Experimental Information Retrieval

Jaime Arguello
jarguell@email.unc.edu

January 11, 2016

Introductions

- Hello, my name is _____.
- However, I'd rather be called _____. (optional)
- I'm in the _____ program.
- I'm taking this course because I want to _____.

Course Overview

TREC 2015
tracks

- Clinical Decision Support
- Input: A narrative describing a medical case
- Output: A ranked list of documents containing information about the (1) diagnosis, (2) tests, and (3) treatment plan associated with the case

TREC 2015
tracks

- Contextual Suggestion
- Input: User profile + location + “Entertain me”
- User profile: ratings on previously recommended activities/venues from different location + demographic info + “trip type”
- Output: A ranked list of suggested activities/venues in the given location

TREC 2015
tracks

- Dynamic Domain
- Input: Query describing complex, multi-faceted topic
- Output: Multiple rounds of results covering the different subtopics of the topic

TREC 2015
tracks

- Live QA
- Input: Natural language question (possibly recommendation-based)
- Example: “My 105 lb. lab mix just ate a whole box of raisin bran and is acting normal. Should I be worried?”
- Output: 1000-character answer in less than a minute

TREC 2015
tracks

- Microblog
- Input: Query
- Real-Time Output: At most 10 relevant and non-redundant tweets per day (as early as possible)
- Email Digest Output: At most 100 relevant and non-redundant tweets per day (at the end of the day)

TREC 2015
tracks

- Task
- Input: Query describing topic
- Task Understanding: Ranked list of up to 1000 key phrases describing different sub-tasks the user is trying to accomplish
- Task Completion: Ranked list of up to 1000 documents with relevant information about sub-tasks
- Ad-hoc search: Ranked list of up to 1000 documents with relevant information

TREC 2015
tracks

- Temporal summarization
- Input: Query describing sudden-onset event/story
- Sequential Update: Sentences describing new and relevant developments related to the event

TREC 2015
tracks

- Total Recall
- Input: Query describing topic + document-at-time relevance feedback
- Output: One document at a time until the system decides to stop

History of TREC

Jaime Arguello
jarguell@email.unc.edu

January 11, 2016

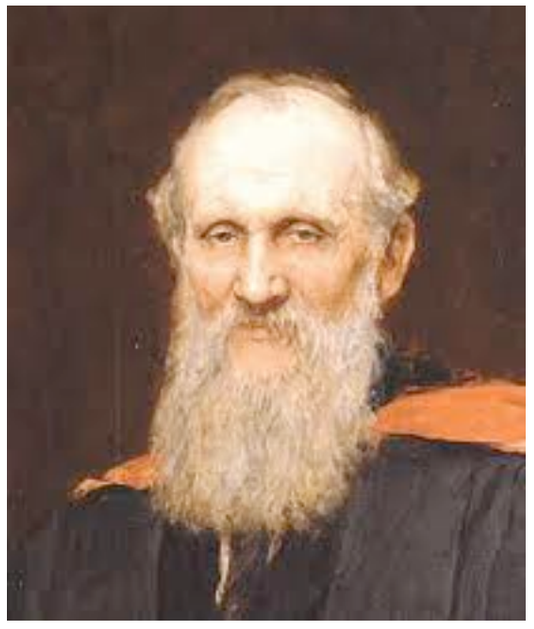
TREC
facts

- Sponsored by NIST and the Department of Defense
- NIST mission statement: To promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.
- First TREC: 1992

TREC

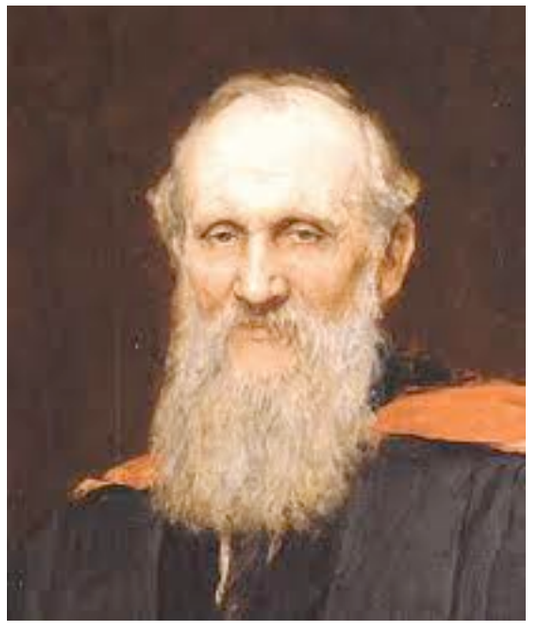
mantra

- “If you cannot measure it, you cannot improve it.”



Lord Kelvin

- “X-rays will prove to be a hoax.”



Lord Kelvin

History of IR
before 1950's

- Information retrieval was a manual process
- Based on controlled vocabularies or subject headings
- E.g.,: "France, History, Middle Ages"
- But, controlled vocabularies are difficult to use and maintain
- Why?

History of IR
before 1950's

- Difficult to use
  - assumes a certain level of expertise (the user must know where to look)
  - the same concept can be expressed in different ways
  - the controlled vocabulary limits the level of granularity
  - the optimal subject heading may be high precision, low recall (too specific, lots of false negatives)
  - the optimal subject heading may be low precision, high recall (too general, lots of false positives)

History of IR
before 1950's

- Difficult to maintain
  - new documents may require new concepts
  - became a real problem after WWII when the rate of new scientific publications increased dramatically
  - requires deep understanding of the collection and the users
  - requires knowing the optimal level of granularity

History of IR
before 1950's

- Uniterm indexes
- Documents indexed using single-term keywords
- E.g., "france", "history", "middle ages"
- Retrieval process: retrieve the set of documents under each heading and take the intersection
- Still no computer in the loop
- But, retrieval became systematic
- Automation became a possibility

History of IR
mid 1950's

- 1954: first system (manual indexing, automatic retrieval)
  - 15,000 bibliographic records
  - indexed using a uniterm index
- 1957: Cranfield experiments (evaluated indexing)
  - 100 documents + a set of information needs + relevance judgements
  - evaluation in terms of precision and recall
  - P: % of retrieved documents that are relevant
  - R: % of relevant documents that are retrieved

History of IR
late 1950's and 1960's

- 1957-1959: Hans Peter Luhn, IBM (automatic indexing)
  - use statistical techniques (term frequency and location) to determine the important "uniterms" directly from the text
- 1960's: Gerard Salton, Cornell (ranked retrieval)
  - index everything in the full-text
  - focus on the important "uniterms" at retrieval time (e.g., tf.idf + vector space model)
  - offload the task of determining what's important to the retrieval model

History of IR
1960's

- The SMART System (Salton *et al.*)
  - natural language queries
  - automatic indexing of the full-text (including important statistics to facilitate ranked retrieval)
  - soft-matching of queries to documents
  - ranked documents in terms of query-document similarity
  - Cranfield-style evaluation

History of IR
1970's

- Computers become more powerful
  - faster
  - more memory
- Lots of commercial systems
  - NLM (MEDLINE), Lockheed Martin (DIALOG), Mead (Lexis Nexis)
  - About 300 public-access systems by 1975
  - Larger collections
  - Simpler retrieval models (not the state-of-the-art)
  - Why?

History of IR
1970's

- Developers are skeptic that statistical methods work
  - lack of a clear evaluation methodology
  - partly due to large collections
  - Cranfield-style evaluation becomes difficult because assessors cannot judge all documents for every test query
  - boolean retrieval places the burden on the user
  - france AND history AND middle AND ages
- 1975: Karen Spark-Jones proposes pooling as a method for test-collection based evaluation

Test-Collection-based Evaluation

overview

- QUERY: parenting
- DESCRIPTION: Relevant blogs include those from parents, grandparents, or others involved in parenting, raising, or caring for children. Blogs can include those provided by health care providers if the focus is on children. Blogs that serve primarily as links to other sites or market products related to children and their caregivers are not relevant.

(TREC Blog Track 2009)

Test-Collection-based Evaluation

overview

- Collect a set of queries (input the system)
- For each query, describe the hypothetical user's information need (unknown to the system)
- For each information need, have human assessors determine which documents are relevant/non-relevant
- Evaluate systems based on the quality of their rankings
- Evaluation metric: quantifies the quality of a ranking with known relevant/non-relevant documents

Test-Collection-based Evaluation
pooling

- Given (any) query, the overwhelming majority of documents are not relevant
- Identify the documents that are most likely to be relevant
- Have assessors judge only those documents
- Assume documents outside of the pool are not relevant

Test-Collection-based Evaluation

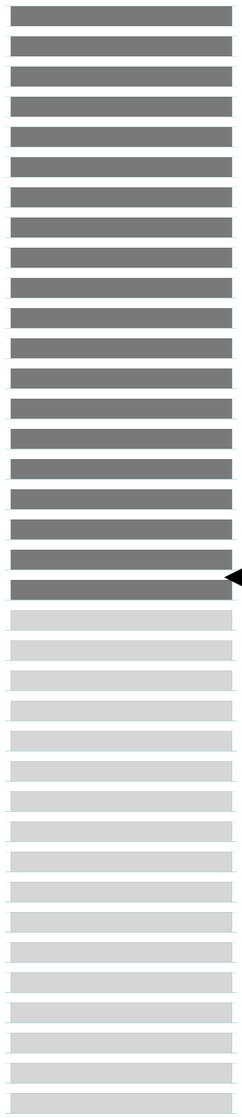
pooling

System A

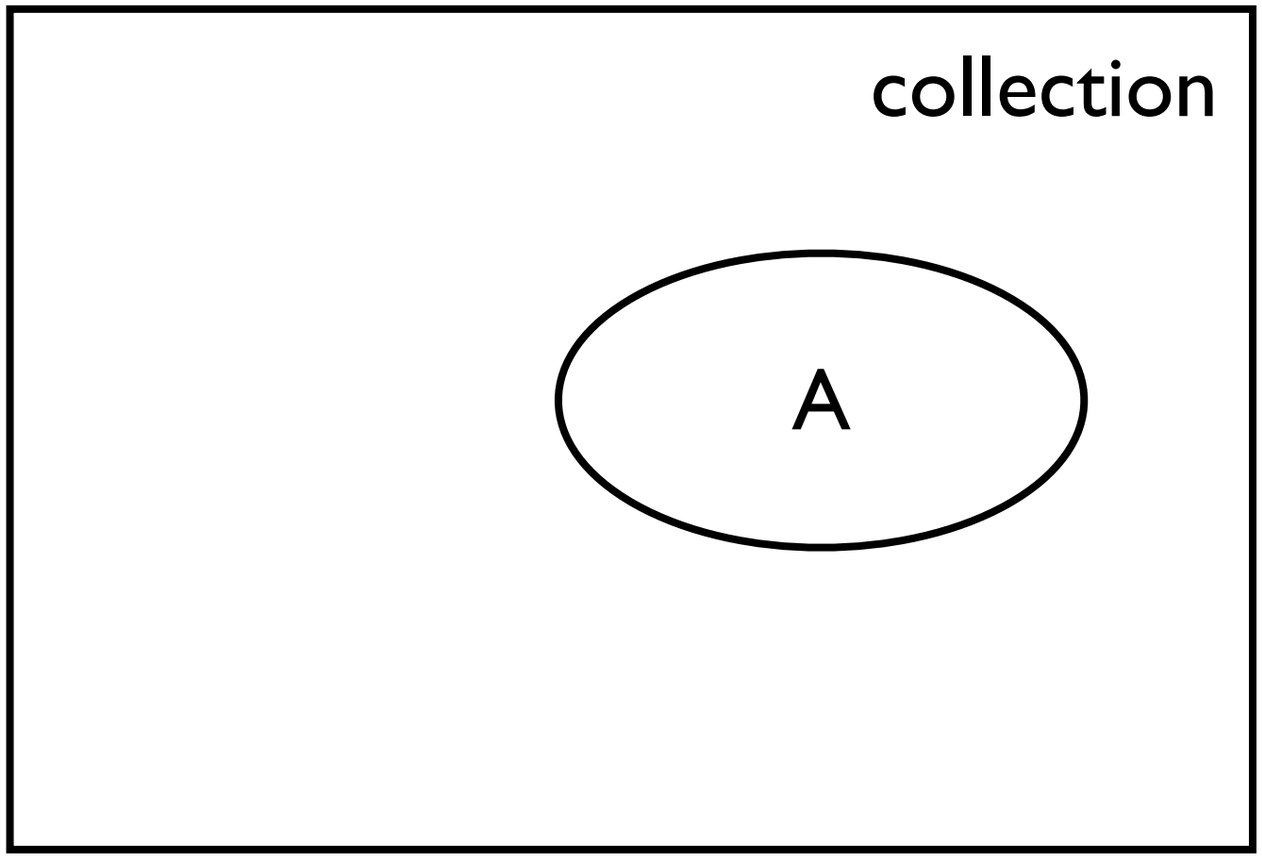
Test-Collection-based Evaluation

pooling

System A



k = 20

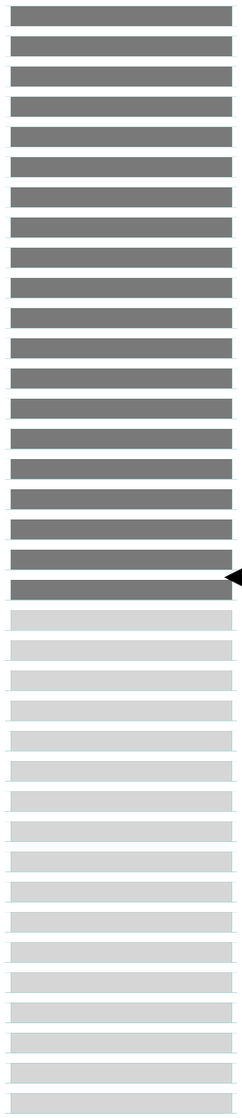


k = depth of the pool

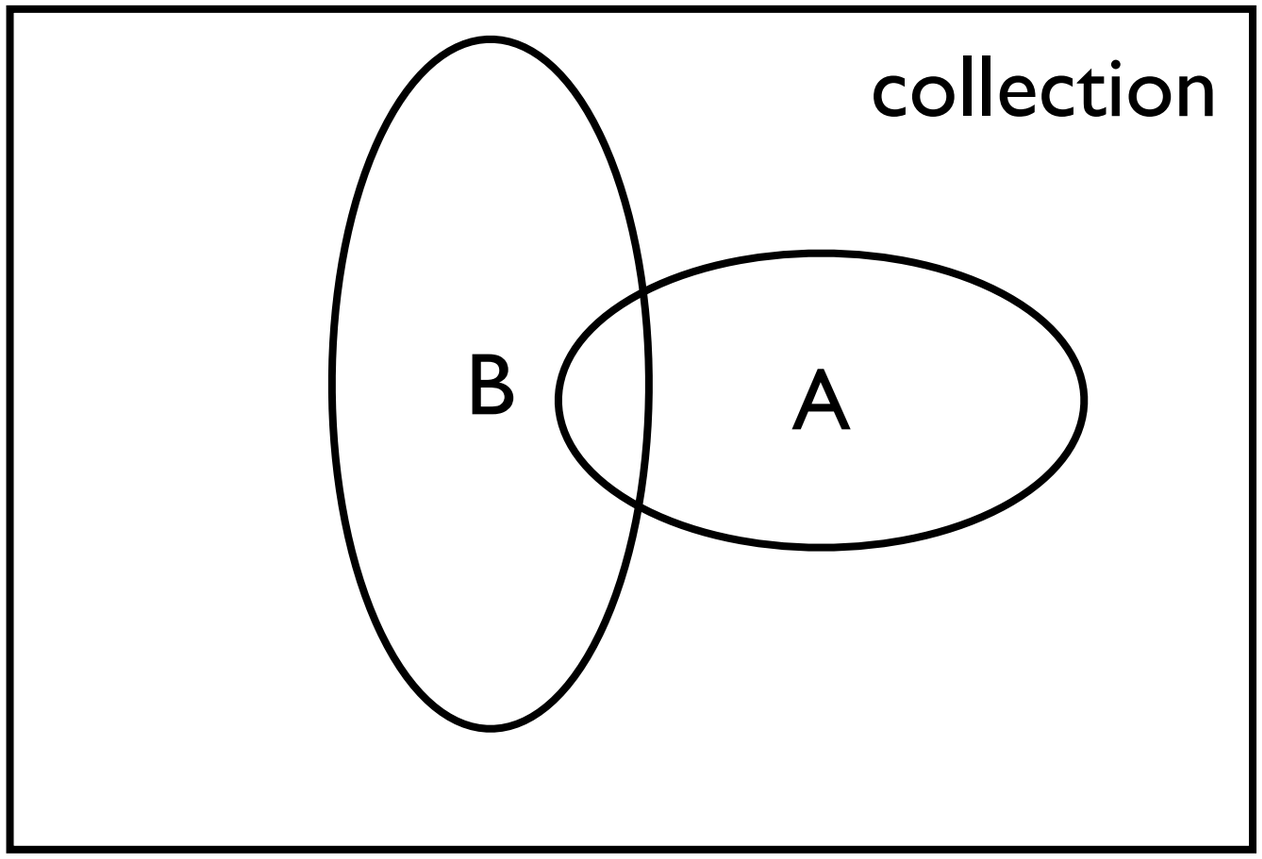
Test-Collection-based Evaluation

pooling

System B



k = 20

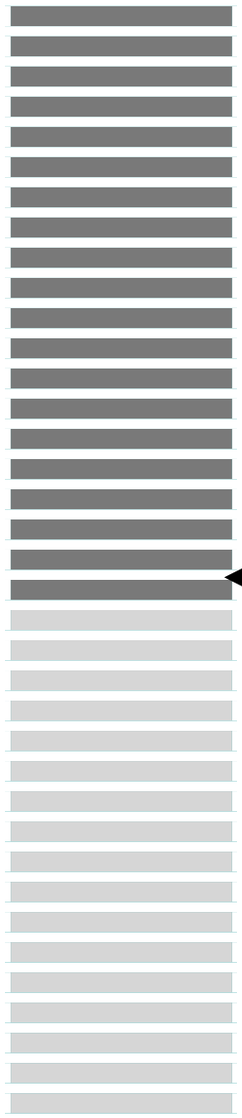


k = depth of the pool

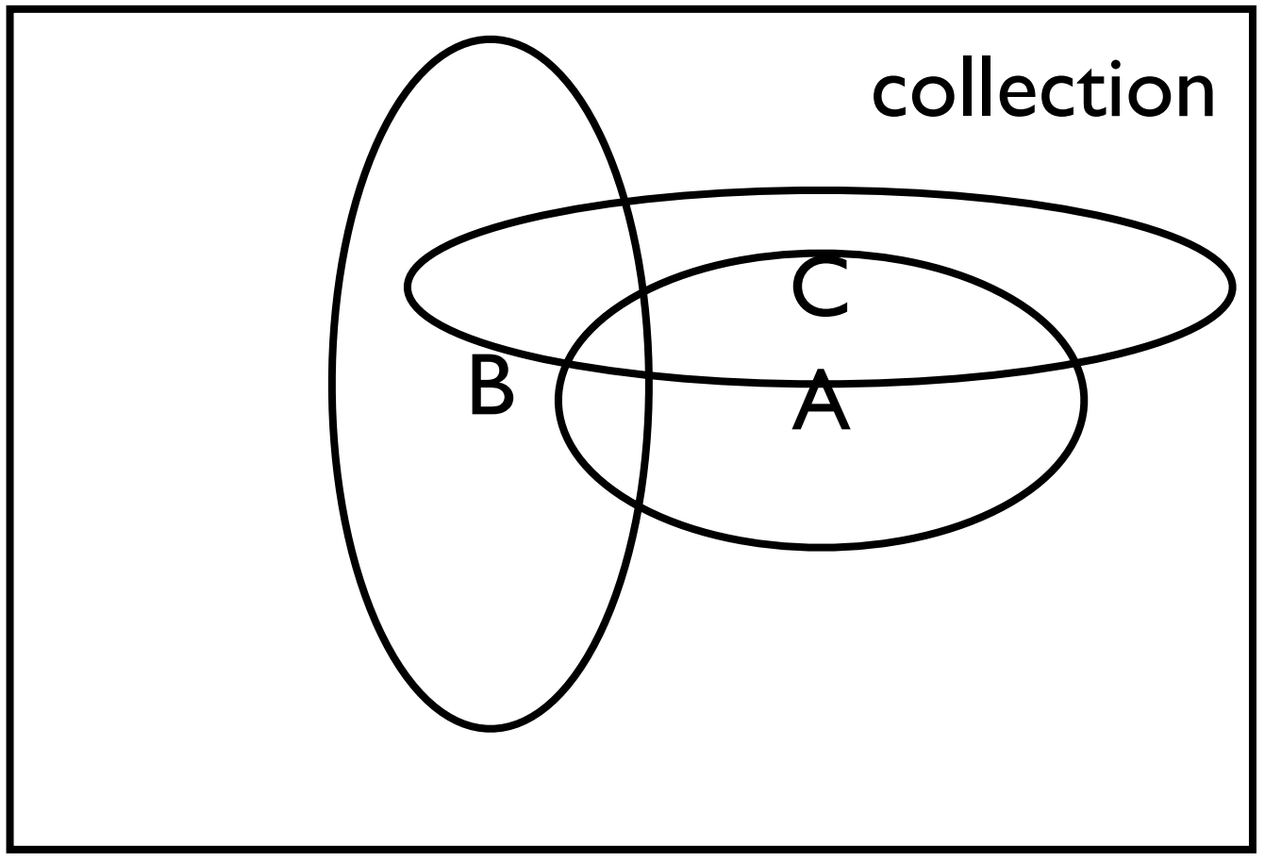
Test-Collection-based Evaluation

pooling

System C



k = 20

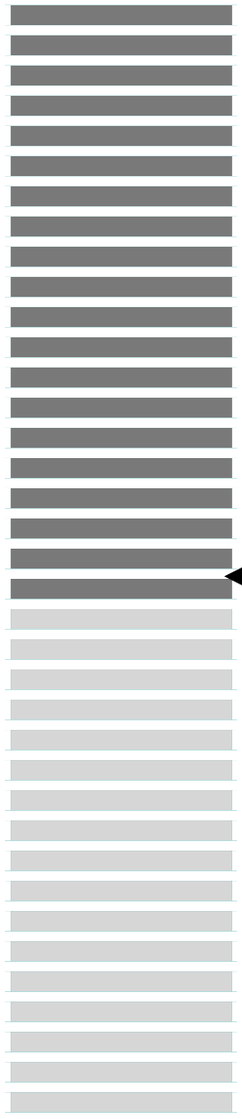


k = depth of the pool

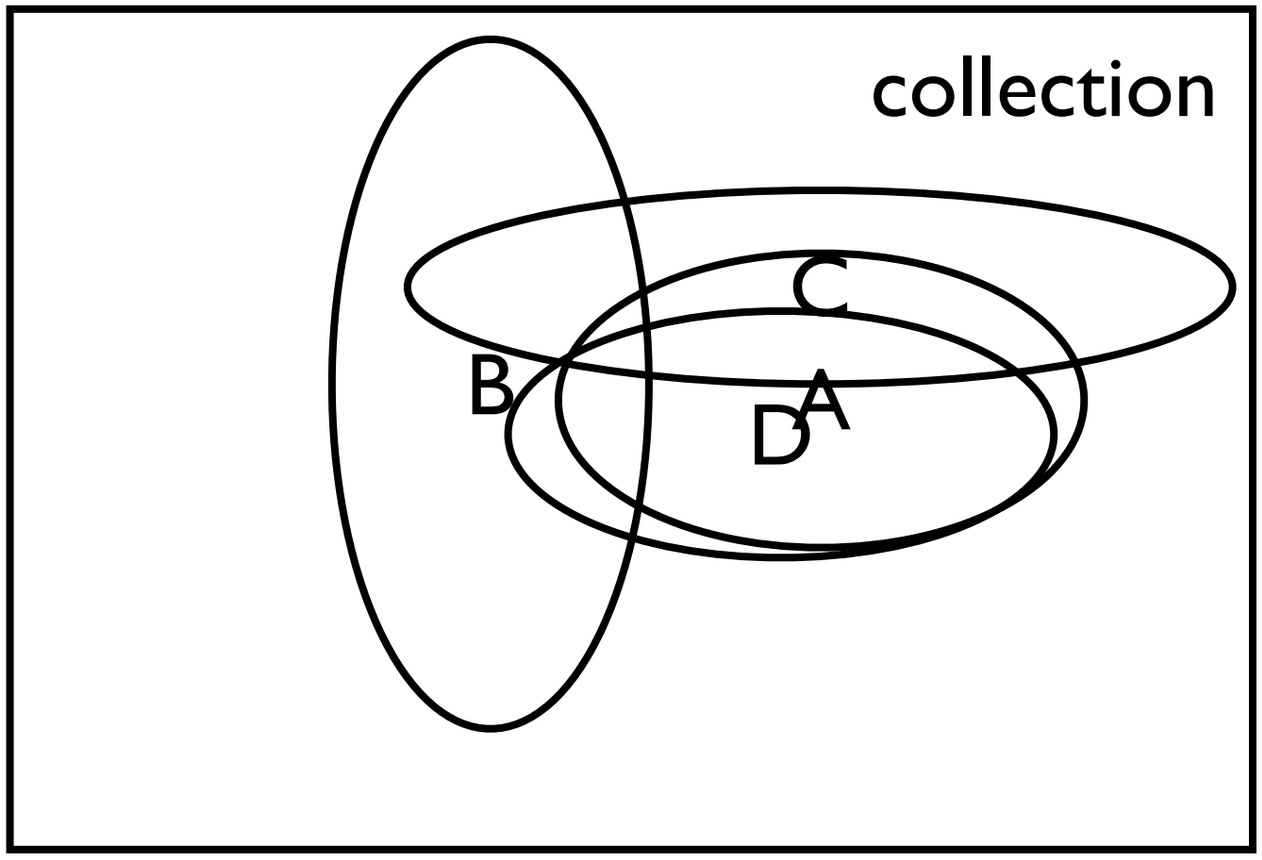
Test-Collection-based Evaluation

pooling

System D



k = 20

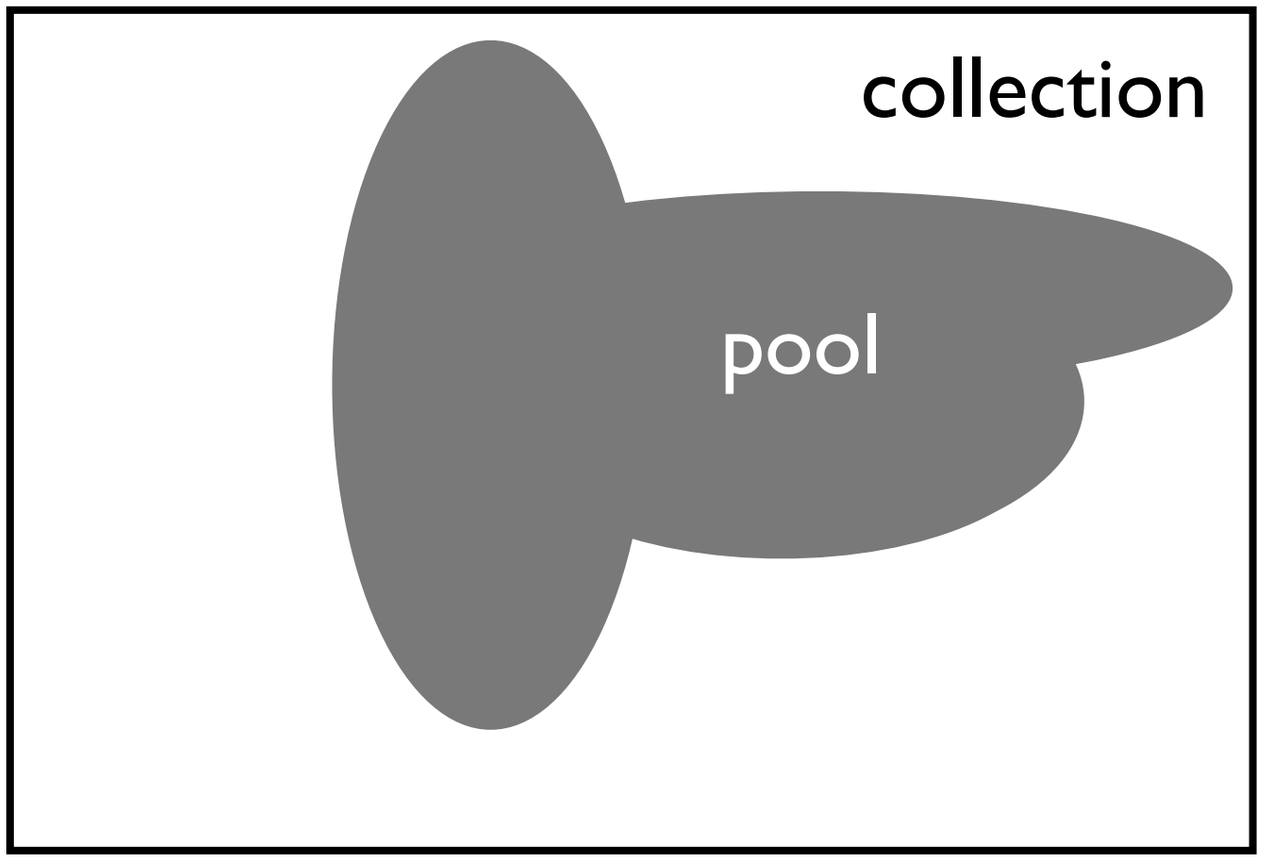


•
•
•

k = depth of the pool

Test-Collection-based Evaluation

pooling



- Documents in the pool are judged; documents outside the pool are assumed to be non-relevant

Test-Collection-based Evaluation
pooling

- Usually the depth of the pool is between 50 and 200 and the number of systems included in the pool is between 10 and 20
- A test-collection constructed using pooling can be used to evaluate systems that are not in the original pool
- However, what is the risk?
- And, how do we mitigate this risk?
- What is the benefit of systems that are in the pool?

History of IR

1945-1990

July 1945
Vannevar Bush's
"As We May Think"
appears in *Atlantic Monthly*

1945

1954
First operational
computer-based IR system
installed at Naval Ordinance
Test Station

1957
First major
IR evaluation study begins at
Cranfield Institute of Technology

1959
Hans Peter Luhn publishes
"Auto-encoding of documents
for information retrieval"

1961
SMART experiments begin

1967
Development of
MEDLINE begins

1972
SDC and Lockheed
begin offering online
IR services to the public

1977
SIRE experiments begin

SIGIR
Special Interest Group
on Information Retrieval

1978
First ACM SIGIR conference

1983
SIRE
goes
commercial

1987
First Message
Understanding
Conference
(MUC)

1989
TIPSTER
program
launched

1990

source: rowe *et al.*, 2010

TREC

1990's

Year	Event Details
1990–1991	Charles Wayne (DARPA) asks Donna Harman (NIST) to help create a new, large test collection for the TIPSTER Program
1991	Donna Harman creates data set with DARPA funding and suggests releasing the data to the public and holding a workshop to compare researchers' IR techniques
1992	First TREC held in Rockville, MD
1993	TREC 2 provides first true baseline performance analysis
1994	TREC 3 expanded to include new tracks
1995	TREC 4 involves official track structure
2000	TREC 9 is first "all-track TREC"

source: rowe *et al.*, 2010

TREC

TIPSTER Program

- 750,000 document collection (largest one to date)
- Two tasks
- Ad-hoc task: known static document collection with unknown queries (hence, ad-hoc)
- Routing task: known static query with unknown streaming documents
- Since the TIPSTER program, TREC has expanded to a wide-range of tasks

TREC

1990 and beyond

- Web search: retrieval from very large collections of web documents
- Multimedia search: retrieval of video and audio documents, not inherently associated with text
- Information extraction: retrieval of answers to the query, not just documents (e.g., question-answering, entity search)
- Domain search: retrieval from a domain-specific collection associated with many external resources (genomics track --- retrieval of biomedical literature)

TREC
Tracks

- Blog track: retrieval of entire blogs that are consistently about the query topic
- Cross-lingual track: retrieval of documents in any language for an English query
- Enterprise track: retrieval of content from various heterogeneous intranet sources
- Filtering track: identifying relevant documents from an input document stream

TREC

Tracks

- Hard track: additional information about the users intent (wants general or detailed information) and level of expertise (novice vs. expert)
- Interactive track: evaluation across multiple user interactions
- Novelty track: retrieval of documents that are relevant and not redundant
- QA track: retrieve answers to a question
- Robust track: evaluation based on robustness rather than average performance
- SPAM track: document spam-detection

TREC
Tracks

- Terabyte track: retrieval from very large collections
- Video track: automatic segmentation, indexing, and retrieval of video

TREC
challenges

- What are three trends that might make pooling a less attractive option for building an evaluation testbed?

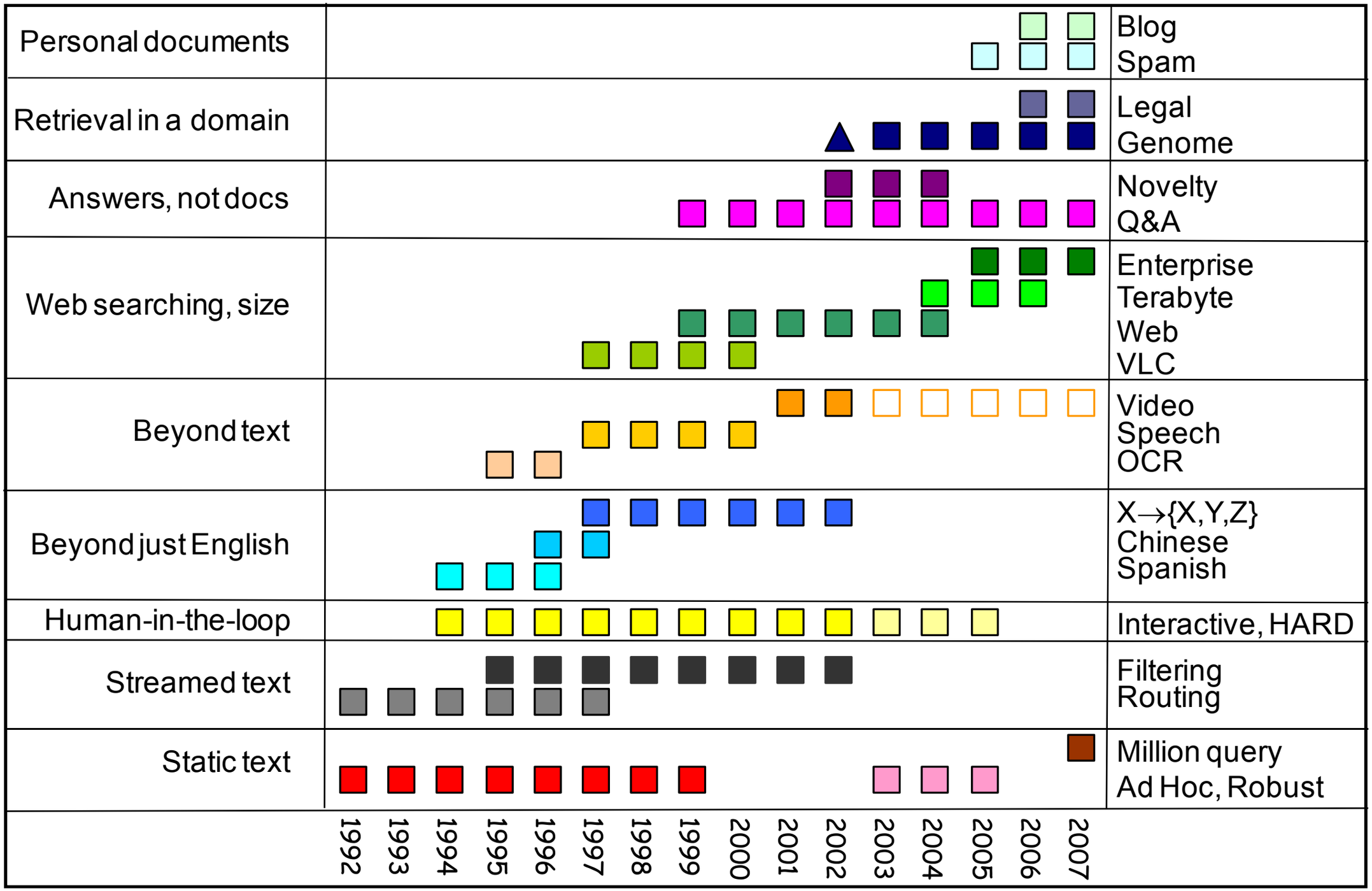
TREC
challenges

- What are three trends that might make pooling a less attractive option for building an evaluation testbed?
  - larger collections
  - more difficult tracks
  - more experimental systems

TREC
challenges

- Million-query track: experimental evaluation using more queries and fewer judgements per query
- Crowdsourcing track: collecting relevance judgements from non-expert, inexpensive assessors

TREC
tasks


Personal documents
Blog
Spam
Retrieval in a domain
Legal
Genome
Answers, not docs
Novelty
Q&A
Web searching, size
Enterprise
Terabyte
Web
VLC
Beyond text
Video
Speech
OCR
Beyond just English
X→{X,Y,Z}
Chinese
Spanish
Human-in-the-loop
Interactive, HARD
Streamed text
Filtering
Routing
Static text
Million query
Ad Hoc, Robust
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007

TREC
format

- What makes a good track task?
  - abstraction of real-world task
  - everything is the same except the system
  - the evaluation methodology and metric(s) must simulate real user behavior to some extent
  - not easy, not impossible
  - the task has relatively simple baselines
  - the adopted metrics should provide hints as to why a system is failing