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

АНО Центр
информационных
исследований

Университетская
информационная система
РОССИЯ

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

Экспериментальные алгоритма поиска/классификации и сравнение с «basic line»

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.

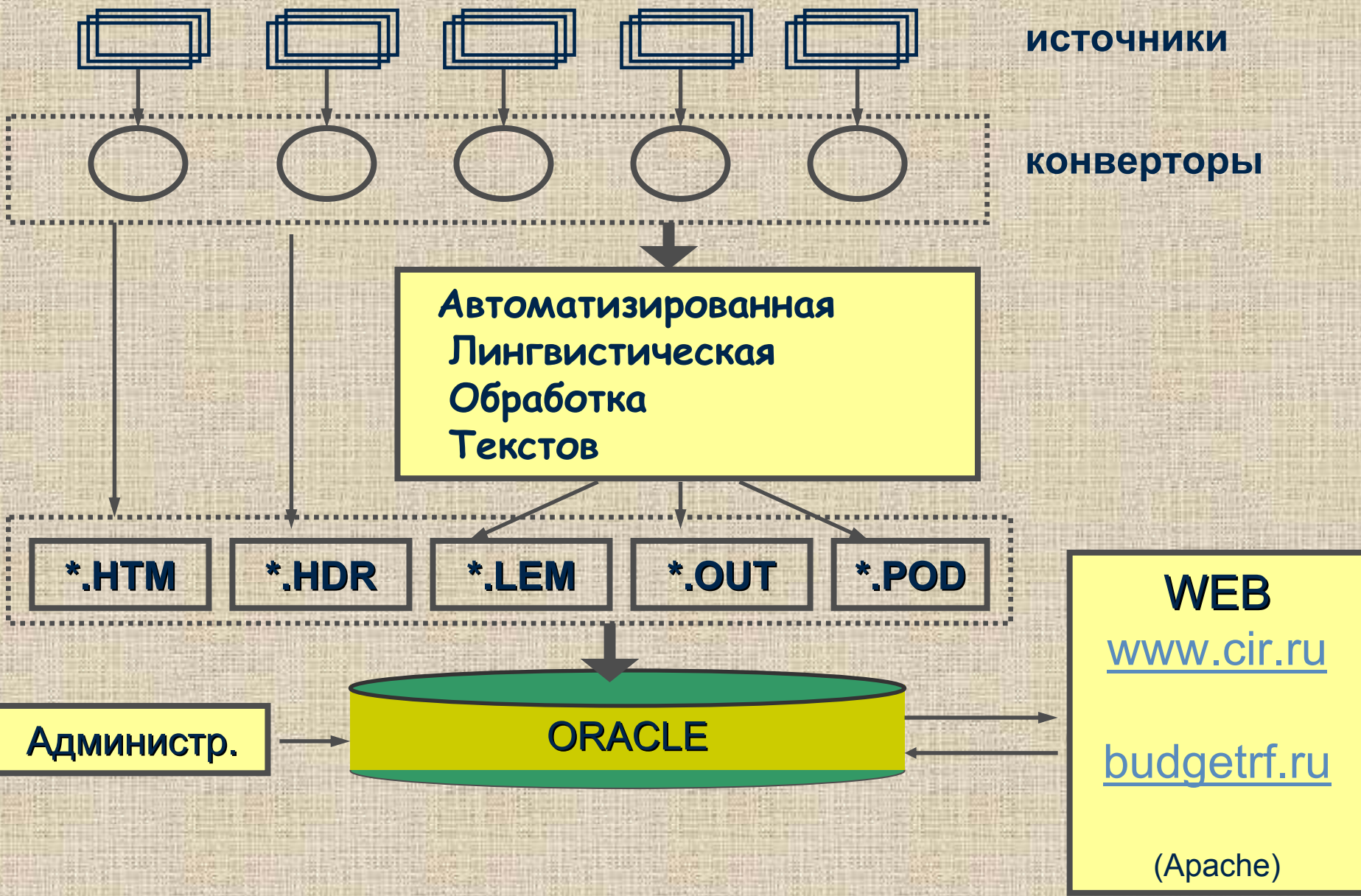
Перепись населения в СССР в 1939 году - предоставлены Университетом Торонто, Канада.

Парламент России

ГосДума РФ. Стенограммы пленарных заседаний |

ГосДума РФ. Информационно-Аналитический бюллетень |

Потоки данных в УИС РОССИЯ



Деловая проза

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
НПА федерального уровня											
СМИ											
НПА регионов											
English											
Стенограммы ГД											
Статистика											
Научные статьи											
Ведомственные НПА											
Интернет											

USER:
BORIS
Доступ: **CIR**
Имя: Пароль:

Регистрация
Забыли пароль?

Поиск по ИСТОЧНИКАМ: нет выбранных коллекций

искать ▶ | очистить | история

добавление условия в запрос

AND | Тезаурус ЦИИ | добавить | список...

спец. атрибуты — свернуть

Выберите коллекции документов, по которым будет производиться поиск. Переход по ссылке с имени коллекции позволяет выбрать атрибуты поиска, специфичные для данной коллекции.

- ☐ ***Все коллекции*** свернуть/развернуть список
  - ☐ ***Издания государственных органов*** список коллекций... описание
  - ☑ ***Средства массовой информации*** свернуть список описание
    - ☑ Аргументы и Факты *(21133 статьи, с 1997 года)* описание
    - ☑ Известия *(53048 статей, с 2000 года)* описание
    - ☑ Финансовые Известия *(1683 документа, с 2001 года)* описание
    - ☑ Ведомости *(51791 статья, с 1999 года)* описание
    - ☑ Комсомольская правда *(41870 статей, с 1999 года)* описание
    - ☑ Независимая газета *(89041 статья, с 1998 года)* описание
    - ☑ Слово *(2502 статьи, с 1999 года)* описание
    - ☑ Сегодня *(17072 статьи, с 2000 года по апрель 2001 года. Издание прекращено)* описание
    - ☑ Региональный пресс-бюллетень агентства ВПС *(11883 статьи, с 1990 года по январь 2001 года. Издание прекращено)* описание
    - ☑ Журнал "Эксперт" *(12205 статей, с 2001 года)* описание
    - ☑ Газета "Поиск" *(2103 статьи, с 2002 года)* описание
    - ☑ Материалы агентства Reuters *(21578 документов)*

Аналитическая работа в УИС РОССИЯ

Файл Правка Вид Избранное Сервис Справка Ссылки

УИС РОССИЯ университетская информационная система РОССИЯ

Главная | Поиск | Статистика РФ | Бюджет РФ | Выборы РФ | Парламент РФ | Соционет/RePEc | OECD | Технологии | Справка/Практикум | УИС РОССИЯ v3 | UIS RUSSIA v3 |

USER: **BORIS**
Доступ: **CIR**
Имя: Пароль:

Регистрация
Забыли пароль?

Поиск по ИСТОЧНИКАМ: Программа "Университеты России". Отчеты

/Термин_расш="ГЕОЛОГИЧЕСКИЕ НАУКИ"

искать ▶ | очистить | история

добавление условия в запрос

AND | Тезаурус ЦИИ | добавить | список...

Атрибуты колл. "Программа "Университеты ..."

Найдено 107 документов. Показано, начиная с 1. (10 док./стр.)

сортировать по релевантности | убрать подсветку

Научная программа "Университеты России", Раздел: 7.2. Экология., отчет за 2002-03 гг., Естественнонаучный институт при Пермском государственном университете (95%)

Эволюция литосферы и формирование современной экологической обстановки

Научная программа "Университеты России", Раздел: 9. "*Геология*.", отчет за 2002-03 гг., Воронежский государственный университет (95%)

Теоретические основы эколого-*геологического* мониторинга - базы создания постоянно действующих моделей экогеосистем.

Научная программа "Университеты России", Раздел: 9. "*Геология*.", отчет за 2002-03 гг., *Геологический факультет* Московского государственного университета им

Анализ результатов запроса по классификатору "Организация" (107 документов):

Организация

+/–		Организация
+ (13)	– (94)	Санкт-Петербургский государственный университет (66)
+ (6)	– (101)	Московский государственный университет им. М.В.Ломоносова (14)
+ (6)	– (101)	Новосибирский государственный университет (35)
+ (5)	– (102)	Воронежский государственный университет (20)
		Томский

УИС РОССИЯ в РОМИП-2004

- ad-hoc для web-коллекции (2)
- ad-hoc для legal-коллекции (2)
- классификация в legal-коллекции (3)

Прежний курс: получение «basic line» с использованием «классических» методов и исследование различных факторов

ad-hoc для web-коллекции

Прогон 1: «Отправная точка»

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, $\beta = 0.4$

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

ad-hoc для web-коллекции

Каждый запрос $Q = w_1 \ w_2 \ w_3 \ldots \ w_m$ 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.*

ad-hoc для web-коллекции

Прогон 2: Влияние расстояния между словами

$$\mathrm{Rank}_D(Q) = \frac{\mathrm{V}_D(Q) + \mathrm{Near}_D(Q)}{2}$$

1Если запрос встречается в заголовке документа (внутри тегов title и h1) в виде подстроки, то . $\mathrm{Near}_D(Q) = 2$

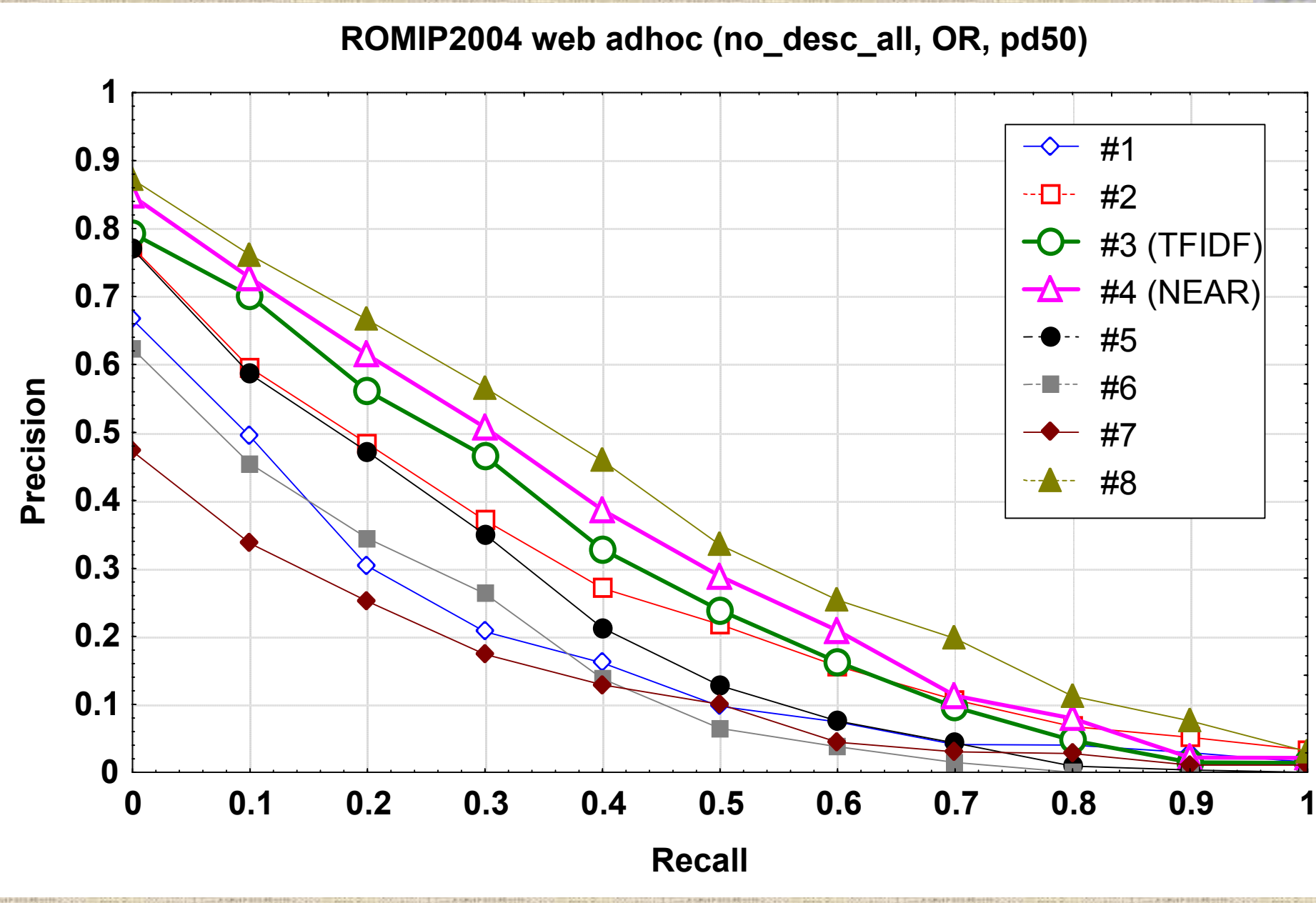
Иначе, если запрос встречается внутри документа в виде подстроки, то .

$$\mathrm{Near}_D(Q) = 1$$

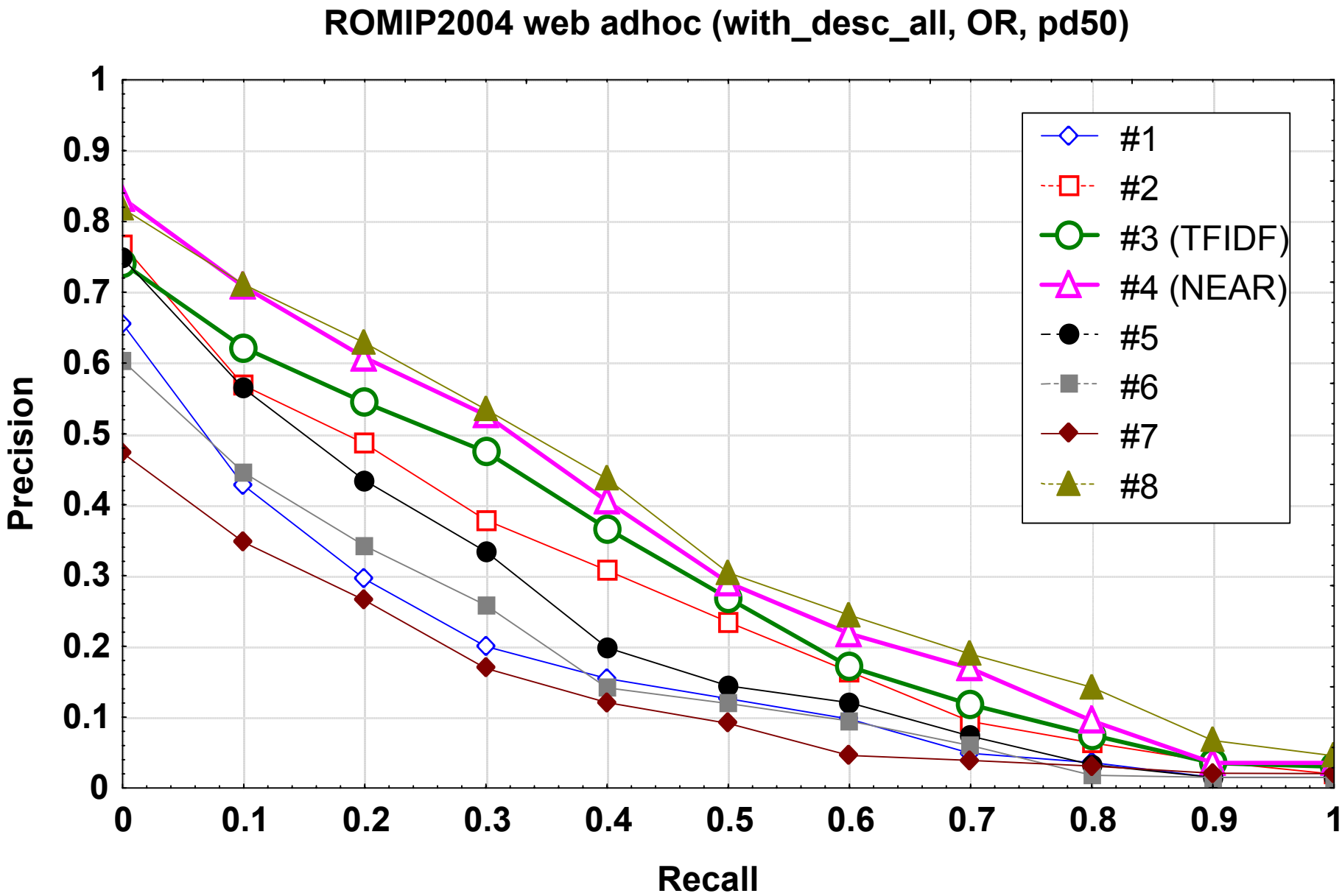
Иначе, производится поиск минимального «куска» документа, в котором содержатся все слова запроса.

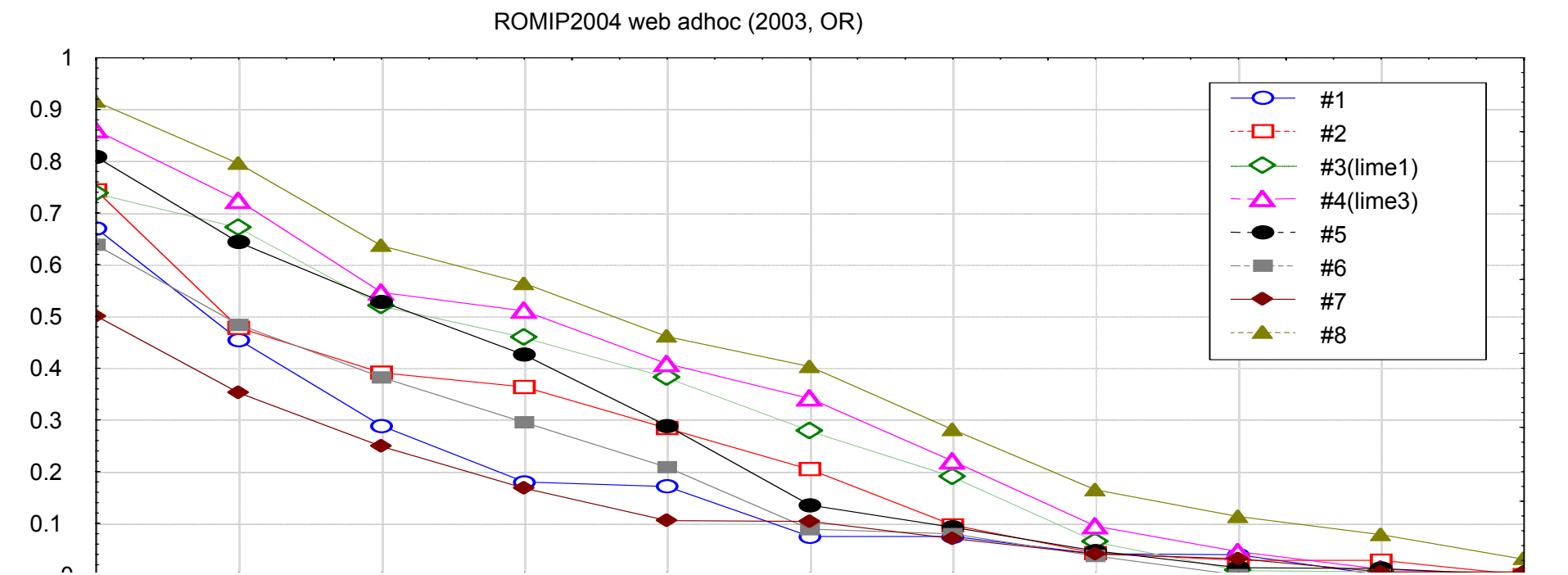
$$\mathrm{Near}_D(Q) = \frac{1}{\ln\left(\lambda_D(Q) - |Q| + 4\right)}$$

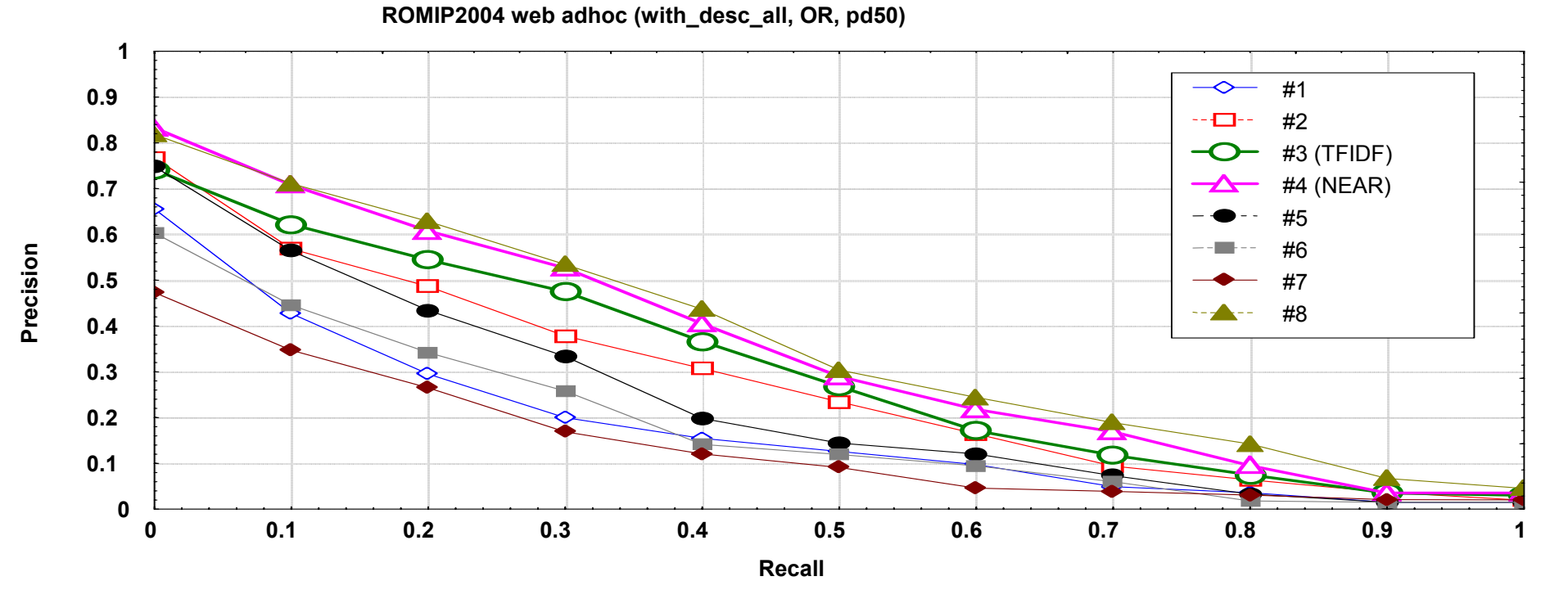
ad-hoc для web-коллекции



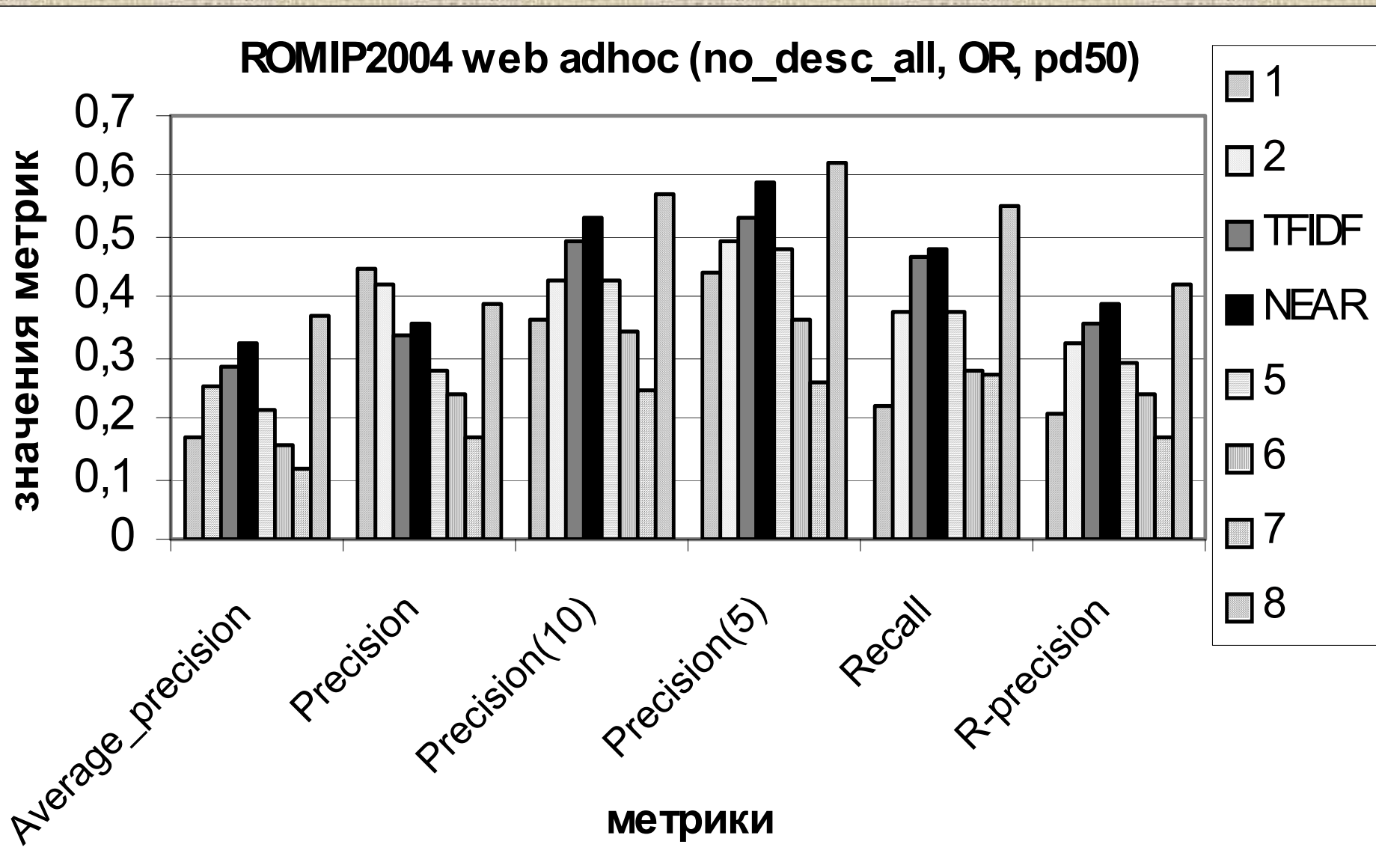
ad-hoc для web-коллекции



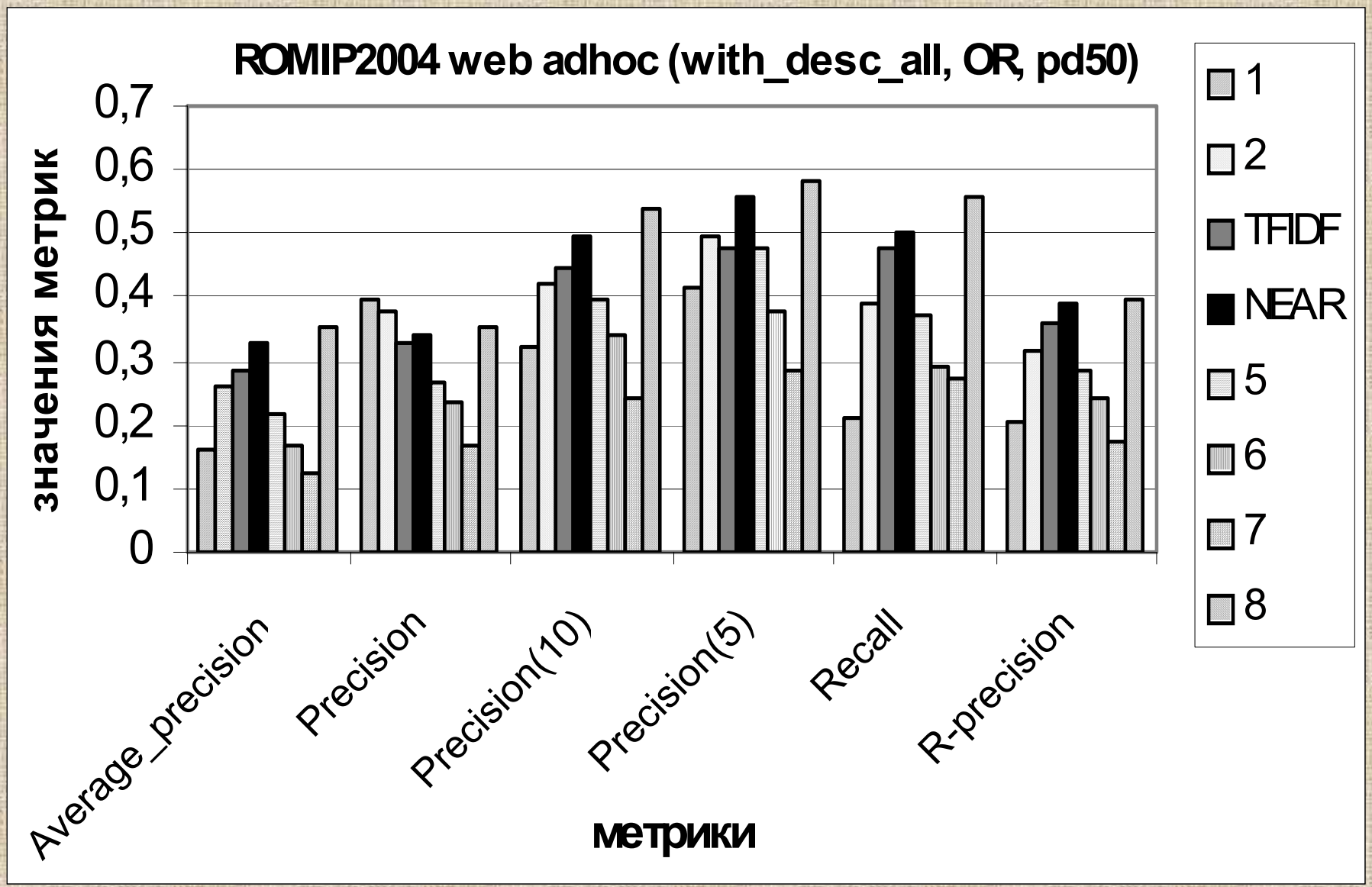

ROMIP2004 web adhoc (2003, OR)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
#1
#2
#3(lime1)
#4(lime3)
#5
#6
#7
#8


ROMIP2004 web adhoc (with_desc_all, OR, pd50)
Precision
1
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
Recall
#1
#2
#3 (TFIDF)
#4 (NEAR)
#5
#6
#7
#8

ad-hoc для web-коллекции



ad-hoc для web-коллекции



ad-hoc для web-коллекции

OR		
1	no_desc	summary
2	no_desc	summary_pd50
3	no_desc_all	summary
4	no_desc_all	summary_pd50
5	no_desc2003	summary
6	no_desc2003	summary_pd50
7	with_desc	summary
8	with_desc	summary_pd50
9	with_desc_all	summary
10	with_desc_all	summary_pd50
11	with_desc2003	summary
12	with_desc2003	summary_pd50

AND		
13	no_desc	summary
14	no_desc	summary_pd50
15	no_desc_all	summary
16	no_desc_all	summary_pd50
17	no_desc2003	summary
18	no_desc2003	summary_pd50
19	with_desc	summary
20	with_desc	summary_pd50
21	with_desc_all	summary
22	with_desc_all	summary_pd50
23	with_desc2003	summary
24	with_desc2003	summary_pd50

ad-hoc для web-коллекции

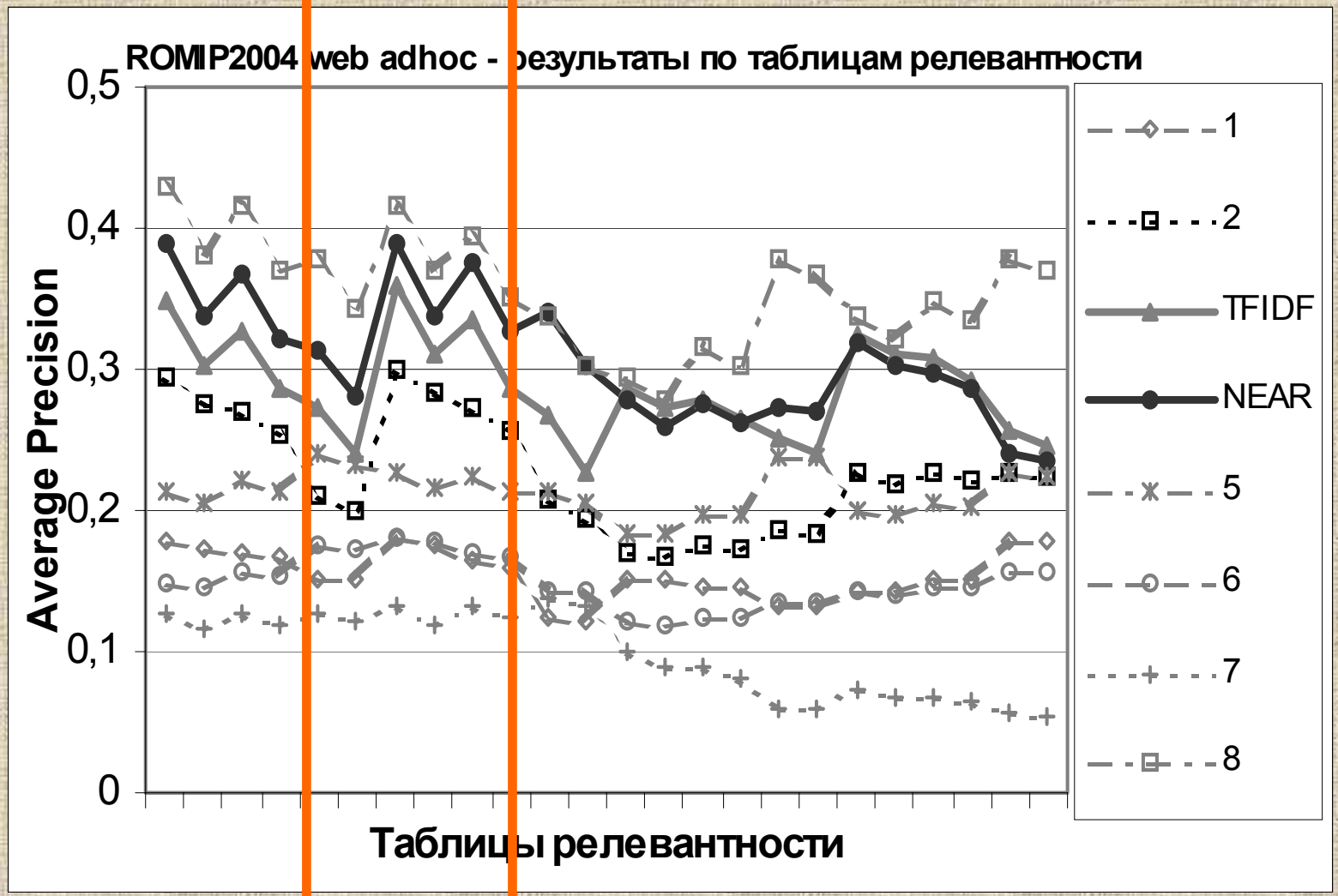


График зависимости Average Precision от таблицы релевантности. Вертикальными линиями отмечены таблицы релевантности, выбранные для анализа.

ad-hoc для web-коллекции

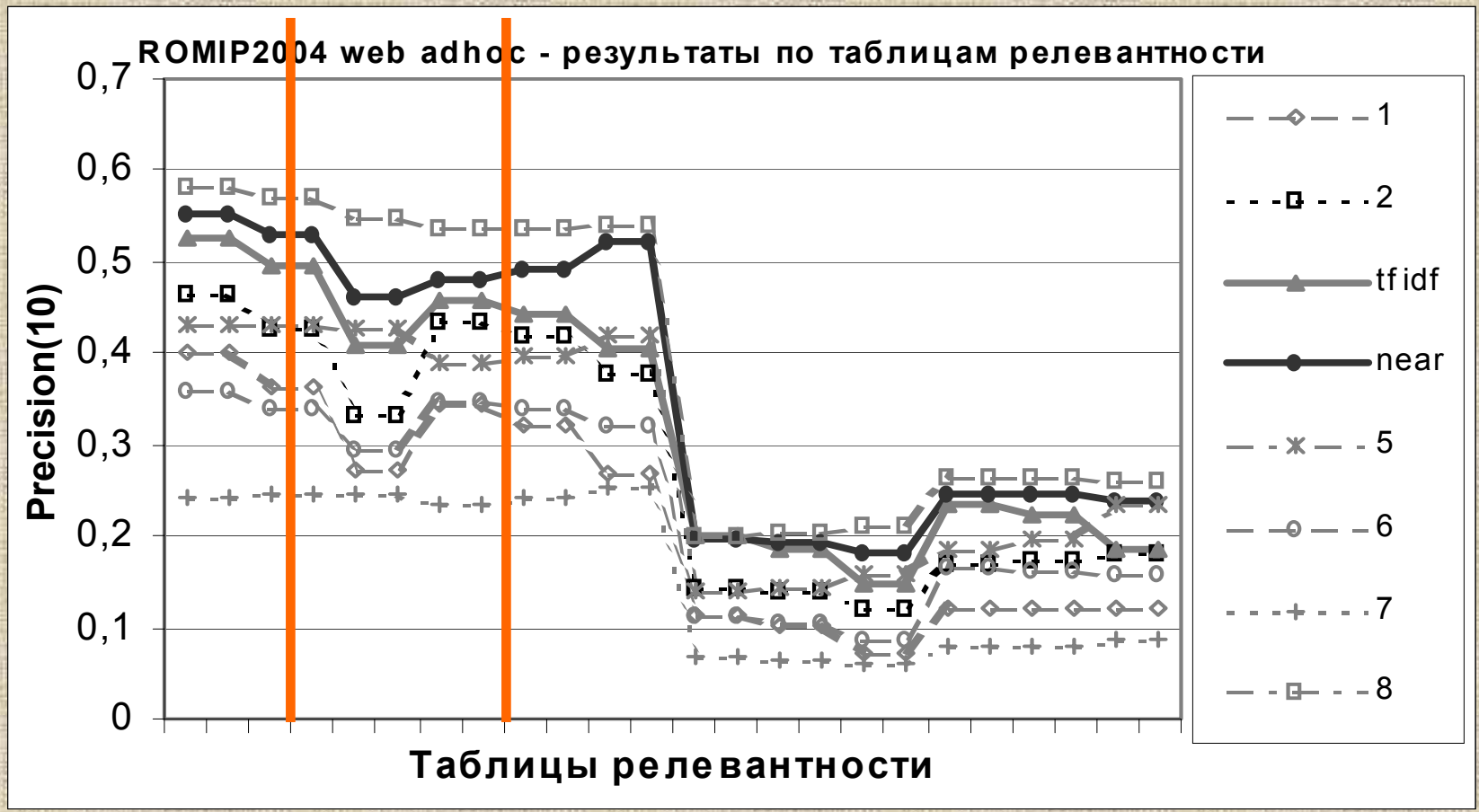


График зависимости Precision(10) от таблицы релевантности. Вертикальными линиями отмечены таблицы релевантности, выбранные для анализа.

ad-hoc для web-коллекции

Выводы по устойчивости:

- **«Места» различных прогонов меняются в основном в пределах одной позиции в зависимости от таблицы релевантности. Поэтому результаты анализа, проведенного нами на двух выделенных таблицах релевантности, можно обобщить на все таблицы.**

- **Точность системы на первых 10 документах существенно зависит от способа объединения оценок разных экспертов.**
  Для матриц «or» Precision(10) существенно выше, чем для матриц «and»

ad-hoc для legal-коллекции

- задача классического ad hoc поиска документов
- около 60 тысяч документов (согласно нашим данным 60015) из коллекции нормативных документов РФ компании «Кодекс»
- 12925 запросов, для каждого из которых необходимо было выполнить поиск и выдать не более 100 документов
- задание и способ его оценки для этой дорожки аналогичны дорожке поиска по web-коллекции
- мы использовали аналогичные методы выполнения заданий (за исключением прогона 2) и анализа результатов

ad-hoc для legal-коллекции

Прогон 2: Повышения веса слов запроса, встретившихся в заголовках

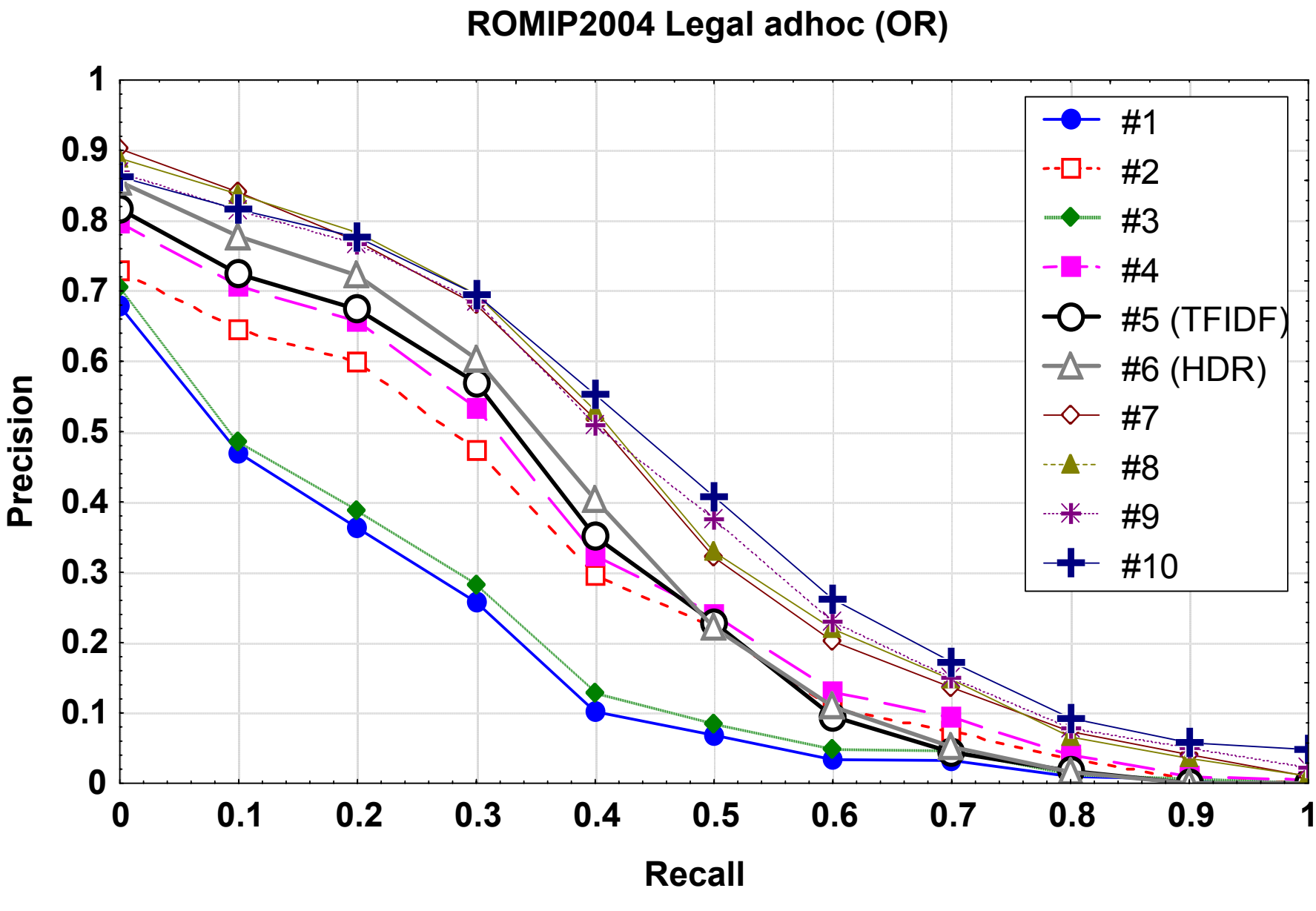
$$\mathrm{Rank}_D(Q) = \frac{\mathrm{V}_D(Q) + \mathrm{HdrFreq}_D(Q)}{2}$$

Ранг по заголовкам равен отношению количества слов запроса, встретившихся в заголовке, к длине запроса:

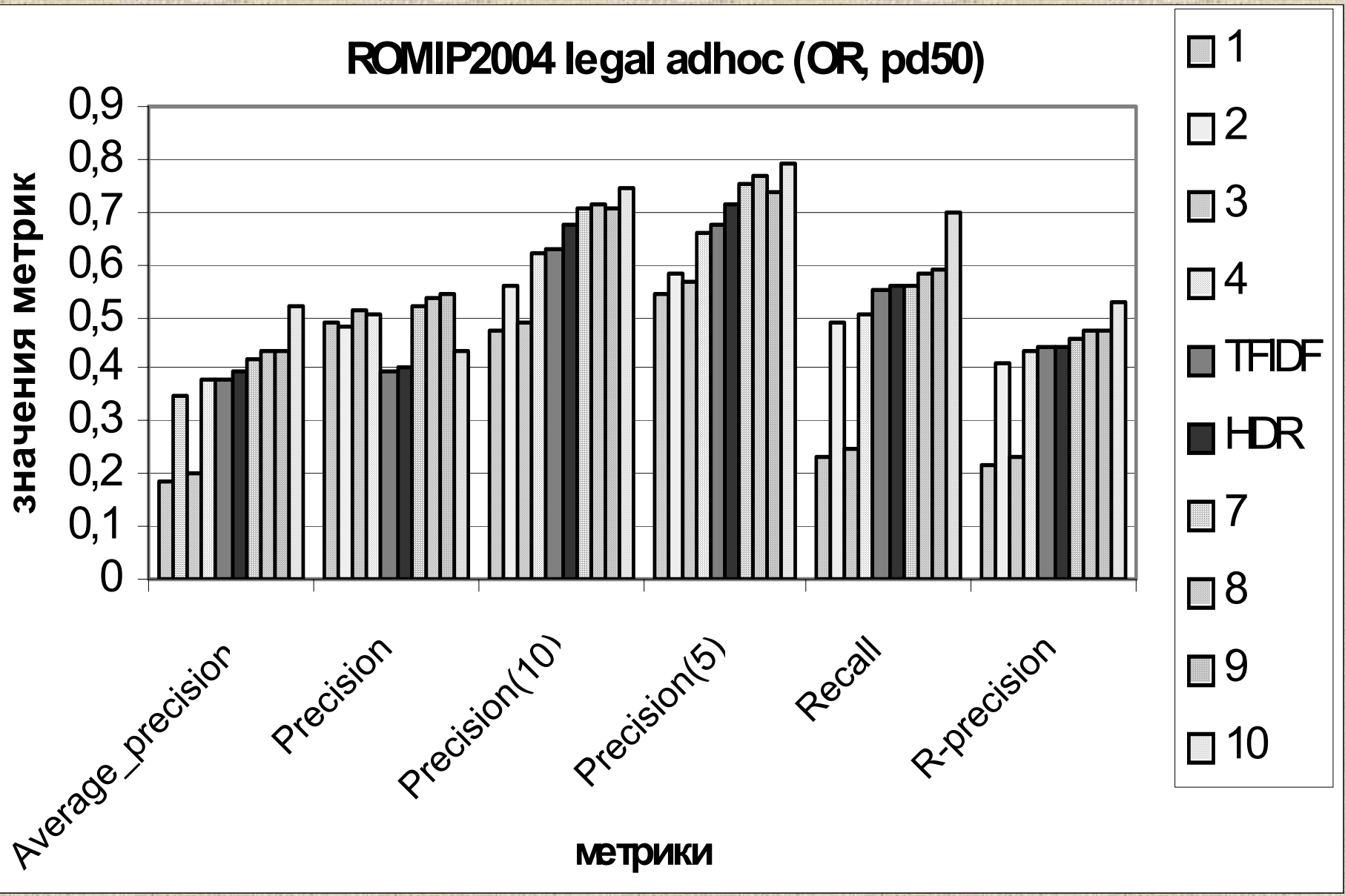
$$\mathrm{HdrFreq}_D(Q) = \frac{|\mathrm{HdrWords} \cap Q|}{|Q|}$$

где HdrWords — множество слов в заголовке документа, — количество слов в запросе.

ad-hoc для legal-коллекции



ad-hoc для legal-коллекции



ad-hoc для legal-коллекции

Результаты :

- **Прогон 1, использующий «классический» метод поиска и ранжирования TF*IDF показывает «средние» результаты**

- **Учёт заголовков повышает качество поиска, но не намного**

ad-hoc для web-коллекции

что такое тюнинг
кинотеатр мечта
все про массаж
породы кошек **классификация**
что такое lol
как приготовить пиццу
как получить российское гражданство
Научная полемика
адрес Птичий рынок **в Москве**
мерседес вито
витамины в пище **человека**
мерседес из германии
что взять в роддом
нобелевская премия
как накачать мышцы
биография петра
Напиток богов амброзия
что такое любовь
самый мощный ваз
болезни сердца

ad-hoc для legal-коллекции

лишение премии
Комментарий к закону об инвестиционной деятельности
о лицензировании отдельных видов **деятелдьности**
поправка к закону о гражданстве РФ
О подготовке и проведению празднования 60-й годовщины Победы
Перечень производств, профессий и работ с тяжелыми и вредными, **особотяжелыми** и особо вредными условиями труда предприятий и организаций
Положение о порядке сдачи квалификационного экзамена и оценки **знанийпретендентов**
Закон Об **аудиторкой** деятельности
устный договор
закон о переписи населения#57-ФЗ
Положение о паспорте гражданина РФ
Об отмене положения о составе затрат
судебные издержки
коммунальная квартира
Об ограничении курения табака и потребления табачных изделий.
уклонение от заключения договора
о страховании гражданской ответственности авиаперевозчиков
федеральный закон о рекламе
Регистрация инвестиционных контрактов

классификация для legal-коллекции

- автоматическая классификация нормативных документов законодательства РФ из БД СПС «Кодекс»
- 183 рубрик -- подмножество большого иерархического рубрикатора нормативных документов
- для обучения процедуры классификации предлагается коллекция из 4496 документов, отрубрицированных по данному классификатору экспертами компании «Кодекс»
- для тестирования предоставлены 55519 документов, для которых необходимо автоматически определить рубрики, к которым эти документы относятся.
- для некоторых рубрик нет документов в коллекции обучения, всего рубрик с ненулевым количеством документов для обучения — 170

классификация для legal-коллекции

Прогон 1: SVM по леммам
Прогон 2: SVM по леммам+терминам

Леммы/понятия, встречающиеся менее, чем в четырёх документах, были усечены.

21746 различных лемм и 1203087 пар лемма-документ для обучающей выборки из 4496 документов.

29918 различных лемм/терминов и 1569958 пар «лемма/термин»-документ.

классификация для legal-коллекции

Прогон 3: Метод машинного обучения, основанный на моделировании логики рубрикатора

- **описание рубрики в виде булевской формулы — запроса к ИПС Элементами формул являются понятия Тезауруса ЦИИ.**
- **Алгоритм строит формулы вида**

$$U = \bigcup_{i=1}^{k} \bigcap_{j=1}^{J_i} t_{i,j}$$

- **Конъюнкции, составляющие формулу, имеют длину от 1 до 3.**
- **Мотивация -- создать алгоритм машинного обучения, который бы моделировал смысл рубрики, составленной человеком, по результатам рубрицирования. Необходимым требованием для данного алгоритма было построение правил описания рубрики, которые можно легко интерпретировать.**

$$U = \bigcup_{i} \bigcap_{j} \left(\bigcup_{k} t_{i,j,k} \setminus \bigcup_{l} t'_{i,j,l} \right)$$

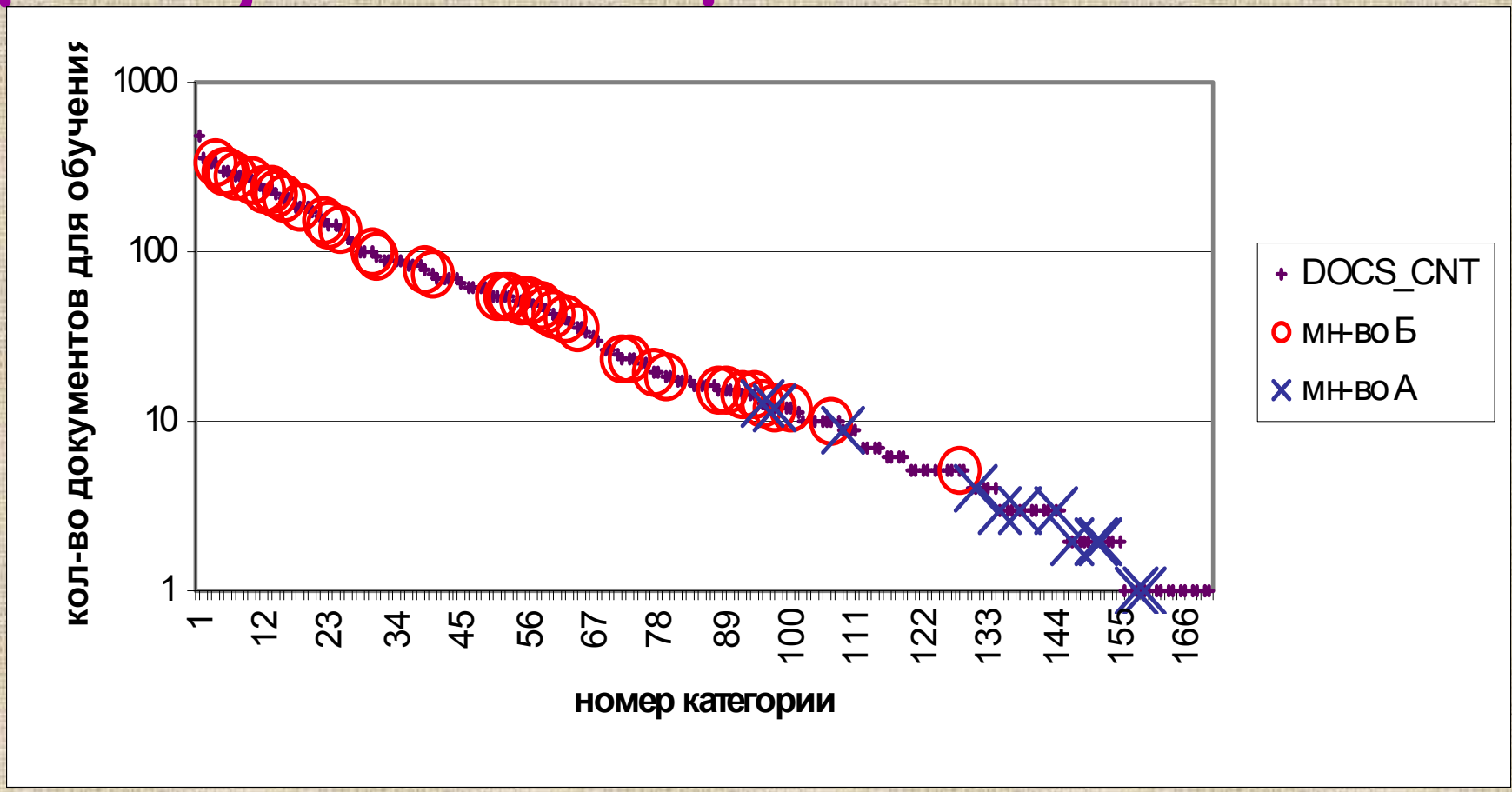
классификация для legal-коллекции

Методика оценки

- большое разнообразие таблиц релевантности (4 варианта) и метрик качества рубрицирования (8 шт.)
- результаты работы систем вычислялись на двух различных подмножествах рубрик, назовём их А (12 шт.) и Б (40 шт.)

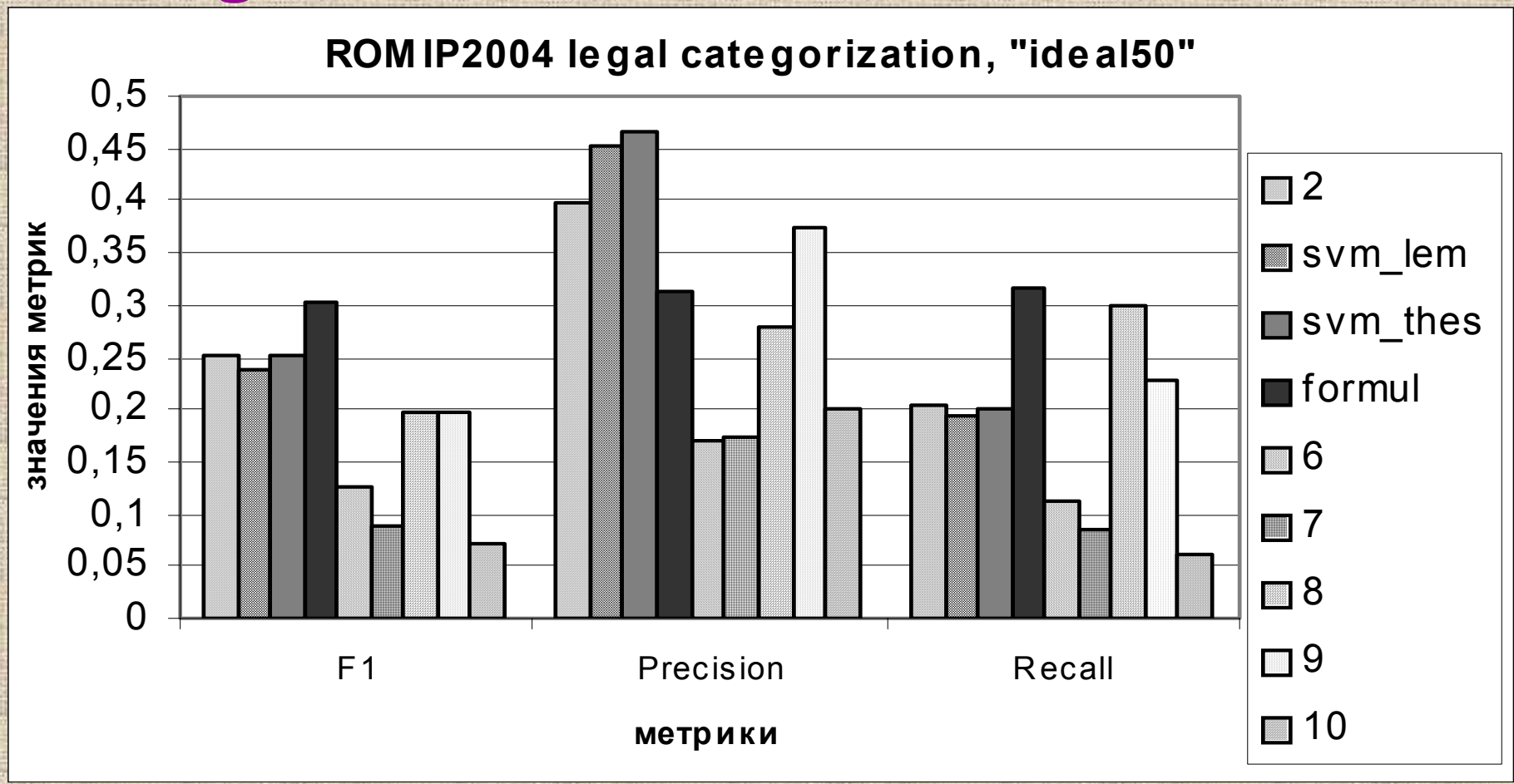
1) “ideal” — оценки экспертов ИС «Кодекс» для А
2) “ideal40” — оценки экспертов ИС «Кодекс» для Б;
3) “and_relevant-minus” — “ideal”; документ считается соответствующем рубрике, если хотя бы один эксперт оценил его как relevant-minus или выше;
4) “or_relevant-minus” — “ideal”; документ считается соответствующем рубрике, если все эксперты оценили его как relevant-minus или выше

классификация для legal-коллекции



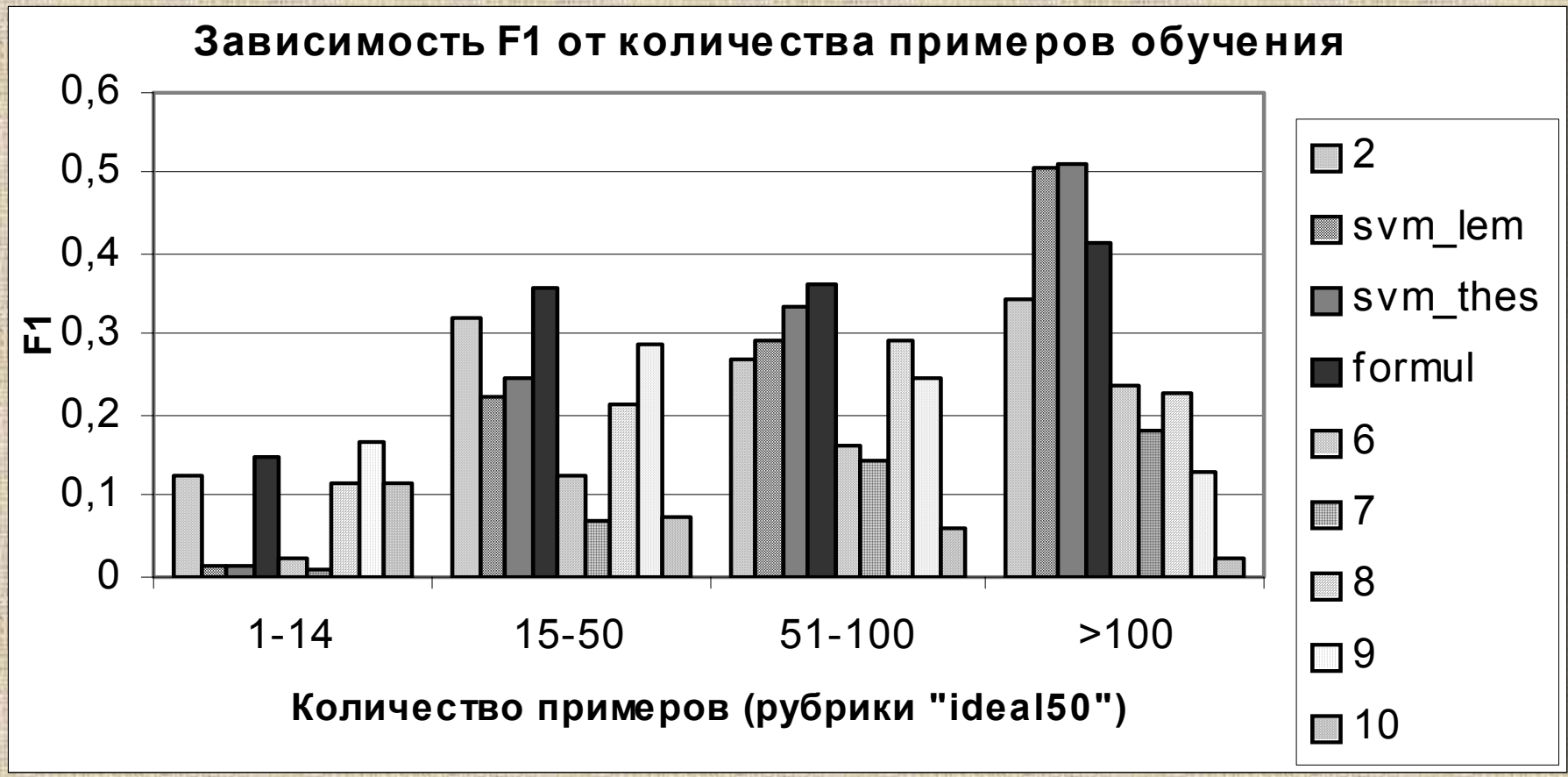
Таблицу релевантности, состоящую из оценок, проставленных экспертами ИС «Кодекс» для рубрик из А ⋃ Б будем обозначать “ideal50”.

классификация для legal-коллекции



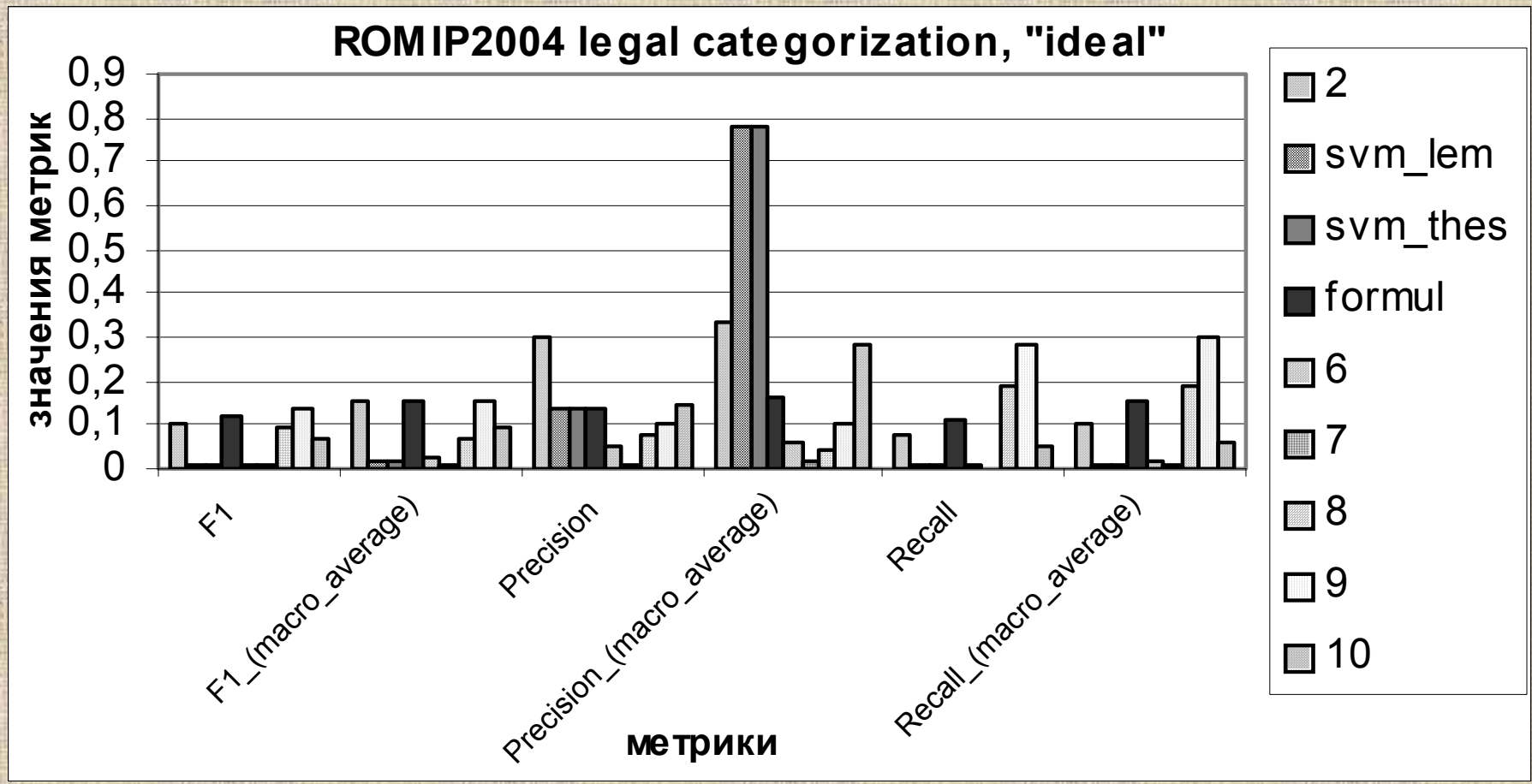
Результаты прогонов участников для таблицы релевантности “ideal50”

классификация для legal-коллекции



Зависимость F-меры от количества примеров для обучения (в среднем для рубрик, частотность которых попадает в указанный интервал)

классификация для legal-коллекции



Результаты прогонов участников для таблицы релевантности “ideal”

Заключение

Предварительный анализ результатов:

- **широко известные, описанные в литературе методы, как TF*IDF для adhoc поиска и Support Vector Machine для классификации с обучением, показывают достаточно высокие результаты;**
- **интересной задачей является разработка либо совершенно новых методов либо гибридных методов, превосходящих по результатам «классические» методы**

УИС Россия

Университетская информационная система

РОССИЯ