Оценка качества информационного поиска

What you can't measure

you can't improve

Lord Kelvin

Мера качества информационного поиска

- Удовлетворенность пользователя – user happiness
  - Скорость ответа важна – легко измеряется
  - Как измерить качество?

Маннинг и др. Введение в информационный поиск – гл. 8.

Картинки из «Advances in Information Retrieval Evaluation” – RUSSIR-2011

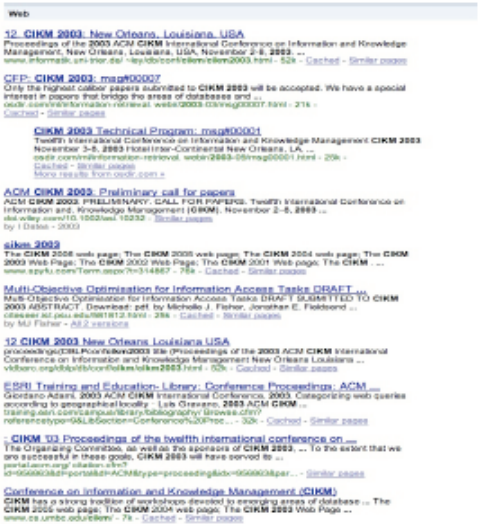
Измерение удовлетворенности

- Приближение: релевантность
- Как измерить
  - Коллекция документов,
  - Коллекция запросов
  - Оценки релевантен/нерелевантен или более подробная оценка

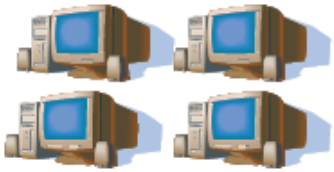
Traditional Experiment

- uses of alternative dispute resolution
- job search vancouver washington
- poem of arrival of columbus

Results



Search Engines



many good docs
I missed/found?

Judges

Эксперименты по оценке качества поиска

- Кренфилдские (Cranfield) эксперименты (1966)
- Text REtrieval Conference (TREC) (1992)
- Исследования основ оценки на базе (TREC) (1998-2001-…)
- NII Test Collection for IR Systems (NTCIR) (1999)
- Cross Language Evaluations Forum (CLEF) (2000)
- Российский семинар по оценке Методов Информационного Поиска (РОМИП) (2003)

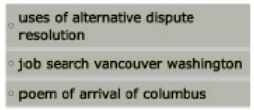
Классическая (Cranfield) процедура оценки

- Составим список запросов и ограничим коллекцию документов
- Для каждой пары запрос/документ выставим экспертную оценку «релевантности»
- Будем рассматривать ответ системы не как последовательность документов, а как множество/последовательность оценок релевантности
- На полученной последовательности/множестве оценок релевантности построим метрики

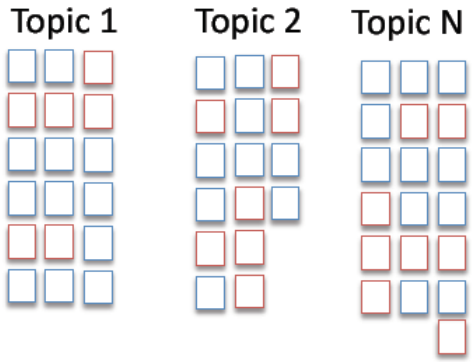
Portable Test Collection

- Document Corpus

- Topics



- Relevance Judgments (QRELs)



Оценка релевантности выдачи

- Информационная потребность выражается запросом
- Релевантность оценивается по отношению к информационной потребности, а не к словам запроса
- Т.е. все слова запроса могут присутствовать в документе, а документ не релевантен

Оценка булевского поиска

- Булевский поиск – не имеет ранжирования (упорядочения)
- Поисковая система разделяет коллекцию на два множества
  - Выдано ответ на запрос – не выдано
  - Эксперты: релевантен – нерелевантен
- Меры качества:
  - Точность
  - Полнота
  - F-мера

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R}$$

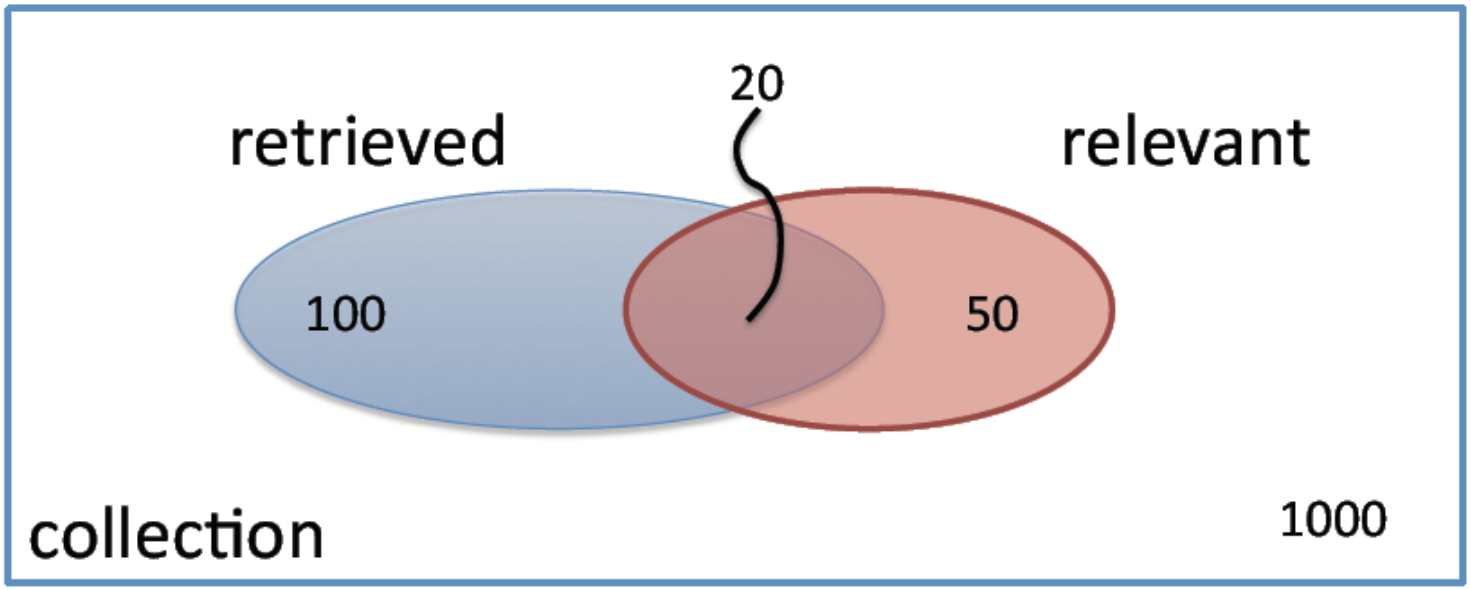
Оценка неранжированного поиска

- **Precision (точность)**: доля релевантных документов в выданных: P(relevant|retrieved)
- **Recall (полнота)**: доля выданных документов среди релевантных докуменов= P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Measuring Boolean Output



Precision = 20/100 = 0.2

Recall = 20/50 = 0.4

Fallout = (100-20)/(1000-50) = 0.08

Полнота/Точность

- Можно получить 100% полноту, но очень низкую точность, если выдать все документы коллекции
- Обычно точность падает, чем больше документов выдано (в хороших системах)

Комбинированная мера: F-мера

- Среднее гармоническое между полнотой и точностью

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- Обычно сбалансированная F-мера:
  - β=1 или α=1/2

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R}$$

Задача

- Эксперт нашел по запросу 200 документов
- Система – 100 документов, 50 из них правильно
- Найти точность, полноту и Fмеру поиска

Оценка ранжированных результатов

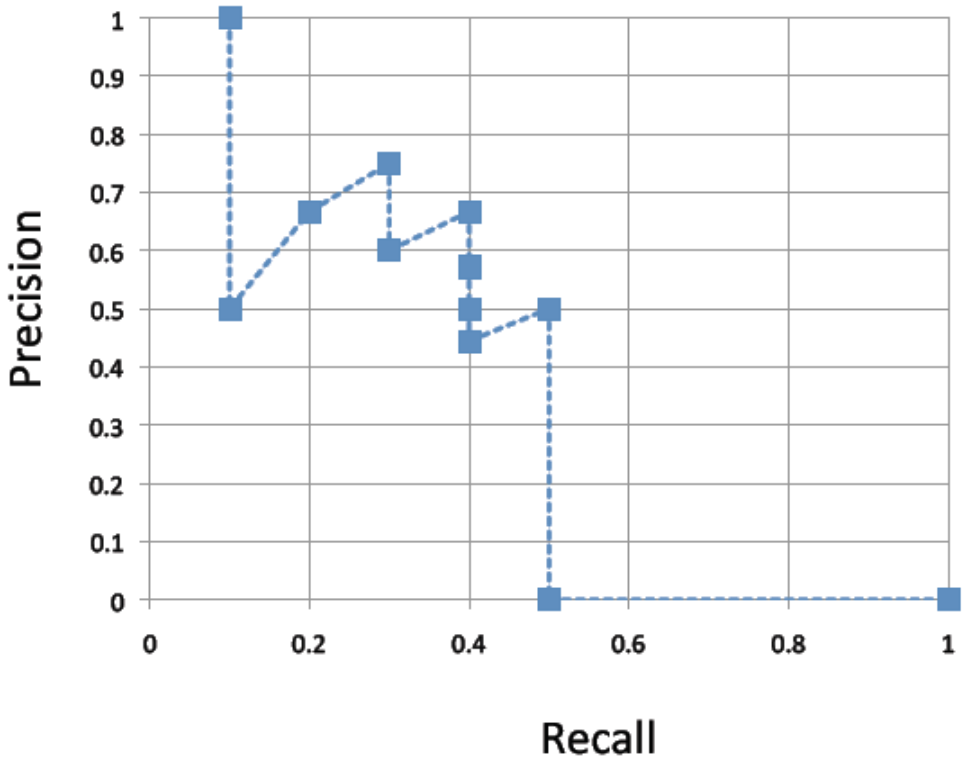
- Современные системы выдают упорядоченные результаты
- Выдача может быть достаточно большой
- Релевантные документы должны выдаваться раньше нерелевантных
- Можно измерять точность на каждом уровне полноты

Ranking Measures

Topic 1

Rank	Rel.	
1	R	retrieved
2	N	
3	R	not retrieved
4	R	
5	N	
6	R	
7	N	
8	N	
9	N	
10	R	
…	…	

Let R=10



Усреднение по запросам

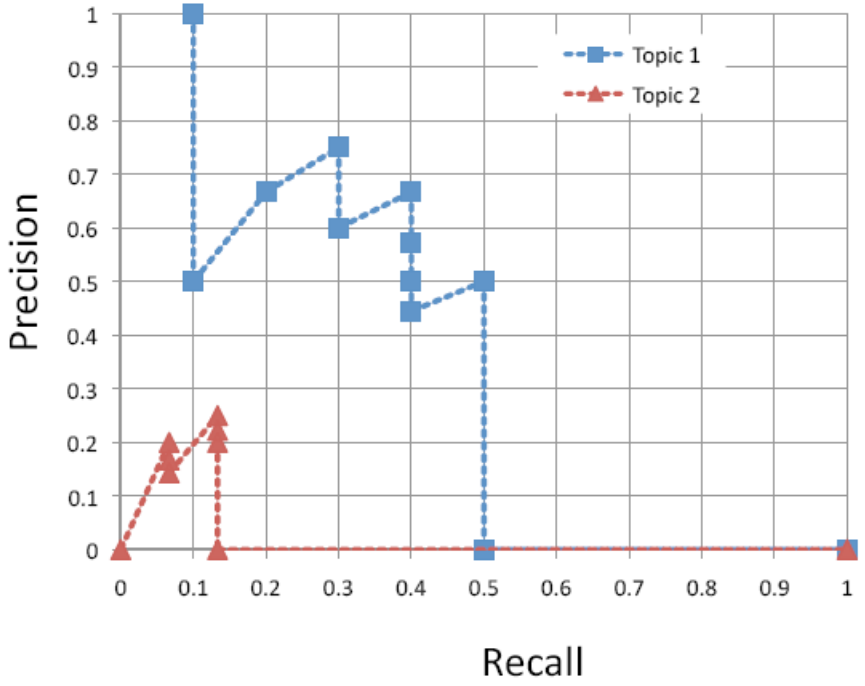
- Кривая полнота-точность для одного запроса не очень интересна
- Нужно построить кривую полнота-точность для совокупности запросов
  - Пока Кривая – это совокупность точек
  - Как интерполировать?

Ranking Measures

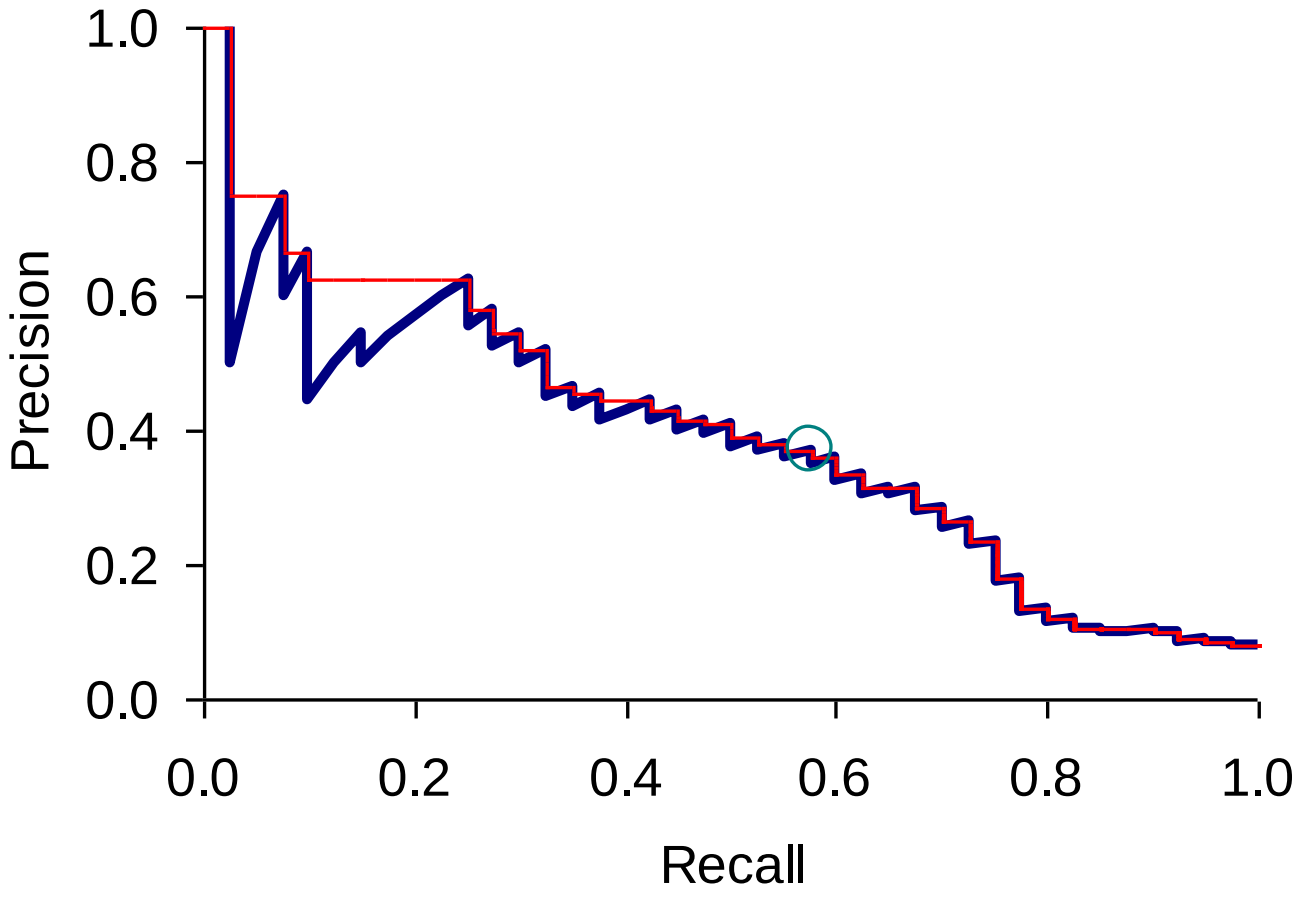
Topic 2

Rank	Rel.	Precision	Recall
1	N	0	0
2	N		
3	N	…	…
4	N		
5	R		
6	N	1/6	4/5
7	N		
8	R	…	…
9	N		
10	N	2/10	2/5
…	…	…	…
∞	R	0	10/10

Let R=15

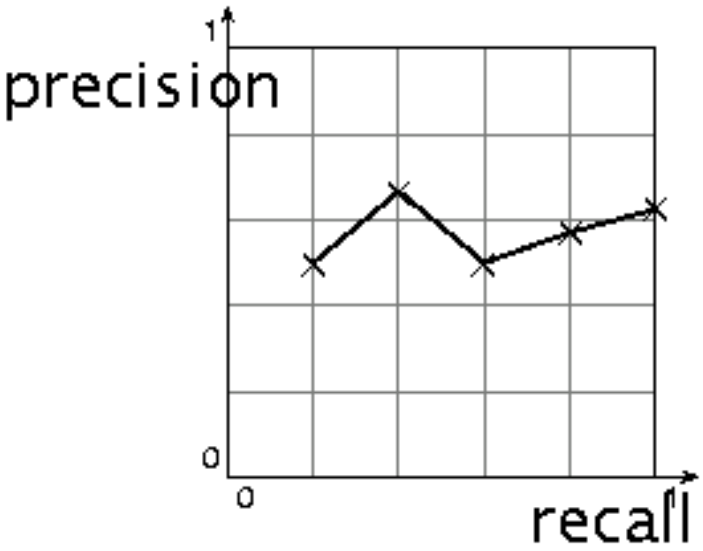


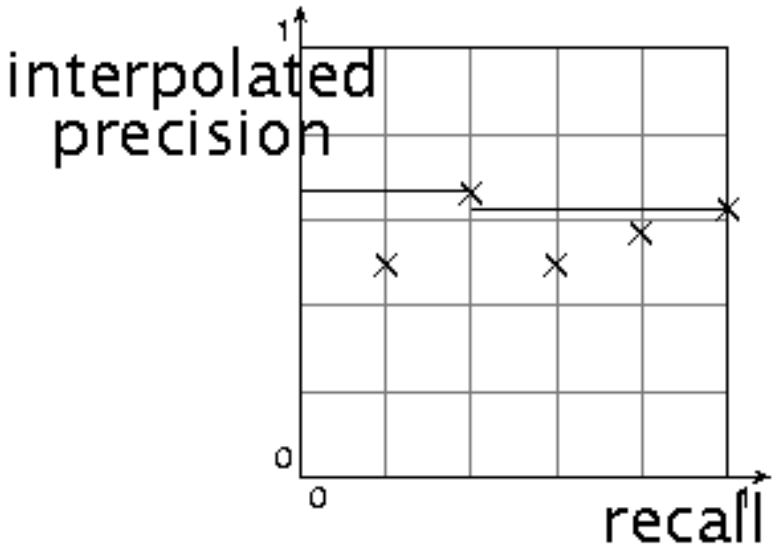
Кривая полнота-точность



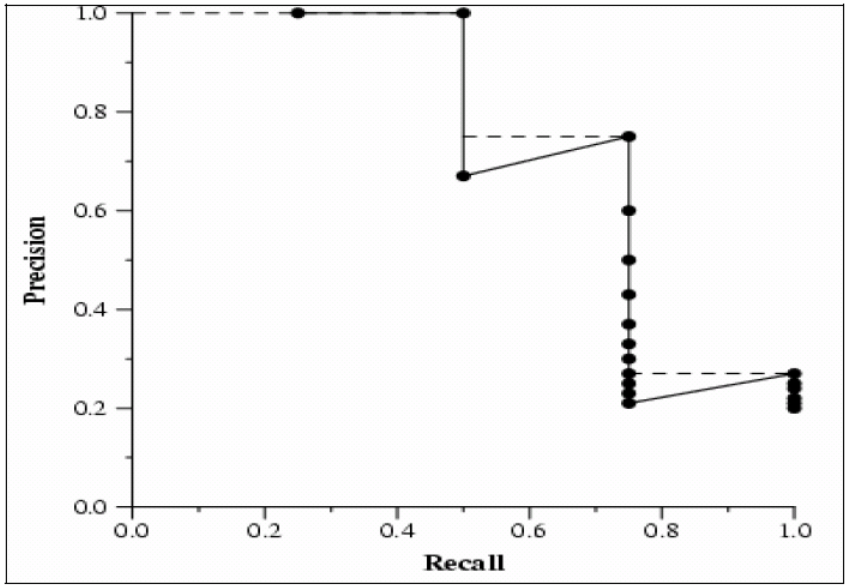
Интерполированная точность

- Идея: Если локально точность возросла с увеличением полноты, то засчитаем ее максимум…
- Т.е. берем максимум точности справа





11-точечный график TREC



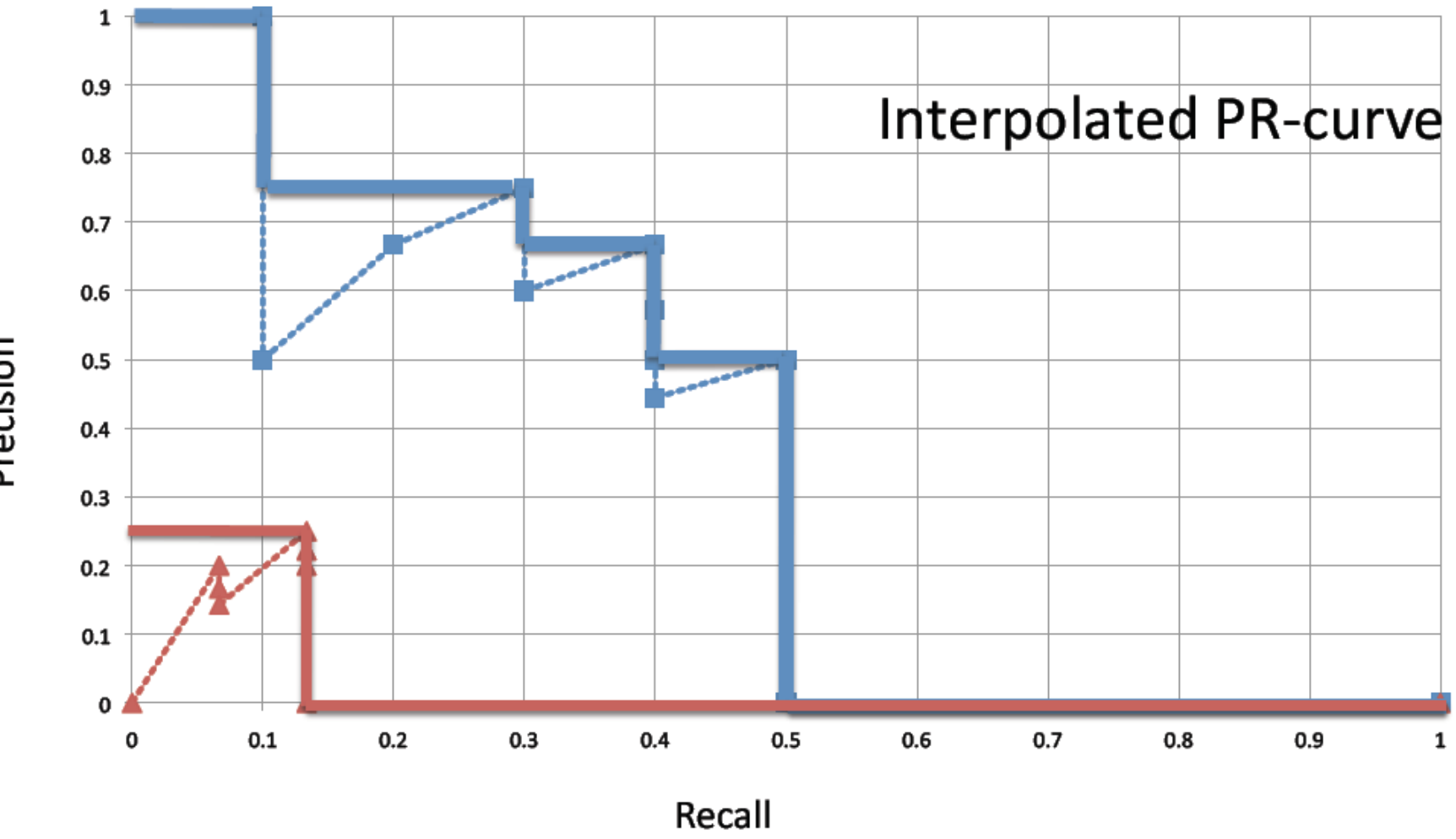
- Значения полноты от 0 до 1 с шагом 0.5
- Интерполяция точности
  - если $r_i > \text{recall}(q_j)$ то

$$p(r_i, q_j) = 0$$

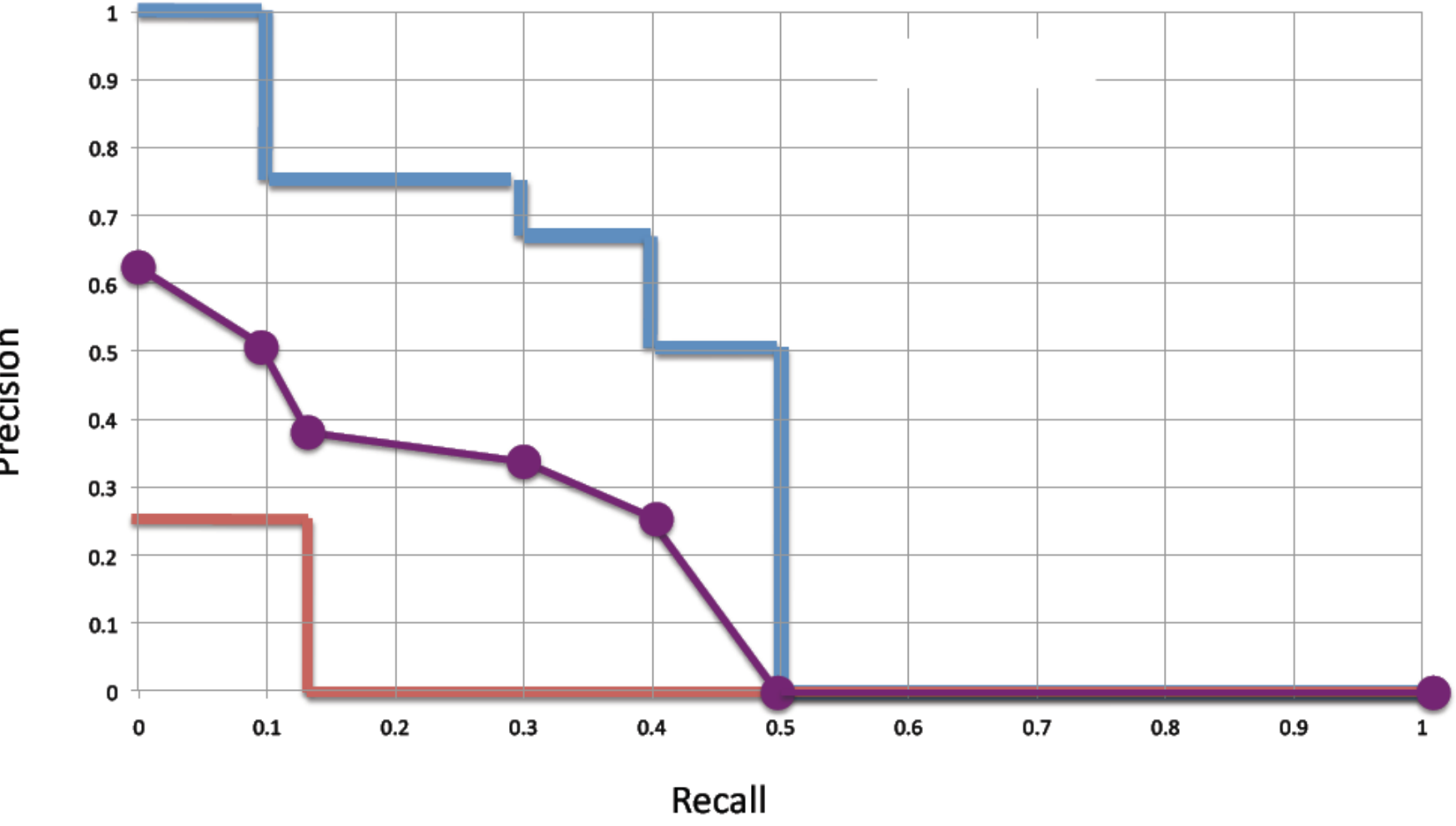
  - если $r_i \le \text{recall}(q_j)$ то

$$p(r_i, q_j) = \max_{n \ge \text{pos}(r_i, q_j)} (\text{precision}(n))$$

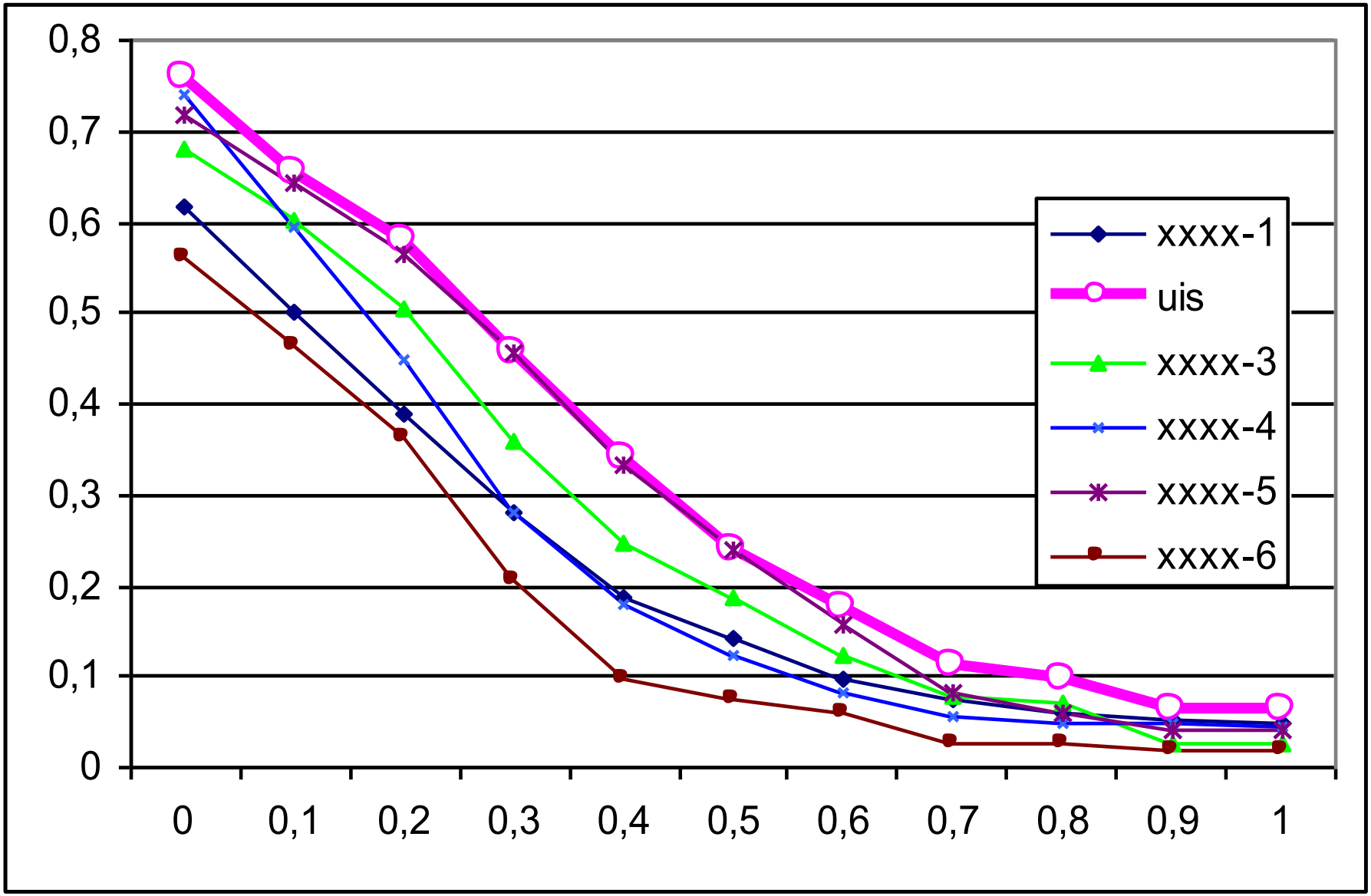
Recall/Precision Graph



Recall/Precision Graph



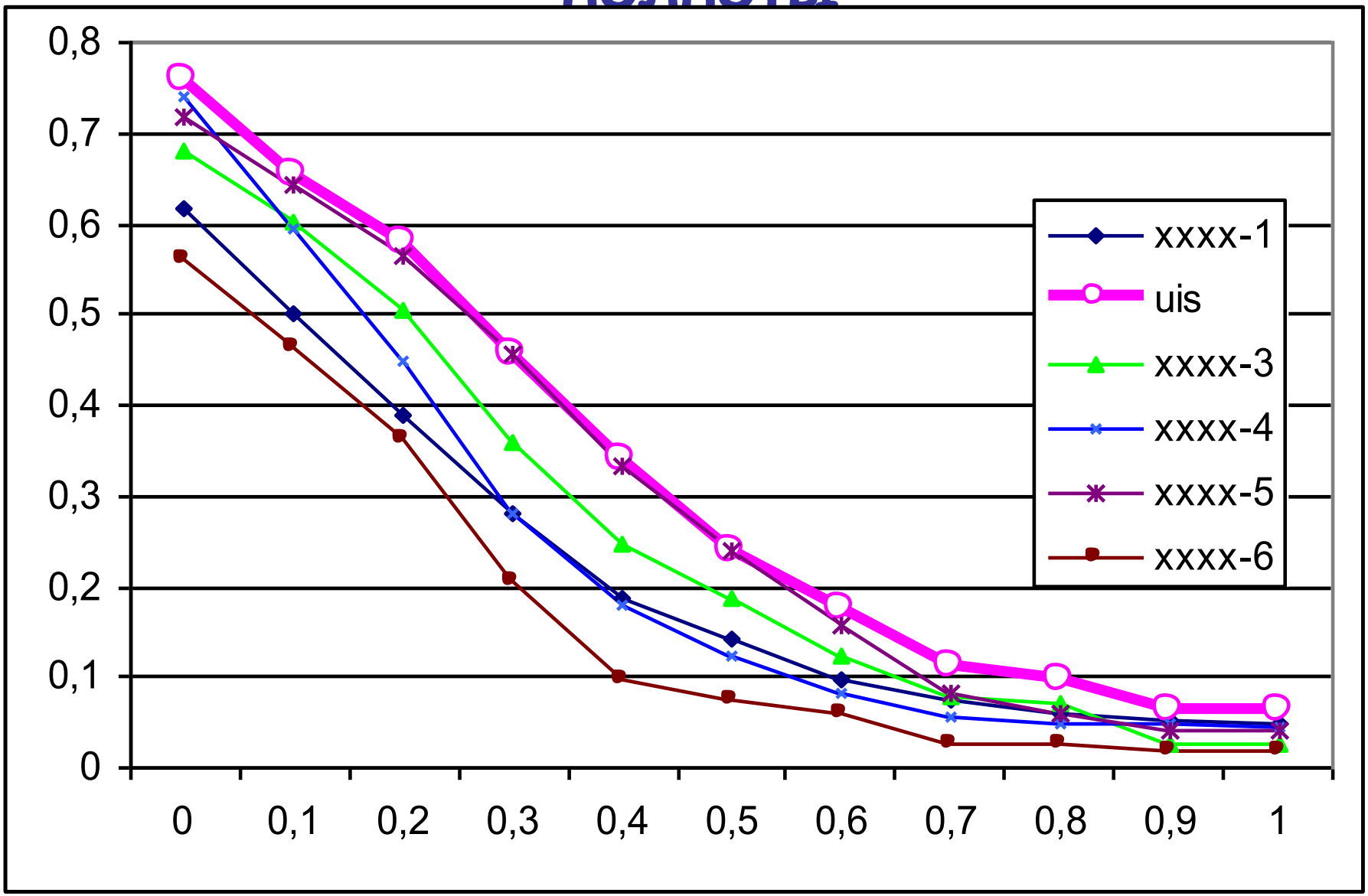
Результаты дорожки Ромип-2008 Legal adhoc, pd35



Получение оценки качества в виде чисел

- Точность в первых n документах: Precision@1, Precision@10
  - Оценка интернет-поиска
  - Плохо усредняется
- Интерполированная средняя точность
  - Имеется 11 значений точности на разных уровнях полноты
  - Используем интерполяцию
  - Можно взять среднее

Интерполированная средняя точность-среднее арифметическое 11 значений полноты



Mean Average Precision (MAP)

- Подсчет точности в тот момент, когда в выдаче релевантный документ
- Суммирование и усреднение (Average precision)
- Нет интерполирования
- Далее усреднение по всем запросам
- (Mean Average Precision)

Average Precision

Topic 1

Rank	Rel.	Precision	Recall
1	R	1/1	1/10
2	N	1/2	1/10
3	R	2/3	2/10
4	R	3/4	3/10
5	N	…	…
6	R	4/6	4/10
7	N		
8	N	…	…
9	N		
10	R	5/10	5/10
…	…	…	…
∞	R	0	10/10

- Average Precision
  - Average of precisions at relevant documents

$$AP = \frac{\frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{6} + \frac{5}{10} + \ldots}{10}$$

Задача

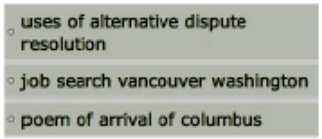
- Эксперт нашел 20 релевантных документов. Система нашла 4 документа в следующей последовательности релевантных и нерелевантных документов:
- RNRNNRRNNNN
- Какова средняя точность поиска – Average Precision

Созданий коллекций для тестирования

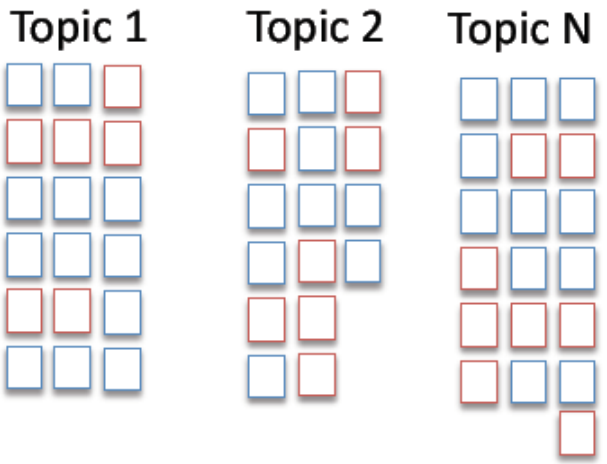
Portable Test Collection

- Document Corpus

- Topics



- Relevance Judgments (QRELs)



Early Test Collections

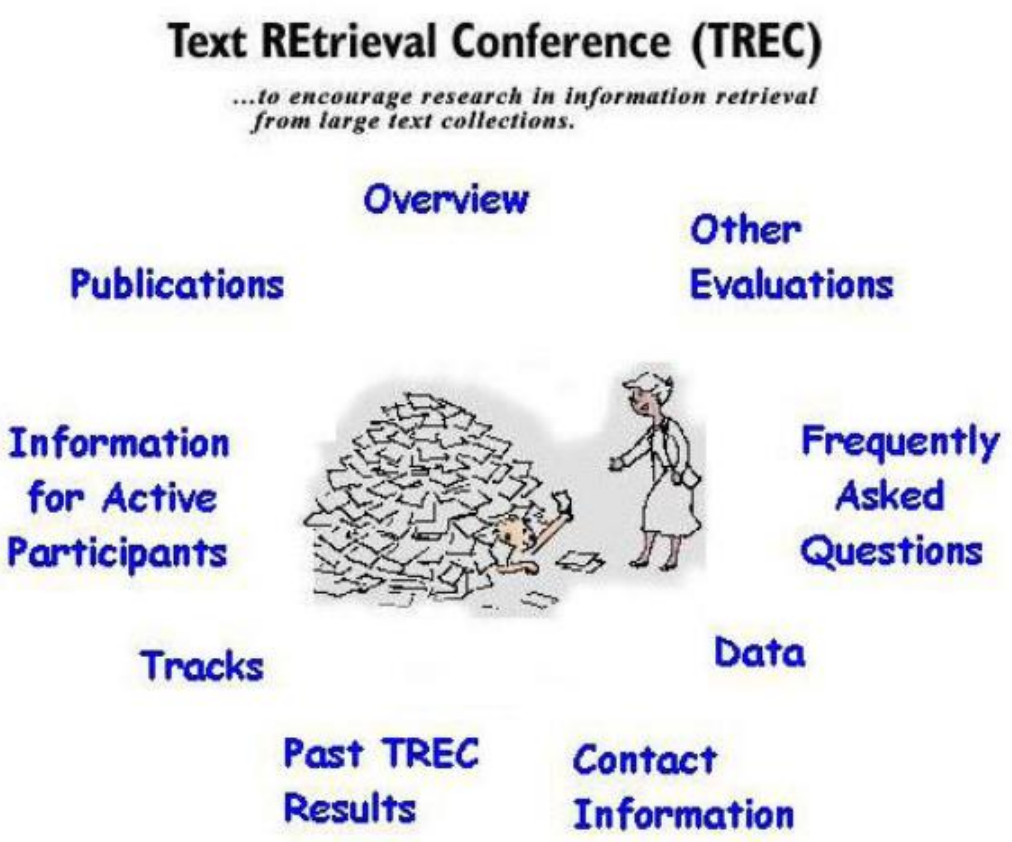
Name	Docs.	Qrys	Year	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945-1963.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	-	A set of abstracts of computer science documents, published in 1959-1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	-	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full text articles from the 1963 edition of Time magazine.

http://ir.dcs.gla.ac.uk/resources/test_collections/

TREC 1992

create test collections for a set of retrieval tasks

standardize evaluation measures



http://trec.nist.gov/images/paper_3.jpg

TREC topics

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

Recent TREC collections

- ClueWeb09 collection
  - about 1 billion web pages in ten languages
  - 5 TB, compressed (25 TB, uncompressed)
  - collected by CMU in January and February 2009
- Other recent TREC collections
  - Collections from wide range of sources
    - Blogs, Twitter, Legal documents, Patents, …
- TREC model copied by others
  - CLEF, INEX, NTCIR, …

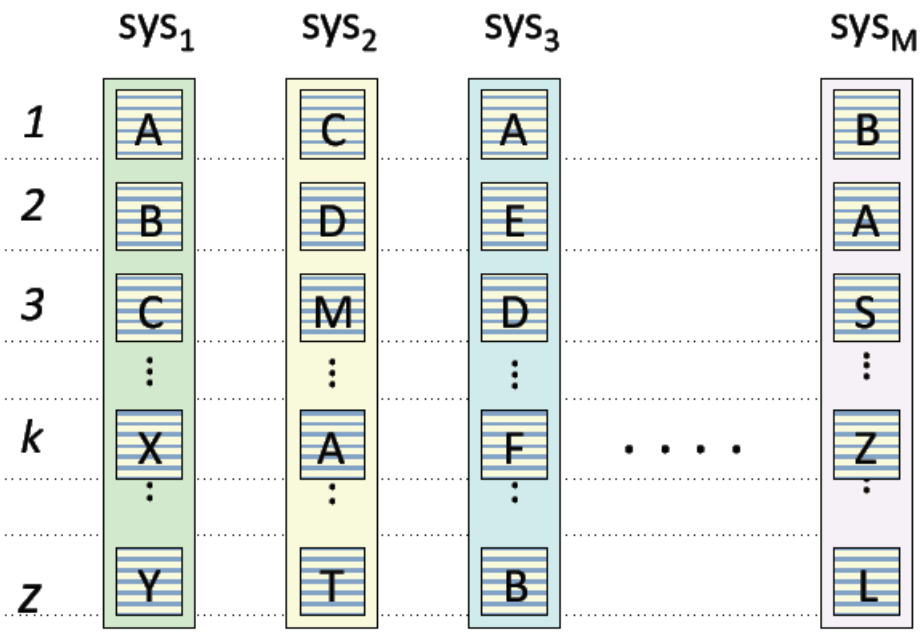
http://lemurproject.org/clueweb09.php/

Пулинг vs. Полнота

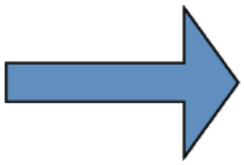
Для каждого запроса:

- Собрать результаты систем участников глубины A
- Выбрать из полученных результатов B первых
- Удалить дубликаты
- Проставить оценки релевантности
- Не оцененные документы считать нерелевантными
- Оценить весь ответ системы (с глубиной A)

Depth-k Pooling

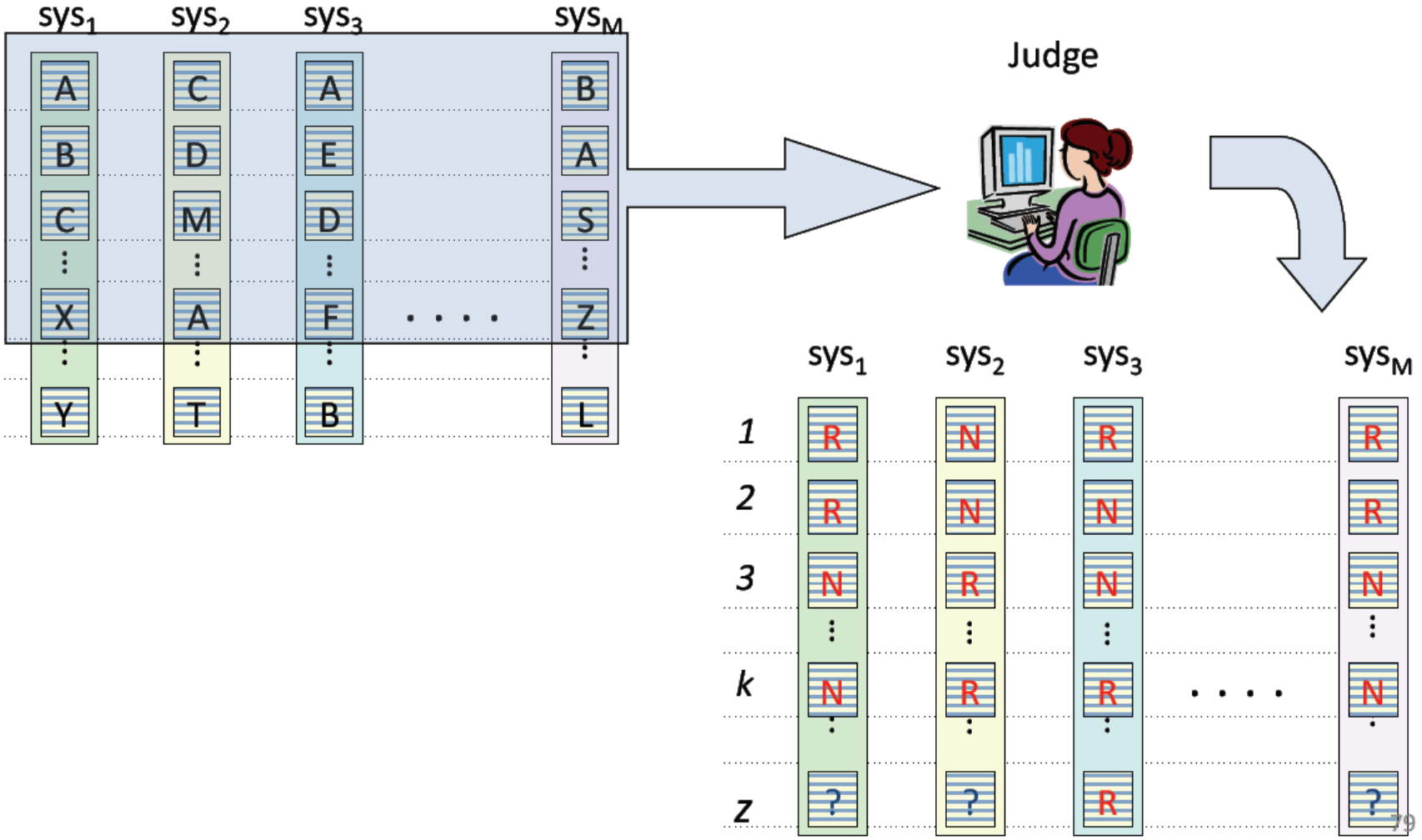


Documents

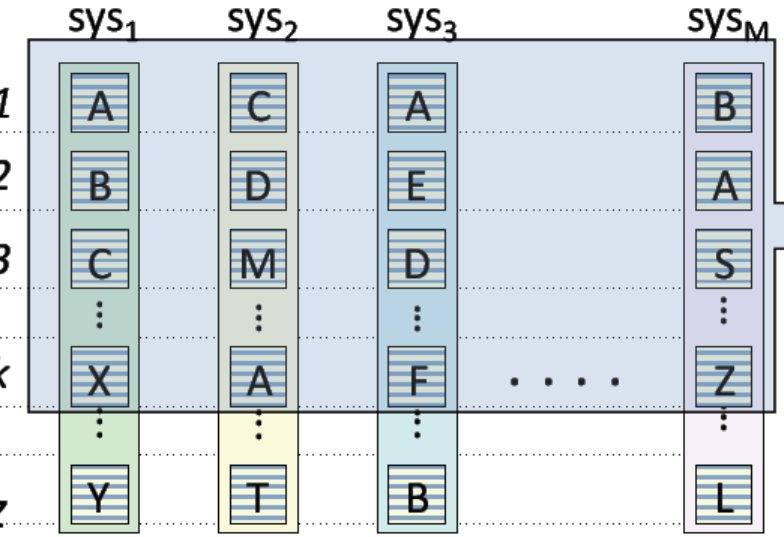


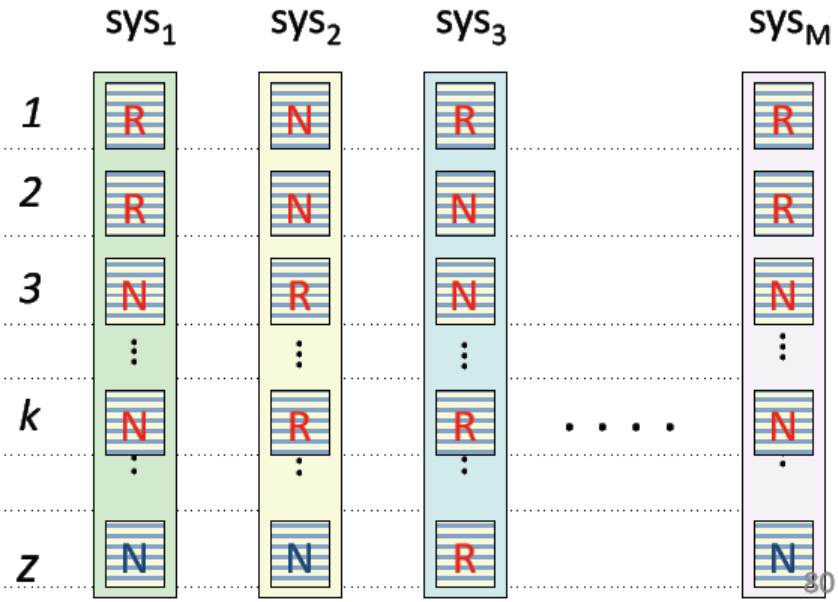
Judge

Depth-k Pooling



Depth-k Pooling





Сложности, связанные с пулингом

- Взаимное усиление систем
- Недооценка систем, не участвовавших в оценке
- Получаемая оценка – оценка снизу

- Но: участники относительно в равных условиях

Критика запросов TReC

- В прошлом:
  - Нерепрезентативность
  - Неоднозначные запросы не включаются
  - 50-100 запросов
  - Запросы со слишком малым или слишком большим количеством запросов не включаются
- В настоящем:
  - Запросы из реальных логов
  - Средней частотности
  - 50 тысяч запросов

NIST Text REtrieval Conference (TREC)

Economic Impact Survey:

"…responsible for approximately ***one-third of an improvement*** of more than 200% in web search products that was observed between 1999 and 2009."

"…total extrapolated benefits were over ***$153 million*** for private, academic, and nonprofit organizations"

Company
Amazon
AOL
Apple
Ask
AT&T Bell Laboratories
Autonomy
Boeing Company
Cirrus Logic
CL Research
Clearwell Systems Inc.
Cleary Gottlieb Steen & Hamilton
Google
Harris Corporation
Hewlett-Packard
IBM
Language Computer Corporation
LCC International
Lucid Imagination
Microsoft
Omniture
Oracle Corp.
Progress Software Corp.
SABIR Research
Sapient Corporation
Sun Microsystems
Texas Instruments
The Echo Nest
Thomson Reuters Corporation

Оценка качества в поисковых машинах

- Полноту невозможно измерить
- K- первых документов
- Релевантные документы должны показываться раньше
- NDCG (Normalized Cumulative Discounted Gain)
- Использование кликов пользователей
  - A/B testing

Шкалы оценок

- В прошлом: TReC – бинарные
- Сейчас TReC:
  - Высоко релевантный
  - Релевантный
  - Нерелевантный

- РОМИП
  - Соответствует
  - Скорее соответствует
  - Возможно соответствует
  - Не соответствует
  - Не может быть оценен

Оценка качества выдачи по небинарным оценкам

- Предположения
  - Лучше, если релевантные документы находятся в начале списка
  - Если есть несколько типов релевантных документов, то лучше, чтобы документы с высокими оценками были раньше в списке
- Существует наилучшее упорядочение расположения оценок от лучших к худшим
- В суммированной оценке выдачи каждая следующая позиция в списке должна давать меньший вклад, чем предыдущая

Оценки для не бинарного случая релевантности

- Cumulative gain

$$CG_\lambda = \sum_{i=1}^{\lambda} g_i$$

- Discounted Cumulative Gain

$$DCG_\lambda = g_1 + \sum_{i=2}^{\lambda} \frac{g_i}{\log i}$$

NDCG (Normalized Cumulative Discounted Gain)

- Нормализация DCG по отношению к лучшему упорядочению по данному запросу

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCGp}$$

Задание

1) Взять 20 новостных документов
2) Обработать морфологическим анализатором
3) Написать вычисление запроса по векторной модели
4) Модель ntc.nnn
5) - запрос – это вектор частот (nnn)
6) - документ – tf.idf + нормализация
7) См. прошлую лекцию