Оценка качества информационного поиска-2

What you can’t measure

you can’t improve

Lord Kelvin

Тест

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum q_i d_i}{\sqrt{\sum q_i^2}\sqrt{\sum d_i^2}}$$

- Если при подсчете косинусной меры сходства запроса с документом забыть разделить на длину документа, то
  - a. Вперед продвинутся документы, в которых запрос больше совпадает с заголовком
  - b. Вперед продвинутся более длинные документы
  - c. Порядок выдачи документов не изменится
  - d. Вперед продвинутся более короткие документы

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

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

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

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>

Оценка ранжированных результатов

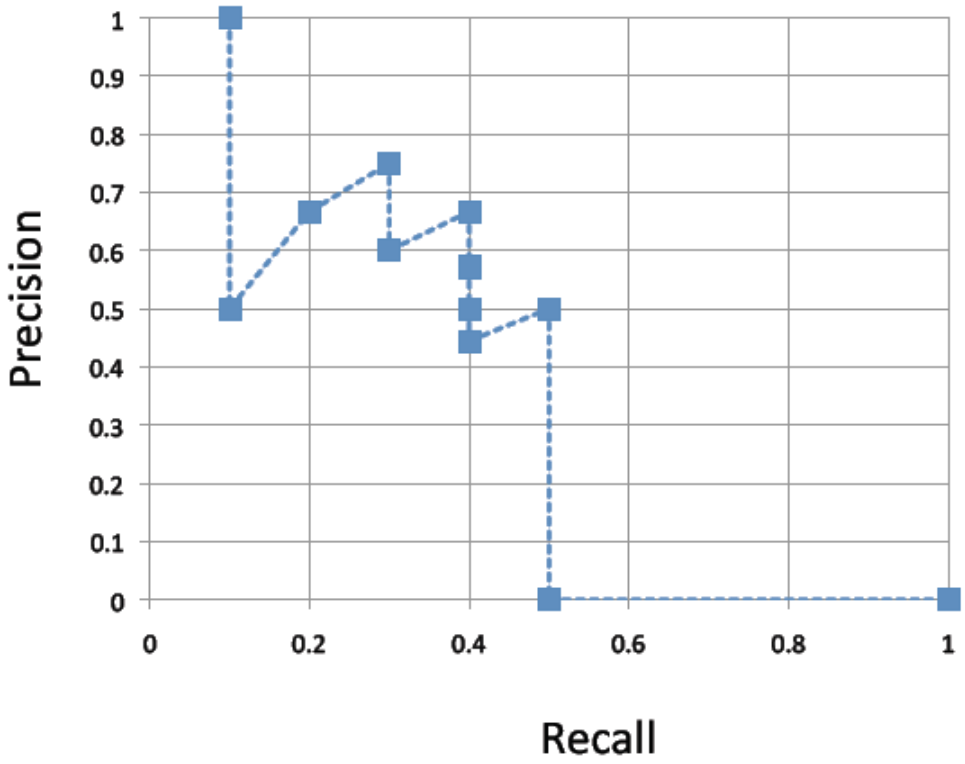
- Современные системы выдают упорядоченные результаты
- Выдача может быть достаточно большой
- Релевантные документы должны выдаваться раньше нерелевантных
- Можно измерять точность на каждом уровне полноты

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

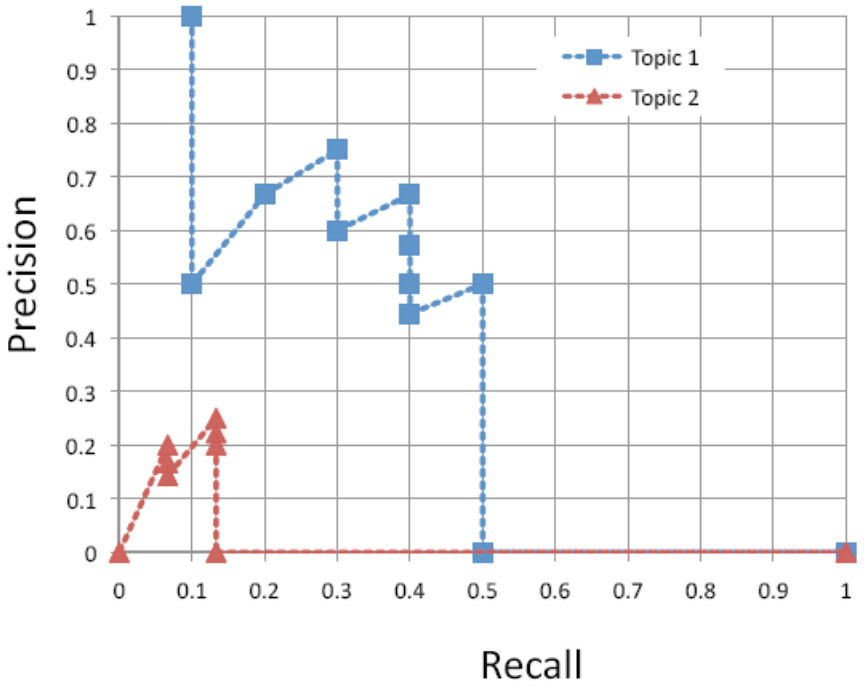
- Кривая полнота-точность для одного запроса не очень интересна
- Нужно построить кривую полнота-точность для совокупности запросов
  - Пока Кривая – это совокупность точек
  - Как интерполировать?

Ranking Measures

Topic 2

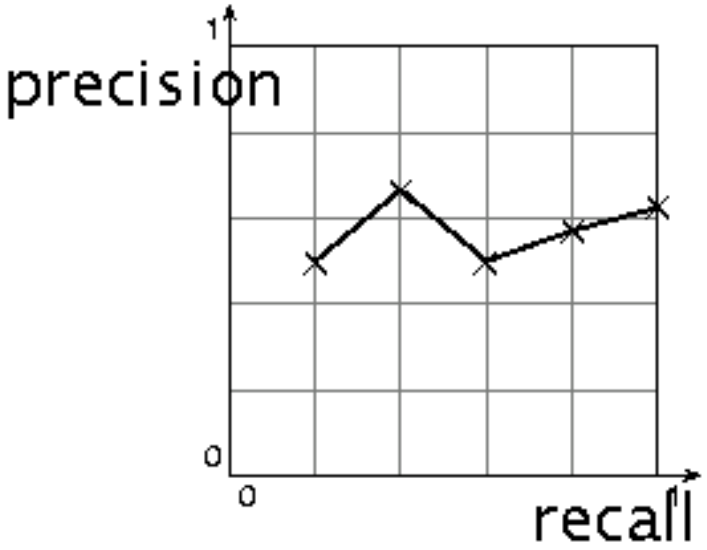
Rank	Rel.	Precision	Recall
1	N	0	0
2	N		
3	N	…	…
4	N		
5	R		
6	N	1/6	4/5
7	N		
8	R	…	…
9	N		
10	N	2/10	2/5
…	…	…	…
∞	R	0	10/10

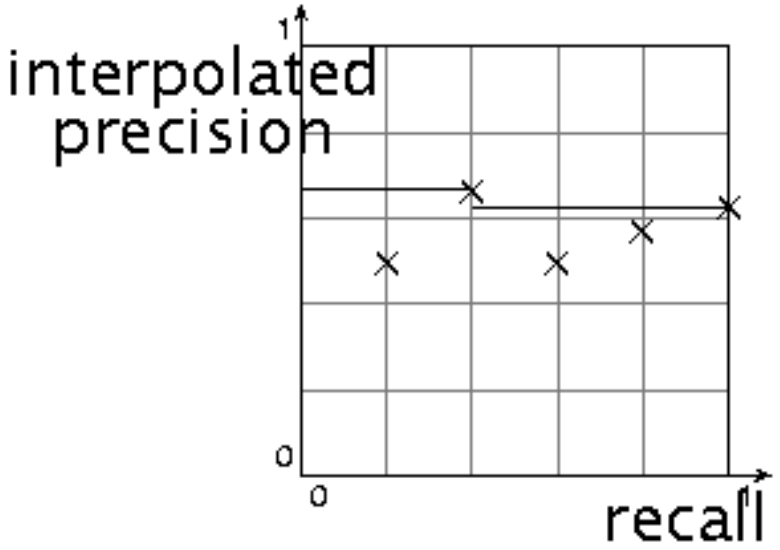
Let R=15



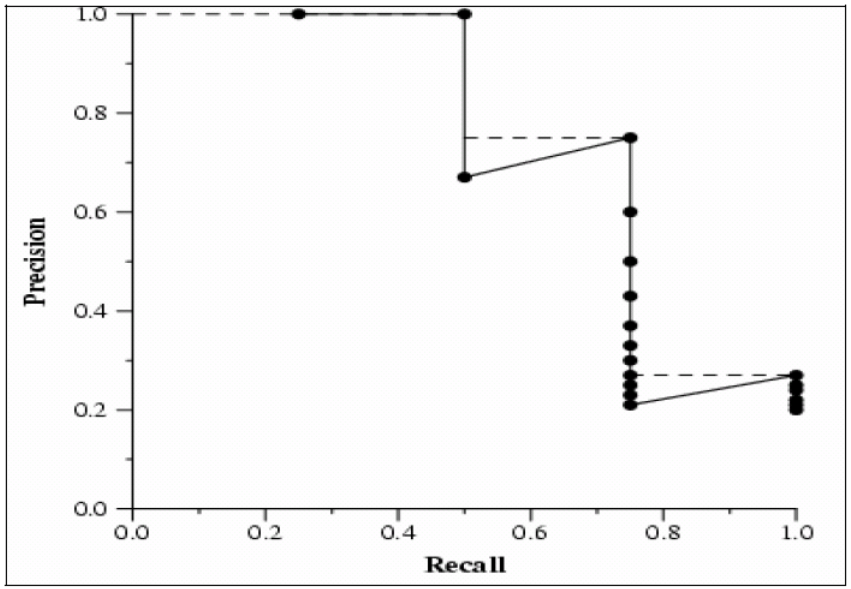
Интерполированная точность

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





Интерполированная точность



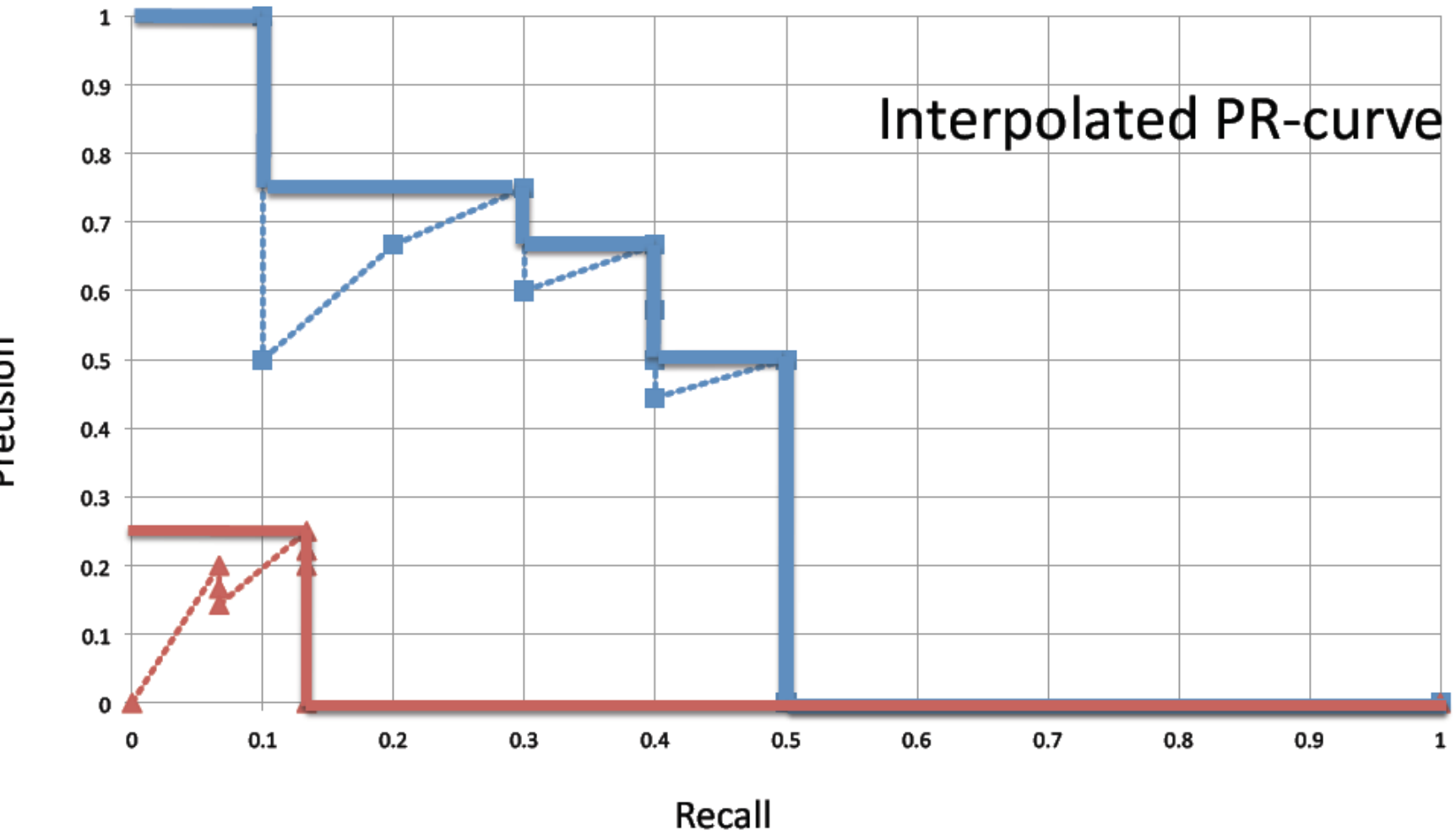
- Значения полноты от 0 до 1 с шагом 0.5
- Интерполяция точности
  - если $r_i > \text{recall}(q_j)$, то

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

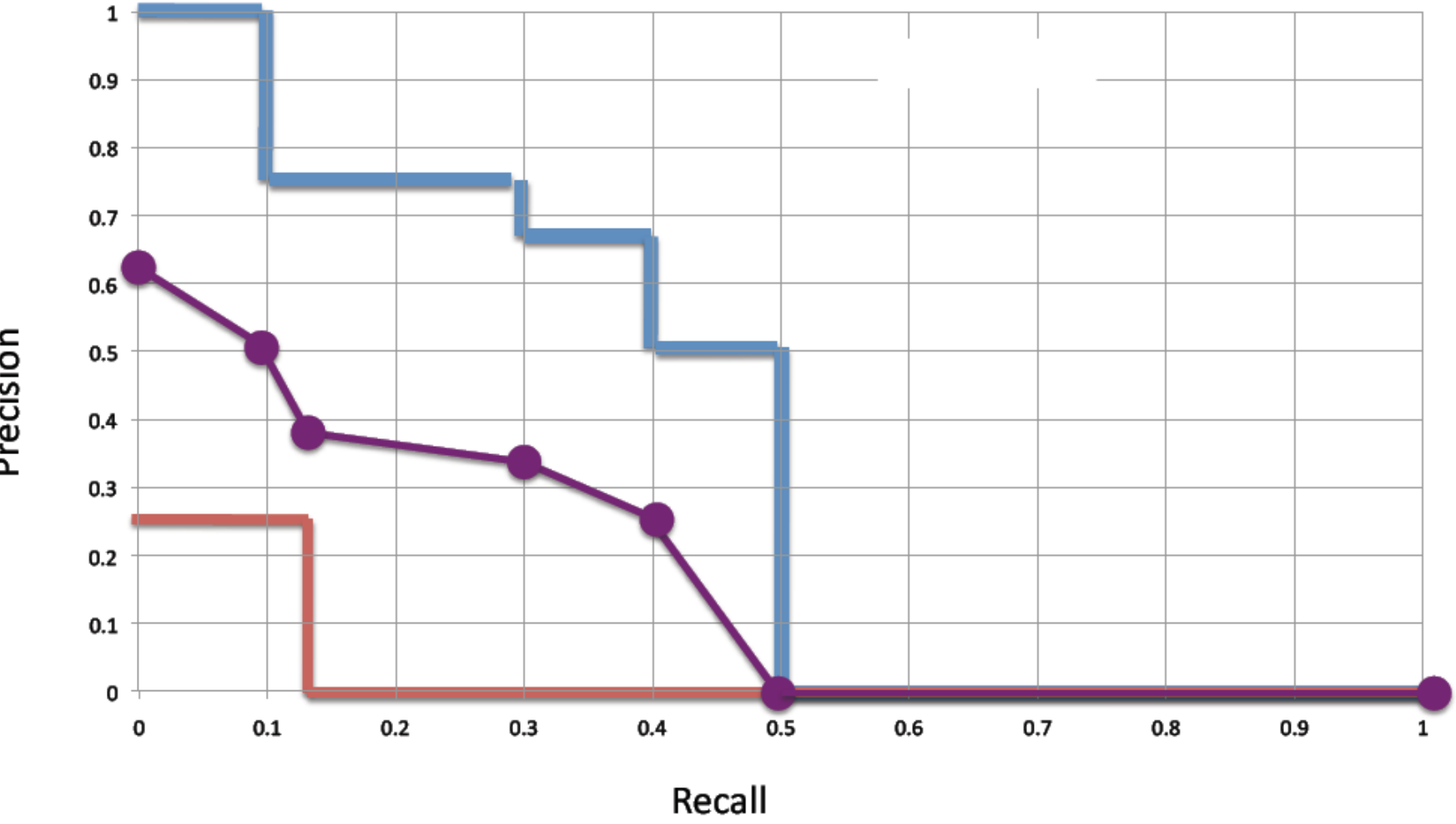
  - Если $r_i \le \text{recall}(q_j)$, то

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

Recall/Precision Graph

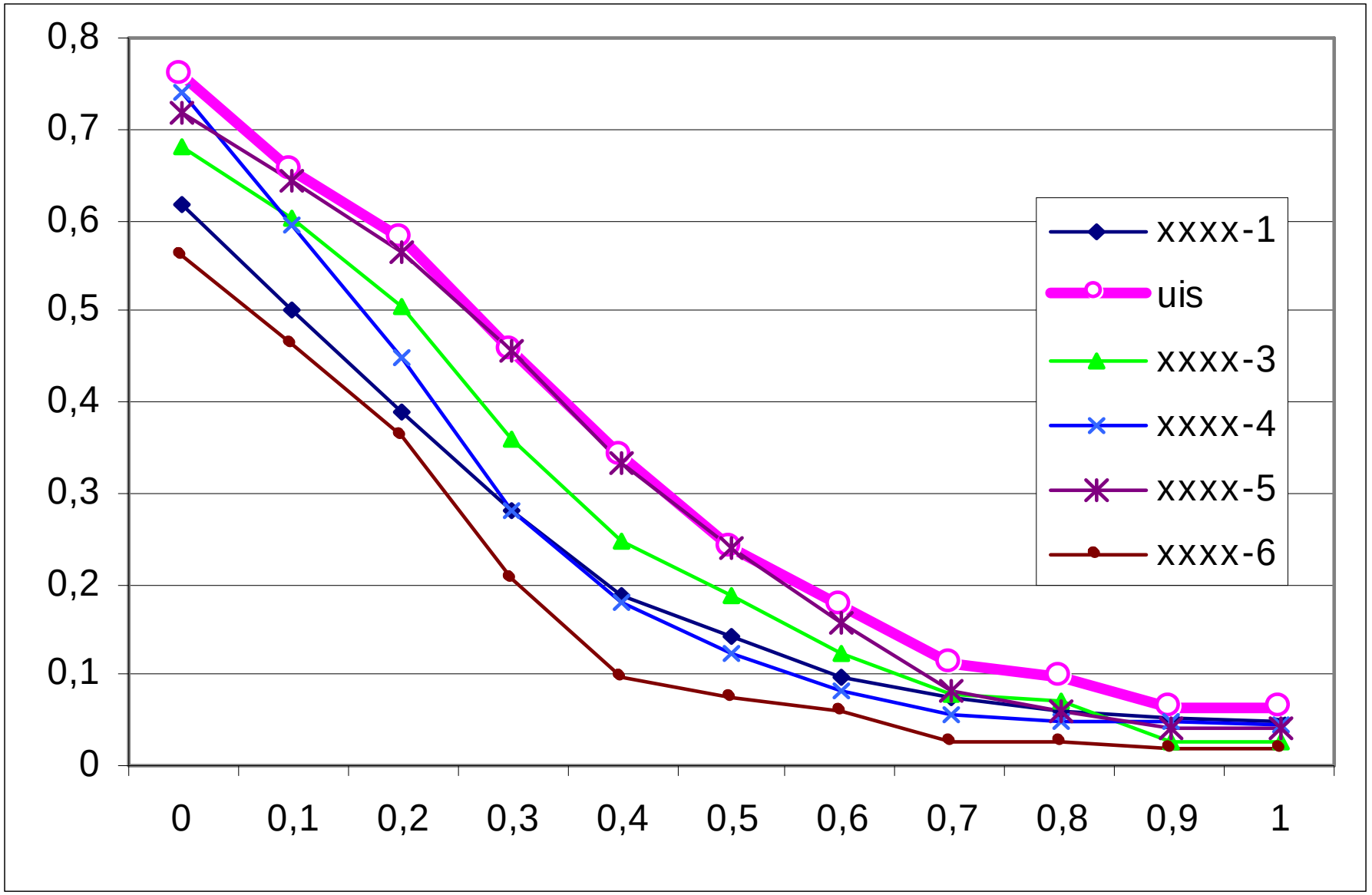


Recall/Precision Graph



Результаты дорожки

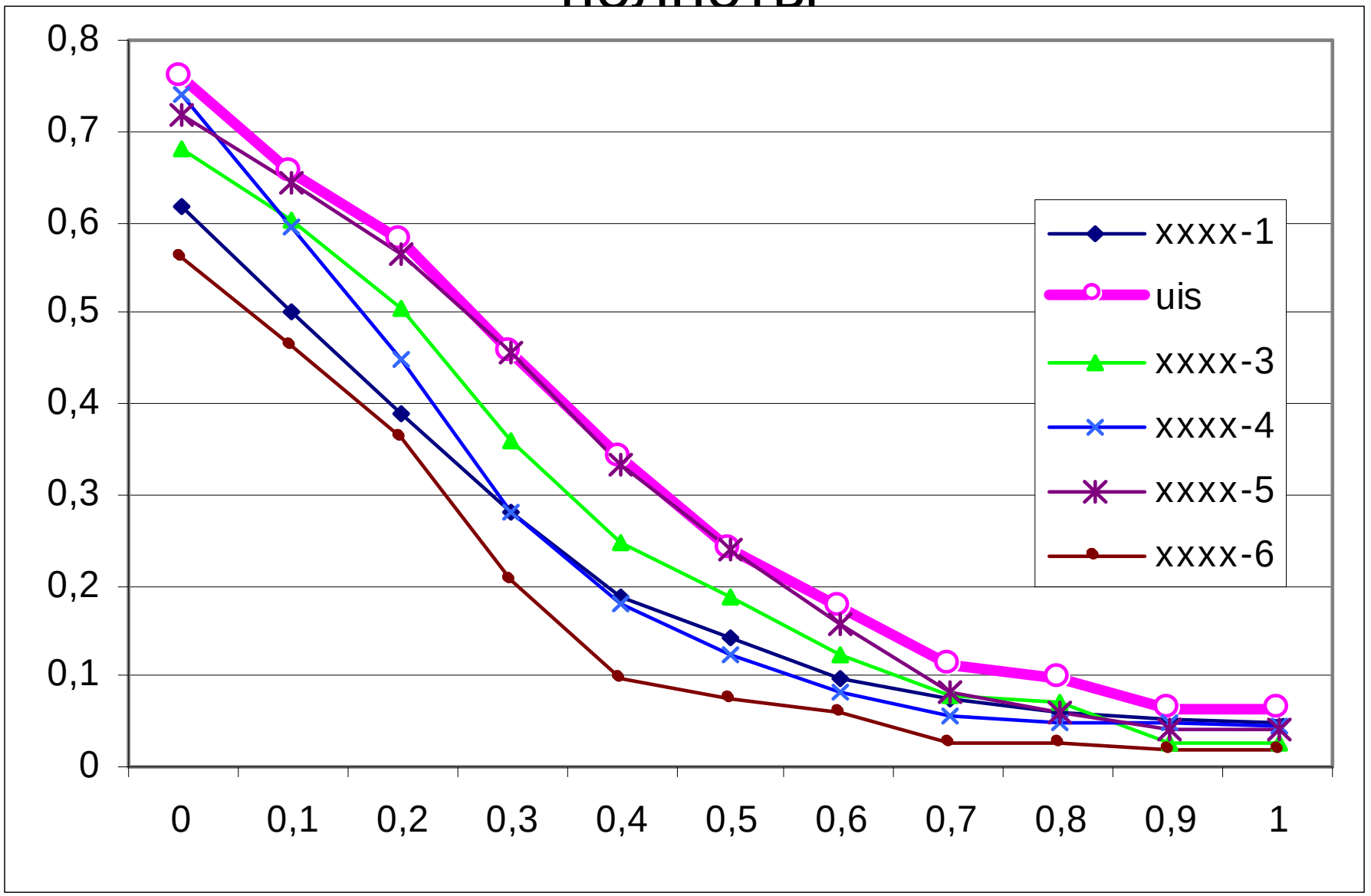
Ромип-2008 Legal adhoc, pd35



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

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

(!) в знаменателе количество релевантных документов, найденных экспертами

Развитие методов тестирование и метрик качества поиска

Пулинг vs. Полнота

В большой коллекции невозможно найти все релевантные документы:

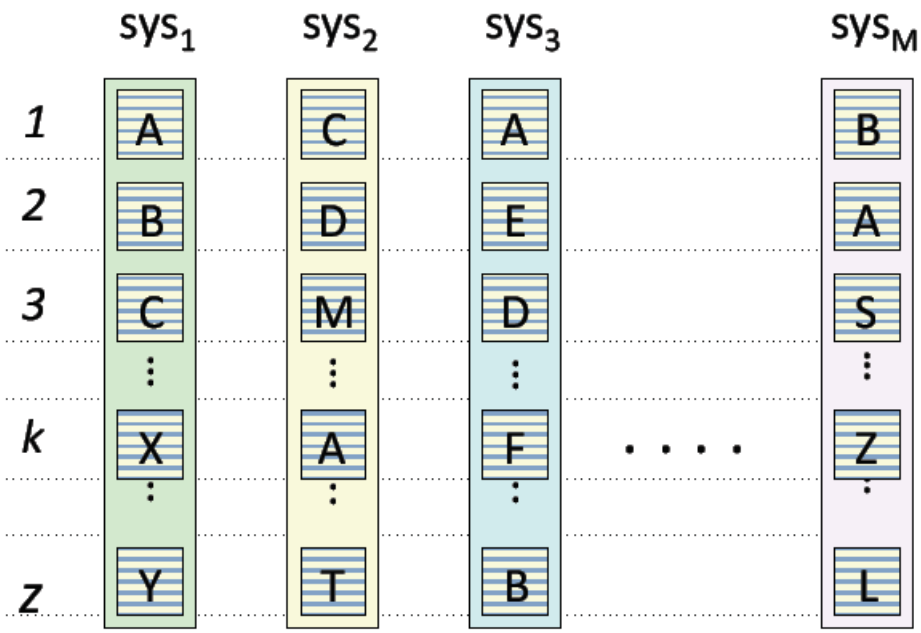
– как оценить полноту?

- предложена процедура пулинга

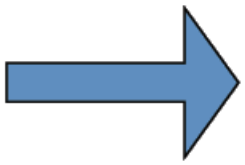
Для каждого запроса:

- 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

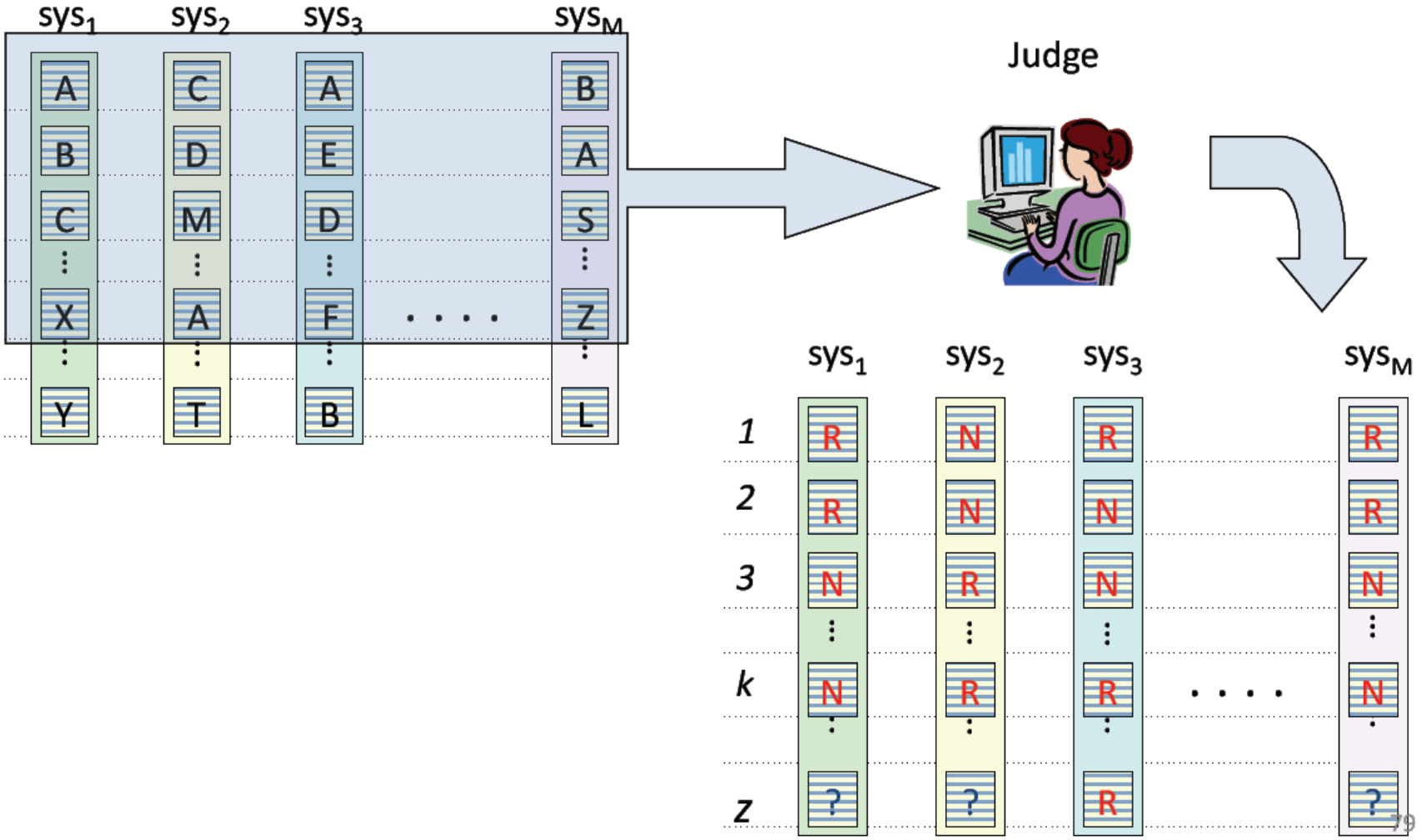


Documents

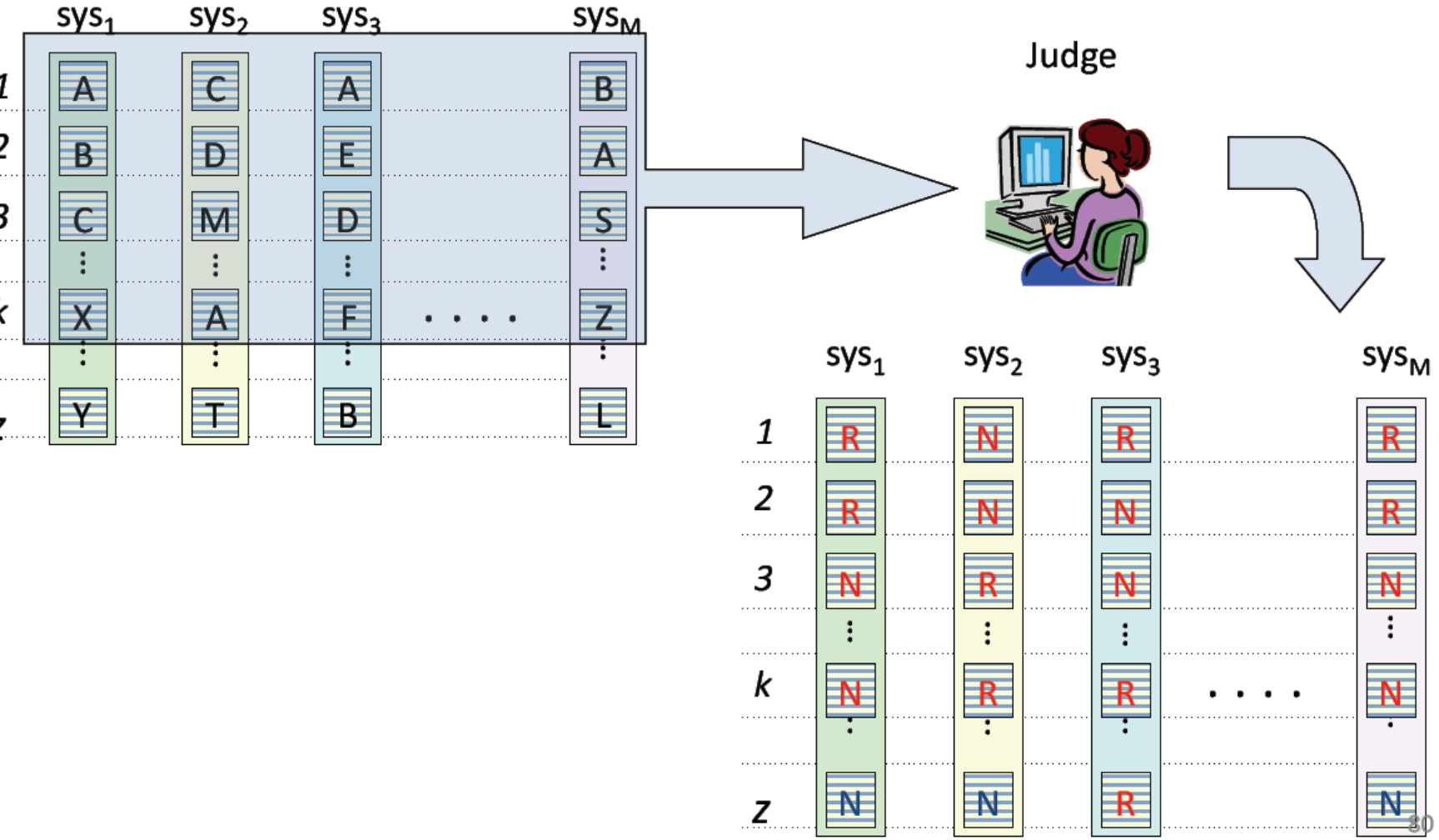


Judge

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

Оценка качества в поисковых машинах (Интернет)

- Полноту невозможно измерить
- K- первых документов
- Релевантные документы должны показываться раньше
- NDCG (Normalized Cumulative Discounted Gain)
- Использование кликов пользователей
  - A/B testing, т.е. использование контрольных и тестовых групп

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

$$DCG_k = \sum_{r=1}^{k} \frac{rel_r}{\log(r+1)}$$

NDCG (Normalized Cumulative Discounted Gain)

- Нормализация DCG по отношению к лучшему упорядочению по данному запросу

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

Метрика bpref

- TREC 2005 terabyte track
- Невозможно создать значимое покрытие возвращаемых документов разметкой, т.е. в выдаче большое количество неразмеченных документов
- Предложена мера bpref
- Пусть r – релевантный документ, n – нерелевантный, R – кол-во размеченных релевантных документов,
- N - кол-во размеченных нерелевантных документов

$$\text{bpref} = \frac{1}{R}\sum_{r}\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}\right)$$

Конференции по тестированию методов

- Проведение специальных конференций по тестированию методов АОТ
  - Независимое тестирование на единых текстовых коллекциях и едином задании
  - TREC (США), CLEF (Европа), NTCIR (Юго-Восточная Азия)

- В России – РОМИП (Российский семинар по методам информационного поиска) (www.romip.ru):
  - 2003 – по н.в
  - Поддержан грантом РФФИ 04-07-90280-в (рук. Добров Б.В.)
  - Участники: Яндекс, Рамблер, Mail.ru, Кирилл и Мефодий, корпорация Галактика, компания RCO, лаб. ЛАИР НИВЦ МГУ

РОМИП
СЕМИНАР

Like 109

Российский семинар по Оценке Методов Информационного Поиска

Целью семинара является создание плацдарма для проведения независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией, а также консолидация сообщества российских исследователей и разработчиков, занимающихся информационным поиском.

Новости

15 января 2013

Все результаты по дорожкам анализа мнений разосланы участникам.

15 декабря 2012

Приглашаем к участию в дорожке по машинному переводу.

3 декабря 2012

Тестовые коллекции разосланы участникам. Обновлены страницы с описанием форматов результатов по каждой дорожке.

9 ноября 2012

Опубликован словарь оценочной лексики в области товаров.

6 ноября 2012

Регистрация закрыта. Началась рассылка данных для обучения.

Текстовый поиск

- Задача поиска по запрос
- Две коллекции: By.Web и KM.RU
- By.Web
  - 550 запросов
  - Глубина пула 20 документов
- KM.RU
  - 100 запросов
  - Глубина пула 50 документов

Текстовый поиск

- *Описание запроса* – понимание того, что искал пользователь и какие ответы ему полезны
- Типы ответов
  - Идеальный (дальше можно не искать)
  - Релевантный+ (полезная но не исчерпывающая информация, *один из многих)*
  - Релевантный- (кусочки полезной информации, не авторитетный источник)

Текстовый поиск

Запрос: *мультик падал прошлогодний снег*

Описание: Цель – найти мультик или детальную информацию о нем. **Идеальный ответ – подробная информация о мультике или страница с видео.**

Релевантный+ - ссылка на страницу с видео**.**

Релевантный- - картинки/фотографии или музыка из мультика, отдельные факты о мультике**.**

Текстовая классификация

- Веб сайты
  - Классификатор сайтов DMOZ
  - Оценка методом общего котла
  - 20 категорий, 4000 пар сайт-категория для оценки
- Веб страницы

Контекстно-зависимое аннотирование

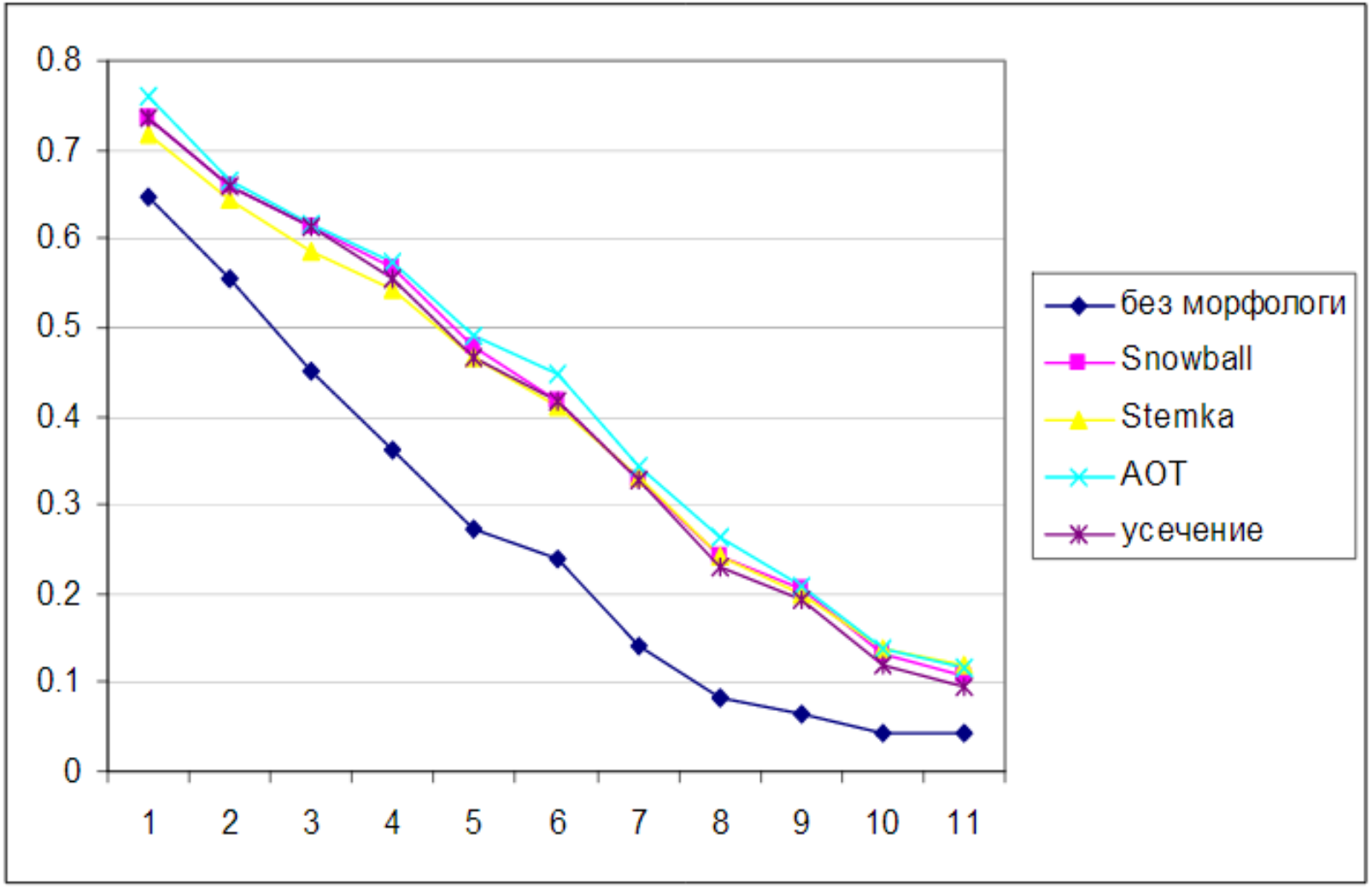
- Составление аннотации документа по запросу
  - *Запрос:* когда состоялась Куликовская битва

  Аннотация1: ... на поле Куликовской битвы осталось...

  Аннотация2: ... Куликовской битвы, которая состоялась в 1380...

  Аннотация 3..... Куликовском ... битва ... месяц *...*
- Для оценки отобраны только релевантные документы
- Оценивается информативность, читабельность

РОМИП-2006 (Губин, Морозов): Влияние морфологии на поиск. Коллекция Legal



Задание 5: Задача-1

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

Задание 5: Задача-2

- При разметке релевантных документов эксперты использовали шкалу от 0 (нерелевантные документы) до 4 баллов.
- При тестировании систем выяснилось, что системы выдали следующие результаты поиска ответов на один и тот же запрос:
- **Система 1**: 4, 2, 3, 1, 2, 0 (и далее 0)
- **Система 2:** 3, 2, 4, 4, 4., далее 0.
-
- Какая система ищет лучше по мере NDCG.