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)

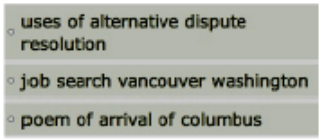
Классическая (Cranfield) процедура оценки

- 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

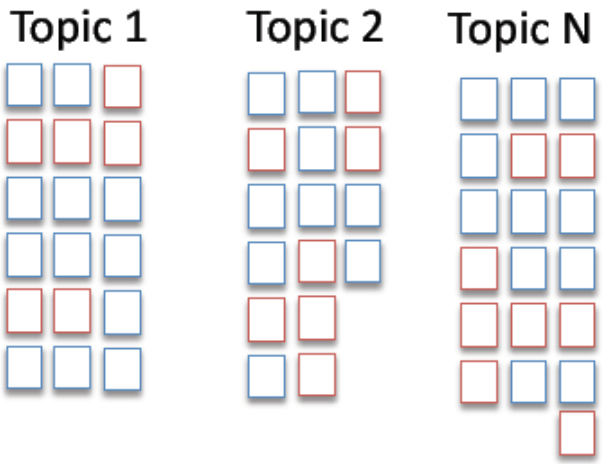
Portable Test Collection

- 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

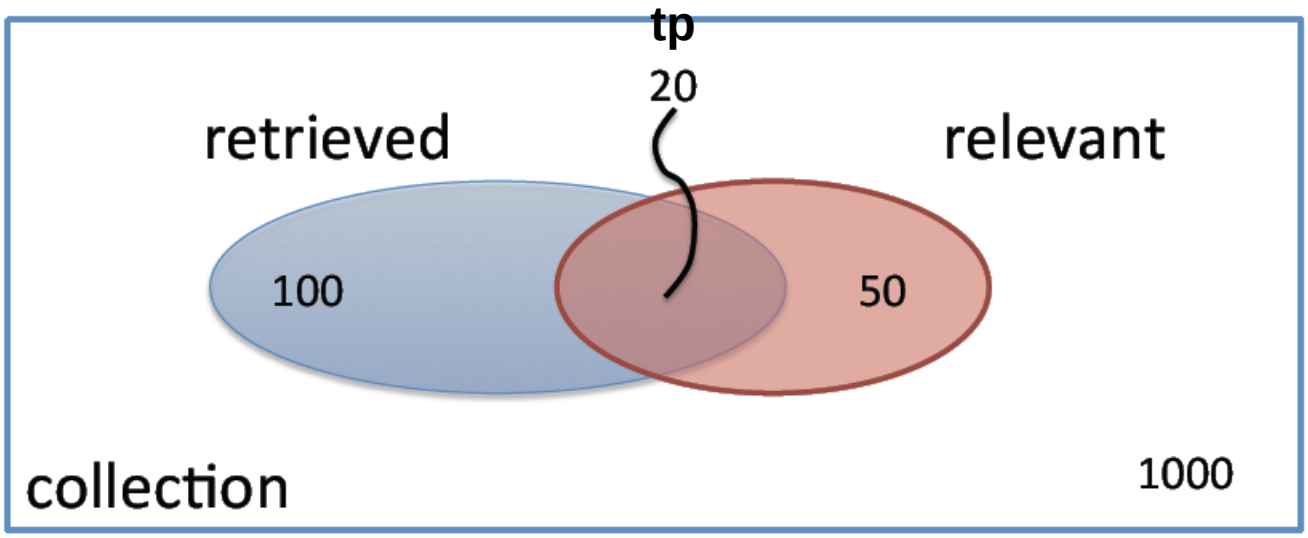
Оценка неранжированного поиска

- **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)
- Нужна комбинированная мера
  - Усреднение не подходит

Комбинированная мера: F-мера

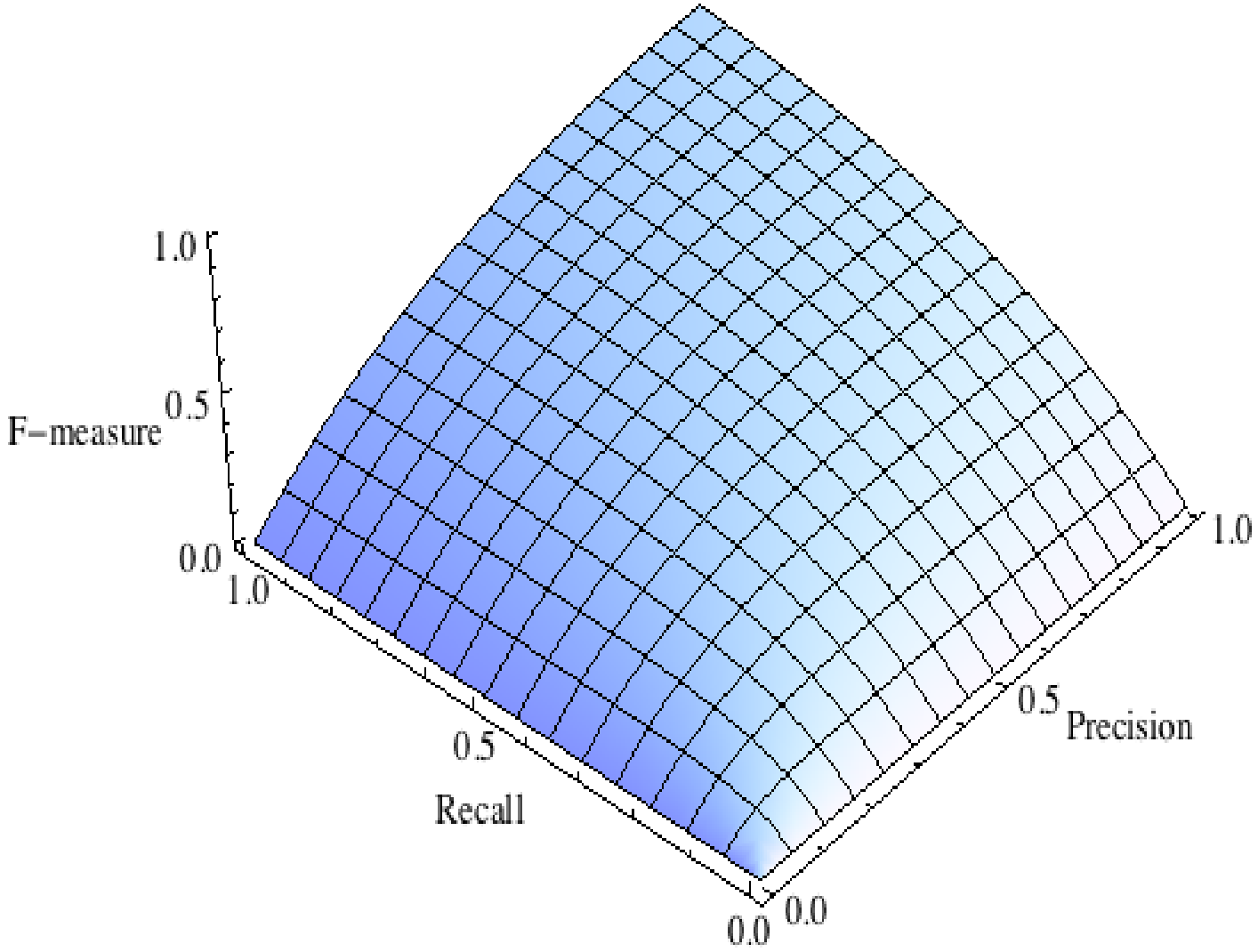
- Среднее гармоническое между полнотой и точностью

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

- Обычно сбалансированная F-мера:
  - $\beta = 1$ или $\alpha = 1/2$

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

Зависимость величины F-меры от точности и полноты

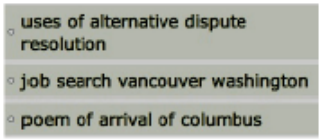


Создание коллекций для тестирования

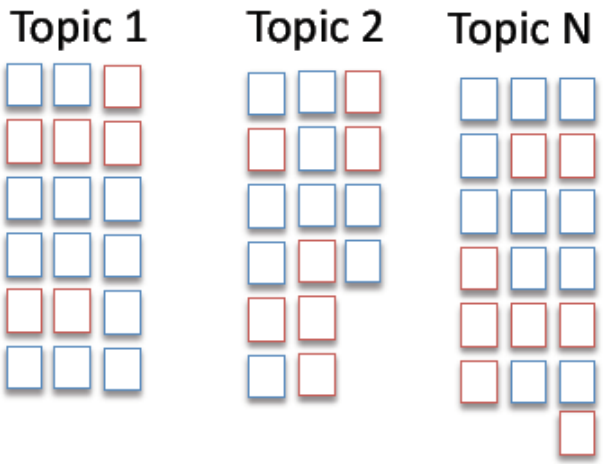
Portable Test Collection

- Document Corpus

- Topics



- Relevance Judgments (QRELs)



Early Test Collections

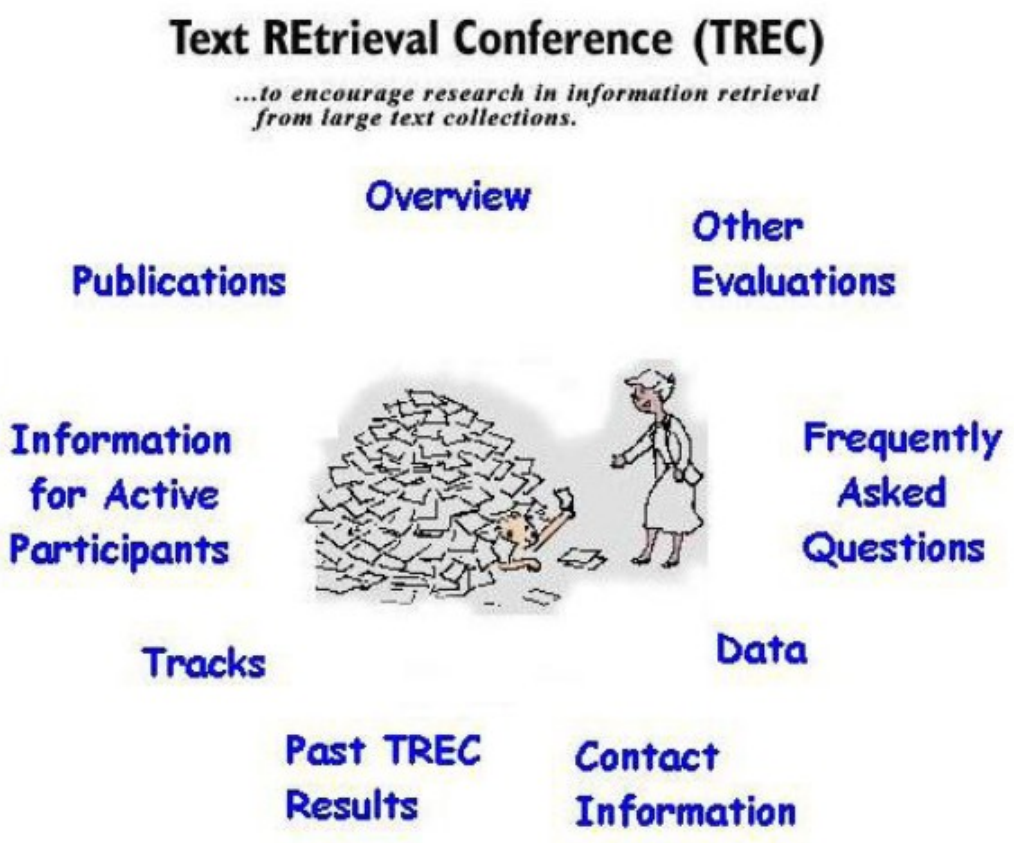
Name	Docs.	Qrys	Year	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945-1963.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
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/

Trec-2018 tracks

- **CENTRE Track:** This is a new track for 2018, which will run in parallel (with somewhat different emphases) in CLEF 2018, NTCIR-14, and TREC 2018. The overall goal of the track is to develop and tune a reproducibility evaluation protocol for IR.
- **Incident Streams Track:** This is a new track for TREC 2018. The Incident Streams track is designed …to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators
- **Precision Medicine Track: T**his track is a specialization of the Clinical Decision Support track of previous TRECs. It will focus on building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments as well as evidence-based literature to identify the most effective existing treatments и др.
-

2010

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

Happy 25th <> TREC !!! 2017

One of the most well-known and celebrated anniversaries is the **25th** wedding **anniversary**, also known as the silver **anniversary**. A milestone this important calls for classy **25th anniversary** gifts, so if you're going the traditional silver route, consider gorgeous jewelry or engraved keepsakes.

25th Anniversary - Gifts.com
https://www.gifts.com/**anniversary**/**25th-anniversary**/ob6bmJ

S. Dumais (2016)

Проблемы TREC и других тестирований

- S. Dumais
  - Исследователи, фонды и т.п. «ищут под фонарем», т.е. используют имеющиеся данные
  - Не всегда ясно, что представляют собой используемая выборка документов и запросов
    - Иногда идет задержка относительно существующих потребностей индустрии и практики
  - Ограниченная
    - Разница между официальными метриками и экспериментами в реальном времени

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

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