KM.RU на РОМИП-2008

Оптимизация параметров поискового алгоритма

Сергей Татевосян,
Наталья Брызгалова
«КМ онлайн»

KM.RU на РОМИП-2008:

- Автоматическая оптимизация параметров на основе оценок экспертов
- Проверка новых поисковых факторов

Общая формула релевантности

$$W = k1*W1 + k2*W2 + k3*W3 + k4*W4,$$

W1 - информационная значимость документа и его вес в коллекции

W2 - информационная значимость ссылок на документ

W3 - учет расстояния между словами запроса в документе

W4 - дополнительные параметры

Новые факторы

W3:

- Встречаемость пар слов из запроса

W4:

- Близость слов из запроса к началу предложения
- Встречаемость точной словоформы из запроса

Дополнительно:

- Применение словаря сокращений
- Не учитывались стоп-слова

Оптимизация параметров

В основе программы-оптимизатора:

- Оценки экспертов
- Количественная мера оценки документа – Bpref-10 + Precision(10)
- Оптимизация методом модифицированного координатного спуска

Особенности функции релевантности

- Функция кусочная
- Функция немонотонная
- Функция обладает заведомо бóльшим числом параметров, чем используется для ее вычисления

=> Проблема: нахождение глобального максимума недостижимо

2 способа оптимизации:

1. Параметры оптимизируются все сразу
2. Оптимизация каждого параметра по отдельности - лучше

Предварительная обработка документов и запросов

Дано:

Неструктурированные запросы

+

Неструктурированная коллекция

Цель:

Структурированные запросы

+

Структурированная коллекция

Причина: работать со структурированной коллекцией гораздо легче

Структурирование данных

- **Структурирование документов:**
- Удаление информационного шума - оформления страниц (рекламные материалы, блоки новостных ссылок и т.д.)
- **Структурирование запросов**:
- Расширение запросов: использовали словарь сокращений
- Исправление опечаток *(эксперименты вне прогонов)*

Прогоны

- **Web**
  - Алгоритм, представленный на портале KM.RU
  - Алгоритм с оптимизированными и новыми параметрами
- **Legal**
  - Цель: увидеть, как алгоритм для веб отработает на произвольной коллекции _со_ссылками_
- **Смешанная коллекция**
  - Цель: увидеть, документ из какой коллекции будет в выдаче 1-м по соответствующему запросу (параллельный поиск)

Эксперименты вне прогонов

- **Поиск по кворуму:**

- коллекция KM
- коллекция BY

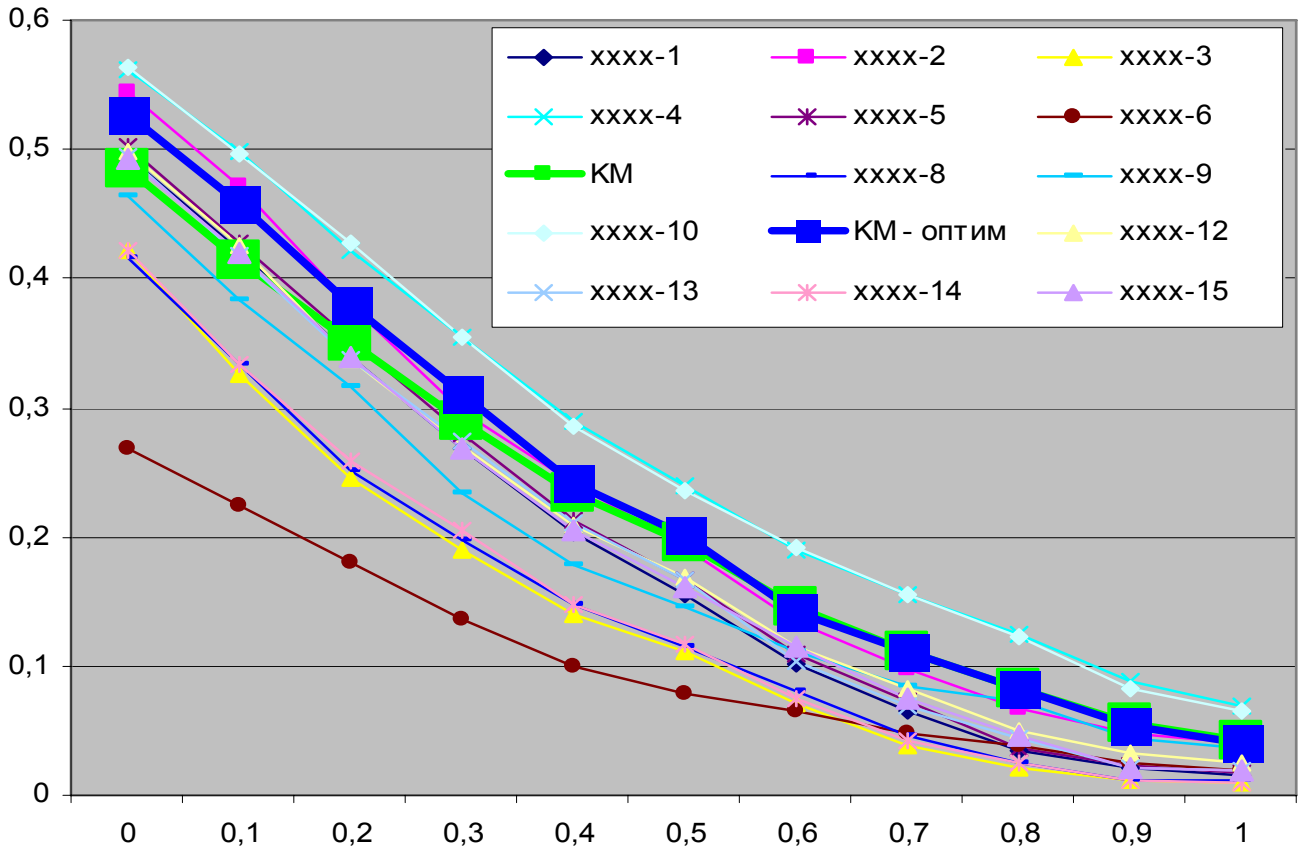
- **Исправление опечаток:**

- коллекция BY

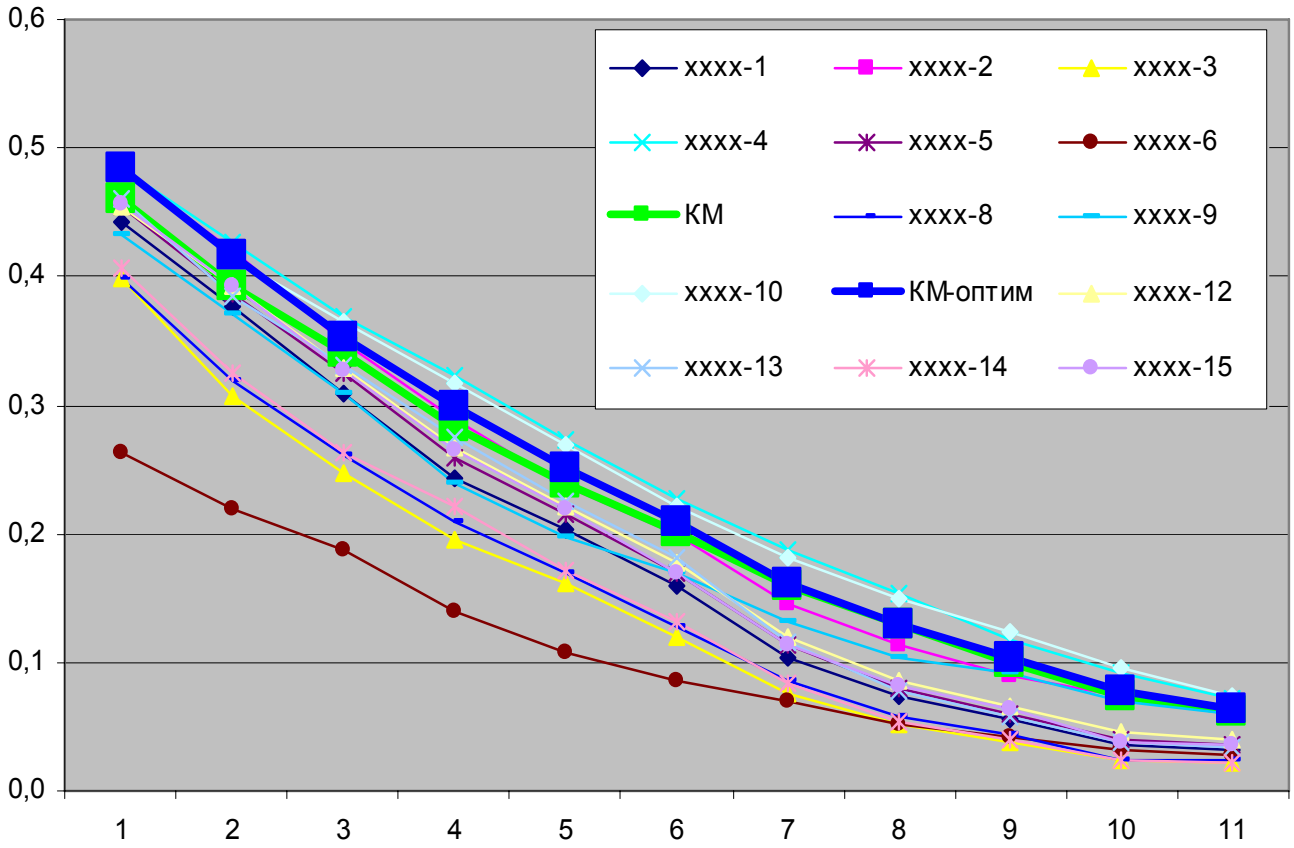
- **Без ссылочного ранжирования:**

- коллекция Legal
- коллекция BY

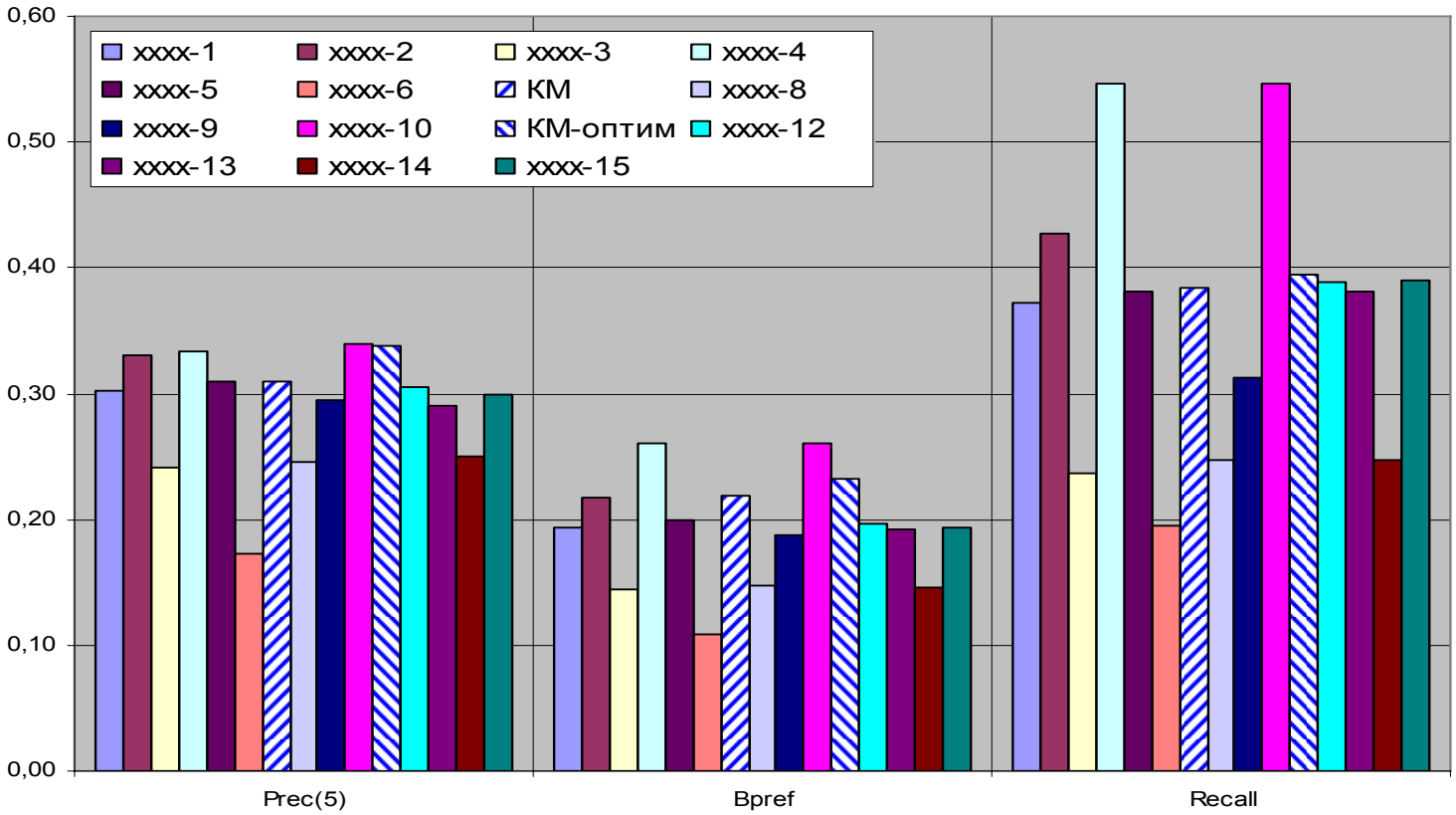
Результаты: Web adhoc, BY, OR



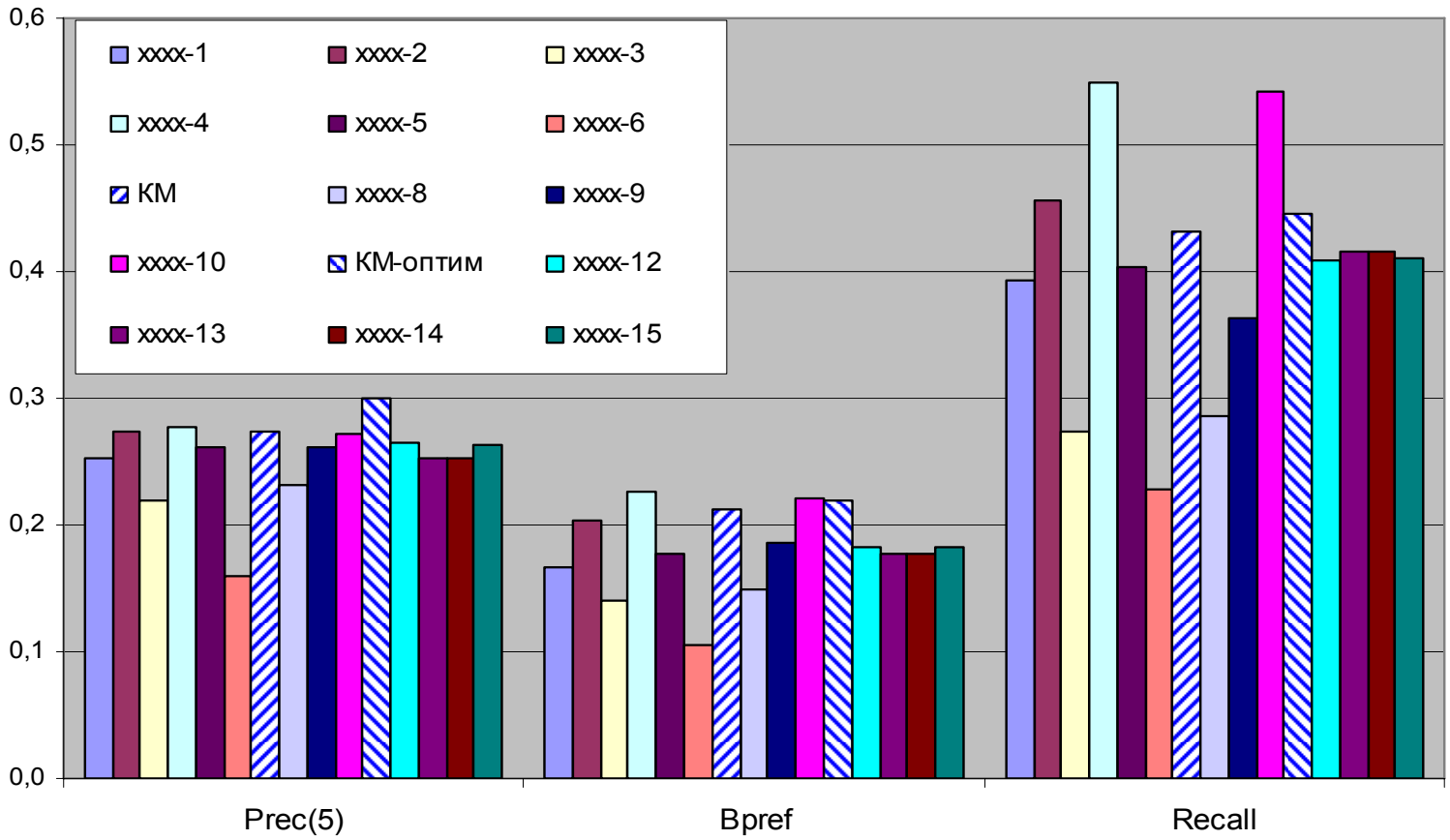
Результаты: Web adhoc, BY, AND



Результаты: Precision(5), bpref, recall для web adhoc, BY, OR



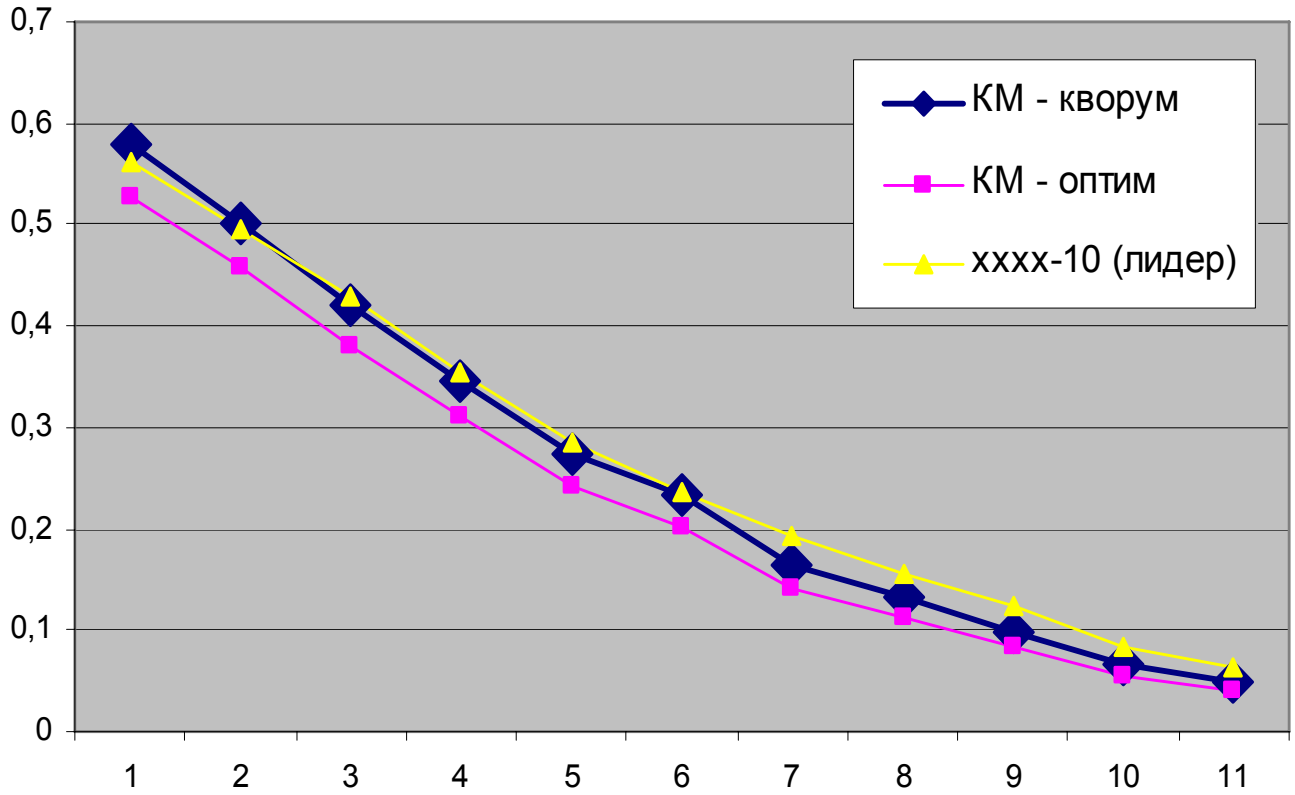
Результаты: Precision(5), bpref, recall для web adhoc, BY, AND



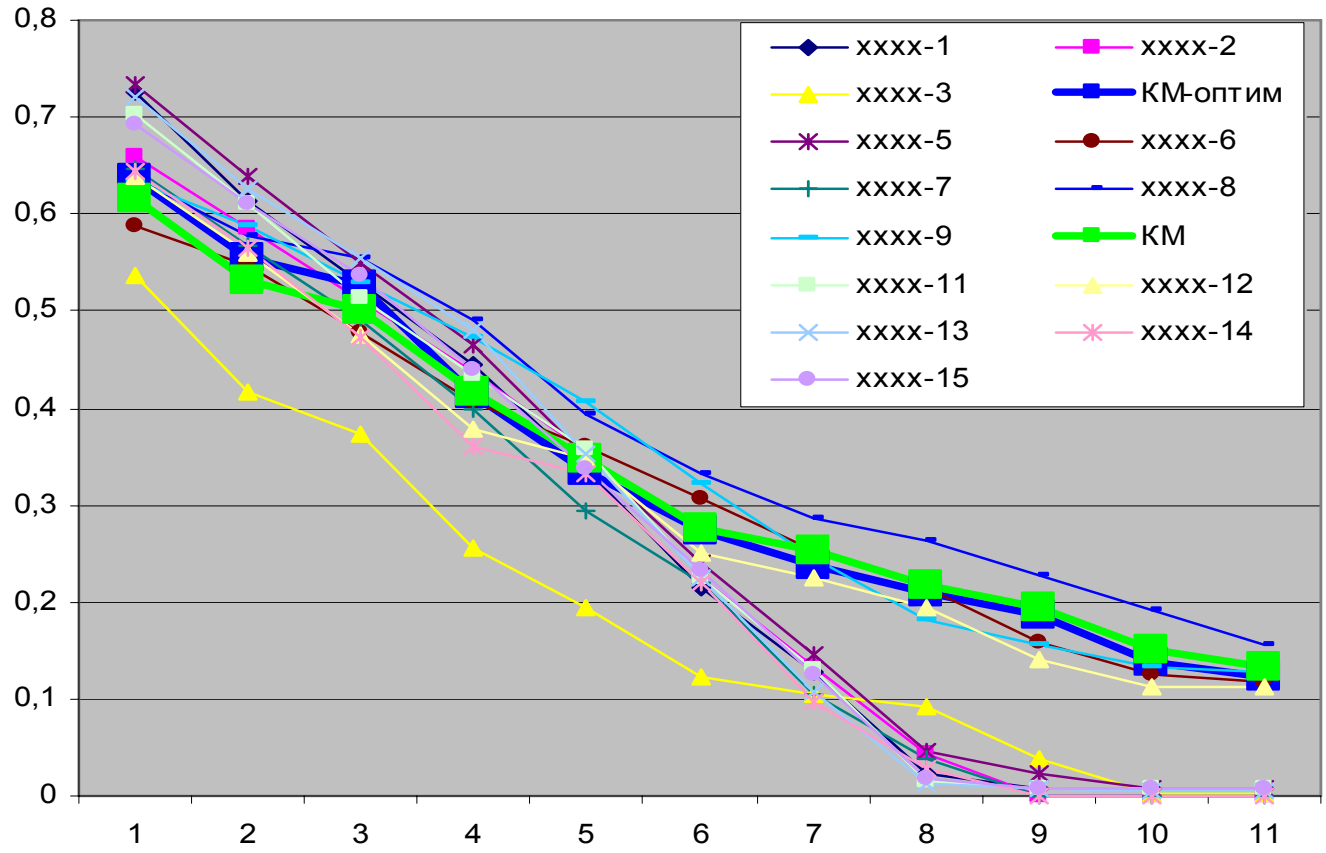
Анализ запросов

- **Случай 1:** запросы, где система находит документы, – показатели Precision(5) на уровне лидера (хорошая работа функции релевантности)
- **Случай 2:** запросы, где система не находит документов. Решение – кворум и исправление опечаток.

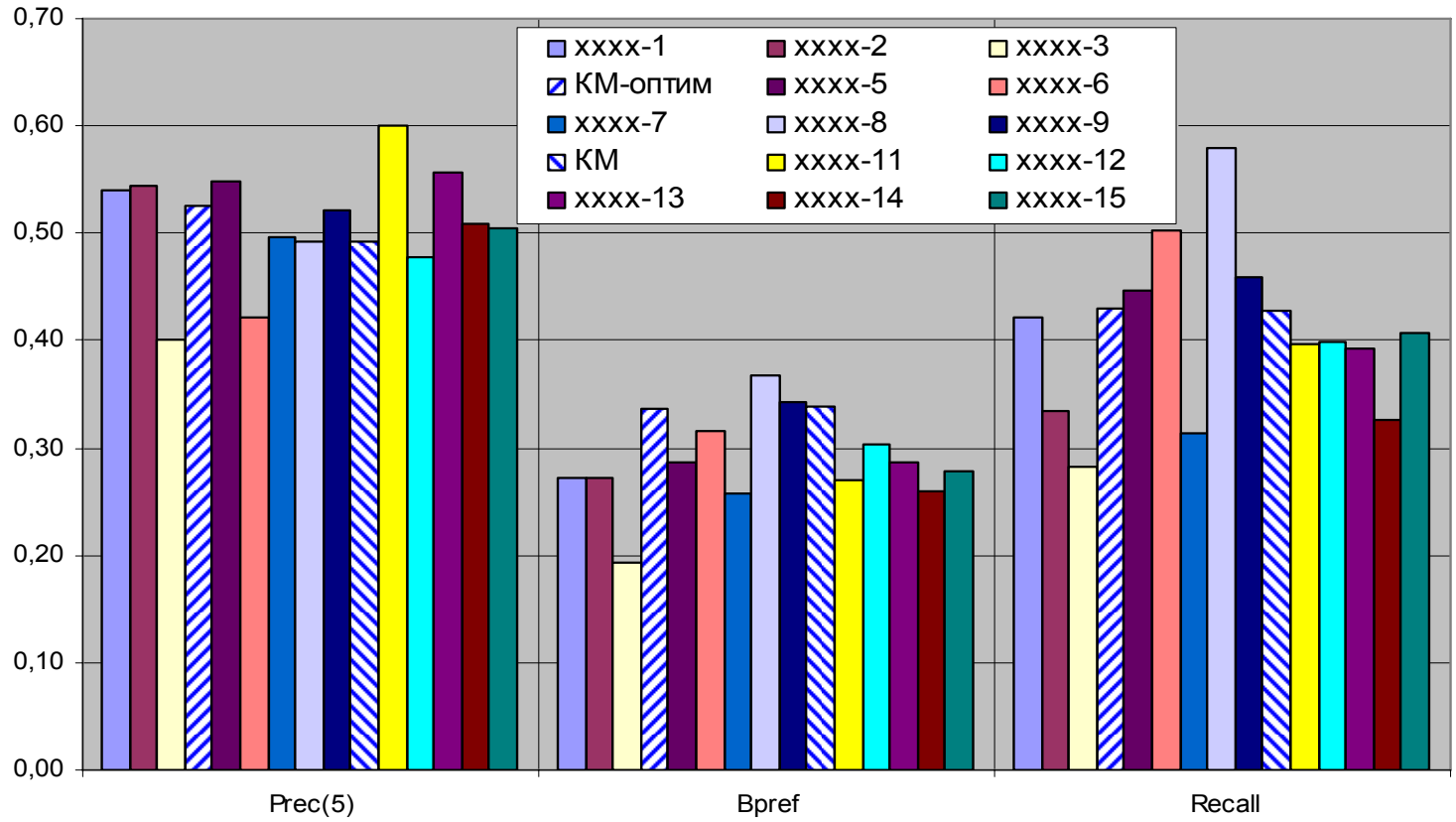
Вне прогонов: BY - добавление кворума и исправление опечаток



Результаты: Web adhoc, KM, OR



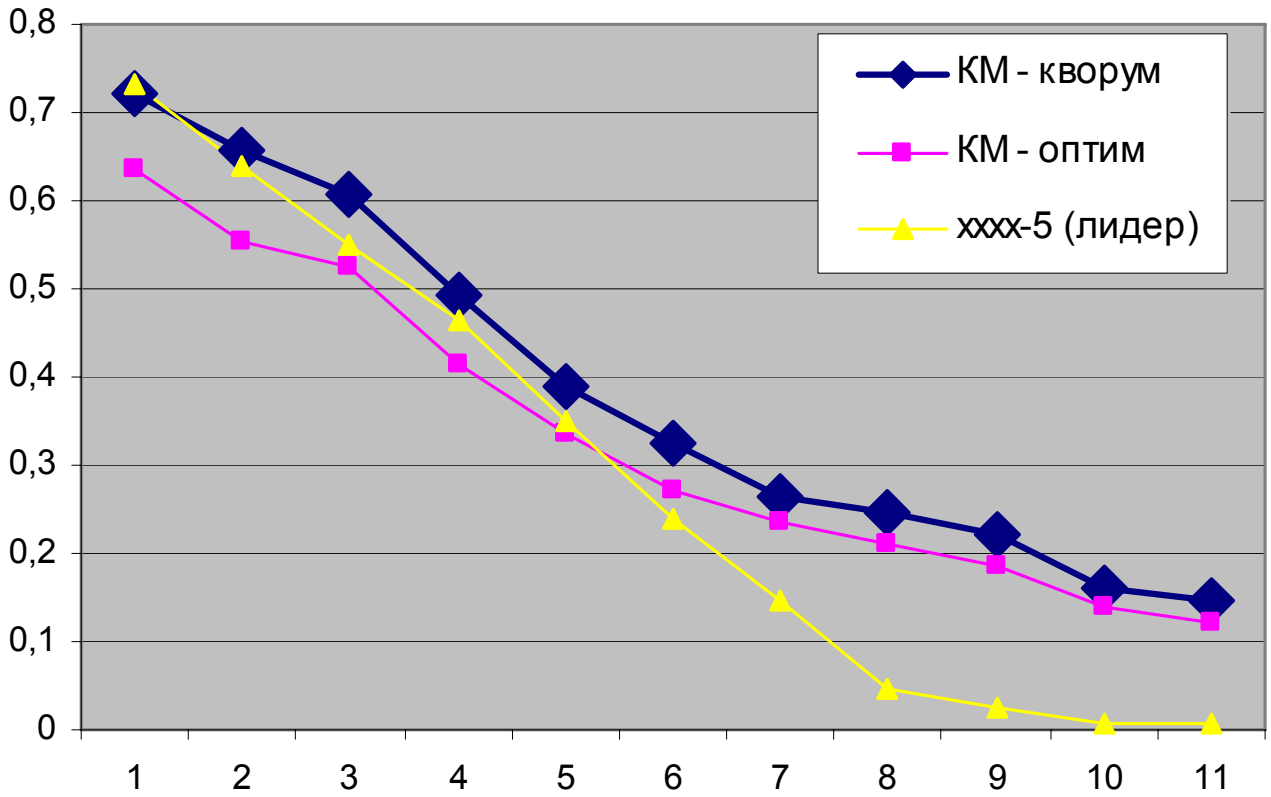
Результаты: Precision(5), bpref, recall для web adhoc, KM, OR



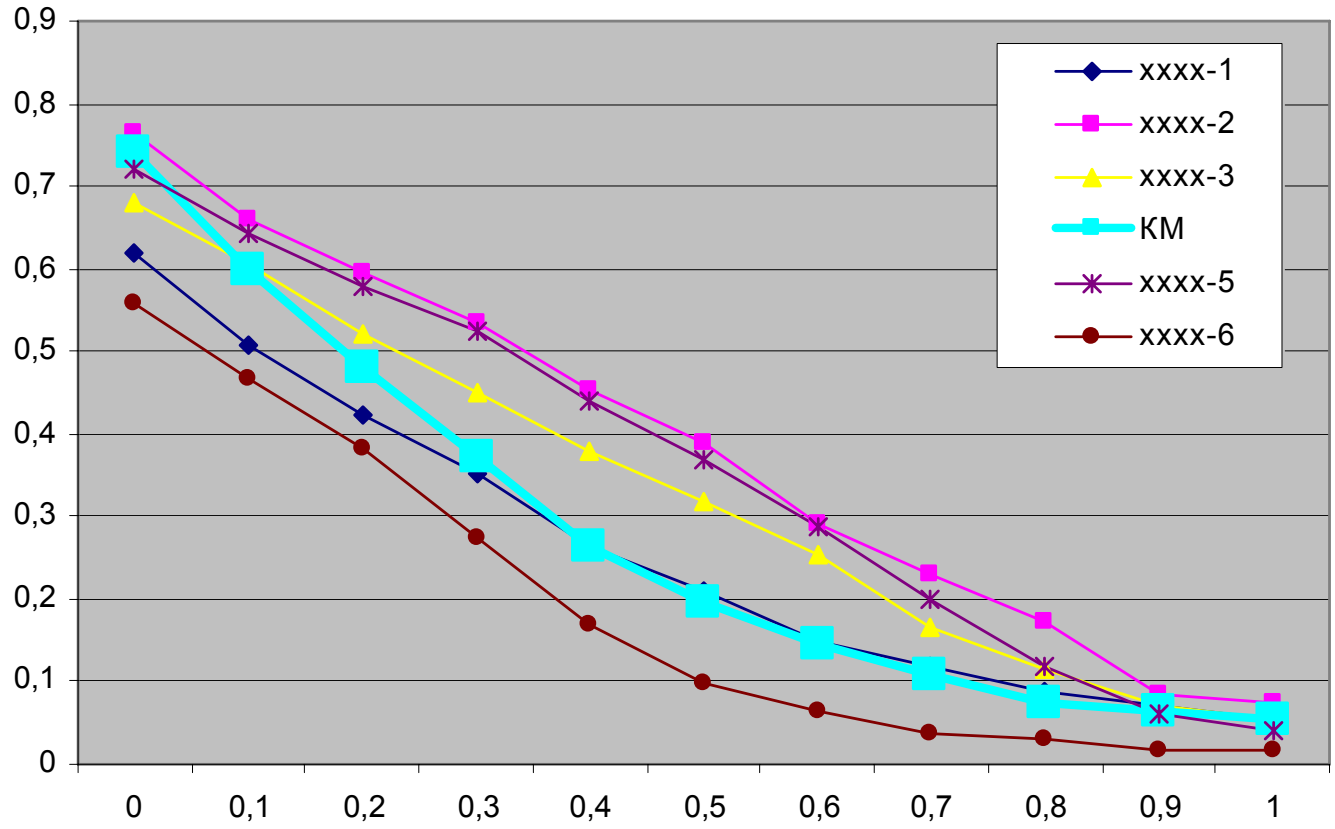
Анализ запросов

- **Случай 1:** запросы, где система находит документы, – показатели Precision(5) на уровне лидера (хорошая работа функции релевантности)
- **Случай 2:** запросы, где система не находит документов. Решение – кворум.

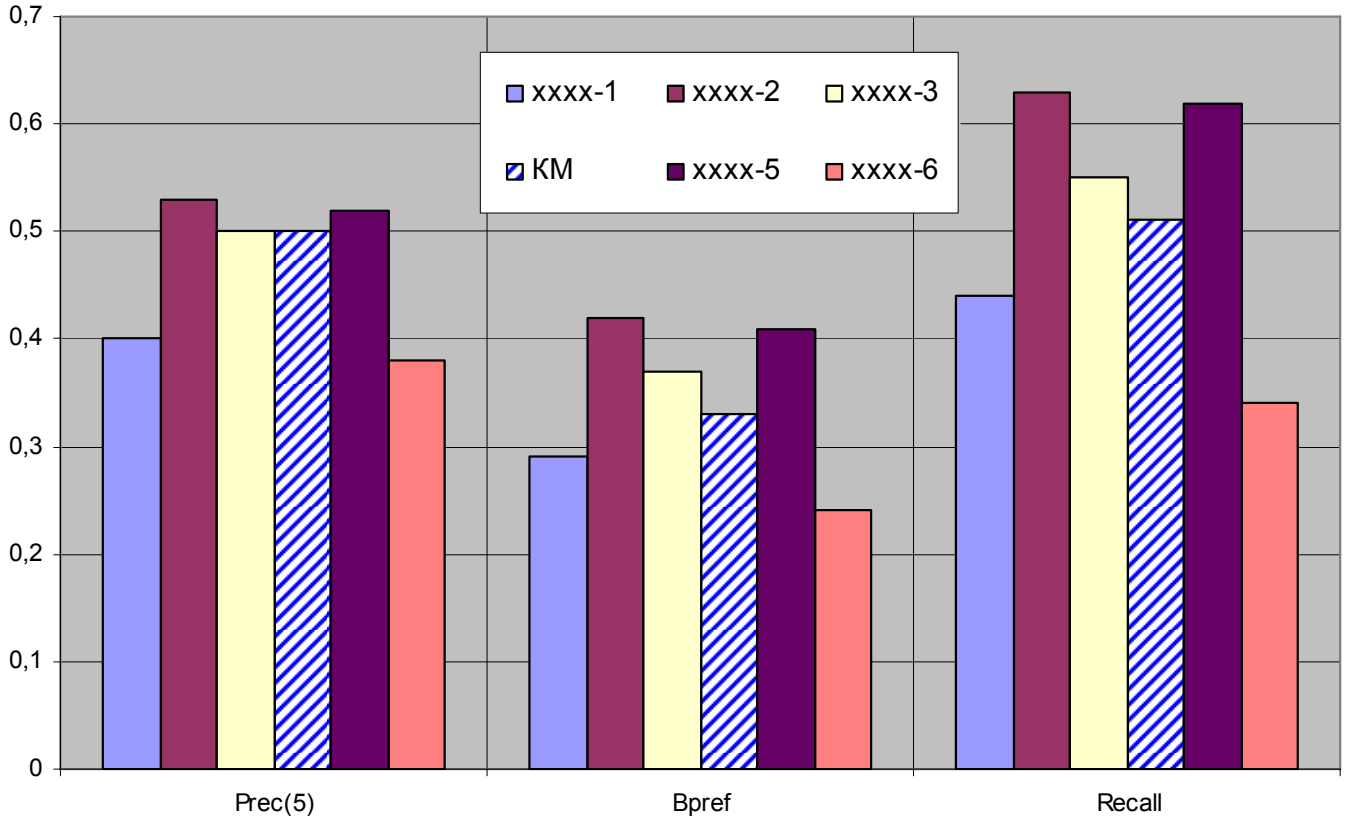
Вне прогонов: KM - добавление кворума



Результаты: Legal adhoc, OR



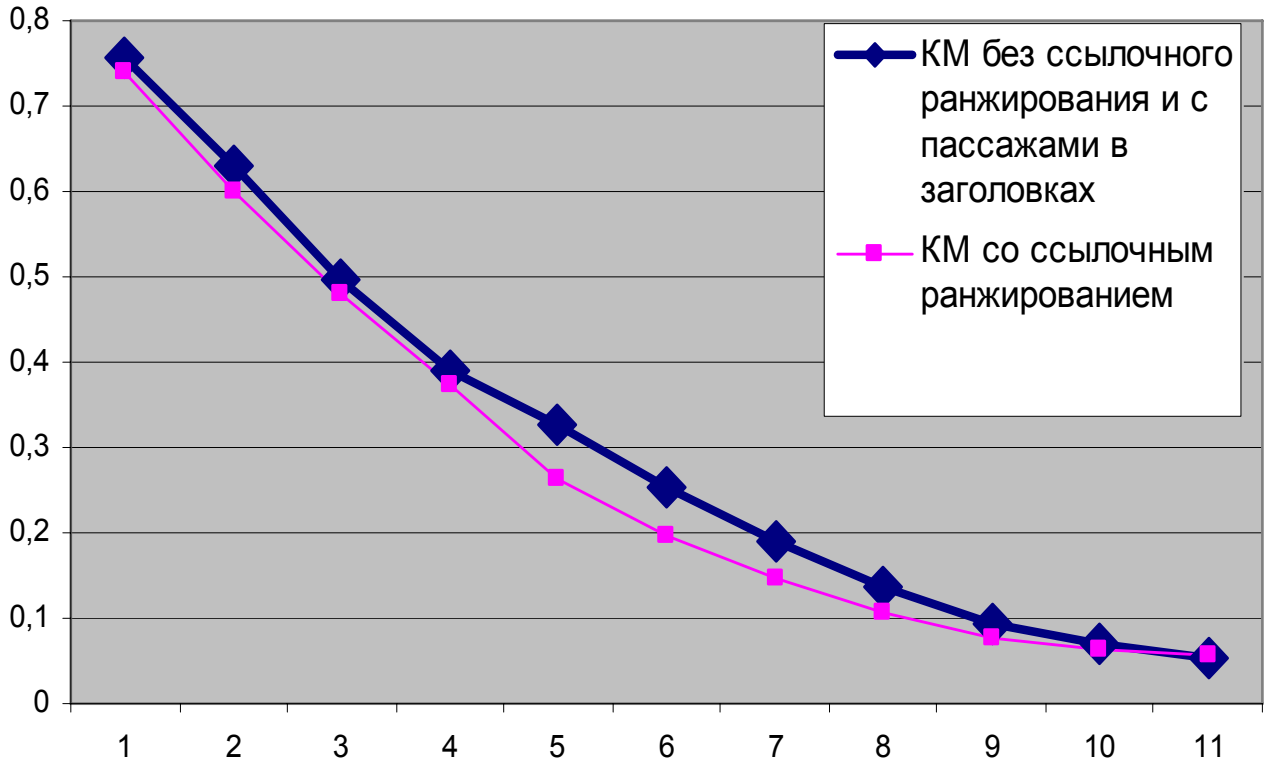
Результаты: Precision(5), bpref, recall для legal adhoc, OR



Анализ запросов

- **Случай 1:** запросы, где система находит документы, – показатели Precision(5) на уровне лидера (хорошая работа функции релевантности)
- **Случай 2:** запросы, где система не находит документов. Решение: 1) без ссылочного ранжирования, увеличение веса заголовков; 2) кворум.

Вне прогонов: Legal – без ссылочного ранжирования и с пассажами в заголовках



В планах: анализ запроса

Синтаксические преобразования:

Браслет из золота = золотой браслет

Расшифровка сокращений (уже работает):

МГУ = Московский государственный университет

Английский vs. русский вариант написания:

БМВ = BMW

В планах: база связанных понятий

Введение базы связанных понятий (проводим эксперименты):

1998

Президент РФ <=> Ельцин

2006

Президент РФ <=> Путин

2008

Медведев <=> Президент РФ, *но* Президент РФ => Ельцин
Президент РФ => Путин

поэтому Медведев ≠> Ельцин, Путин

Итоги

- Оптимизация коэффициентов дает прирост качества
- Новые параметры - пары слов и близость слов к началу предложения – работают
- Эксперименты на Web вне прогонов (поиск по кворуму, исправление опечаток) – график на уровне лидера
- Статистика – количественная оценка запроса и документа, языковые характеристики – качественная оценка

Спасибо за внимание!