранжирующая модель



Анализ данных

Шаг 1. Преобразование данных
(логарифмирование, нормировка)

Шаг 2. Сокращение размерности



Список факторов

Номер фактора	Название фактора	Часть документа
13	stream length (Длина потока)	Заголовок
15	stream length (Длина потока)	Весь документ
16	IDF (Обратная частота документа)	Тело документа
23	sum of TF (Сумма частот слов)	Заголовок
48	sum of stream length normalized TF (Сумма длин потока нормированная на частоту слова)	Заголовок
57	max of stream length normalized TF (Максимальное значение длины потока нормированное на частоту слова)	Якорь (содержимое тега гиперссылки)
118	LMIR.DIR (Языковая модель DIR)	Заголовок
122	LMIR.JM (Языковая модель JM)	Якорь
126	Number of slash in URL (Число слешей в URL-адресе документа)	
131	SiteRank (Рейтинг сайта)	
132	QualityScore (Показатель качества)	
133	QualityScore2 (Показатель качества 2)	

Туннельная кластеризация

Для каждого кластера выбирается эталонный объект.

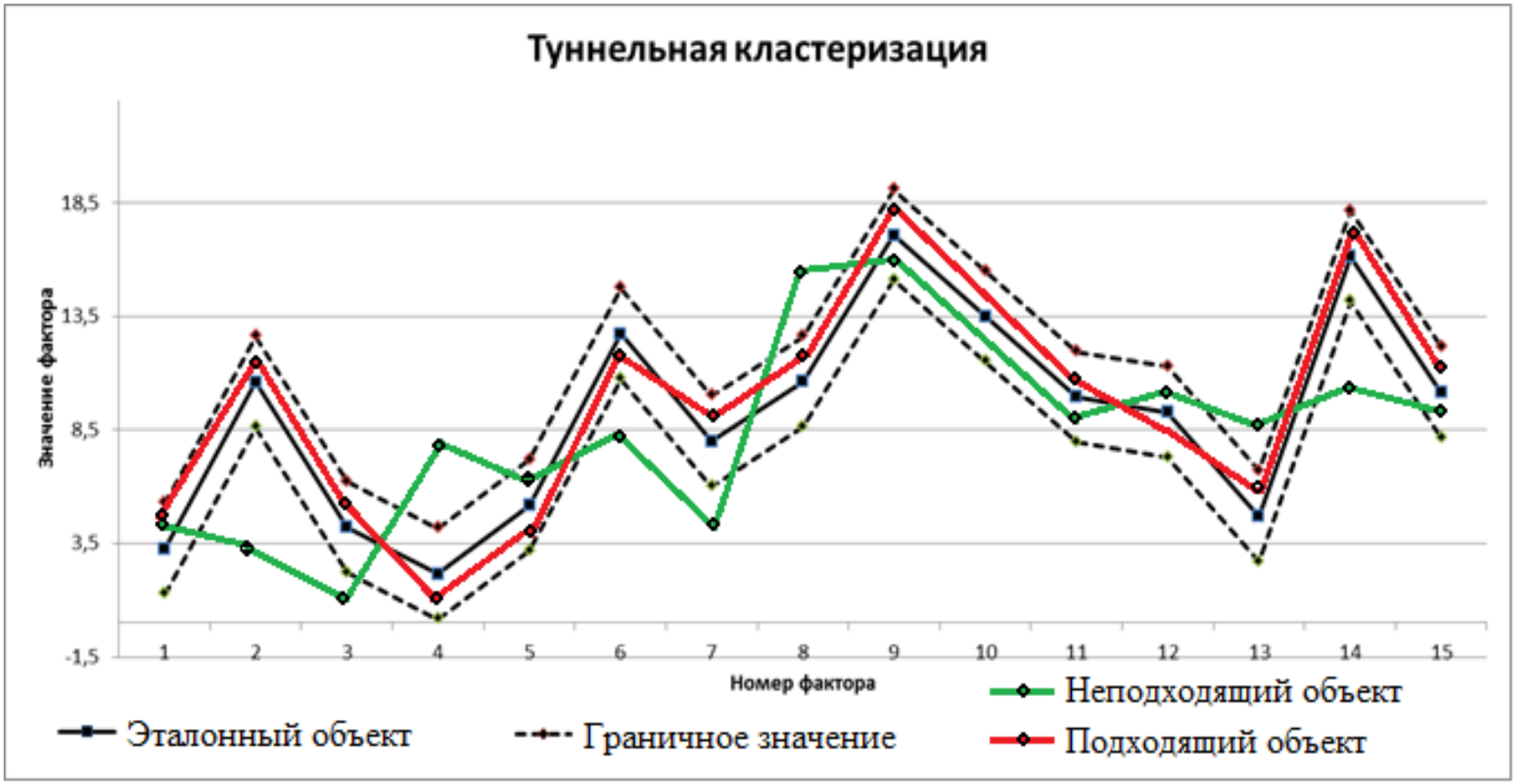
Правило кластеризации: $\forall k \epsilon M: \left|\alpha_{ik} - \alpha_{jk}\right| \leq \varepsilon \Rightarrow$ объекты i и j принадлежат одному кластеру.

Где:

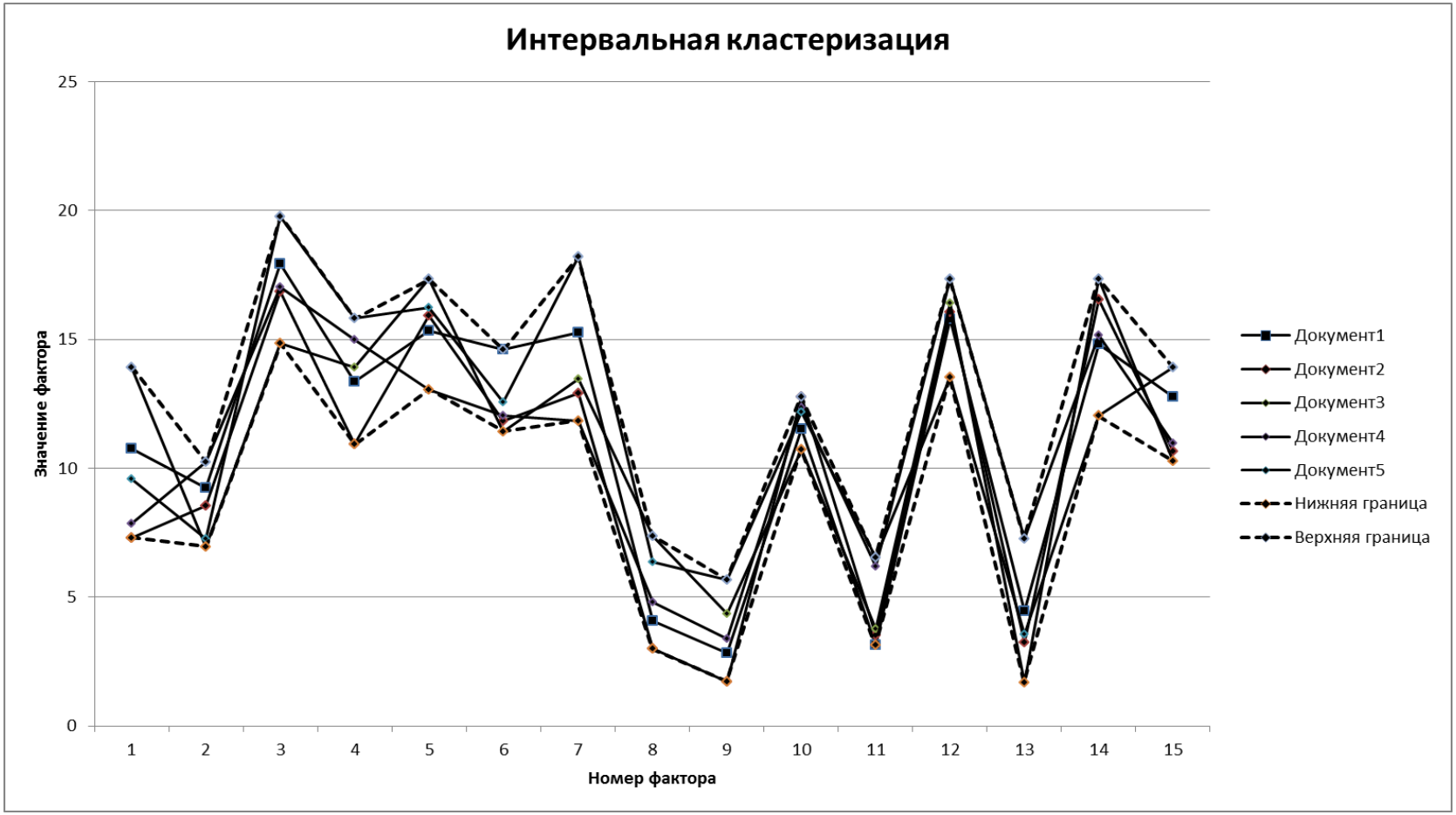
- k – фактор, M – множество факторов,
- i – эталонный объект,
- j - любой другой объект.

Если разность значений факторов эталонного и текущего объектов меньше ширины ε-полосы, то объекты принадлежат одной и той же группе.

Туннельная кластеризация



Интервальная кластеризация



Классификация подходов к ранжированию

1. *Поточечный подход* - каждой странице выставляется некоторое значение релевантности запросу, по которому происходит сортировка всех веб-страниц.
2. *Попарный подход* - попарное сравнение страниц между собой по степени релевантности.
3. *Списковый подход* – сравнение наборов веб-страниц

Алгоритм суперпозиции надпороговых процедур

Все *объекты* представляются как *точки в N-мерном пространстве.*

Результаты применения алгоритма

Название теста	Процент запросов с точностью ранжирования				Средняя точность ранжирования
	<40%	40%-60%	60%-90%	>90%	
Тест 1	35%	13%	12%	40%	65%
Тест 2	10%	8%	17%	65%	84%

Тест 1: отличить релевантные страницы (значение релевантности “3” и “4”) от нерелевантных страниц

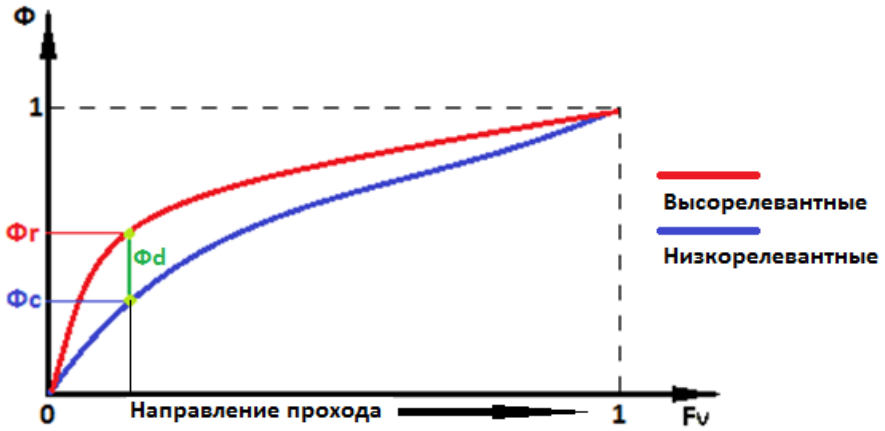
Тест 2: отличить высоко релевантные страницы от всех остальных.

Сложность алгоритма – $O(n \cdot k \cdot \log(k))$, где n – количество объектов, k – количество факторов.

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ

Основная идея алгоритма – нахождение эффективной границы *b* для каждого фактора таким образом, что объекты из множества P могут быть распределены по двум множествам: множеству релевантных объектов Q и множеством остальных объектов T.

Шаг 1. Решающее правило.



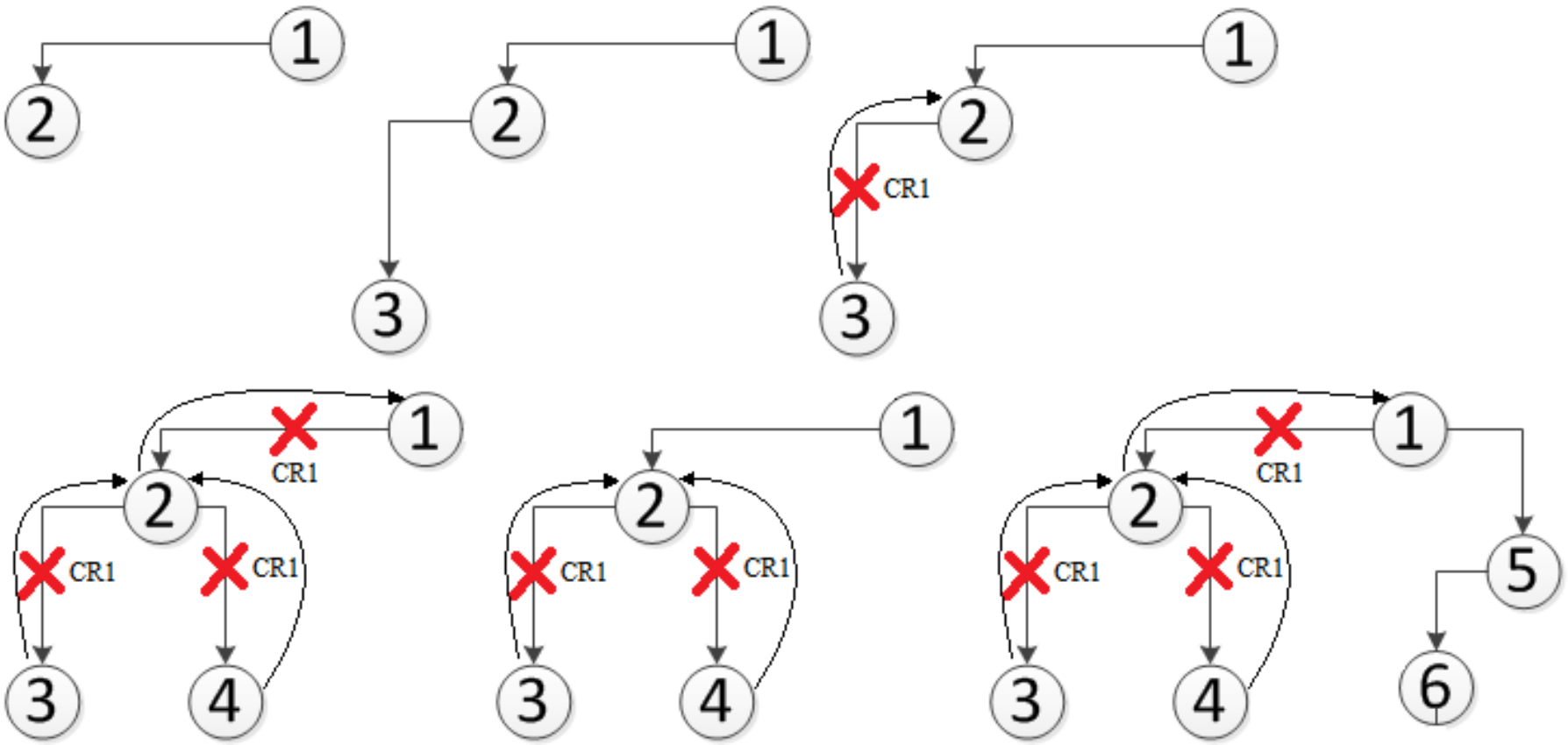
$b = Fv: \Phi_d \rightarrow max$

Шаг 1. Эффективные границы *b* вычисляются для всех факторов из F, после чего они сортируются по возрастанию значения Φ_d

Шаг 2. Фильтрация данных.

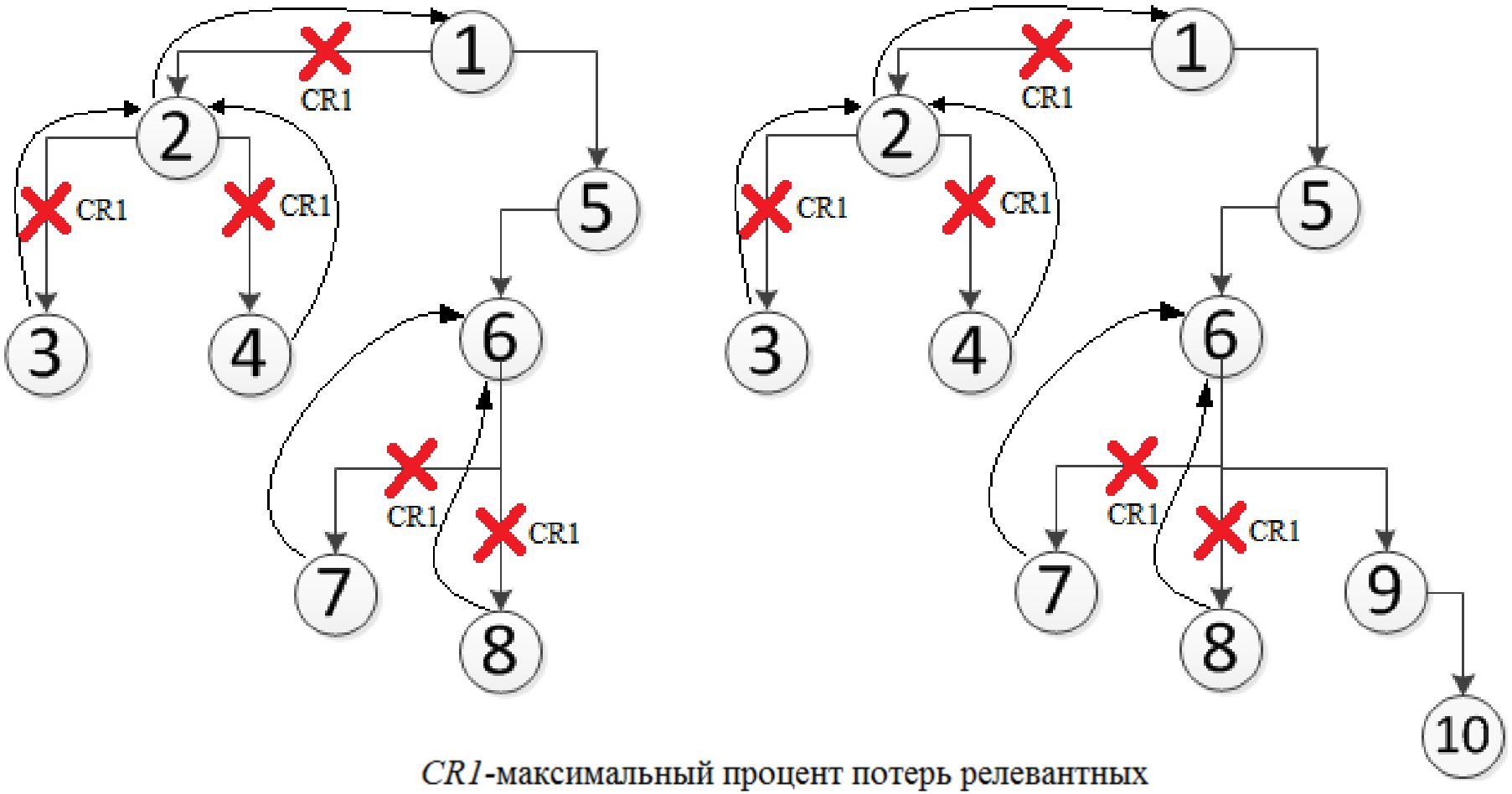
Шаг 3. Повторение шага 1,2

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ



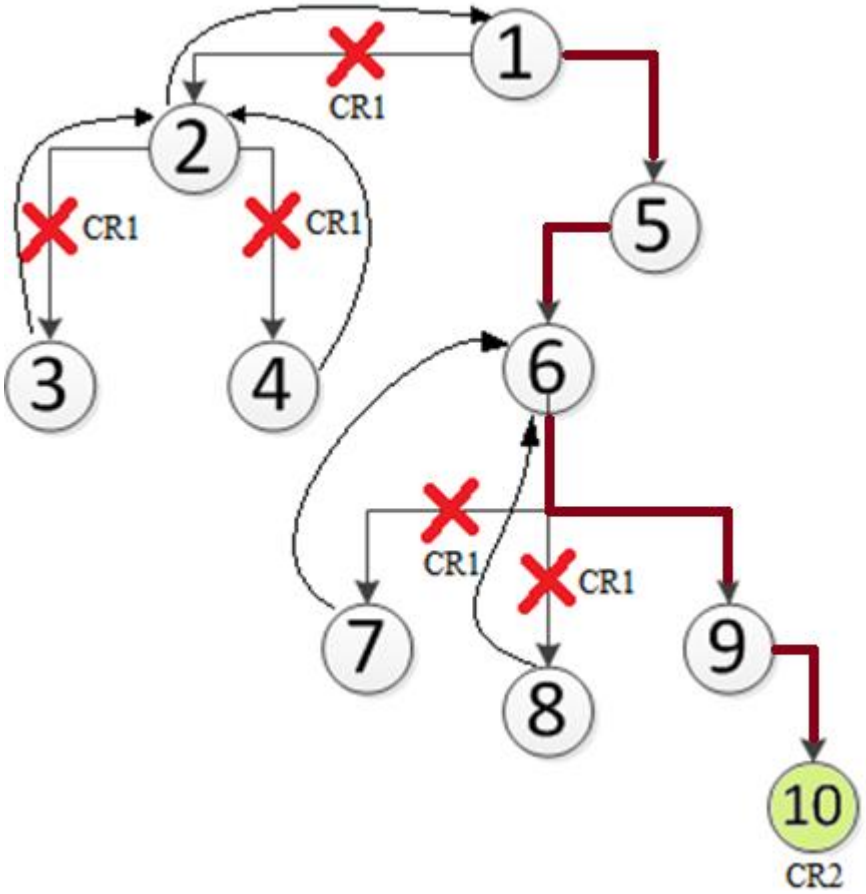
CR1-максимальный процент потерь релевантных объектов от их первоначального числа

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ



CR1-максимальный процент потерь релевантных объектов от их первоначального числа

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ



CR1-максимальный процент потерь релевантных объектов от их первоначального числа

CR2 – процент релевантных объектов среди остальных после проведения очередного шага процедуры фильтрации

Тестирование алгоритма

Статистика по LETOR Dataset

Всего объектов	235000	Объектов в тестовой выборке	16627
Число релевантных объектов	6012	Число релевантных объектов в выборке	1404
Процент релевантных объектов*	2,56	Процент релевантных объектов*	8,44

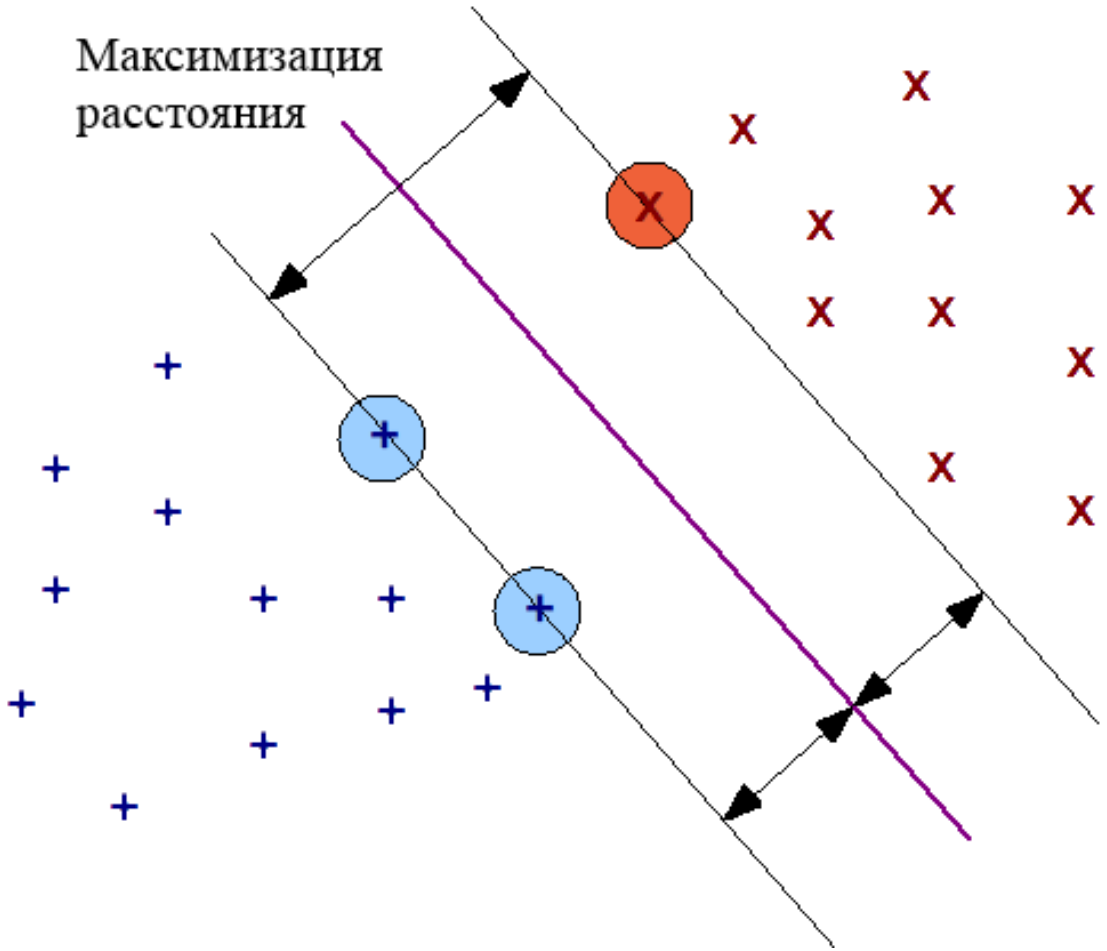
Сравнение результатов до и после работы алгоритма

Минимальный процент релевантных объектов до фильтрации	*7,12*
Максимальный процент релевантных объектов до фильтрации	*10,2*
Средний процент релевантных объектов до фильтрации	**8,44**
Минимальный процент релевантных объектов после фильтрации	*26*
Максимальный процент релевантных объектов после фильтрации	*75,1*
Средний процент релевантных объектов после фильтрации	**41,3**

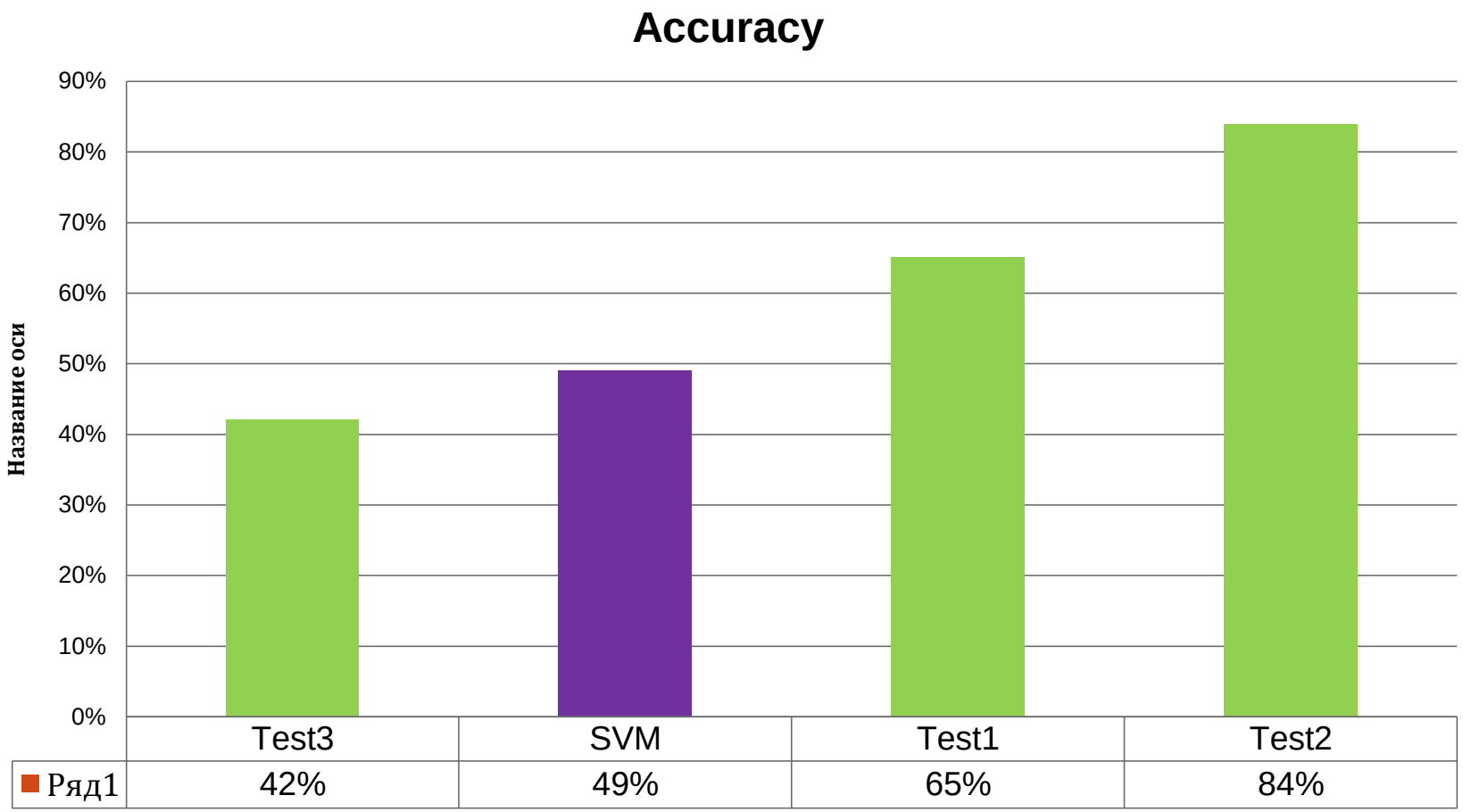
*В данном случае релевантными были признаны объекты, имеющие релевантность 3 и 4.

Теоретическая сложность алгоритма – $O\,(k * (n + k))$, где n – число объектов, k – число факторов

Метод опорных векторов (SVM)



Результаты


Accuracy
Название оси
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Test3
SVM
Test1
Test2
Ряд1
42%
49%
65%
84%

Совершенствование алгоритма

Причина	Решение
Человеческий фактор	Переход к вероятностной модели
Недостаточное количество факторов	Добавление неиспользованных факторов

Результаты

- Проведена процедура выявления существенных факторов
- Проведена кластеризация запросов
- Разработан алгоритм определения релевантности страницы
- Разработано программное обеспечение, реализующее все стадии выполнения исследования

Выводы

- Экспериментально подтверждена эффективность рассматриваемого подхода, определены несколько направлений развития алгоритма;
- Результаты, полученные в результате работы алгоритма, не являются окончательными, но уже могут быть применены в решении практических задач;
- Разработана серия методов, способных серьезно упростить и ускорить процесс обучения ранжированию и автоматизации информационного поиска.

Спасибо за внимание!

Приложение 1. Регрессионный анализ

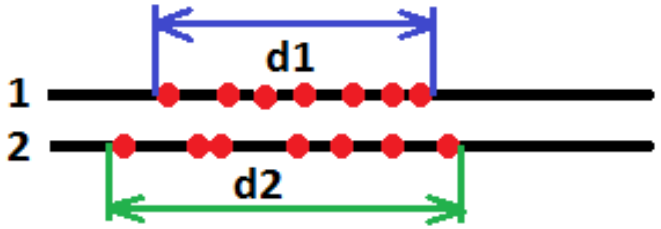
Номер фактора	Коэффициент регрессии	Стандартная ошибка	t - статистика	Название фактора	Часть документа
13	3,3	0,563	6,500	Длина потока	Заголовок
15	-0,4	0,043	-12,314	Длина потока	Весь документ
16	0,3	0,093	3,285	Мера IDF	Тело документа
23	-0,8	0,157	-6,490	Сумма частот слова	Заголовок
48	1,7	0,167	10,205	Сумма длин потока нормированная на частоту слова	Заголовок
57	-0,1	0,035	-6,264	Максимальное значение длины потока нормированное на частоту слова	Якорь
118	-10,3	2,313	-4,779	Языковая модель информационного поиска с помощью распределения Дирихле	Заголовок
122	12,9	1,925	6,359	Языковая модель информационного поиска с помощью сглаживания Джелинека-Мерсера	Якорь
126	-1,9	0,147	-13,082	Число слешей в URL-адресе документа	
131	0,2	0,028	8,139	Рейтинг сайта	
132	0,1	0,030	3,173	Показатель качества	
133	-0,1	0,036	-5,182	Показатель качества 2	

Алгоритм суперпозиции надпороговых процедур

Все *объекты* представляются как *точки в N-мерном пространстве.*

Алгоритм состоит из 2 шагов.

Шаг 1. Поиск наиболее “кучного показателя”



Алгоритм суперпозиции надпороговых процедур

Шаг 2. Процедура последовательного исключения объектов с помощью применения надпороговых процедур. Построение дерева исключений.