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

АНО Центр
информационных
исследований

Университетская информационная система
РОССИЯ

УИС РОССИЯ:
поиск и классификация в РОМИП 2007

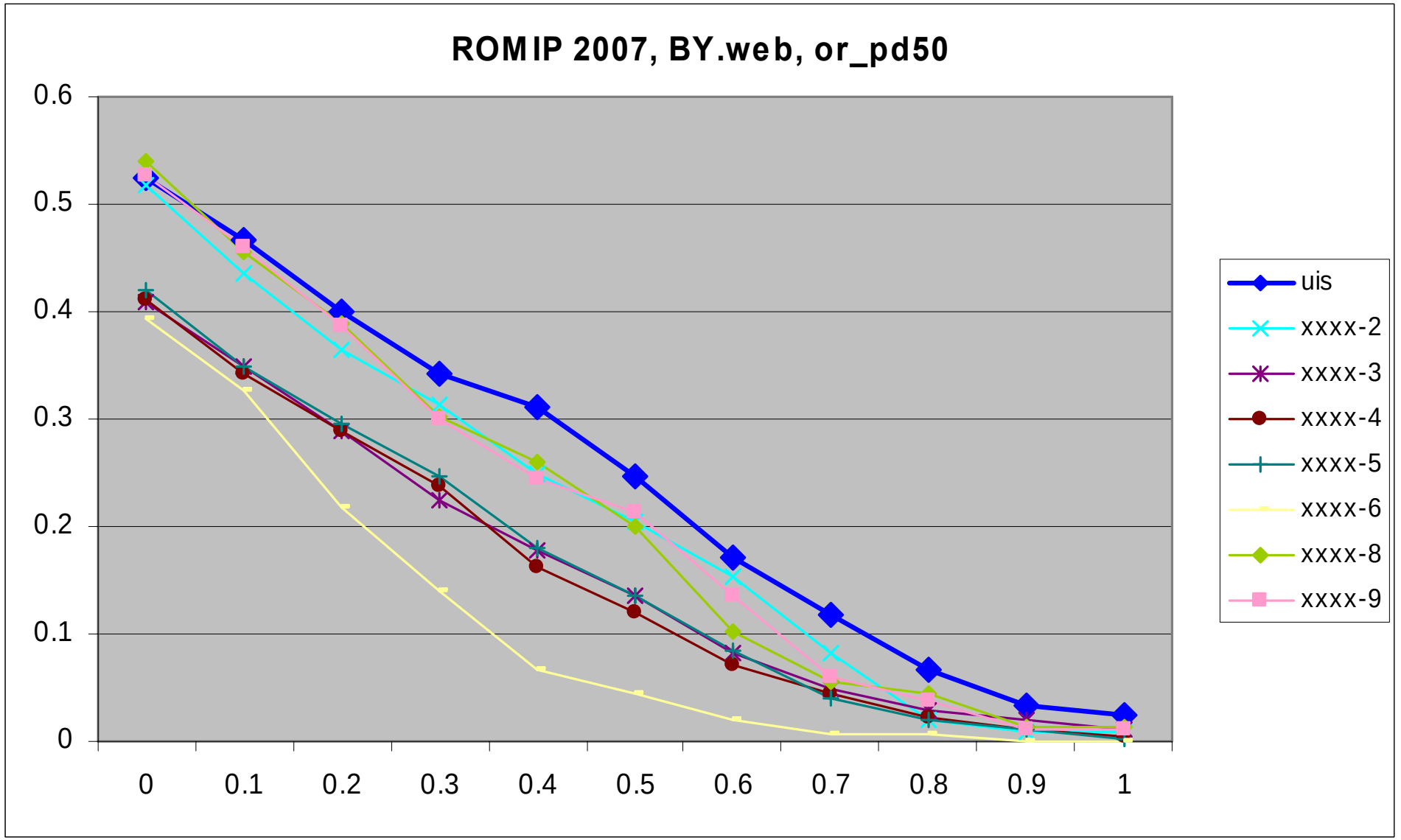
Агеев М.С., Добров Б.В., Красильников П.В.,
Лукашевич Н.В., Сидоров А.В., Штернов С.В.

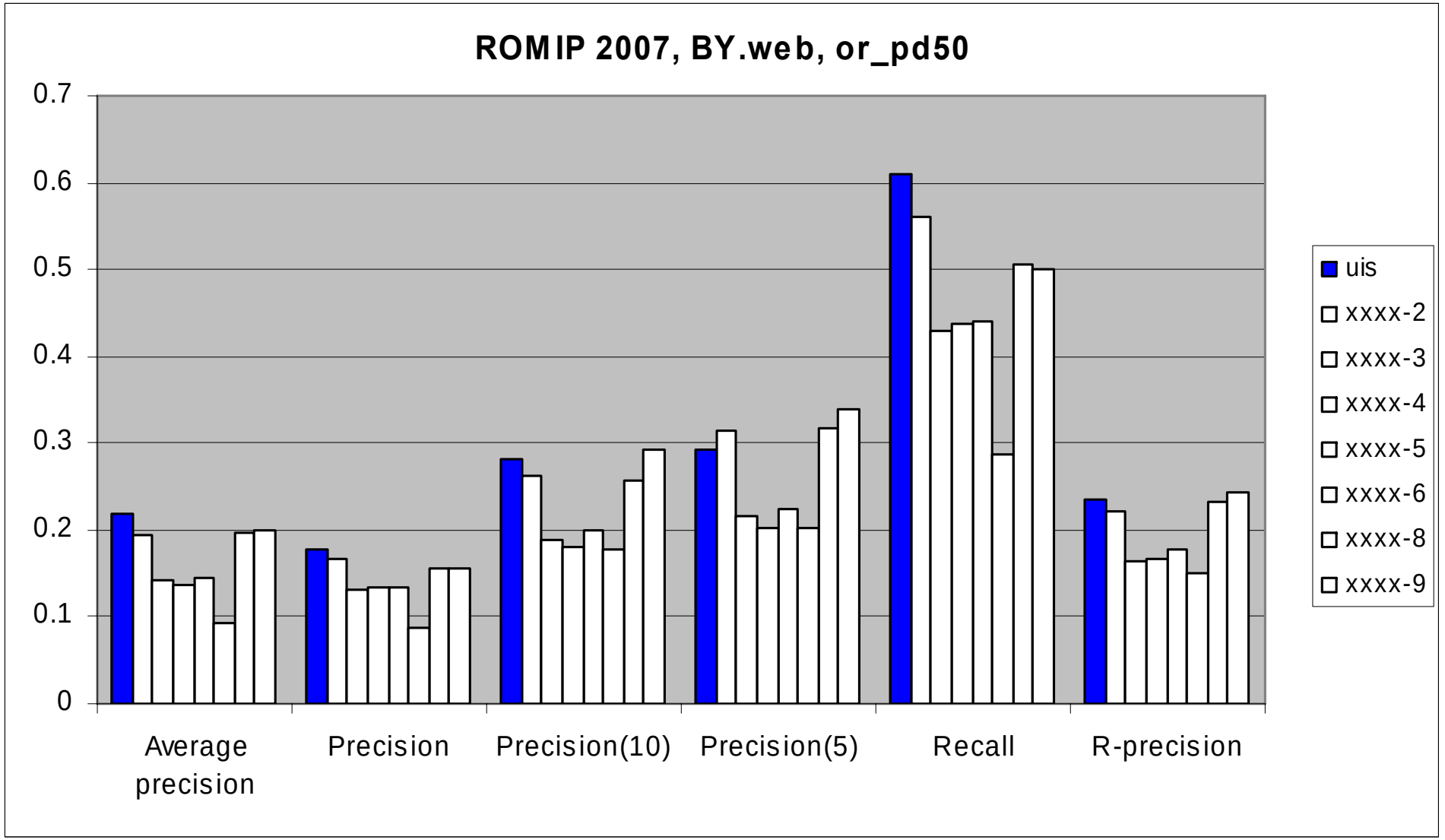
Ad-hoc поиск документов

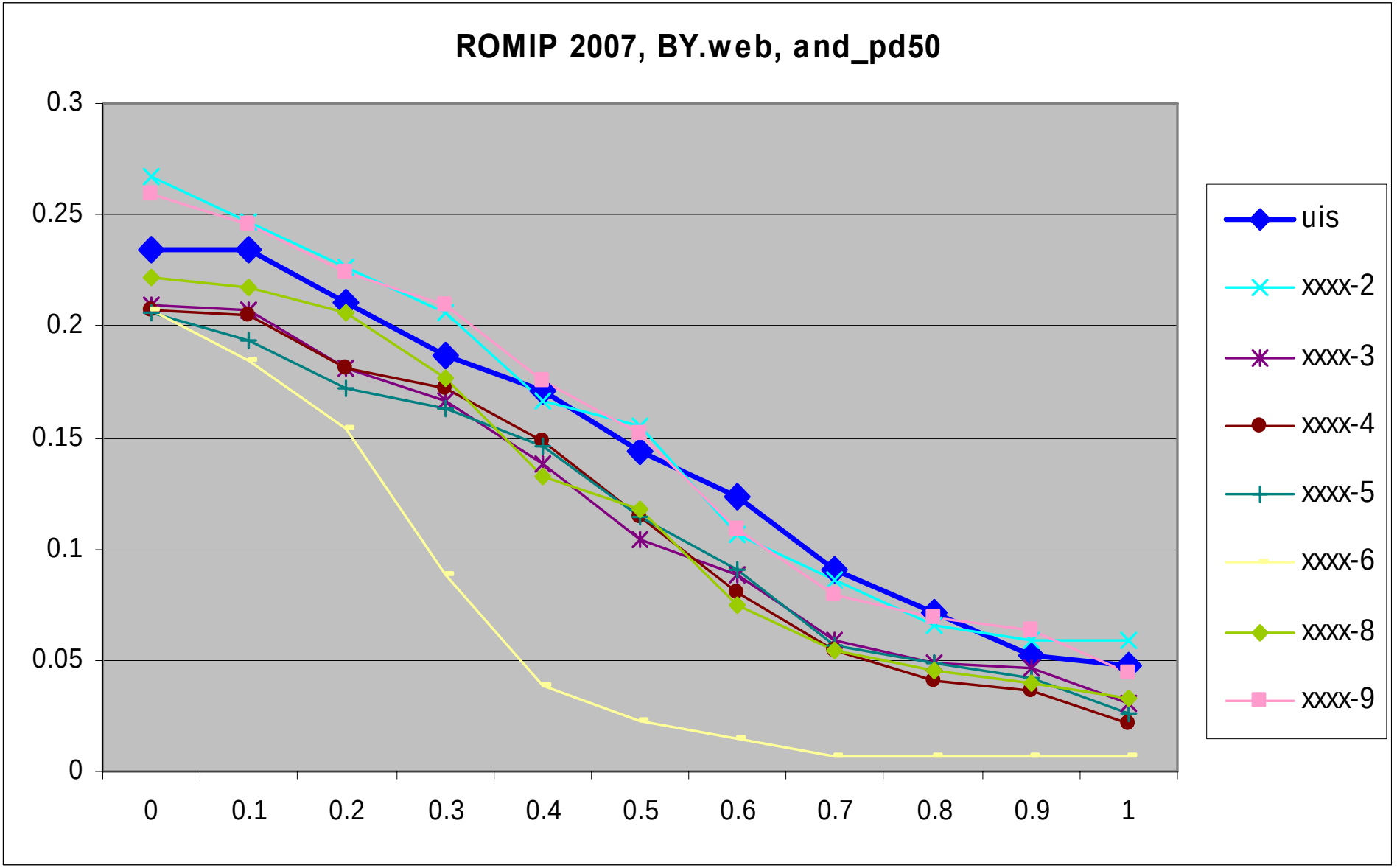
- **Поиск по коллекции нормативных документов “Legal2007”**
  - **348410 документов из БД Кодекс**
  - **14797 запросов из лога запросов к правовому разделу портала www.kodeks.ru**
- **Поиск по коллекции «Беларусский интернет» “BY.web 2007”**
  - **1525585 документов (выборка из страниц домена .by из индекса yandex.ru)**
  - **19627 запросов (смесь запросов из лога yandex.ru , km.ru, оцененных запросов РОМИП 2003-2006)**

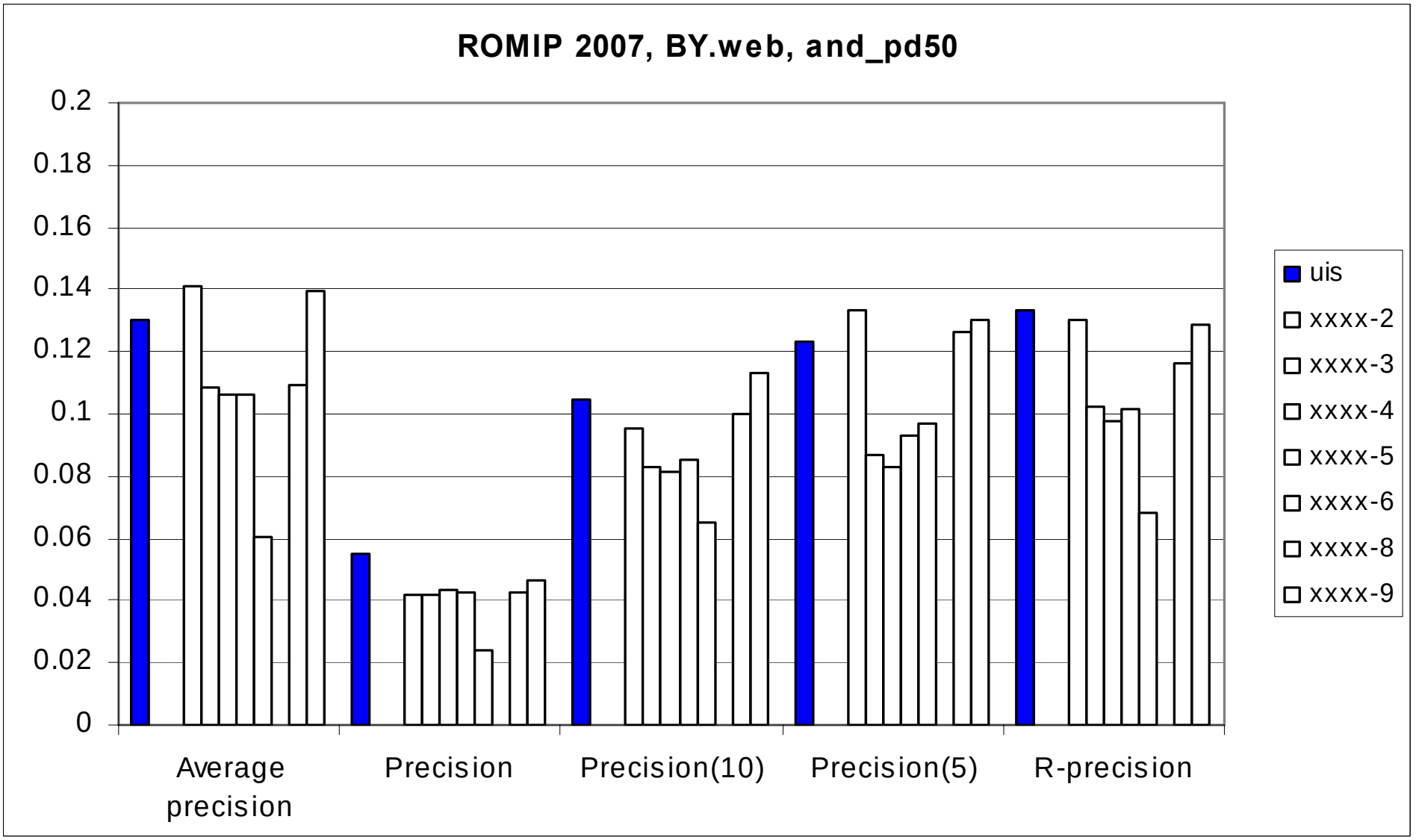
УИС РОССИЯ в РОМИП 2003-2005…2007

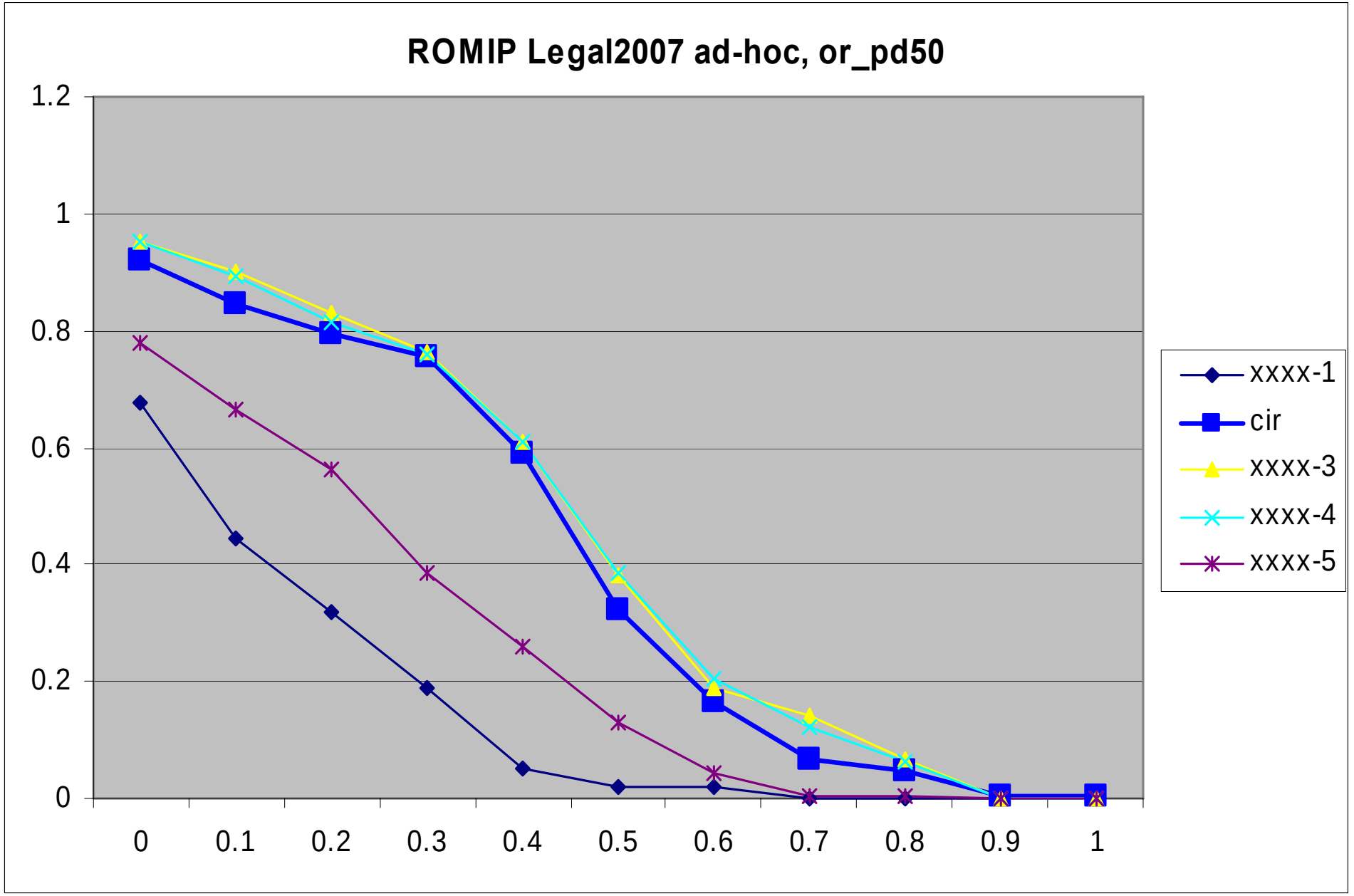
- **Технологии:**
  - **2003-2004: таблицы Oracle, bag-of-words (*)**
  - **2005: (*) + sorted lists: экспериментальные прогоны**
  - **2007: новый индексатор и поисковик (sorted lists)**
- **Эксперименты:**
  - **Учет факторов:**
    - **TF*IDF: (2003-2005)**
    - **минимальное окно: 2005**
    - **кворум: 2005**
    - ***близость по парам: 2007***
  - **Подбор коэффициентов:**
    - ***Красильников П. В. «Воспроизведение лучших результатов ad hoc поиска семинара РОМИП»***

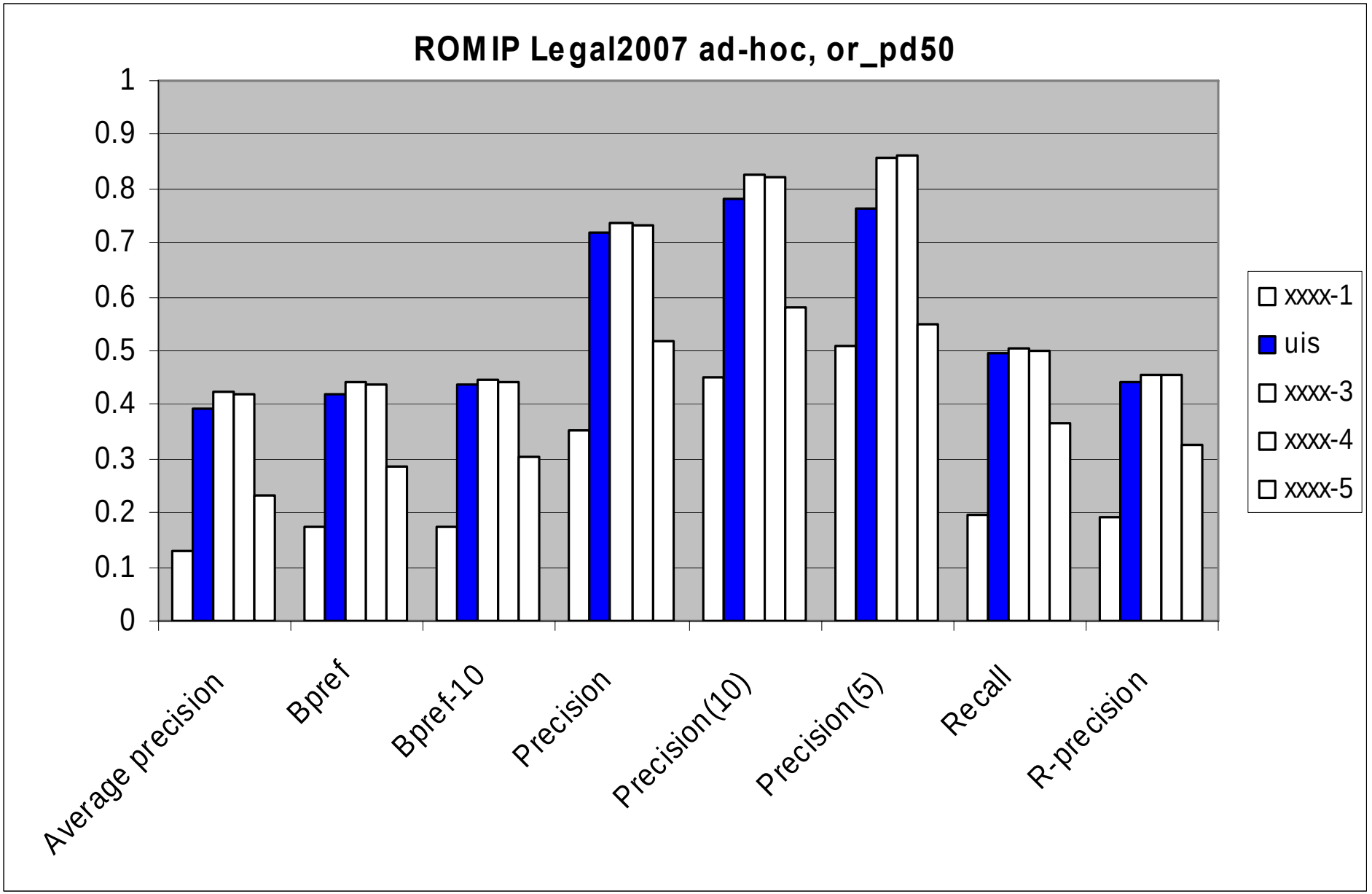

ROMIP 2007, BY.web, or_pd50
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
uis
xxxx-2
xxxx-3
xxxx-4
xxxx-5
xxxx-6
xxxx-8
xxxx-9

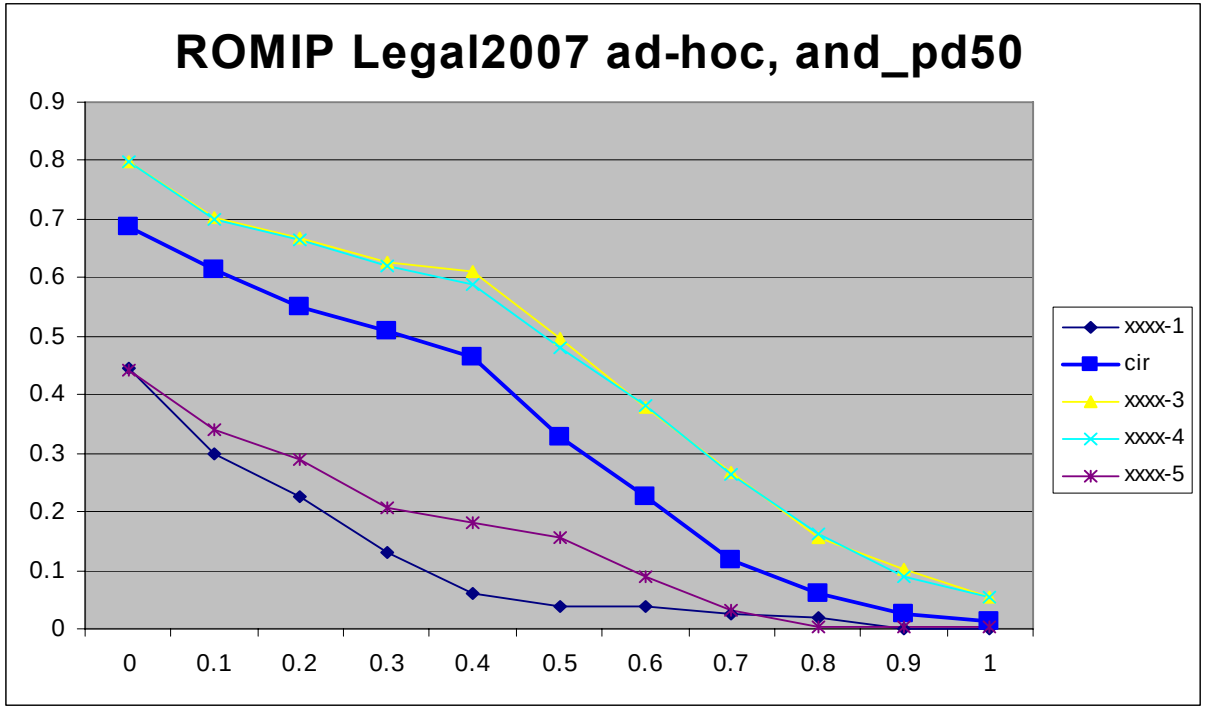

ROMIP 2007, BY.web, or_pd50
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Average precision
Precision
Precision(10)
Precision(5)
Recall
R-precision
uis
xxxx-2
xxxx-3
xxxx-4
xxxx-5
xxxx-6
xxxx-8
xxxx-9

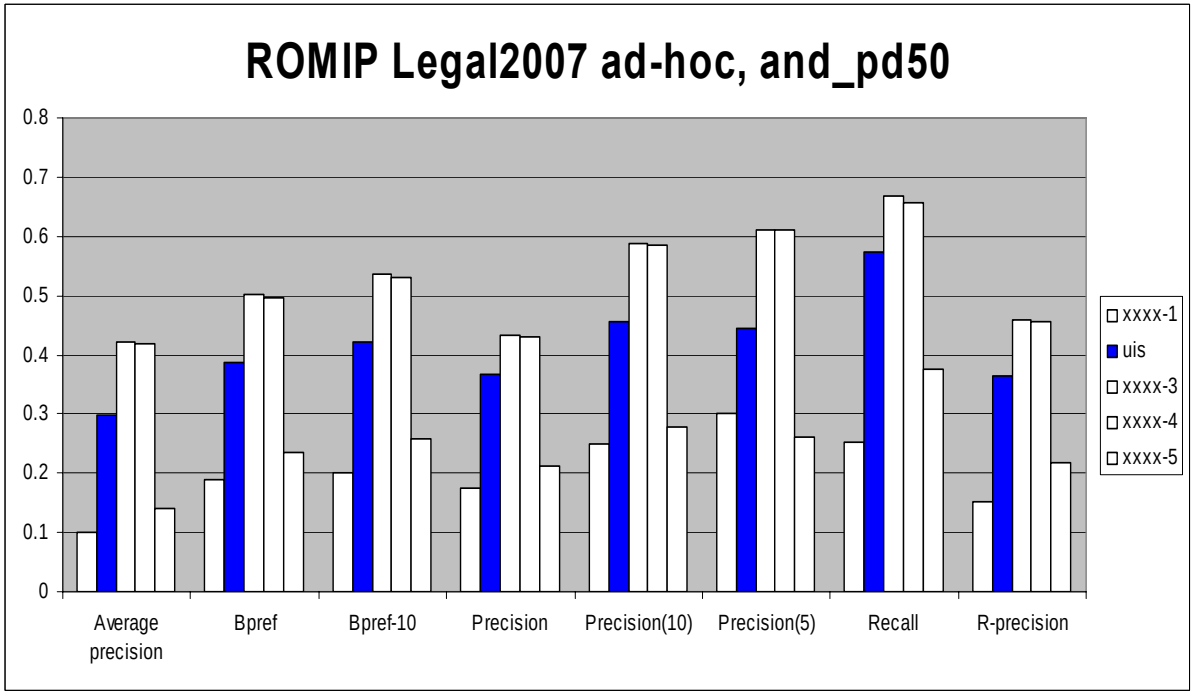

ROMIP 2007, BY.web, and_pd50
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
uis
xxxx-2
xxxx-3
xxxx-4
xxxx-5
xxxx-6
xxxx-8
xxxx-9


ROMIP 2007, BY.web, and_pd50
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Average precision
Precision
Precision(10)
Precision(5)
R-precision
uis
xxxx-2
xxxx-3
xxxx-4
xxxx-5
xxxx-6
xxxx-8
xxxx-9


ROMIP Legal2007 ad-hoc, or_pd50
1.2
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
xxxx-1
cir
xxxx-3
xxxx-4
xxxx-5


ROMIP Legal2007 ad-hoc, or_pd50
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Average precision
Bpref
Bpref-10
Precision
Precision(10)
Precision(5)
Recall
R-precision
xxxx-1
uis
xxxx-3
xxxx-4
xxxx-5


ROMIP Legal2007 ad-hoc, and_pd50
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
xxxx-1
cir
xxxx-3
xxxx-4
xxxx-5


ROMIP Legal2007 ad-hoc, and_pd50
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Average precision
Bpref
Bpref-10
Precision
Precision(10)
Precision(5)
Recall
R-precision
xxxx-1
uis
xxxx-3
xxxx-4
xxxx-5

Анализ ad-hoc

- Лучшие результаты получены по метрике Average Precision, or_pd50
  - По этой же метрике и обучали алгоритм
- Результаты по P(10) и/или and_pd50 заметно хуже
  - Обучаться по этим метрикам?
  - Проанализировать найденные документы!
  - Можно ли улучшить одновременно по всем метрикам?
- Замешивание оценок по КМ и смешанной коллекции – как интерпретировать результаты?

Ad-hoc: выводы

- Участие в РОМИП'2007 позволило отладить новые технологические модули и проверить эффективность формулы ранжирования
- Предложенная в статье «Воспроизведение лучших результатов…» формула ранжирования показывает хорошие результаты поиска
- Дальнейшие планы:
  - Учет новых факторов: relevance feedback, структура (заголовки, ссылки)
  - Анализ отдельных запросов – поиск «равномерного» улучшения