Алгоритмы ранжирования и их применение в задачах информационного поиска

Алескеров Ф.Т.

Митичкин Е.О.

Швыдун С.В.

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

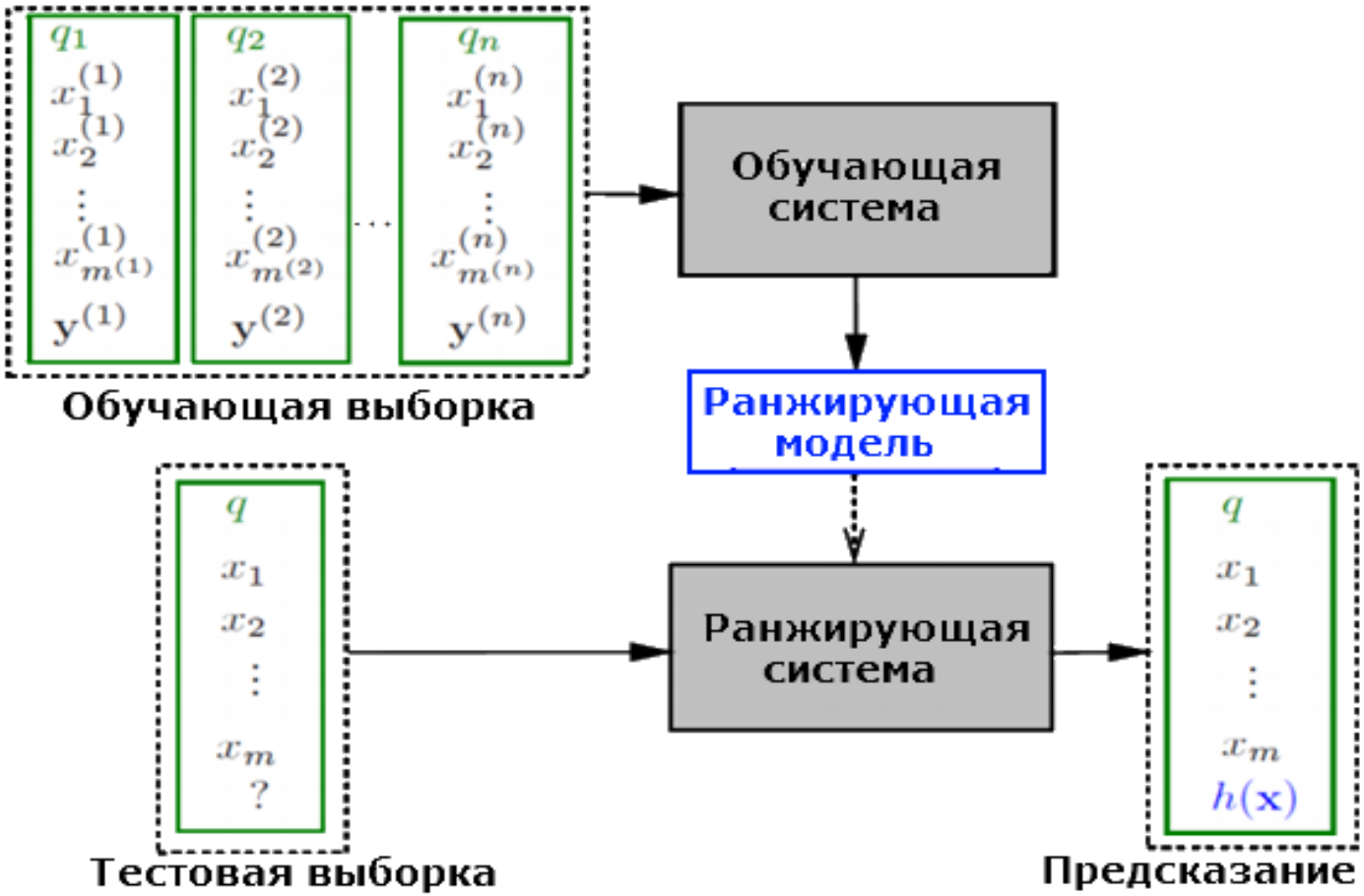
Область исследования

Обучение ранжированию (Learning to rank) – 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

Принцип работы поисковых систем

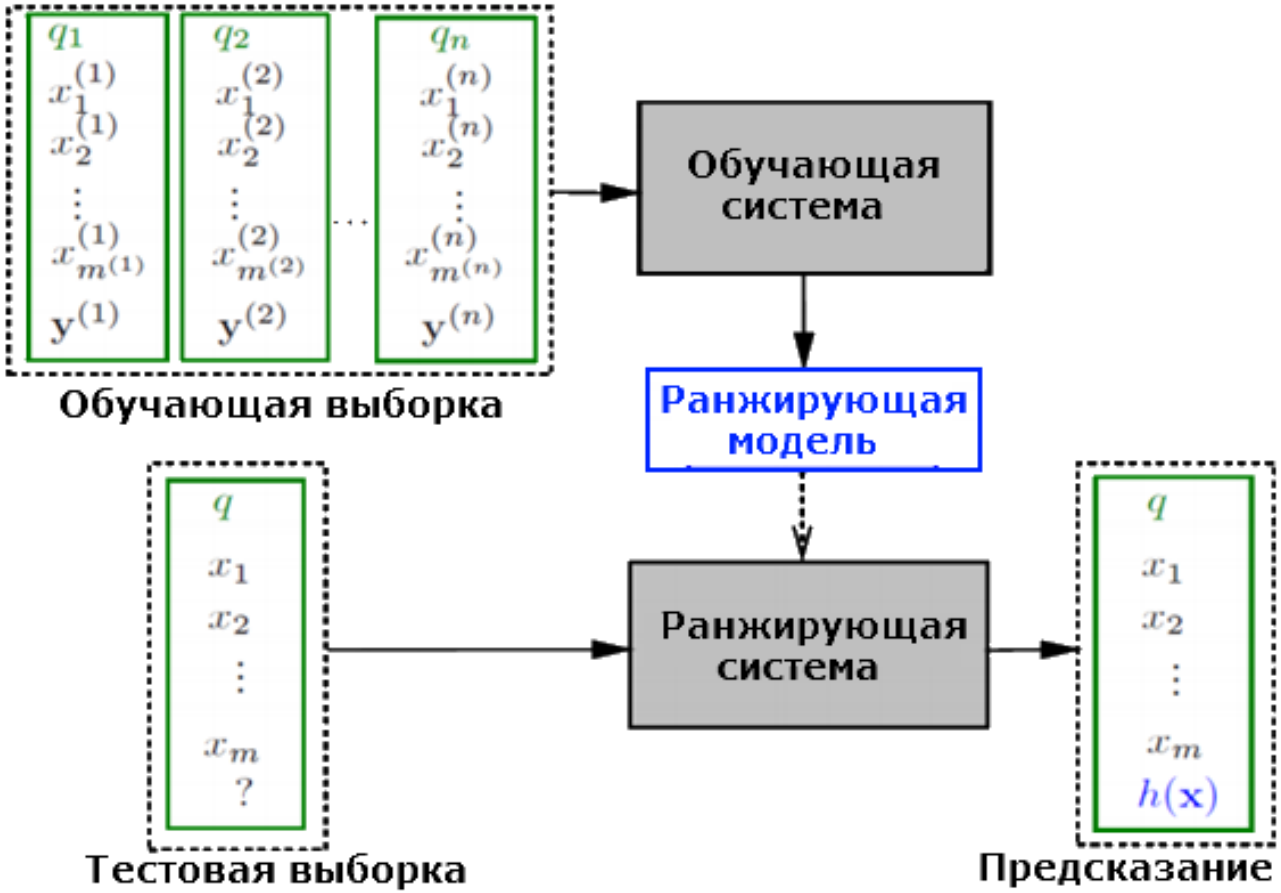


Структура обучающей выборки

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	$R(D_N)$	$\{f_1(D_N),\ldots, f_N(D_N)\}$

$$\{D_i\} \Rightarrow \{R, F\}$$

Принцип работы поисковых систем

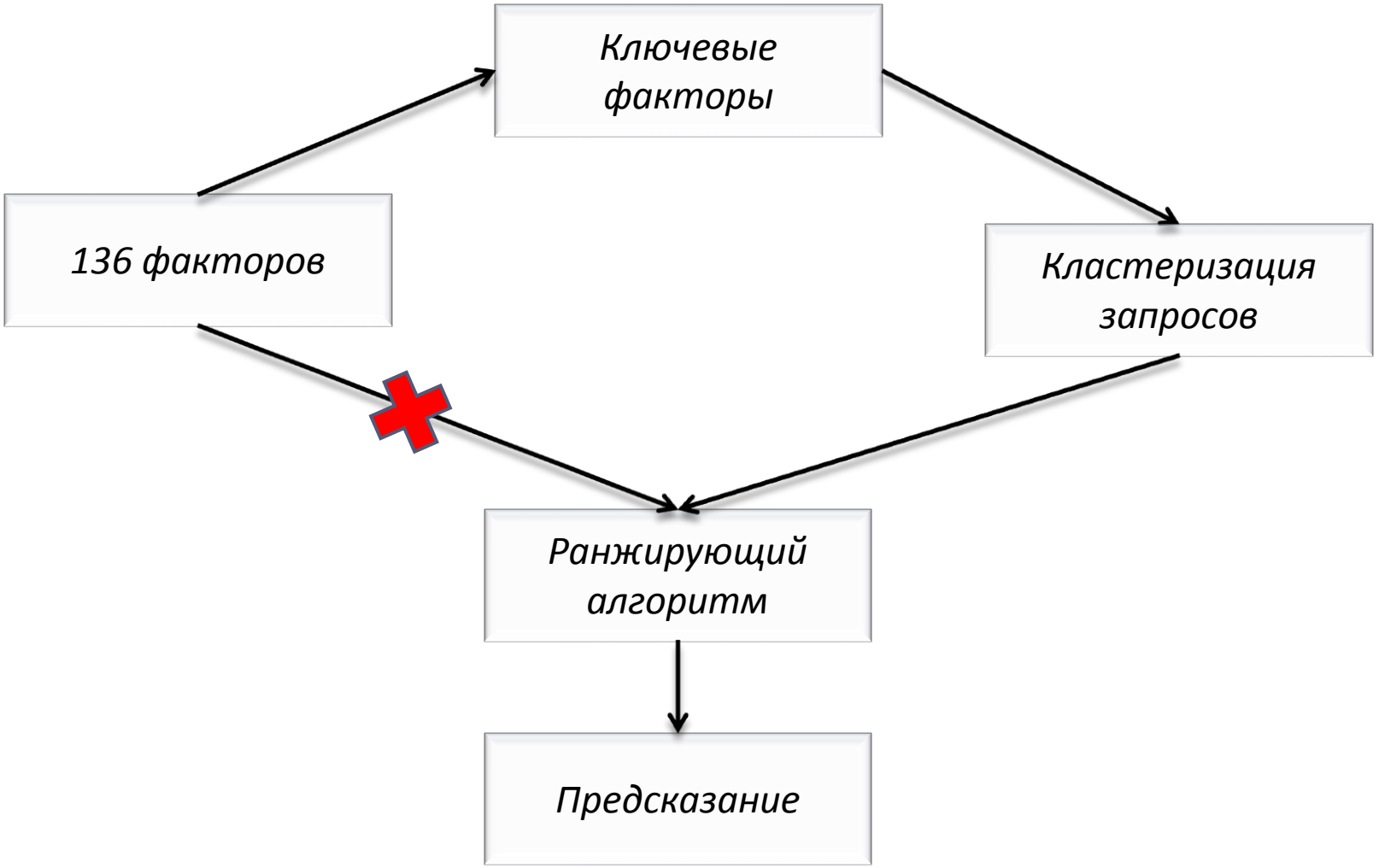


Данные LETOR 4.0

- 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;
- 3 типа факторов: зависящие от запроса, описывающие объект поиска и агрегирующие.

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

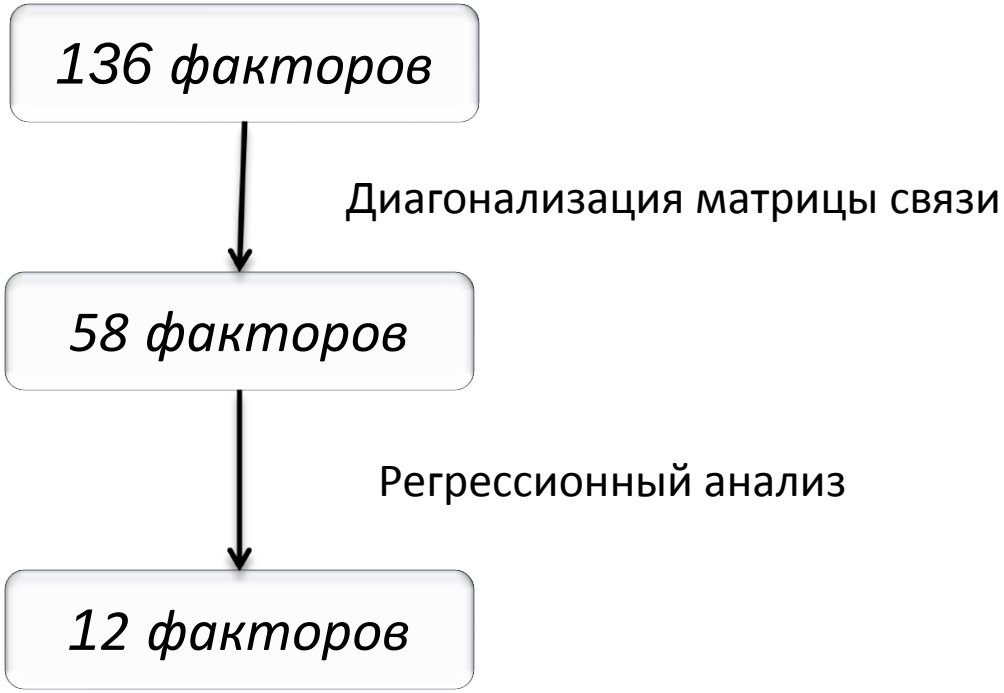
Ранжирующая модель



Анализ данных

Шаг 1. Преобразование данных
(логарифмирование, нормировка)

Шаг 2. Сокращение размерности



Список факторов

Номер фактора	Название фактора	Часть документа
13	stream length	title
15	stream length	Whole document
16	IDF	body
23	sum of term frequency	title
48	sum of stream length normalized term frequency	title
57	max of stream length normalized term frequency	anchor
118	LMIR.DIR	title
122	LMIR.JM	anchor
126	Number of slash in URL	
131	SiteRank	
132	QualityScore	
133	QualityScore2	

Туннельная кластеризация и кластерный анализ

Предпосылки для проведения кластеризации:

- Данные имеют разнородную структуру;
- Релевантные объекты могут описываться разным набором факторов в зависимости от предметной области;

Вначале для каждого кластера выбирается эталонный объект.

Правило кластеризации:

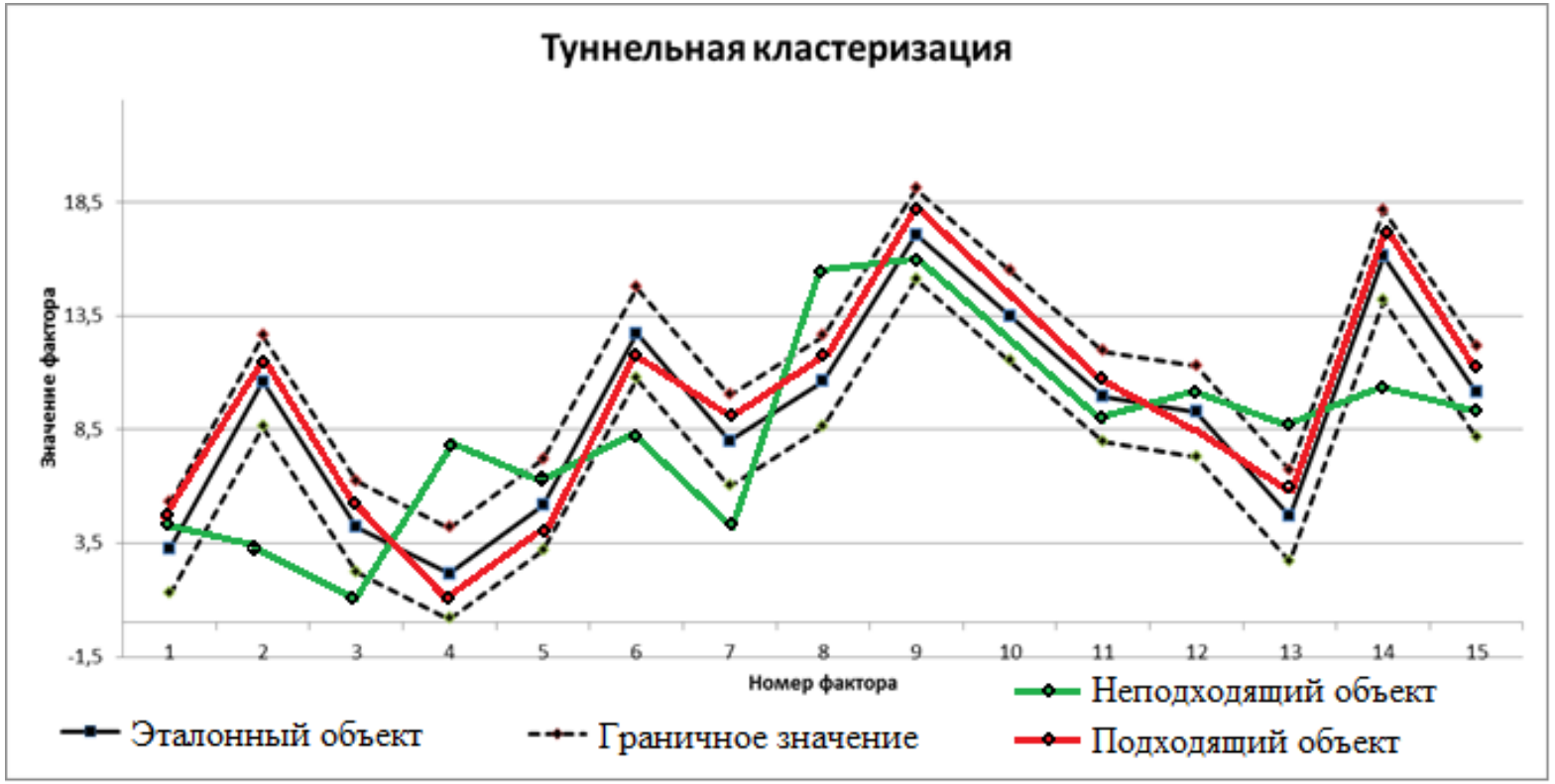
$$|\qquad|\qquad.$$

Где:

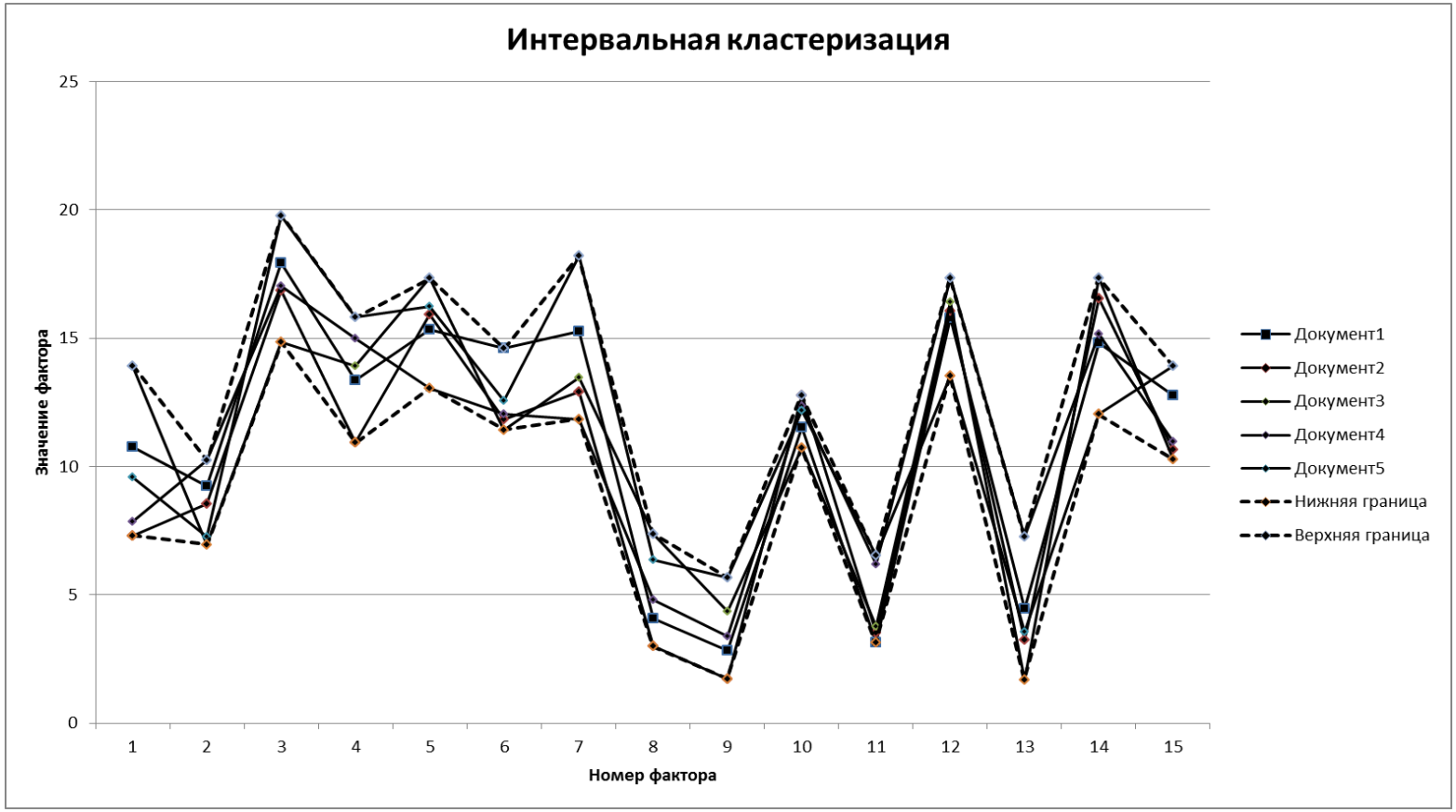
- k – фактор, M – множество факторов,
- *i* – эталонный объект,
- *j* - любой другой объект.

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

Туннельная кластеризация



Интервальная кластеризация



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 – $O(n*k^2)$, где n – количество объектов, k – количество факторов.

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ

Релевантным называется объект, релевантность которого равна 3 или 4. Эти объекты образуют множество релевантных объектов Q.

Множество P – это множество, состоящее из всех объектов запроса, то есть *Q* P;

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ

F – множество факторов, N_f – мощность множества F;

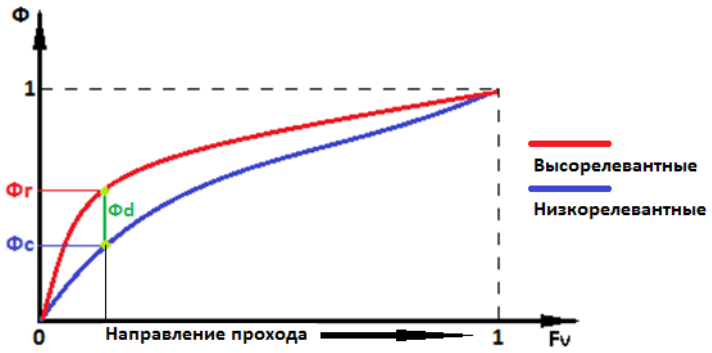
Z – дерево применения факторов, определяющее последовательность применения фильтров и их параметры;

R – множество коэффициентов регрессии, подсчитанной на шаге обработки данных;

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ, шаг 1

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

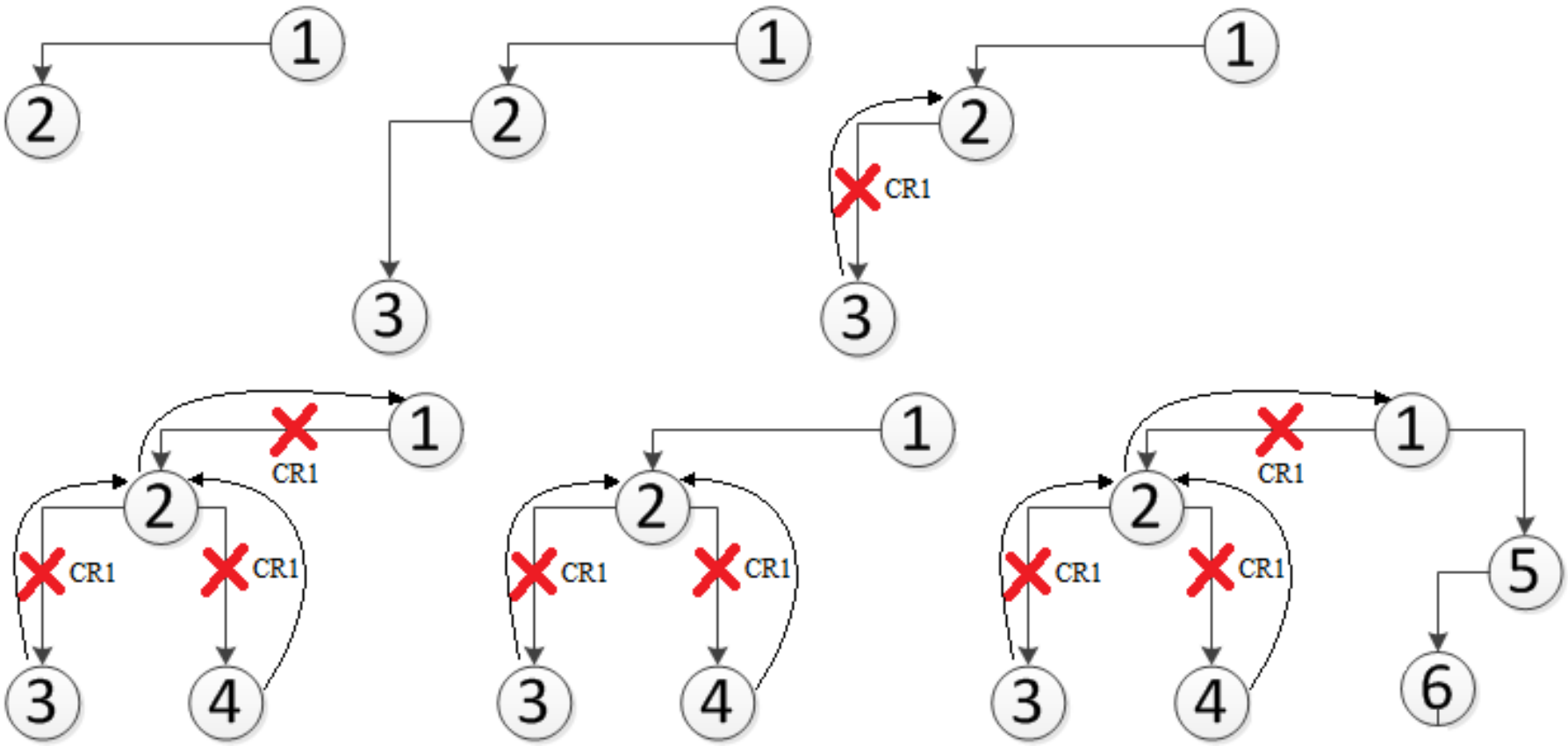
Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ, шаг 2

Шаг 2. Фильтрация данных

После того как эффективная граница *b* вычислена и определен соответствующий фактор *f,* инициируется процедура фильтрации. Формируется множество RES.

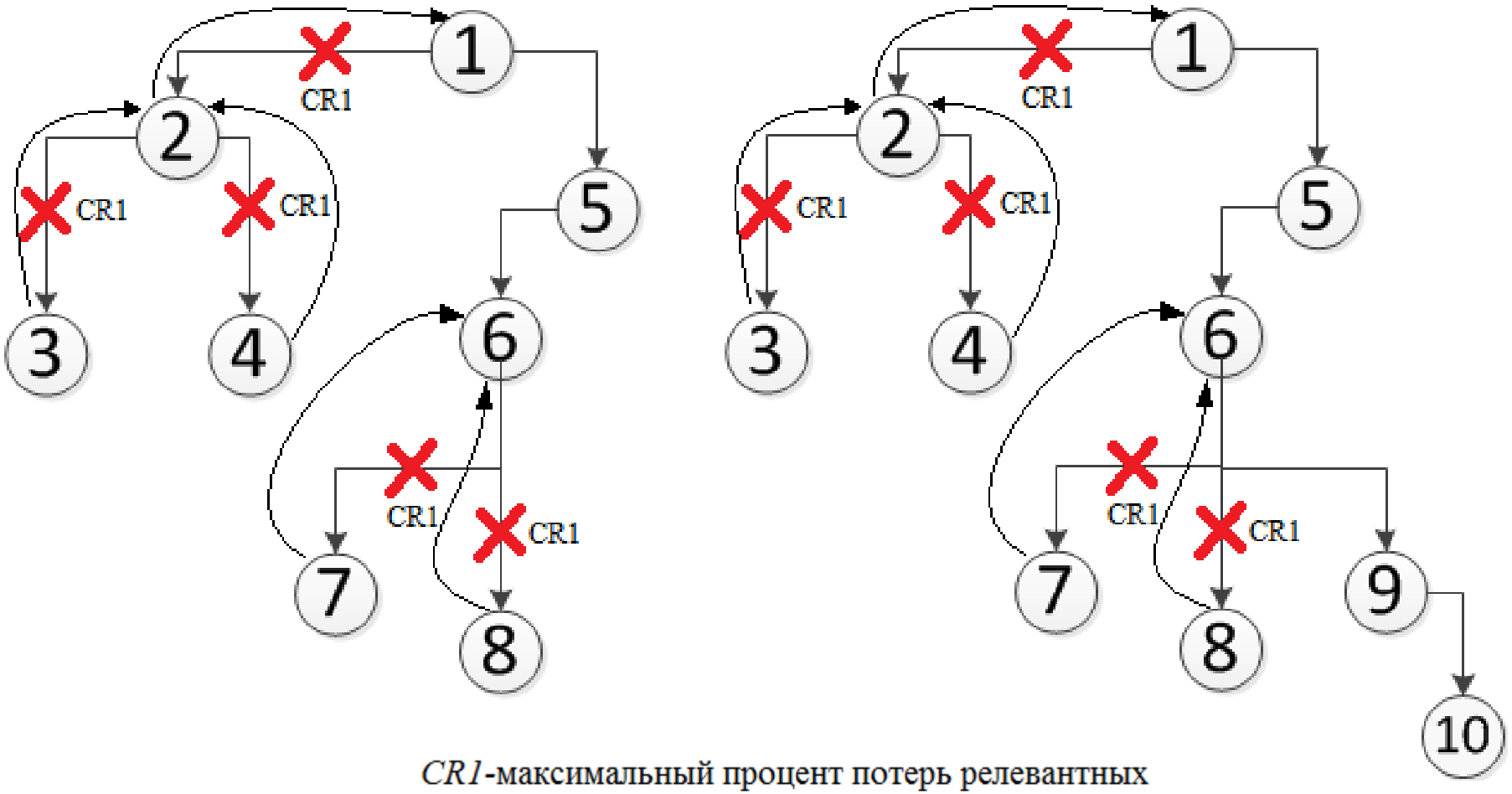
Фактор *f* заносится в множество Z (граф переходов) и исключается из дальнейшего рассмотрения, а алгоритм возвращается к шагу 1.

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ, шаг 2



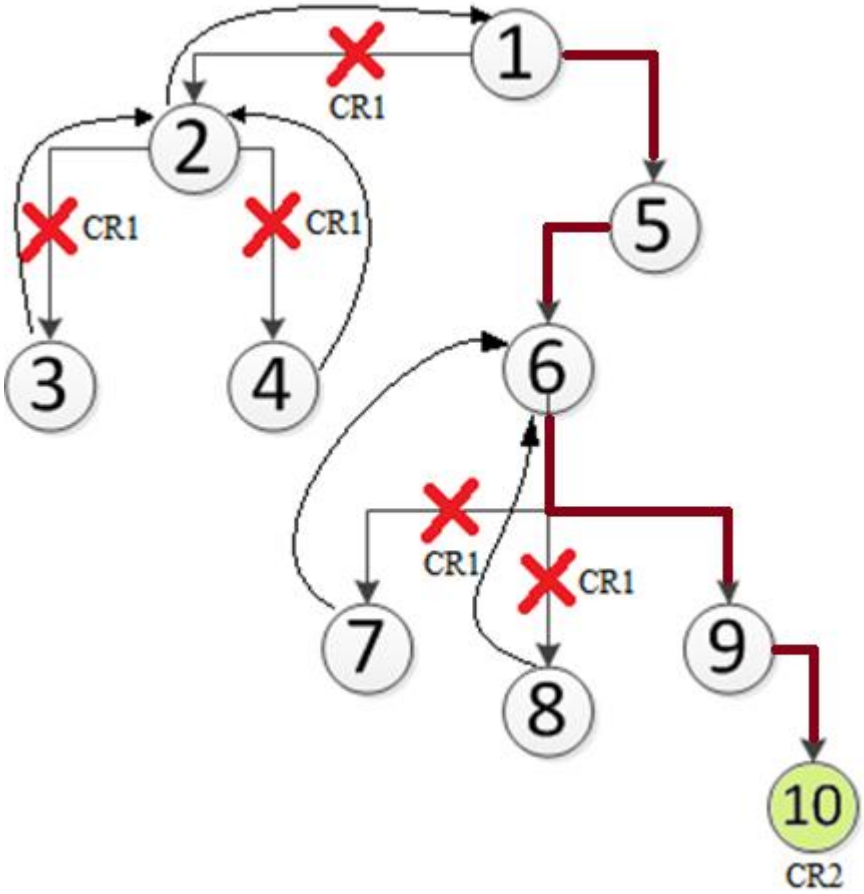
CR1-максимальный процент потерь релевантных объектов от их первоначального числа

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ, шаг 2



CR1-максимальный процент потерь релевантных объектов от их первоначального числа

Алгоритм надпороговой суперпозиции и его применение к нахождению эффективных границ, Построение дерева



CR1-максимальный процент потерь релевантных объектов от их первоначального числа

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

Шаг 3. Остановка алгоритма и преобразование результатов

Возможные усовершенствования

- Повышение «гибкости» итеративной процедуры
- Усовершенствование алгоритма отсечения ветвей дерева решений и перебора в возвратом

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 – ( ), где n – число объектов, k – число 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

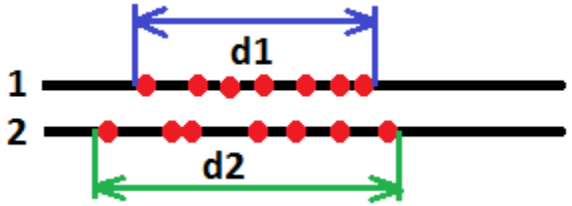
Номер фактора	Коэффициент регрессии	Стандартная ошибка	t - статистика	Название фактора	Часть документа
13	3,3	0,563	6,500	Длина потока	Заголовок
15	-0,4	0,043	-12,314	Длина потока	Весь документ
16	0,3	0,093	3,285	Мера IDF	Тело документа
23	-0,8	0,157	-6,490	Сумма частот слова	Заголовок
48	1,7	0,167	10,205	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.