TREC Style Evaluation

Donna Harman

National Institute of Standards and Technology

NIST
National Institute of Standards and Technology

Outline

I. why do we need TREC style evaluations
II. history and basic framework
III. TREC (and other TREC-like evaluations)
IV. how to use the test collections/how to build your own test collection
V. what are the limitations of current TREC style evaluations and where could we head in the future

Why TREC style evaluations

Users are not always around

Users are hard to control

Users are not sheep!!

- They are unpredictable, making it difficult to design an experiment that actually measures what you expected
- They are not homogeneous; they come to a task with different levels of knowledge, and they work/learn at different speeds, making user variation a major statistical problem requiring lots of users
- They are EXPENSIVE!!

When are they important?

- For interface design
- To identify critical issues in an information access task
- For operational system testing, such as pinpointing the needs for training
- To verify results from user simulation studies

When are they not necessary (or even negative)

- During initial system testing
- When tight controls are needed on variables, especially to see interactions
- When there needs to be many repetitions of a given experiment, usually with small changes in variables
- When there is a critical need for statistical significance testing (often difficult to get enough users)

History and basic framework

Early information access

- Before the web (1992) and before information was electronically available, most information access was via the library with librarians using indexed versions of journals/book lists (such as Index Medicus, Engineering Index, card catalogs, etc.)
- These indexes were manually produced, usually following (different) guidelines

Some manual indexing issues

- What terms to use to describe an article?
- How many terms to use?
- Should the terms be grouped into phrases rather than just single terms?
- Should the terms be selected from a controlled list?
- Should the terms be expanded using a thesaurus?
- Etc.

Cranfield experiments

- Designed and led by Cyril Cleverdon, head librarian at the College of Aeronautics, Cranfield, England in the 1960s
- Goal: To learn what makes a good set of indexing terms (descriptors)

Cranfield 2 indexing schemes

- Manual
  - four different types of indexing descriptors
  - three levels of exhaustivity (31, 25, and 13 descriptors)
- "automatic" indexing using the terms from abstracts and titles

What to measure

- How well the four descriptor types and three levels of exhaustivity (12 experiments) plus the "automatic" versions functioned when used as the descriptors in a search by a librarian
- To make the results statistically sound, he would have needed to do many searches involving a LOT of librarians
- So instead he simulated the task by creating a ***test collection***

His user simulation

- User model: researcher wanting all documents relevant to their question
- Documents to be searched: 1400 abstracts from recent papers in aeronautical engineering
- Questions were gathered from authors of the papers, asking for the basic problem the paper addressed and also supplemental questions that could have been put to an information service

Getting the correct answers

- Graduate students spent a summer checking the ~225 questions against all 1400 abstracts to find "possible" answers
- This was then filtered by authors
  - Complete answer to a question
  - High degree of relevance, necessary for work
  - Useful as background
  - Minimal interest, historical interest only
  - No interest

Final Cranfield test collection

- 1400 abstracts
- 225/221?? questions
- A list of abstracts for each question that are the correct answers (relevant documents for that question), broken into the 5 levels of relevance/non-relevance; note that ALL of the abstracts had manual relevance judgments

Cranfield experiment

- Librarians manually searched the abstracts for each question, using each of the 33 indexing descriptor combinations
- Recall and precision used as the metrics
- Results: single terms were best but the "automatic" indexing worked astonishingly well; this result led to major IR research
- Since the test collection was NOT based on the specific indexing methods used, it was infinitely reusable

Cranfield Paradigm

- Faithfully model a real user application, in this case searching appropriate abstracts with "real" questions judged by questioner
- Have "enough" documents and queries to allow significance testing on results
- Build the collection BEFORE the experiments in order to prevent human bias and to enable infinite reusability
- Base the metrics on how a user would see the results, i.e., intuitive metrics

SMART Test Collections

Name	#docs	# Q	#relevant	Comments
ADI/IRE-3 (1965)	82/780	35/34	4.9/17.4	Built by students
Cranfield (1967+)	1398/200 /424	225	7.2/4.7/ 6.4	Aeronautics
ISPRA (1967)	1268	48	17	Lib. Science
MEDLARS (1967-1970)	273/450 /1033	18/30	4.8+/ 9.2/23.2	Medical abstracts
OPHTH (1970)	853	30	30?	Specific MED
TIME (1970)	425	83	8.7	Full text
CACM (1982) ISI/CISI	3204 1460	52/64 76/112	15.3 49.8	CS real users bibliometrics

TREC (and other TREC-like evaluations)

Continuation in TREC

- In 1990 DARPA asked NIST to build a new test collection for the TIPSTER project
- User model: intelligence analysts
  - Large numbers of full text documents from newspapers, newswires, etc.
  - "formatted" queries called topics in TREC
  - High recall users meaning that "complete" relevance judgments were needed

TIPSTER Disk 1 and 2

Source	Size (MB)	documents	comments
Wall Street Journal, 1987-89 1990-92	267 242	98,732 74,520	
Associated Press newswire, 1989 1988	254 237	84,678 79,919	errors, repeats
Federal Register 1989 1988	260 209	25,960 19,860	Very long texts
Computer Selects articles (Ziff-Davis)	242 175	75,180 56,920	Different domain
DOE abstracts	184	226,087	Diverse domain

Sample TREC-2 Topic

<top>
<head> Tipster Topic Description
<num> Number: 104
<dom> Domain: Law and Government
<title> Topic: Catastrophic Health Insurance

<desc> Description:
Document will enumerate provisions of the U.S. Catastrophic Health Insurance Act of 1988, or the political/legal fallout from that legislation.

<narr> Narrative:
A relevant document will detail the content of the U.S. Medicare act of 1988 which extended catastrophic illness benefits to the elderly, with particular attention to the……..

……continued

Sample TREC-2 Topic, continued

<con> Concept(s):
1. Catastrophic Coverage Act of l988, Medicare Part B, Health Care Financing Administration
2. catastrophic-health program, catastrophic illness, catastrophic care,
acute care, long-term nursing home care
3. American Association of Retired Persons, AARP, senior citizen, National Committee to Preserve Social Security and Medicare
<fac> Factor(s):
<nat> Nationality: U.S.
</fac>
<def> Definition(s):
</top>

Sample TREC-3 Topic

<top>

<num> Number: 396

<title> sick building syndrome

<desc> Description:
Identify documents that discuss sick building syndrome or building-related illnesses.

<narr> Narrative:
A relevant document would contain any data that refers to the sick building or building-related illnesses, including illnesses caused by asbestos, air conditioning, pollution controls. Work-related illnesses caused by the building, such as carpal tunnel syndrome, are not relevant.

</top>

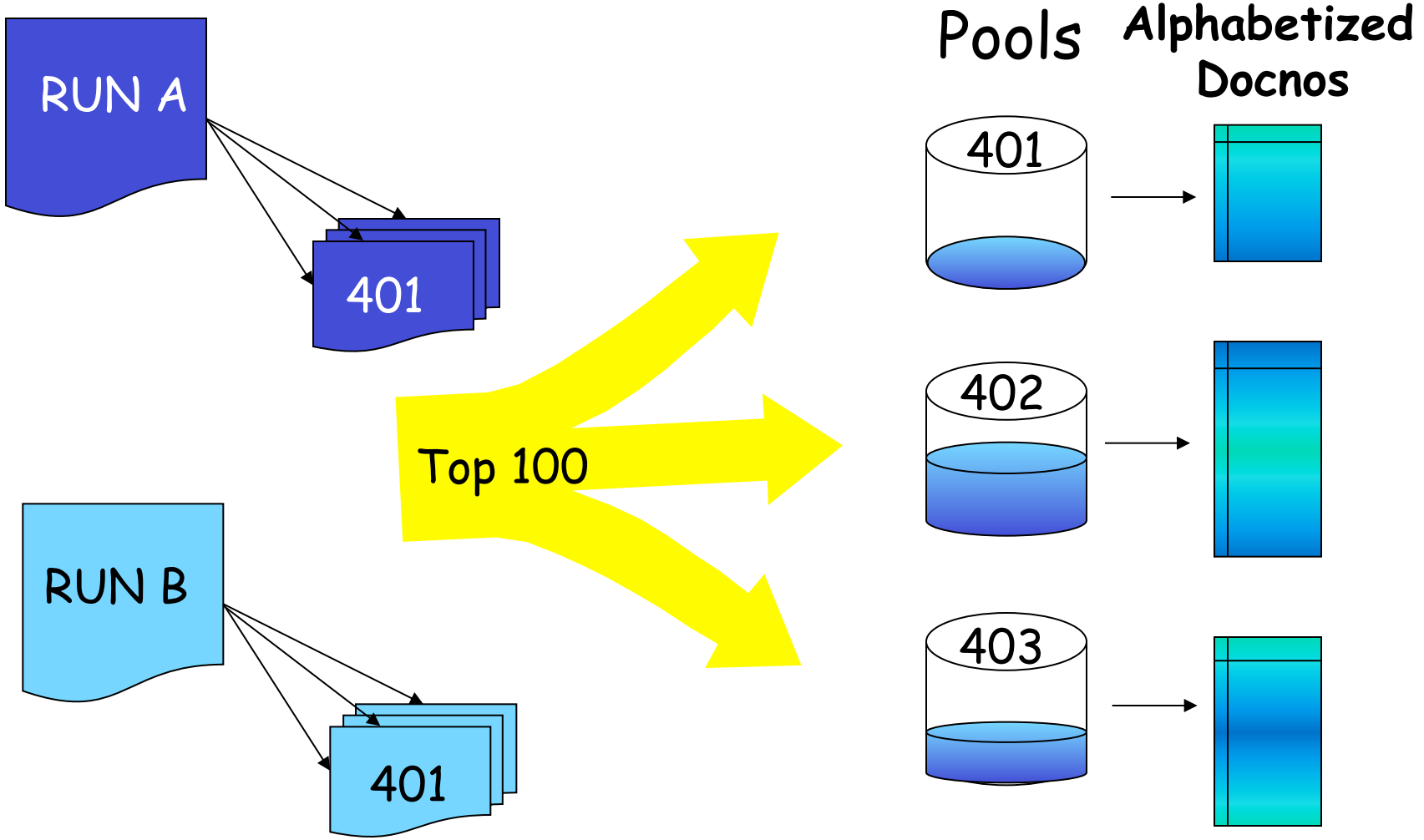
Relevance Judgments

Three possible methods for finding the relevant documents

FOR EACH TOPIC:

- Full relevance judgments on all 2GB of documents
- Relevance judgments on a random sampling of the document collection
- Relevance judgments on the sample of documents selected by the various participating systems
  - This method is known as the pooling method, and had been used successfully in creating the NPL and INSPEC collections.

Pooling


RUN A
401
RUN B
401
Top 100
Pools
Alphabetized Docnos
401
402
403

NIST
National Institute of Standards and Technology

What is relevant?

- Back to the user model (plus pragmatics)
- A document is relevant if you would use it in a report in some manner
- This means that even if only one sentence is useful, the document is relevant
- "Duplicates" also relevant as it would be very difficult to define and remove these

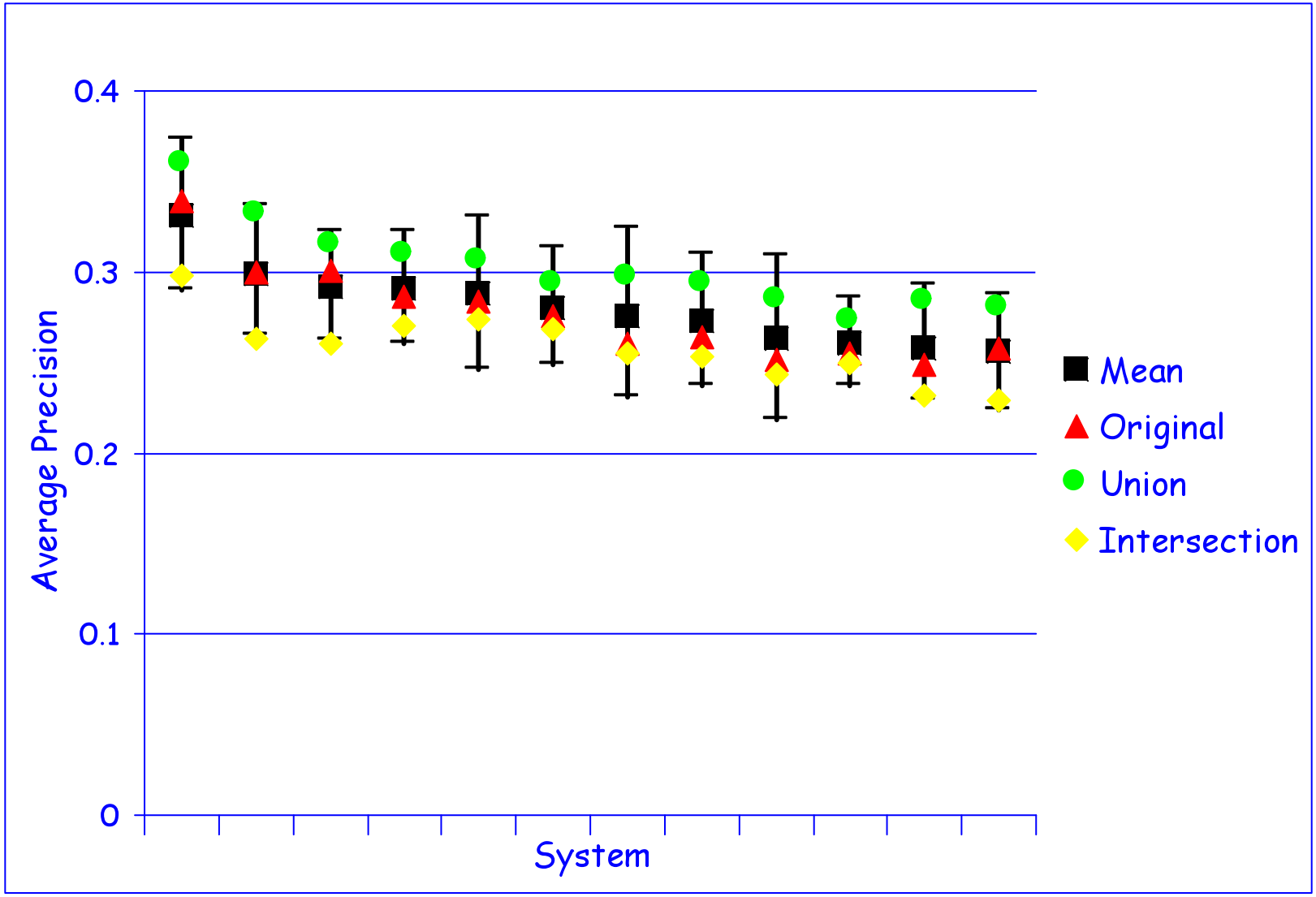
Relevancy FAQs

1) How do you know you have "all" the answers if not everything is judged??
   a. If documents that are not judged are automatically declared non relevant, isn't this biased against new systems, either not in the pool or "majorly" different in methodology?
2) These are manual judgments and there is known to be large variations of opinion; doesn't this make the results "unstable"?

How complete is relevant set?

- TREC-3 study: documents beyond rank 100 added to the pool for judgment
  - Some additional relevant documents found, however not enough to effect system ranking
  - topics with many relevant tend to have even more relevant documents
- Study by Zobel [SIGIR-98]: TREC ad hoc collections not biased against systems that do not contribute to the pools

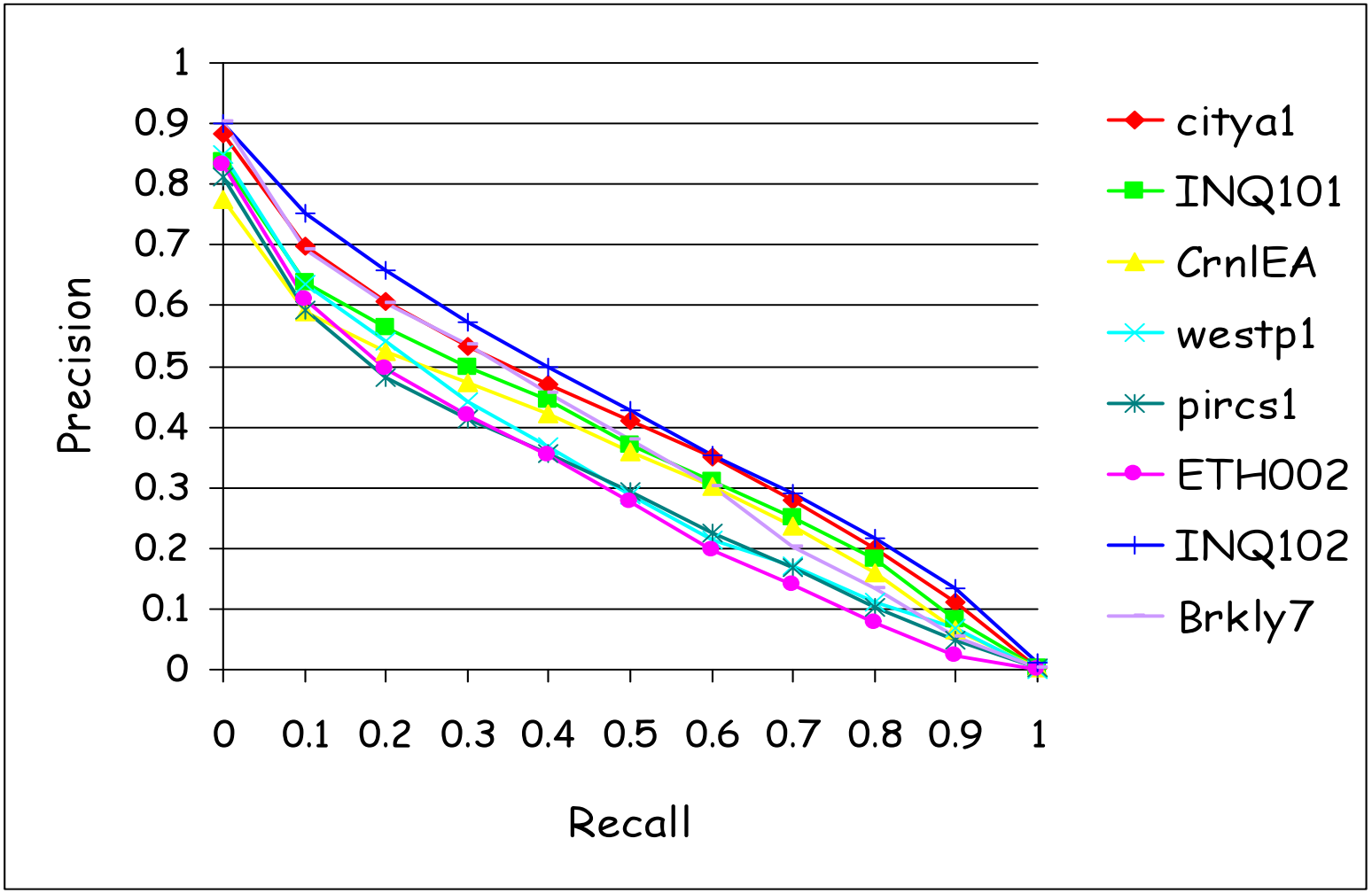
Stability of relevance judgments



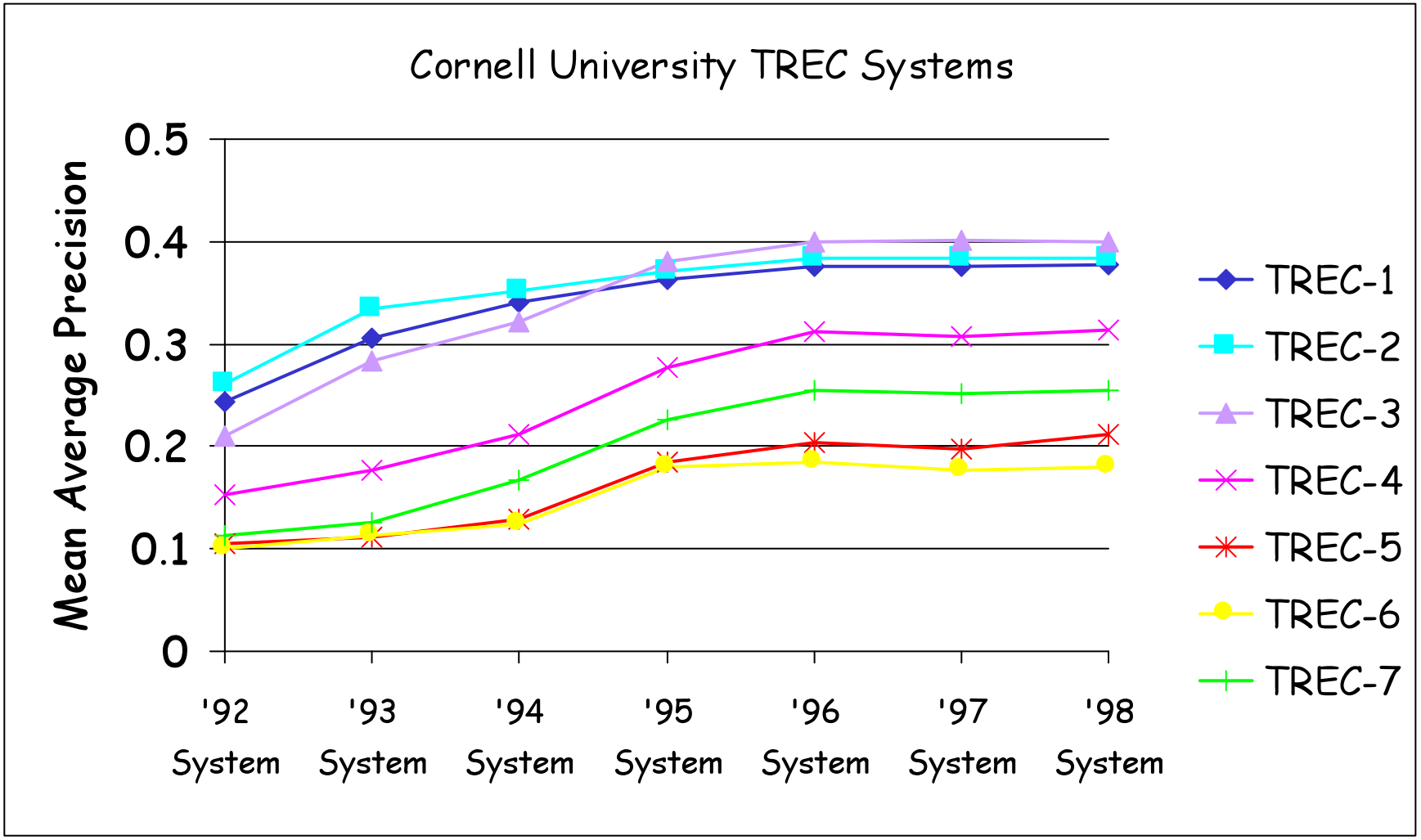
Other Relevancy issues

- Relevancy is time and user dependent
  - Learning issues, novelty issues
  - User profile issues such as prior knowledge, reason for doing search, etc.
- TREC picked the broadest definition of relevancy for several reasons
  - It fit the user model well
  - It was well-defined and thus likely to be followed
  - Thousands of documents must be judged quickly
  - This creates a collection which can then be subset

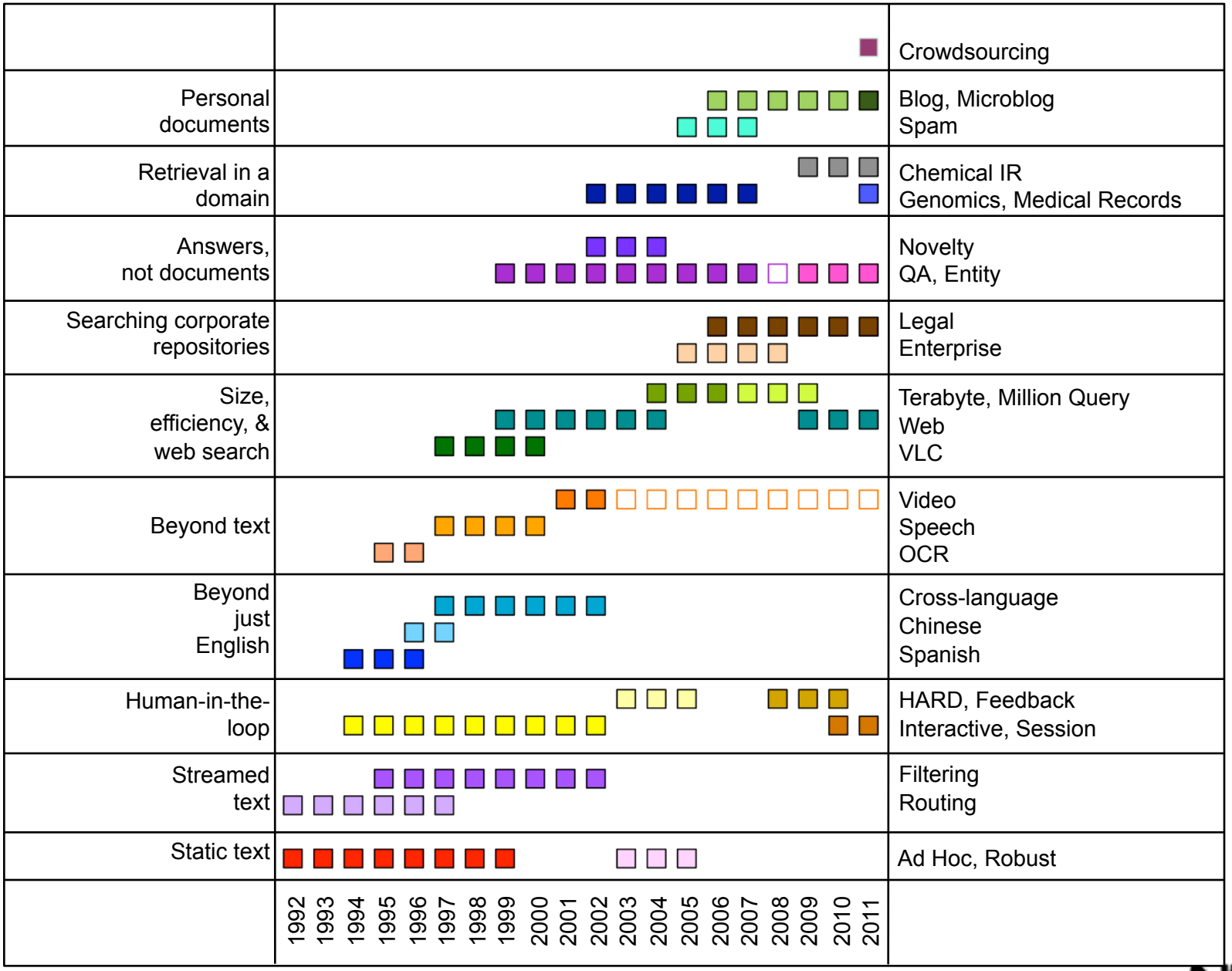
TREC-3 Ad Hoc runs



Performance improvements



TREC TRACKS


Crowdsourcing
Personal documents
Blog, Microblog
Spam
Retrieval in a domain
Chemical IR
Genomics, Medical Records
Answers, not documents
Novelty
QA, Entity
Searching corporate repositories
Legal
Enterprise
Size, efficiency, & web search
Terabyte, Million Query
Web
VLC
Beyond text
Video
Speech
OCR
Beyond just English
Cross-language
Chinese
Spanish
Human-in-the-loop
HARD, Feedback
Interactive, Session
Streamed text
Filtering
Routing
Static text
Ad Hoc, Robust
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

TREC-8 Cross Language Track

- User model: ad hoc search for documents written in a variety of languages using topics in one language
  - documents:
    - ~700,000 news articles from the same time period (French, German, & Italian from Swiss newswire SDA, German Swiss newspaper Neue Zürcher Zeitung, English articles from AP newswire)
  - Topics/relevance judgments done in a distributed mode by four countries in their native language
    - 28 topics (7 from each country) translated to all four languages; relevance judgments done within each language

TREC Genomics Track

- User model: medical researchers working with MEDLINE and full-text journals
- Topics
  - Started with a user survey looking for questions
  - Included topics based on 4 generic topic type templates and instantiated from real user requests
    - e.g., *What is the role of DRD4 in alcoholism?*
      How do HMG and HMGB1 interact in hepatitis?
- System response
  - ranked list of up to 1000 passages (pieces of paragraphs)
  - each passage a contribution to the answer
  - Usual difficulties with passage evaluation

TREC Legal Track

- Very dependent on user model
  - modeled after actual legal discovery practice with topics and relevance judgments done by lawyers
- **Documents:** 7 million messy XML records on tobacco
- **Topics:** (hypothetical) complaints (3 in 2008) with multiple requests to produce documents (topics) per complaint (45 topics in 2008)
- **Relevance judgments:** from pool created using sampling, over 500 per topic by law students
- **Metrics:** set retrieval, F at K (optimal cutoff)

TREC Web Tracks

- Initially used ad hoc user model, just scaled up to 100 gigabytes
- Also tried homepage finding, etc. where the goals (metrics) were early success
- Then scaled to 426 gigabytes (0.5 TB)
  - Judgments unlikely to be complete
  - Possible bias in relevant documents towards those that use of title words
- ClueWeb09 has 25 TB

Other domain/task models

- NTCIR patent retrieval tasks:
  - Patent tasks: Japanese and English patents, use of rejected patents as relevant documents, use of passages, patent classification of research papers
- ImageCLEF (2008 for example)
  - Photographic retrieval task: 20,000 color photos with captions in English or German; real search requests, results judged on relevancy and diversity
  - Medical retrieval task: 66,000 radiological images, topics must include at least 2 specific "axes" such as anatomical region, disease, etc.

The How-tos

1) How to participate in these evaluations
2) how to use their test collections in your own experiments
3) how to design and build your own test collection

TREC cycle

- Meetings are held each year in the middle of November at NIST
- Tracks for the next year are decided at that meeting and online discussions of the guidelines happen over the next couple of months
- Groups sign up in early February to participate in a track and receive information on the data
- Results are due sometime in mid-summer
- There is wide variation in data, the result format, deadlines, etc. across the tracks

How to participate

- Respond to the call for participation, get some background on TREC and also on (possible) earlier runnings of the track
- Carefully read the guidelines and join in the discussion to improve them
- Get the data, run the experiments, turn in results by deadline to allow attendance at TREC
- Write a notebook paper on those experiments for the November meeting
- Do further analysis of your results for the meeting and for later publications

Other TREC-like evaluations

- CLEF (labs run on a yearly schedule with meetings in Europe towards the end of September; work in CLIR, images, INEX, etc.)
- NTCIR (run on a 18-month cycle, with meetings in Tokyo in Dec or May; work in CLIR, patents, QA, MT, Geotemporal, etc.)
- FIRE (run on a yearly schedule with meetings in India; work in CLIR for Indian languages)
- Note there are similar things for NLP, MT, video, etc, etc, etc.

For more information, see

- TREC: trec.nist.gov
- CLEF: www.clef-initiative.eu
- NTCIR: research.nii.ac.jp/ntcir/
- FIRE: www.isical.ac.in/~fire/
- Note that these web sites host the publications, current meeting information, and also where to get the test collections for use outside of the evaluations

Using existing test collections

Using existing test collections

- The advantage of using an existing test collection is not just the cost savings but the fact that there is training data, and results to compare with, and publications using the data
- Existing test collections from all of these evaluations are generally available: see the home site of these evaluations for info
- It is critical to read the full set of information about these test collections to understand their limitations

What are important issues here

- Does the user model on which the test collection was based "match" the user model of your experiment so the results are applicable?
- For cases where there are multiple test collections for a given user model (such as the TREC ad hoc task), are you using the best one?
- For example, the TREC ad hoc collections from TRECs 7 and 8 are generally considered the best ones to work with; similarly some of the earlier collections for given evaluations are less desirable than later ones.

What about other collections

- For non-English ad hoc collections, or ones for CLIR research, check out CLEF, NTCIR, FIRE and the 2002 TREC Arabic ones
- For other areas, such as patents (NTCIR), image retrieval (CLEF), video (TRECvid), structured data (INEX), look at those web sites
- In using any test collection, however, it is CRITICAL to read as much as you can find about this collection because often there are unexpected interactions between the collection and your experiment that need to be recognized

What about using the older collections such as TIME, CACM

- This is generally a very bad idea!!
- All of them but TIME are abstracts rather than full text; we have moved beyond this
- As a learning exercise, it is OK to use the TIME collection, however any conclusions drawn from that collection need to be tested on the newer, larger collections
- In particular, it is unlikely that you will get a paper accepted using only the older collections; ideally it is best anyway to work with multiple collections to fully test ideas

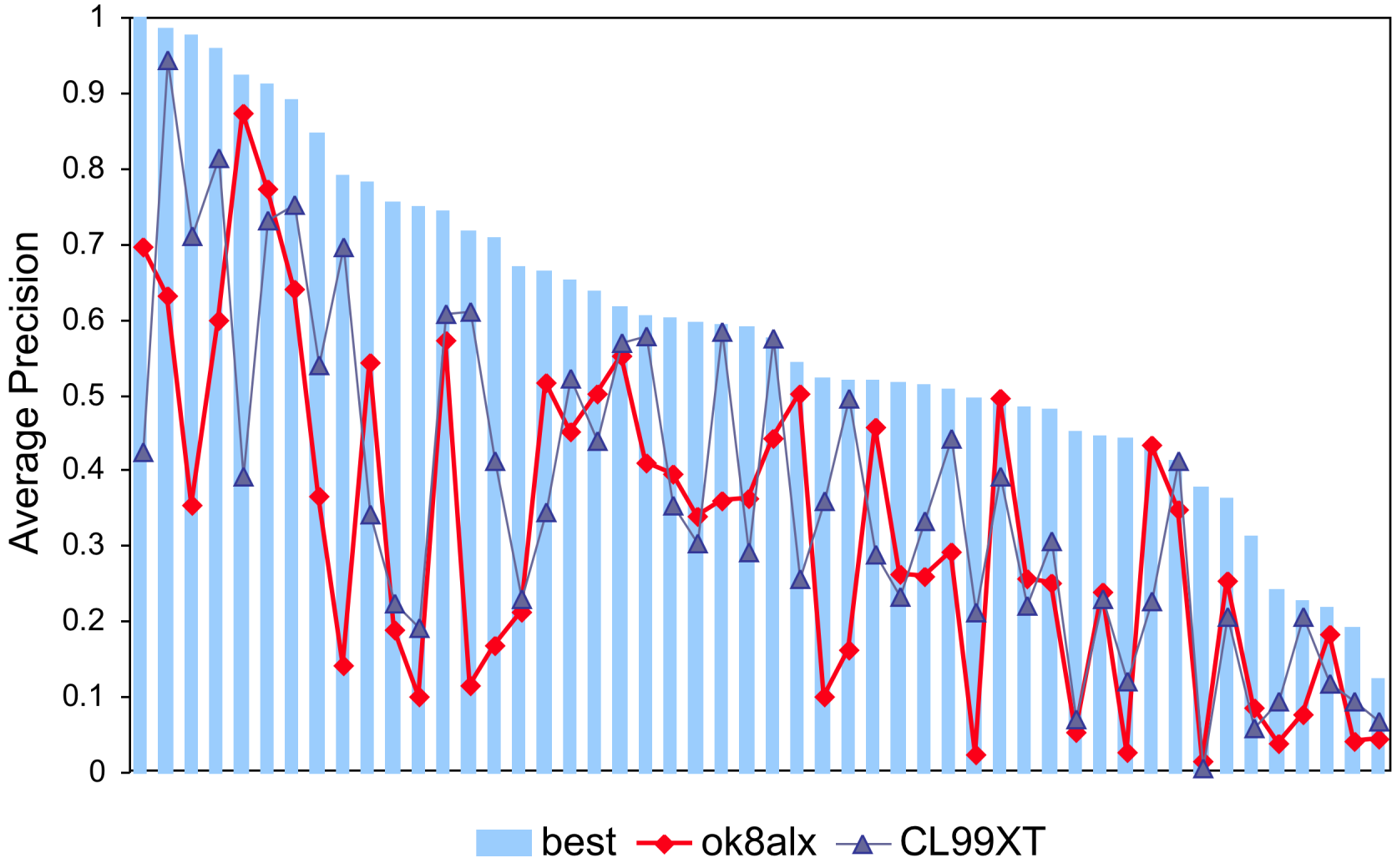
Running experiments

- Again it is critical to first read what others have done with the collection to help identify issues with the collection and your test area
- It is important to use a good baseline from previous work in comparing your results
- Baseline results that differ greatly from previous work need to be analyzed to see where the differences occur (coding errors or different system settings?)

Reporting results

- In reporting your work, it will be most useful if "complete" details are given so that others will know what was done, why it was done, how well it worked, and have some idea of why it worked (or didn't)
- Statistical analysis is important
- Also looking at the actual results
  - Which "pieces" of your new technique worked best
  - Check out some individual topic/question results such as looking at those that did NOT improve and why

Average Precision per Topic



Building new test collections

Building a new test collection

- This is harder, more time-consuming and more costly than you think!!
- It is absolutely CRITICAL that some real task be modeled in building a test collection
- This ensures that results are applicable to at least one task
- But most importantly it allows a natural and consistent approach to selecting testing methods and metrics

Documents

- What types of documents fit the user model; what are their characteristics?
- What can you actually obtain and can you get permission to let others use them?
- Do the documents need reformatting?
- How many documents are needed and if you are "subsetting" some natural set, how do you pick a representative subset?

Topic creation (some choices)

- Natural sources (likely the best)
  - Search logs (will require edits)
  - Other natural sources (FAQs, etc.)
- Working with surrogate users
  - Recruit based on user model
  - Try to align with their own interests
  - Have clear guidelines as to what is wanted
  - Likely some interaction with the data is necessary but must be done carefully

Topic creation (more choices)

- Central creation or distributed creation; this affects your costs and the variety of questions, but how do you keep "control" of the topic quality across sites
- Distributed creation is critical when multiple languages are involved; also if a wide variety of topics is important
- Costs rise steeply as the number of topics increase but too few topics will limit statistical analysis

Relevance assessments

- The user model MUST dictate the definition of relevant
- Is it reasonable to use graded relevance assessments; how do you define the grades clearly?
- Central or distributed across "teams"; how to create consistent judgments if done in "teams"?

Cost issues for assessing

- Largest financial cost in building test collections
  - reduce by short cuts/automatic methods from "found" data
  - distribute to participants
  - crowdsource
  - minimize judgments based on some type of sampling

Analysis of test collection

- How complete is it; this matters if you are going to distribute it for re-use
- Are the variations in relevance assessment large enough to effect relative results?
- How to insure "proper" understanding by later "outsiders"; good documentation is critical even without distribution

What are the limitations of current TREC style evaluations and where could we head in the future

Some outstanding issues

- Current pooling methods do not scale to the terabyte collections; new metrics are needed or else the collections cannot be considered reusable for recall
- Test collections in new areas such as personalization, medical records, etc. are increasingly difficult to build due to privacy issues

More outstanding issues

- There is not enough connection between users and TREC-type of evaluation
  - No interaction between system/user
  - Limited types of tasks being modeled
  - Etc., etc. etc.
- Ongoing work in "user simulation" such as the TREC "Sessions" track or the new "Contextual Suggestion" Track; other work in CLEF, FIRE

Other outstanding issues

- Statistical analysis/understanding results not well done currently (my opinion)
- Results not improving for basic retrieval, other things (CLIR) have hit the wall
- How do we move forward; where does TREC-style evaluation need to go??

TREC-style ad hoc experiments need to continue

- Scores still not good; we know from the RIAO workshop that there are "easy" things that could be done (probably on a per query basis) to improve results significantly
- There are many different information access needs that are basically traditional ad hoc retrieval; specific tasks, long queries, etc.
- However we need to think more about web/ mobile applications: where the action is!

ClueWeb09

- If we were going to do ad hoc retrieval, where do we get "enough" topics?
- How do we get relevance judgments; is it possible to sample and still have "reusable"?
- Is reusable important; how do we reconcile the fact that users only look at the top ranks (the web user model) with the need for reusability of a collection?

Specific subsets of Web

- Rose & Levinson, WWW2004: other subsets such as urls, products
- Clough et. al., SIGIR09 poster: diversity
- Downey et. al., CIKM'08: user interaction with rare vs common queries
- Bendersky & Croft, WSCD'09: long queries

User simulation

- Lin & Smucker (SIGIR 2008) suggested that Cranfield is only one model for user simulation
- We have log studies, plus examples of feature tables from log studies to provide some reality
- Can we convince student interns at search engine companies to investigate phenomena that will allow user simulation outside the company
- ?

Further reading

- Historical Cranfield: Cyril Cleverdon's talk at SIGIR 1991
- TREC book (Voorhees and Harman), MIT Press 2005
- Information Retrieval Evaluation (Harman), Morgan/Claypool series, 2011
- The various TREC, CLEF, NTCIR, etc. websites
- SIGIR, ECIR, CIKM, WWW, WSCD etc.

Thanks!!

Questions???