Coopetition in IR Research

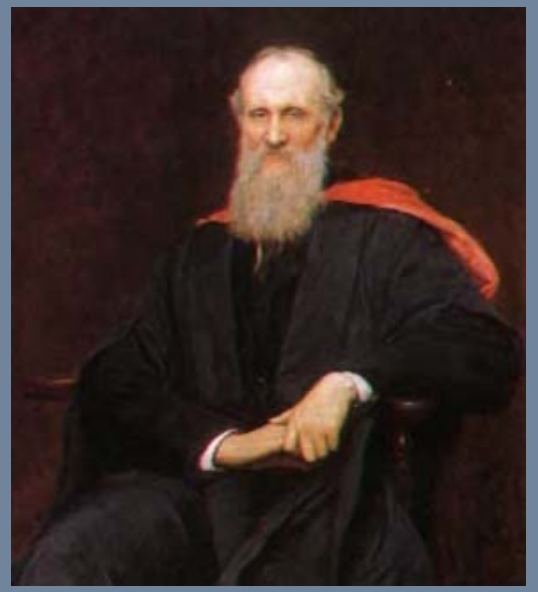
Ellen M. Voorhees

National Institute of Standards and Technology
U.S. Department of Commerce

National Institute of Standards and Technology

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.

NIST mission statement



If you can not measure it, you can not improve it.

Lord Kelvin

Dr. Kelvin

Symbol: K

SI unit of
TEMPERATURE

Candela

Symbol: cd

SI unit of
BRIGHTNESS

Monsieur Kilogram

Symbol: kg

SI unit of
MASS

Major Uncertainty

Symbol: ?

ERROR!

Meter Man

Symbol: m

SI unit of
LENGTH

The Mole

Symbol: mol

SI unit of
AMOUNT OF SUBSTANCE

Ms. Ampere

Symbol: A

SI unit of
ELECTRIC CURRENT

Professor Second

Symbol: s

SI unit of
TIME

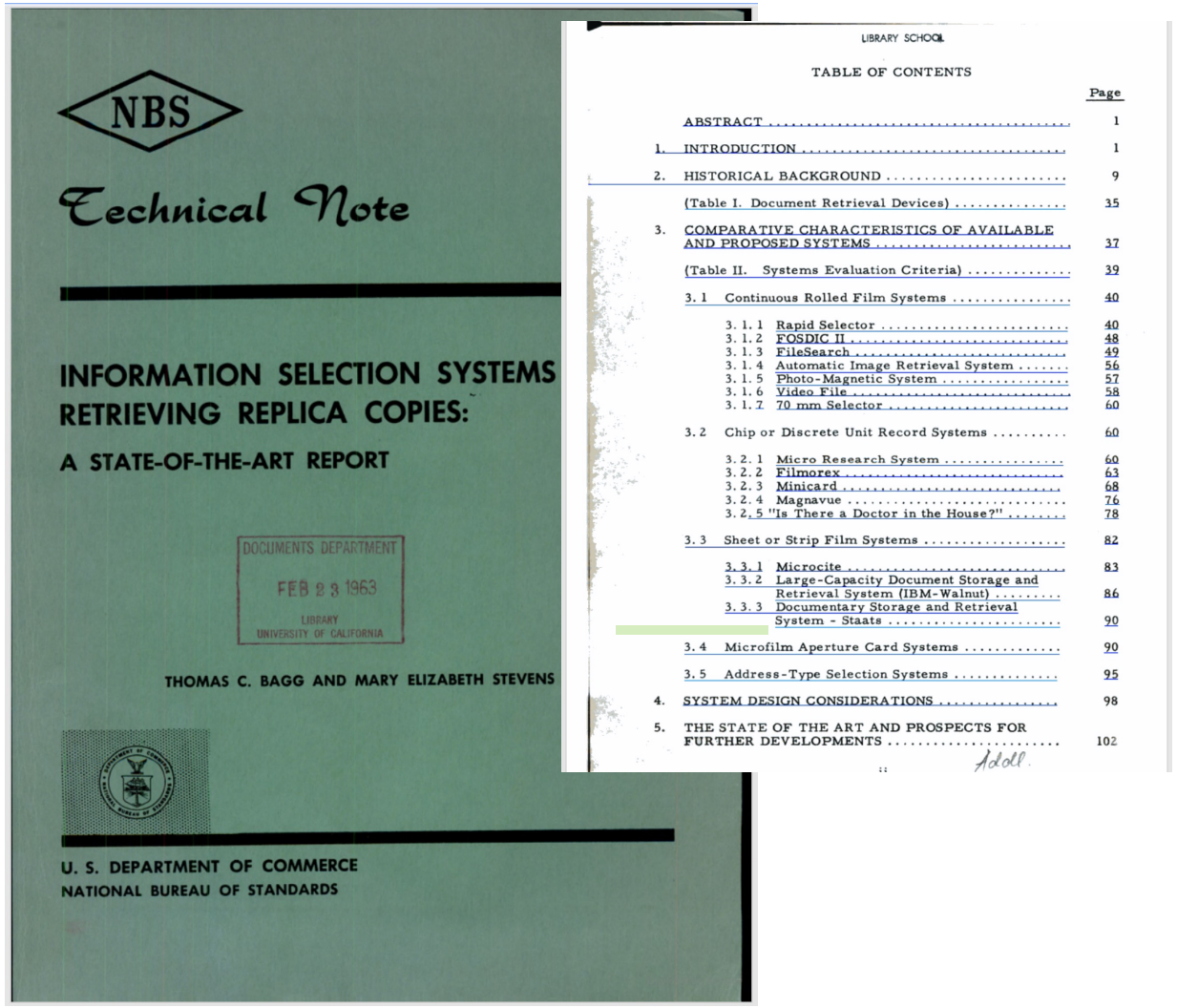
Human Language Technology Measurement

Ida Rhodes and the Problem with 'Water Goats'

2

BY KATIE RAPP ON MARCH 16, 2016 ELECTRONICS, INFORMATION TECHNOLOGY, MATHEMATICS AND STATISTICS, NIST GENERAL

NIST (NBS) has a long history of language evaluation work



NBS

Technical Note

INFORMATION SELECTION SYSTEMS RETRIEVING REPLICA COPIES:

A STATE-OF-THE-ART REPORT

THOMAS C. BAGG AND MARY ELIZABETH STEVENS

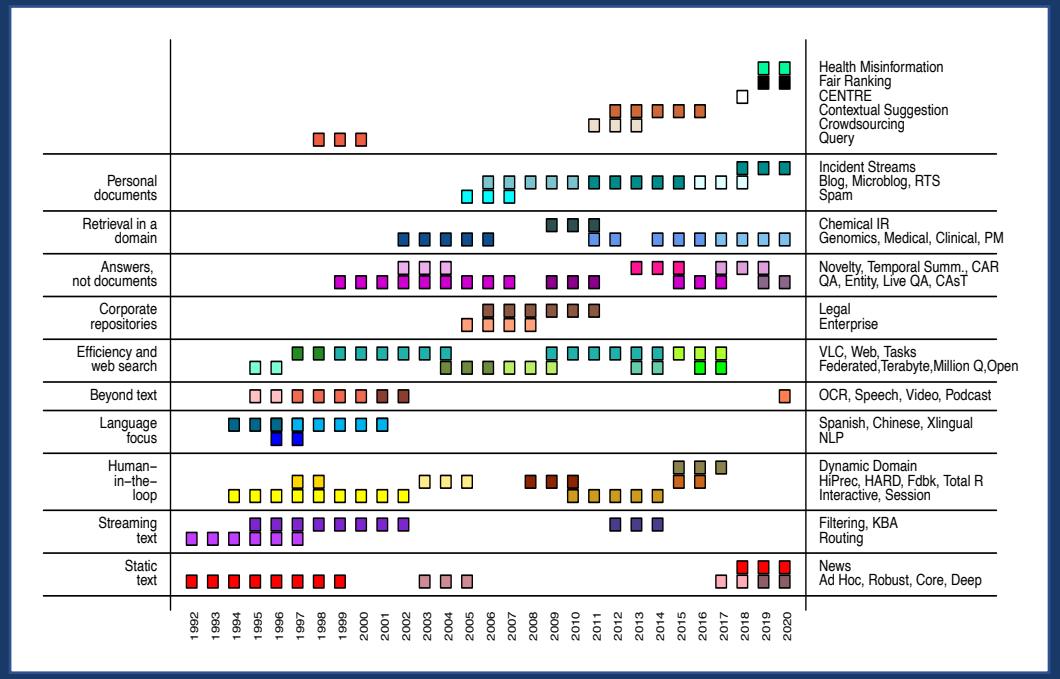
U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

LIBRARY SCHOOL

TABLE OF CONTENTS

Text REtrieval Conference (TREC)

Series of community evals (“coopetitions”) that build research infrastructure.



- pioneered use of “pooling” for building large collections
- built > 150 test collections for dozens of search tasks
- hundreds of participant teams world-wide
- premier venue for determining research methodology

Coopetition?

Competing may give you a bigger piece of the pie …

- Benchmark: standardized task
- Coopetition: “cooperative competition” (Wikipedia)
- In IR, built through events such as TREC, CLEF, NTCIR, FIRE, INEX…

Coopetition?

…while cooperation makes the whole pie bigger.

- Wikipedia: “cooperative competition”
- In IR, known by a variety of other names:
  - challenge problems
  - community evaluations
  - shared tasks
- Exemplified by TREC, CLEF, NTCIR, FIRE, INEX…

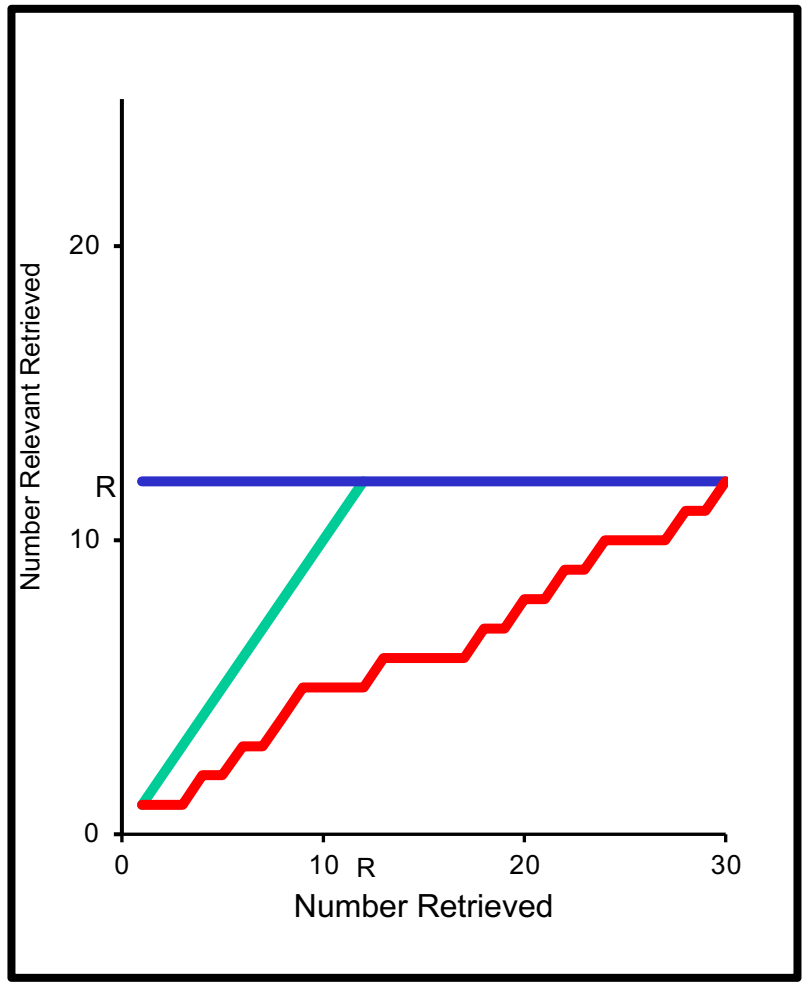
Cranfield Paradigm

Cyril Cleverdon

a test collection

- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
- Test collections
  - set of documents
  - set of questions
  - relevance judgments

Cranfield Paradigm



- Retrieval system response to a question is a ranked list of documents.
- The ideal output is a list with all relevant documents ranked before any non-relevant document.
- Easy to compute a variety of different evaluation measures from a ranked list if you know the set of relevant documents

Cranfield Paradigm

Defines “core competency” of IR: retrieve relevant before non-relevant

Different effectiveness metrics provide different abstractions of real-world task(s).

Abstracted task is a necessary but not sufficient proxy for real task.

Significance tests suggest whether a change is worthwhile.

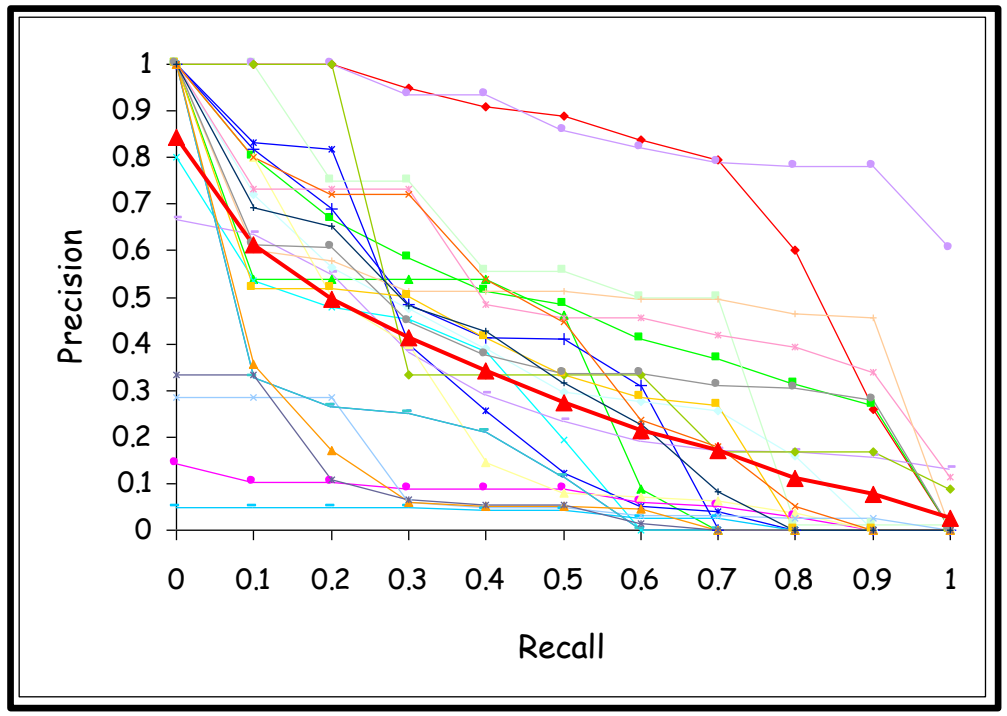
What is a good search result?

Search is
inherently a user activity

Different users have different
(conflicting) criteria for success

There is no single Truth

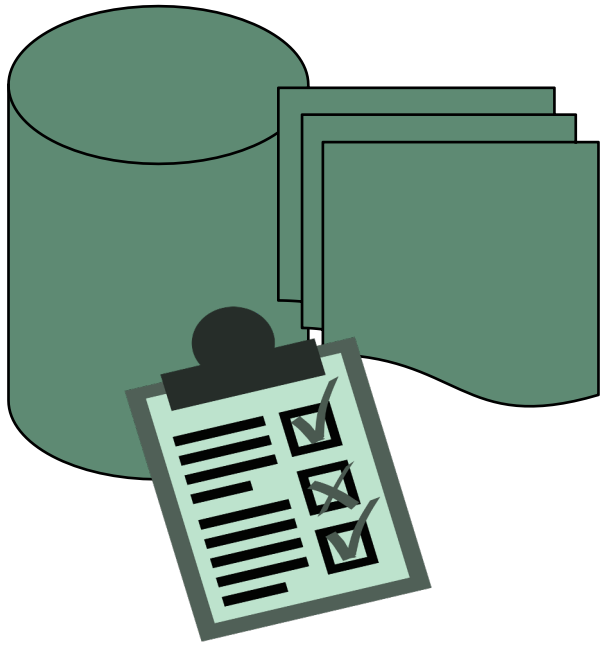
Scoring Ranked Lists



Interpolated precision at standard recall points for 25 individual topics from the best TREC-7 ad hoc run, plus the average curve over all 50 topics in TREC-7 test set for that run (heavy red line).

- Variance across topics is large---larger than differences between systems
- Signal to be measured is small relative to noise
- This variability is why Cranfield averages scores over many topics

Rationale for Cranfield

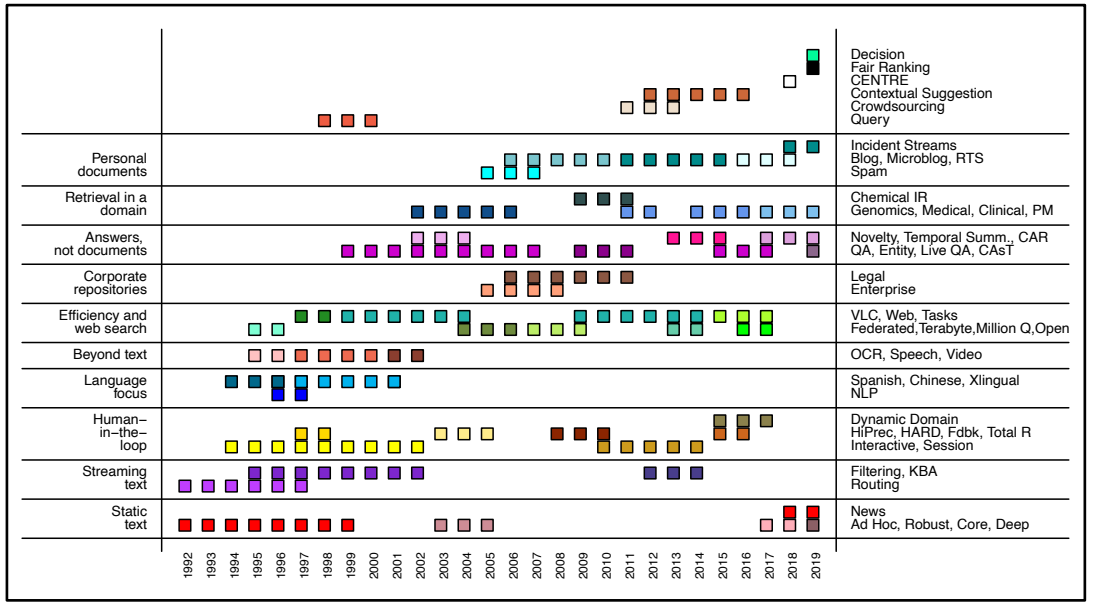


Sufficient fidelity to
real user tasks to be informative

General enough to be broadly applicable,
feasible, relatively inexpensive

Lose realism to gain
control over variables:
more experimental
power at lower cost

Coopetitions in IR

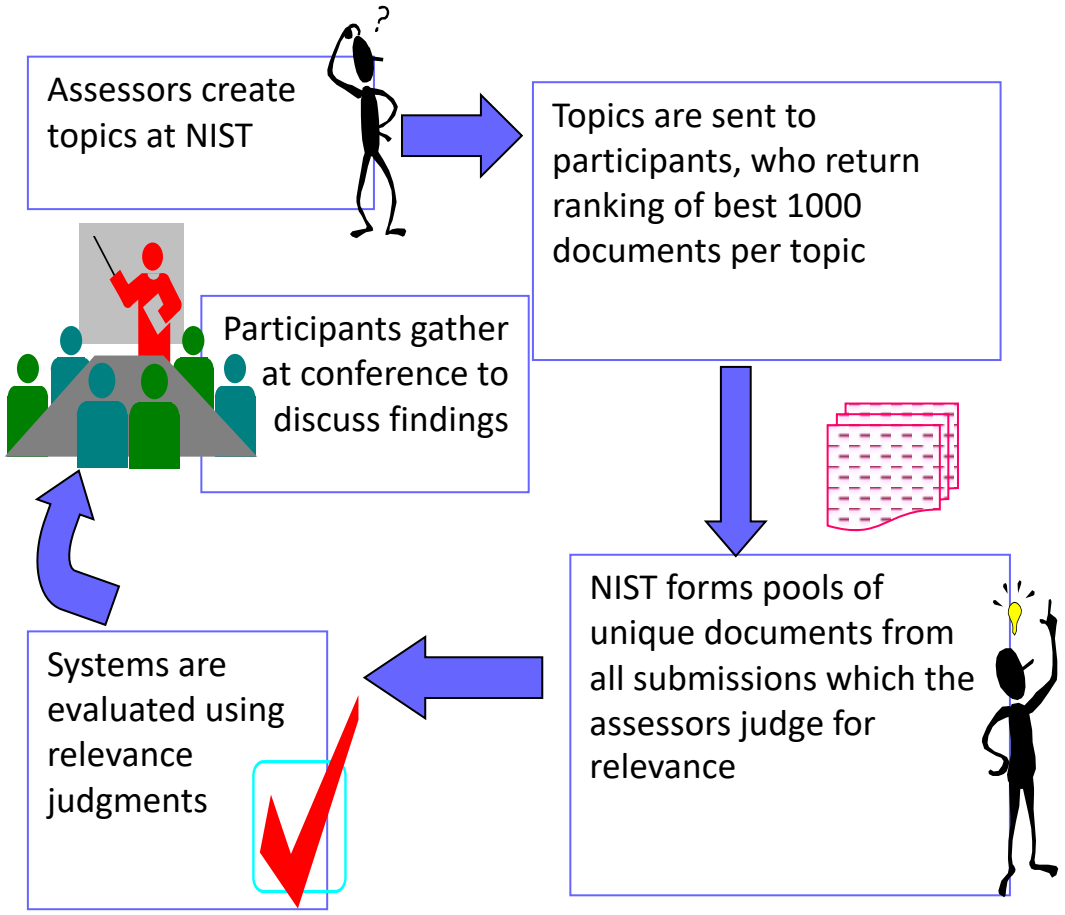


TREC tracks through the years

- Competitions are where the infrastructure that enables Cranfield is created
- Infrastructure includes
  - test collections
  - task definitions (with measures)
  - methodology

Coopetitions in IR

- Organizers provide appropriate data set, generally “documents” and “topics”.
- Participants use their own systems to create results that are returned to organizers.
- Some of the results get annotated by humans.
- Annotations used to score results; scores returned to participants.
- Data, annotations, results, scores archived for future use.



Benefits of Coopetitions

Improve the state-of-the-art

Establish the research methodology

Form/solidify a research community

Facilitate technology transfer

Amortize the costs of infrastructure

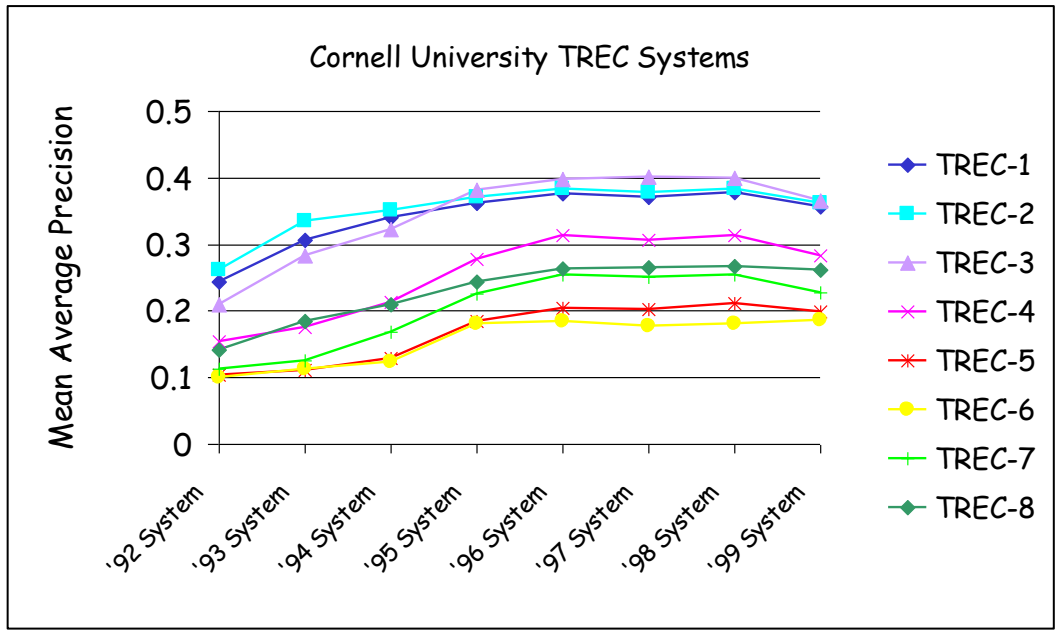
Risks of Coopetitions

Community overfitting to single dataset/task

Poor task abstraction

Method conformity

State of the Art



The TREC data revitalized research on information retrieval. Having a standard, widely available, and carefully constructed set of data laid the groundwork for further innovation in the field. The yearly TREC conference fostered collaboration, innovation, and a measured dose of competition (and bragging rights) that led to better information retrieval.

Hal Varian, Google Chief Economist
March 4, 2008

- Improve and document the state of the art
  - scores doubled in early years of TREC
  - improvement driven largely by the test collections constructed
- Risk: entire community overfits
  - construct multiple collections for a task
  - evolve task

Research Community

- Infrastructure enables research in an area, which attracts critical mass
- Risk: poor task abstraction
- Good eval task:
  - abstraction of real-world task so variables affecting performance can be controlled…
  - …but must capture salient aspects of real task or exercise is pointless
  - metrics must accurately predict relative effectiveness on those aspects
  - adequate level of difficulty
  - best if measures are diagnostic

This project [the TREC Legal track] can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.

Magistrate Judge Paul W. Grimm
Victor Stanley v. Creative Pipe

Research Methodology

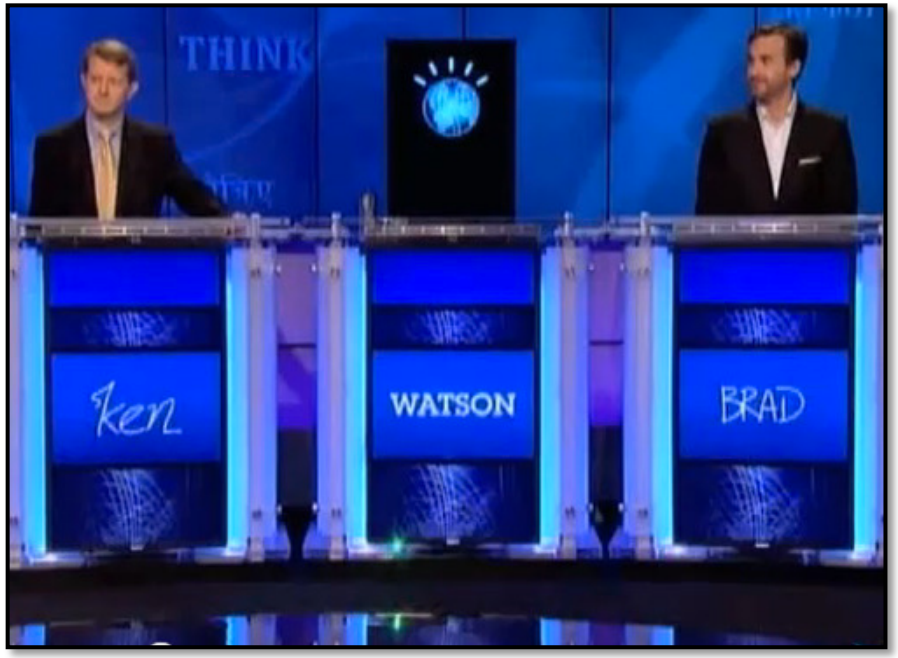
Image: arielrobin/Pixabay

TREC is an annual benchmarking exercise that has become a de facto standard in Information Retrieval evaluation.

Stephen Robertson
Microsoft
SIGIR 2007

- Way to do research in area
  - measures
  - experimental design
- Standardize evaluation measures
  - e.g., interpolation
  - e.g., measure parameter values
- [Risk: data set abuse
  - e.g., compare MAP on TREC web collections
  - e.g., relevance == popularity]

Technology Transfer



TREC has proven to be a valuable forum in which IBM Research has contributed to an improved understanding of search, while at the same time the insights obtained by participating in TREC have helped to improve IBM's products and services.

Alan Marwick, et al.
IBM chapter of the TREC book
2005

- Effective approaches adapted and incorporated into variety of methods
  - e.g., weighting schemes
  - transfer among all sectors
- Risk: conformity of techniques
  - don't want to standardize too early

Costs of Infrastructure

- Building infrastructure does require resources
  - financial
  - researcher time
- Shared infrastructure amortizes the costs over entire community
- [Risk: costs of participation
  - individual can save participation costs assuming others participate so infrastructure gets built
  - minimize coopetition-only costs such as reporting formats]


WRALtechwire

HOME ADVERTISE MOBILE

Updated Jan. 17, 2011 at 6:03 a.m.

RTI study: Search engines' success driven in big part by fed support

Post a comment

By RICK SMITH, LTW Editor

(Note: This post has been updated to restore a link to the NIST search engine report)

The Skinny blog is written by Rick Smith, editor and co-founder of Local Tech Wire and business editor of WRAL.com.

RESEARCH TRIANGLE PARK, N.C. – Where would Google, Bing and other search engines be today without federal support? Well, you might be surprised to learn that for all the genius attributed to their creators, the feds had a big role to play in today's taken-for-granted global instantaneous search capability.

So concludes a new study out from RTI International.

It's common knowledge that without research and development support from the Pentagon and the federal government, the Internet might not even exist today.

Much lesser known, however, is the role of the National Institute of Standards and Technology – or NIST - in the development of search and retrieval tools.

Brent Rowe, a senior economist at RTI and lead author of the new 137-page report, points out that NIST, has invested more than $29 million in search tool development and has shared that knowledge with Google among others.

The end result is hundreds of millions in benefits – and acknowledgment from the private sector that the NIST support has helped them.

NIST's logo for TREC

Image 1 of 2 Next Image...

On The Web

- Search engine report
- WRAL Tech Wire at Twitter - sign up today for latest news alerts

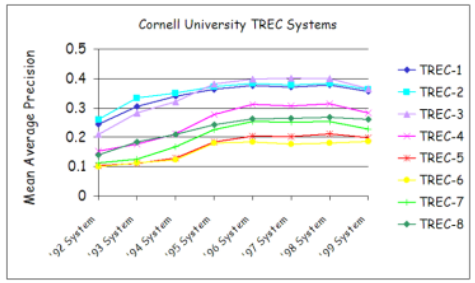
Site Search

Search

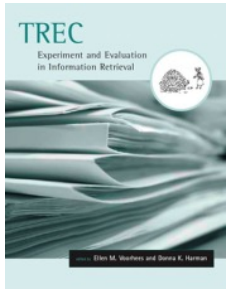
In other words, for every $1 NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers…These responses [to RTI's survey of IR researchers] suggest that the benefits of TREC to both private and academic organizations go well beyond those quantified by this study's economic benefits.

RTI International
Economic Impact Assessment of NIST's Text Retrieval Conference (TREC) Program
December 2010

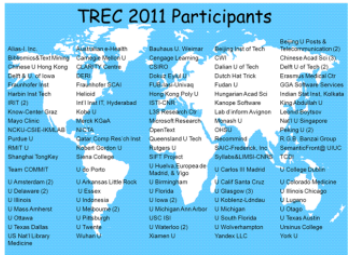
Coopetitions



Improve the state-of-the-art



Establish the research methodology



Form/solidify a research community

Facilitate technology transfer

…for every $1 invested in TREC, at least $3.35 to $5.07 in benefits accrued…

Amortize the cost of infrastructure

A good benchmark task:

- abstraction of real-world task so variables affecting performance can be controlled...
- ...but must capture salient aspects of real task or exercise is pointless
- metrics must accurately predict relative effectiveness on those aspects
- adequate level of difficulty
- best if measures are diagnostic