МГУ им. М.В.Ломоносова
Научно-исследовательский
вычислительный центр

АНО Центр
информационных
исследований

Университетская информационная система
РОССИЯ

Оптимизация параметров алгоритма поиска на основе анализа оценок экспертов

Михаил Агеев, Борис Добров

RCDL'2005, Российский семинар
по оценке методов информационного поиска
Ярославль, 2005

УИС РОССИЯ (www.cir.ru)

Файл Правка Вид Избранное Сервис Справка Ссылки

Университетская информационная система РОССИЯ

О проекте

Университетская информационная система РОССИЯ (УИС РОССИЯ) создана и поддерживается как база электронных ресурсов для исследований и образования в области экономики, социологии, политологии, международных отношений и других гуманитарных наук и с 2000 года открыта для коллективного доступа университетов, вузов, научных институтов РФ и специалистов.

подробнее...

Полный список коллекций | Академический сервис | Партнеры | Участники | Зеркала | Как к нам пройти

Поиск по ресурсам УИС РОССИЯ

Поиск по источникам: все коллекции (Уровень доступа = FREE)

искать ▶

Расширенный поиск

Изменить уровень доступа (для зарегистрированных пользователей)

Имя: Пароль:

Об уровнях доступа | Зарегистрироваться | Забыли пароль? | Справка/Практикум

Новые ресурсы

- 12.11.2003
  - Социально-экономическое положение России (сентябрь 2003 г.)
- 11.11.2003
  - Краткосрочные экономические показатели Российской Федерации. (сентябрь 2003 г.)
- 04.11.2003
  - Краткосрочные экономические показатели Российской Федерации. (август 2003 г.)
  - Социально-экономическое положение России. (август 2003 г.)
- 30.09.2003
  - Социально-экономическое положение России (июль 2003 г.)
  - Краткосрочные экономические показатели Российской Федерации. (июль 2003 г.)

Новости

- 10.02.2003 10-14 февраля 2003 года в НИВЦ МГУ проведен российско-американский семинар "Новые технологии в поддержку государственного управления. Роль университетов в создании национальной информационной инфраструктуры". Программа семинара прилагается.

Ресурсы Университетской информационной системы РОССИЯ

Интегрированная коллекция | Бюджетная система России | Статистика России | Выборы в России | Парламент России | Соционет/Repec | Ресурсы зарубежных организаций

Бюджетная система России

Информационно-аналитический комплекс для изучения бюджетной системы РФ. Формируется из открытых первоисточников, предоставляемых органами государственной власти, научными институтами, аналитическими центрами. Содержит бюджетную статистику с 1995 года. Включает тематические публикации научных журналов, центральных СМИ. Представлены материалы профильных учебных курсов Экономического факультета МГУ им. М.В. Ломоносова.

Интегрированная коллекция

Полный список

Документы государственных органов | Издания исследовательских центров | Научные издания | Коллекции зарубежных организаций | Социологические опросы

Статистика России

(для зарегистрированных пользователей)

Интегрированная коллекция статистических и аналитических материалов характеризует социально-экономическое развитие Российской Федерации и регионов в ретроспективе с 1996 года. Формируется на базе первоисточников - публикаций Госкомстата России, Минэкономразвития, других государственных организаций, а также изданий независимых аналитических центров.

Реляционная база данных

Соционет / RePEc

Research Papers in Economics - библиотека библиографических описаний информационных ресурсов, создаваемых специалистами по общественным наукам (экономике, социологии, политологии) во всем мире. Включает архивы электронных публикаций, оглавления онлайновых журналов, каталоги новых поступлений библиотек, планы издательств и др. - более 200 тысяч записей. **СОЦИОНЕТ** - библиографическое описание материалов по общественным наукам на русском языке.

Выборы в России

Интерактивная карта выборов | Выборы Президента РФ (1996, 2000 год) | Выборы в ГосДуму РФ (1993, 1995, 1999 годы) | Выборы в субъектах РФ | Административно-территориальное деление РФ (СОАТО)

Ресурсы зарубежных организаций

Организация экономического сотрудничества и развития (OECD) - OECD Health Data. 2002. 30 countries.

База данных по системам здравоохранения в 30 странах. Поддерживается Организацией экономического сотрудничества и развития. Включает 1200 показателей. По некоторым показателям временные ряды прослеживаются с 1960 года.

Перепись населения в СССР в 1939 году - предоставлены Университетом Торонто, Канада.

Парламент России

ГосДума РФ. Стенограммы пленарных заседаний |

ГосДума РФ. Информационно-Аналитический бюллетень |

Информационные ресурсы УИС РОССИЯ
750.000 документов (около 8 Гб)

	Источник	Период	Количество
Нормативные д-ты	НТЦ Система	1993-…	70,000
Межд. договоры	НТЦ Система	1993-…	2,300
Стенограммы ГД	Аппарат ГД ФС РФ	1994-…	144,000
Статистические данные	Госкомстат РФ, Межгос. Стат. Ком. СНГ	1998-…	54,000
Материалы СМИ	«АиФ» - «Эксперт»,	199(7)-…	391,000
Аналитические материалы	Минэкономразвития, ЦБ, РЕЦЭП, БЭА, ИПН	1996-…	22,000
Научные издания	Вестник МГУ, «Соц.исследования»	Экономика 1998-…	5,300
Социологические опросы	Линейные распределения Мониторинга ВЦИОМ	2000	240

УИС РОССИЯ в РОМИП

- 2003
  - adhoc, коллекция web, TF*IDF
  - +эксперимент по учету точной формы слова
- 2004
  - adhoc, коллекции web, legal, TF*IDF
  - +эксперимент по учету расстояния между словами
  - +эксперимент по повышению веса заголовков
  - text categorization, legal
- 2005
  - adhoc, коллекции web, legal, TF*IDF
  - +эксперимент по учету расстояния между словами и «мягкий» поиск
  - +оптимизация параметров алгоритма на основе оценок РОМИП'2004

Компоненты известных алгоритмов поиска и ранжирования документов

- поиск документов, содержащих все слова запроса
- расширение запроса с учетом морфологической вариации слов
- исключение незначимых слов (стоп-слов) из запроса
- поиск документов, частично удовлетворяющих запросу (т.е. содержащих не все слова запроса)
- расширение запроса пользователя тематически близкими словами/терминами
- учет относительной частоты встречаемости слов запроса в найденном документе (чем чаще слово встречается - тем лучше)
- учет относительной частоты встречаемости слов запроса в документах коллекции (более редкие слова имеют больший вес)
- учет формы слова, которое встретилось в документе (лучше, если слово встретилось в документе в той же форме, что и в запросе)
- учет расстояния между словами запроса в документе (лучше, если слова запроса находятся рядом)
- учет структуры документа (слова в заголовке имеют больший вес)
- учет ссылок между документами (на авторитетный документ часто ссылаются)

«Отправная точка» (TF*IDF)

Вес каждой леммы документа :

$$\mathrm{TFIDF}_D(l) = \beta + (1-\beta)\cdot tf_D(l)\cdot idf_D(l)$$

где “term frequency” – учет частотности леммы в документе:

$$tf_D(l) = \frac{\mathrm{freq}_D(l)}{\mathrm{freq}_D(l) + 0.5 + 1.5\cdot\frac{dl_D}{\mathrm{avg\_dl}}}$$

$freq_D(l)$ - частотность леммы l в документе, dl_D – мера длины документа, avg_dl – средняя длина документа, $\beta = 0.4$

$$\mathrm{idf}(l) = \frac{\log\left(\frac{|c|+0.5}{df(l)}\right)}{\log(|c|+1)}$$

«Отправная точка» (TF*IDF)

Каждый запрос $Q = w_1 \quad w_2 \quad w_3 \ldots \quad w_m$ представлялся в виде формулы

$$L(Q) = L(w_1) \,\&\, L(w_2) \,\&\, L(w_3) \,\&\, \ldots \,\&\, L(w_m),$$

где $L(w) = l_1(w)$ **OR** $l_2(w)$ **OR** $\ldots$**OR** $l_q(w)$, $l_k(*)$ - леммы морфологического разбора слова.

Тогда оценка релевантности документа D для запроса *Q вычисляется по формуле:*

$$V_D(Q) = \frac{\sum_{i=1}^{N} \sum_k \left(\theta_{ik} \cdot \mathrm{TFIDF}_D(l_{ik}(w_i)) \right)}{\sum_{i=1}^{N} \sum_k |\theta_{ik}|}$$

где $\theta_{ik} = \theta_i = 1.0$ — *“вес” леммы в запросе — равен весу, устанавливаемому для соответствующего слова запроса.*

Учет дополнительных факторов

$$\mathrm{Rank}_D(Q) = \frac{FF-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{V_D(Q) + \mathrm{Near}_D(Q)}{2}$$

|Q| — **количество слов запроса**

$\mathrm{Rank}_D(Q)$ — **ранг соответствия документа запросу**

FF — **количество слов запроса, которые содержатся в данном документе**

$V_D(Q)$ — **ранг, вычисленный по формуле TF*IDF**

$$\mathrm{Near}_D(Q) = \frac{1}{\ln\left(\lambda_D(Q) - |Q| + 4\right)}$$

$\lambda_D(Q)$ — **длина минимального «куска» документа, в котором содержатся все слова запроса**

Оптимизация алгоритма поиска

1. Поиск с использованием классической векторной модели TF*IDF
2. Расширение алгоритма TF*IDF
   - ранжирование с учетом расстояния между словами
   - поиск документов, содержащих не все слова запроса
3. Оптимизация параметров, входящих в предложенный алгоритм:

$$\mathrm{Rank}_D(Q) = \mathrm{TFIDF}_D(Q) \cdot \frac{\mathrm{FF}}{|Q|} + \beta \cdot \mathrm{Near}_D(Q) \cdot 2^{-\alpha \cdot (|Q| - \mathrm{FF})}$$

β - относительный вес кучности слов запроса

α - зависимость веса кучности от количества найденных слов

4. Оценка результатов на материалах РОМИП

Оптимизация параметров: перебор на 4 классах функций

Оптимизация параметров производилась для следующих классов функций ранжирования:

1) $$\mathrm{Rank}_D(Q) = \frac{\mathrm{FF}-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{\mathrm{V}_D(Q) + \beta \cdot \mathrm{Near}_D(Q)}{1+\beta}$$

— однопараметрическое семейство функций с параметром $\beta > 0$, задающим относительный вес кучности слов запроса;

2) $$\mathrm{Rank}_D(Q) = \frac{\mathrm{FF}-1}{|Q|} + \frac{1}{|Q|} \cdot \frac{1+\beta}{\frac{1}{\mathrm{V}_D(Q)} + \frac{\beta}{\mathrm{Near}_D(Q)}}$$

— однопараметрическое семейство функций с параметром $\beta > 0$, задающим относительный вес кучности слов запроса;

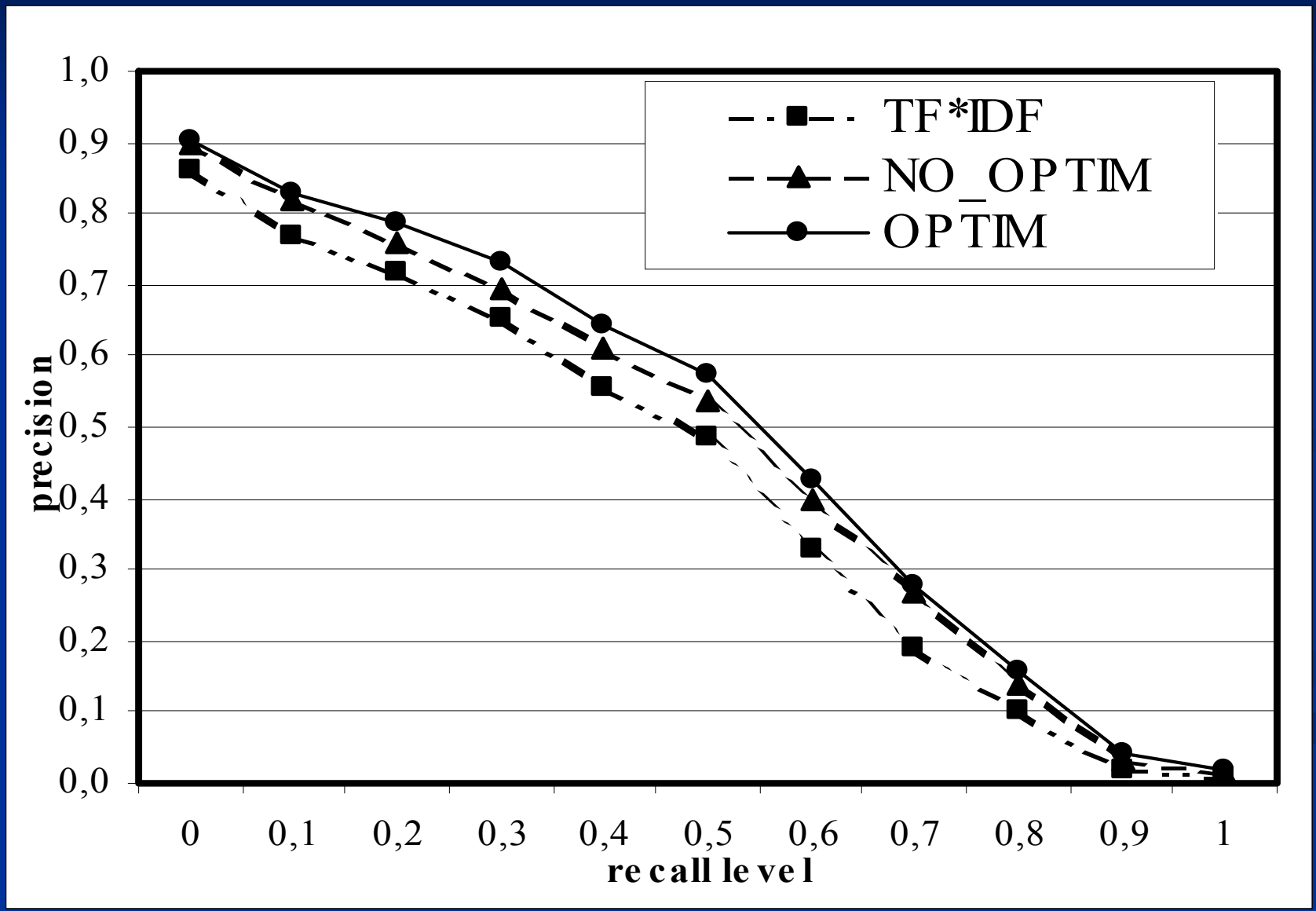
3) $$\mathrm{Rank}_D(Q) = \mathrm{V}_D(Q) \cdot \frac{\mathrm{FF}}{|Q|} + \beta \cdot \mathrm{Near}_D(Q) \cdot 2^{-\alpha \cdot (|Q| - \mathrm{FF})}$$

— двухпараметрическое семейство функций с параметрами $\beta > 0$ (относительный вес кучности слов запроса) и $\alpha > 0$ (зависимость веса кучности от количества найденных слов);

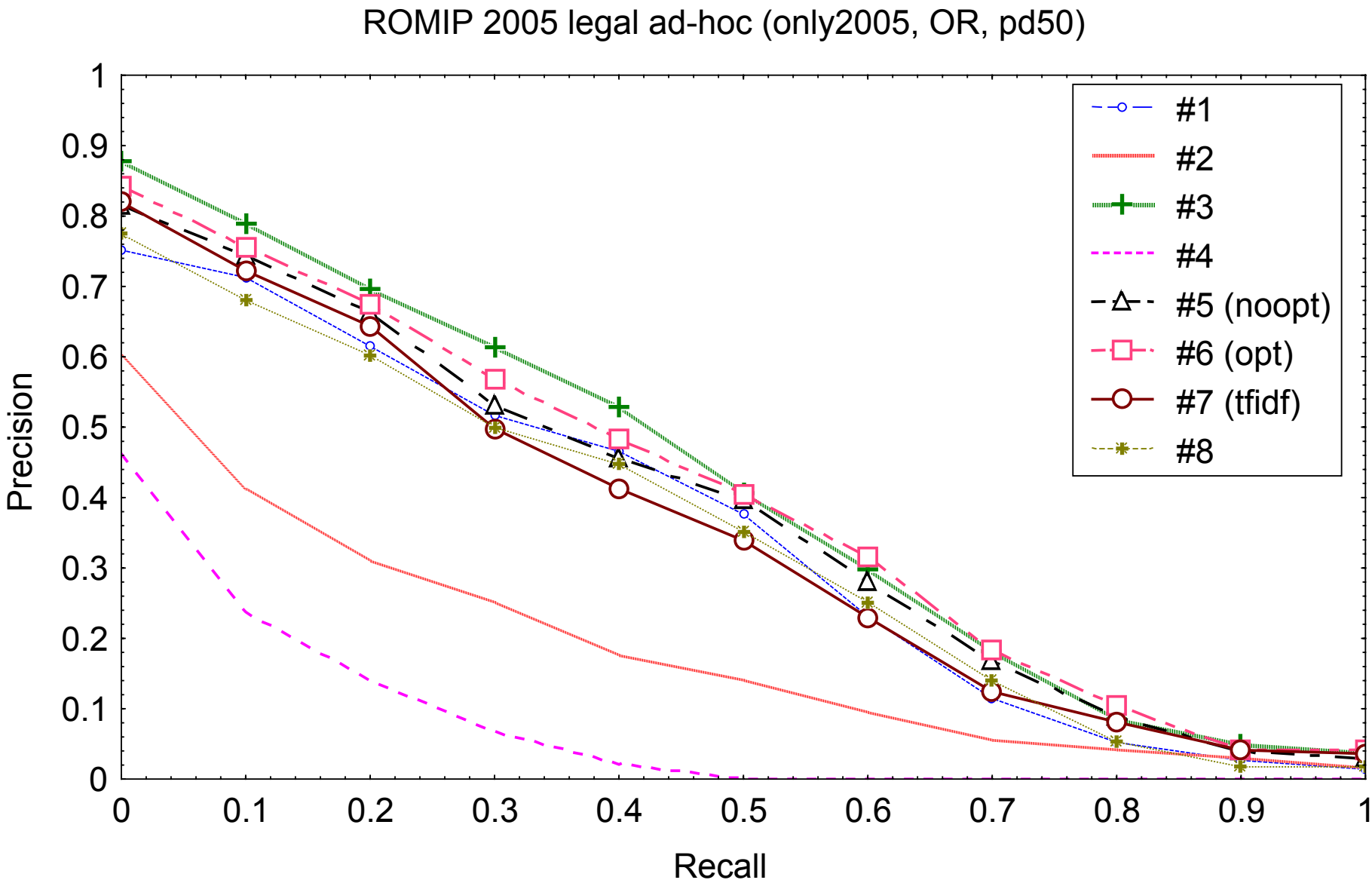
4) $$\mathrm{Rank}_D(Q) = \frac{\mathrm{V}_D(Q) + \beta \cdot \frac{1}{\ln^{\gamma}\left(\lambda_D(Q) - |Q| + 4\right)} + \alpha \cdot \mathrm{FF}}{1 + \beta + \alpha \cdot |Q|}$$

— трёхпараметрическое семейство функций с параметрами $\beta > 0$ (относительный вес кучности слов запроса), $0 < \alpha \le 1$ (вес количества найденных слов запроса) и $\gamma > 0$ (жесткость функции ранжирования по кучности слов).

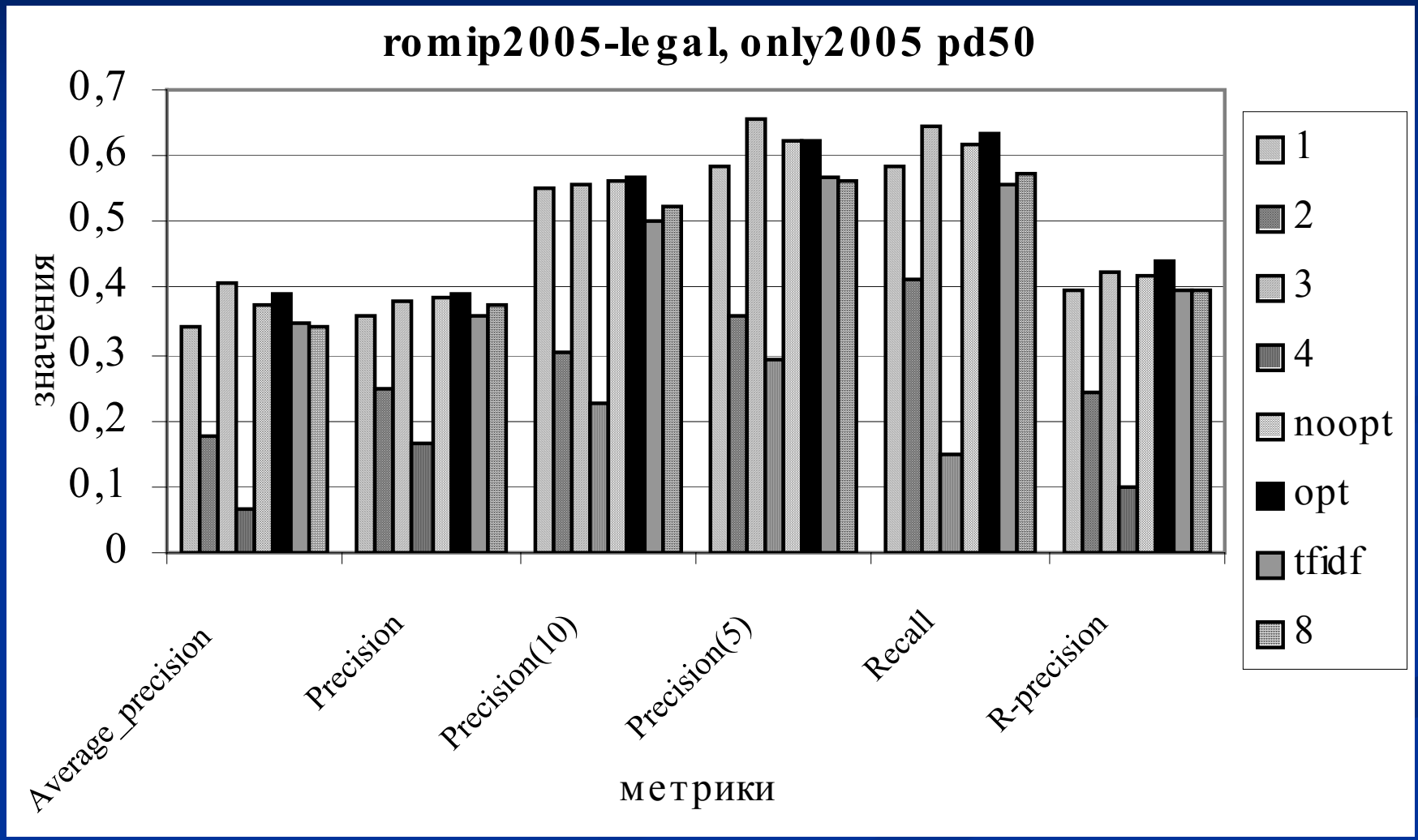
Результаты: оценка на «тренировочных» запросах (РОМИП'2004, legal-adhoc, OR-pd50)



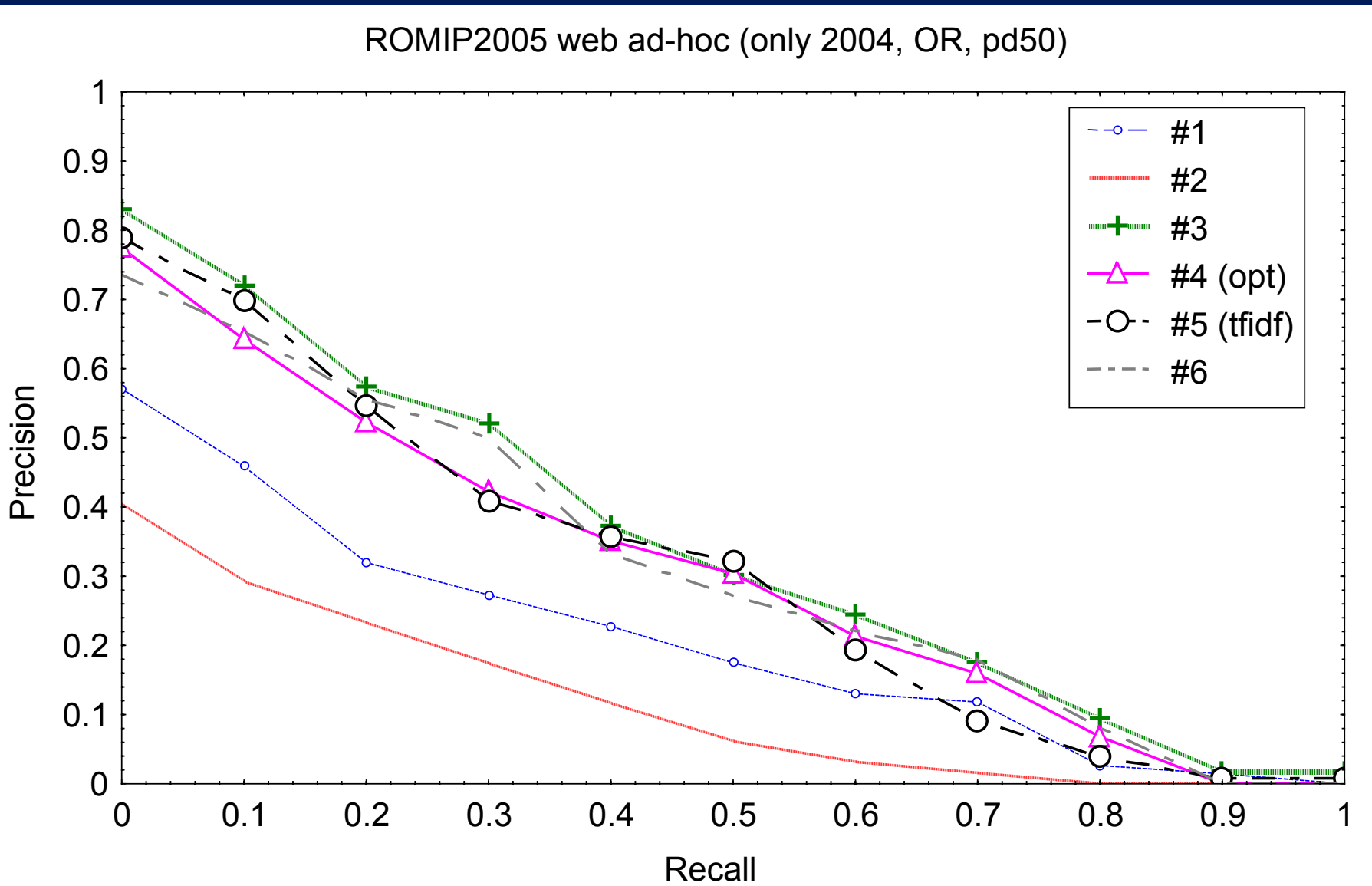
Результаты: оценка на новых запросах (РОМИП'2005, legal-adhoc, OR-pd50)



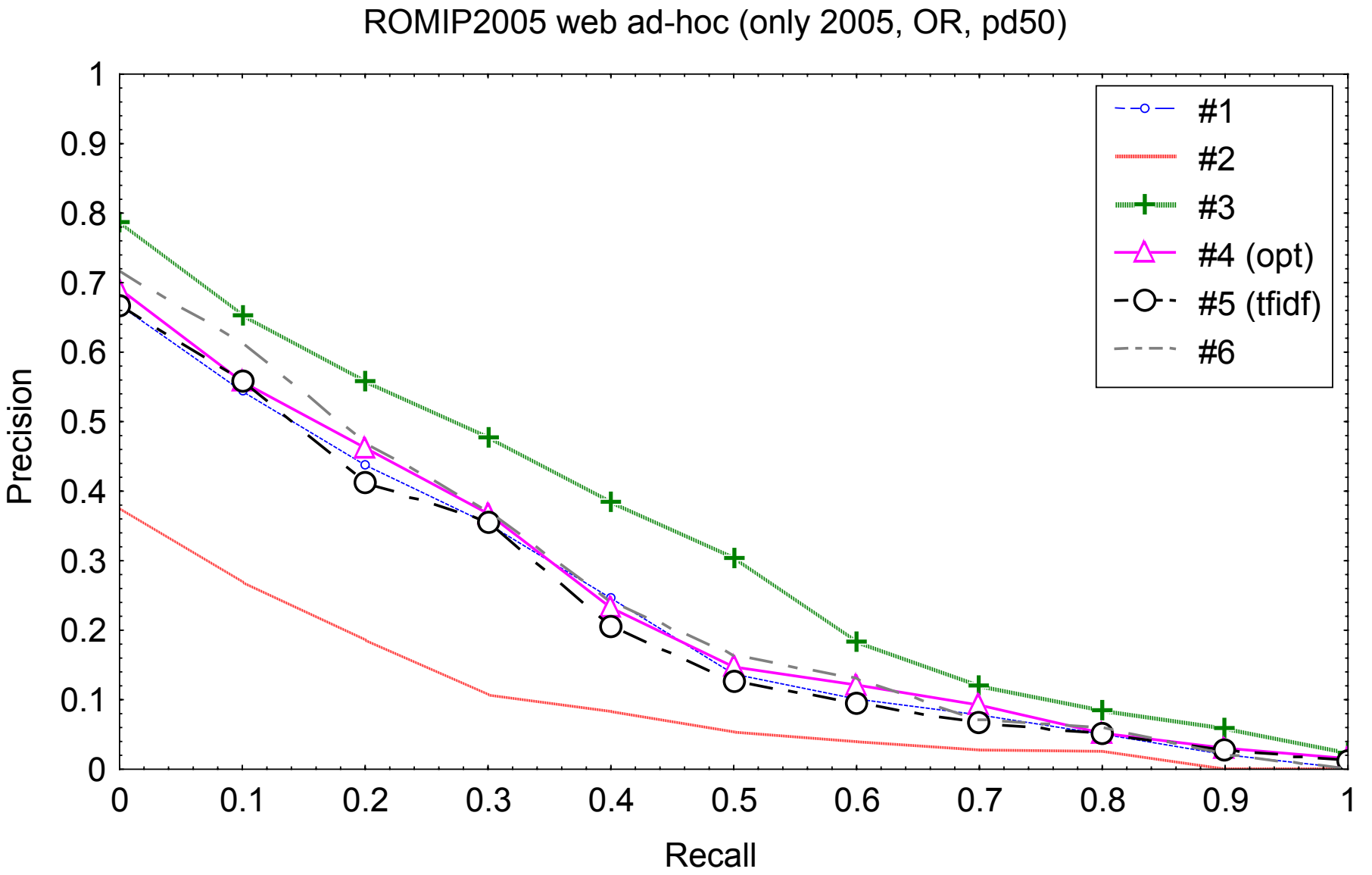
Результаты: оценка на новых запросах (РОМИП'2005, legal-adhoc, OR-pd50)



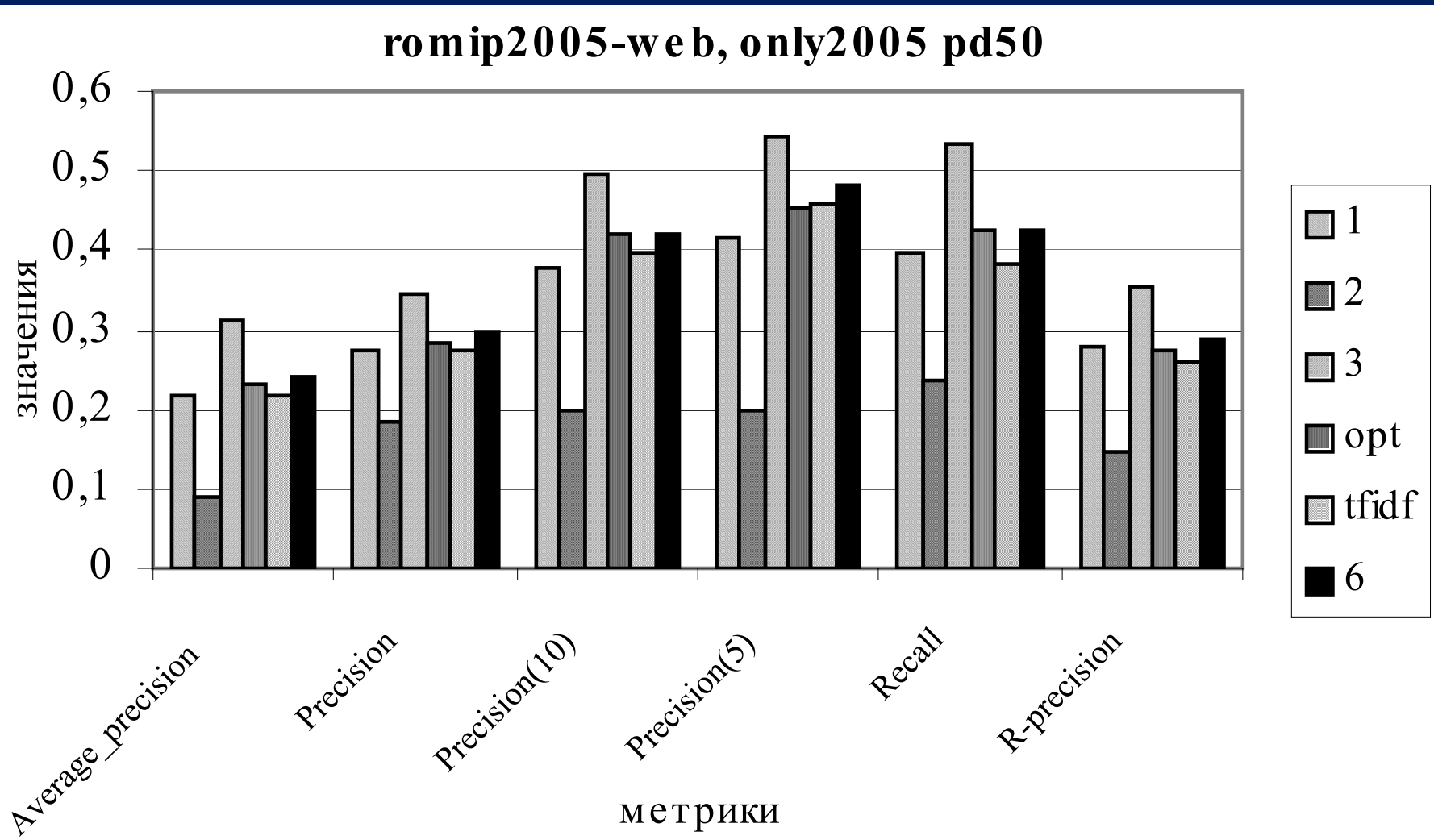
Результаты: оценка на тренировочных запросах (РОМИП'2004, web, OR-pd50)



Результаты: оценка на новых запросах (РОМИП'2005, web-adhoc, OR-pd50)



Результаты: оценка на новых запросах (РОМИП'2005, web-adhoc, OR-pd50)



Заключение

- Предложенный метод автоматической настройки параметров алгоритма поиска может быть использован для повышения эффективности поисковых машин

- В дальнейшем мы планируем развивать данный подход в направлении расширения набора параметров, влияющих на работу поискового алгоритма