TREC: Improving Information Retrieval through Evaluation

Ellen Voorhees

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

NIST

- An agency of the Dept. of Commerce
  - one of only two federal labs to have its own authorization, annual appropriation, and be headed by Presidential appointee
  - non-regulatory, non-defense
- Mission

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 Fast Facts

- 2 main locations:

  Gaithersburg, MD & Boulder, CO
- ~3000 employees + visiting researchers
- Operational units
  - 9 laboratories
  - Baldridge National Quality Program
  - Manufacturing Extension Partnership
  - Technology Innovation Program
- FY2010 budget ~ $1 billion (all sources)

Information Retrieval

- Academic field researching content-based access to data developed for humans rather than computers
  - now also known as "search" (e.g., web search engines)
- Largest early influence was from the library community:
  - "find documents that satisfy the user's information need"
  - implies (relatively) large scale, many domains, varied population

Main Components

- How will the content (text & other media) be represented? [*indexing*]
- How will the information need (query) be represented? [*query language*]
- How will respective representations be matched? [*retrieval model*]
- How effective is the search?

Earliest Approach

- Electronic analog to physical punch cards
  - librarians assign “index” terms to document
    - generally index terms taken from a controlled vocabulary
    - modest number of terms per document
  - information need represented as Boolean combination of index terms
  - set of documents that satisfy Boolean expression are retrieved

Boolean Retrieval

- Precise definition of what exactly should be retrieved
- But,
  - manual indexing is expensive
  - even if affordable, searchers frequently don't use the terms the indexers selected for a document in their queries
    - granularity issues
    - differences in opinion
    - searchers are trying to describe what they don't know

Landmark Studies

- Cranfield experiments (mid-1960s)
  - reached “preposterous” conclusion that using the words of a document itself was at least as good as using fancy indexing schemes
- SMART-MEDLINE experiments
  - reached same conclusion in larger test comparing SMART’s vector space processing to Library of Medicine’s indexing using MeSH terms

Vector Space Model

- The set of words in a document collection define the dimensions of a vector space
  - do some slight processing to remove very highly frequent words (stop words) and conflate word forms to a common stem
- Represent both documents and (free text) queries as points in this space based on the words that occur in them
- Use the cosine of the angle between the vectors induced by two texts as their

Bag of Words

Czechs Play Indoor Soccer for More Than Four Days Straight for Record

Twenty Czechs broke the world record for indoor soccer last month, playing the game continuously for 107 hours and 15 minutes, the official Czechoslovak news agency CTK reported.

Two teams began the endeavor in the West Bohemian town of Holysov on Dec. 13 and ended with the new world record on Dec. 17, CTK said in the dispatch Monday.

According to the news agency, the previous record of 106 hours 10 minutes was held by English players. The Czechs new record is to be recorded in the Guinness Book of World Records, CTK said.

agency(2); began; bohemian; book; broke; continuously; ctk(3); czechoslovak; czechs(3); days; dec(2); dispatch; ended; endeavor; english; game; guinness; held; holysov; hour(2); indoor(2); minutes(2); monday; month; news(2); official; play(3); previous; record(7); reported; soccer(2); straight; teams; town; twenty; west; world(3)

Final Document Representation

(modern vector space system)

agent 2.80; begin 1.55; bohem 6.01; book 2.63; brok 2.60; continu 1.55; ctk 13.35; czechoslovak 4.36; czech 11.65; day 1.38; dec 4.34; dispatch 4.12; end 1.36; endeavor 5.03; engl 3.51; game 3.14; guin 5.40; held 2.05; holysov 10.75; hour 3.44; indoor 8.13; minut 4.38; monday 2.12; month 1.34; new 2.62; offic .98; play 4.82; prev 1.89; record 5.80; report 1.04; socc 9.16; straight 3.61; team 2.86; town 2.86; twent 3.91; west 2.14; world 3.85

czechoslovak offic 8.64; prev record 6.09; record world 13.69; day straight 6.41; agent new 6.69; ctk report 8.51; book guin 7.15; agent ctk 7.67; begin team 8.20

Advances over Simple Bag-of-Words

- New models/extension to existing models
  - fuzzy Boolean, probabilistic, language modeling…
  - while different theoretical underpinnings, better models remarkably similar in practice
  - good term weighting crucial
- Query expansion
  - massive expansion from “blind” feedback & other sources
- Learning approaches
  - with advent of large repositories of both text and user interactions, statistical machine learning approaches viable

What about NLP?

- Where is the Natural Language Processing (NLP) in this discussion?
  - response 1: IR is a form of NLP (though not NLU)
  - response 2: to date, broad-coverage, large-scale NLP not robust enough and too slow (expensive) to be useful for generic IR
    - recognition of collocations, normalization of word variants is helpful
    - "deep" NLP has proved to be useful in other tasks such as (factoid) question answering
    - how best to exploit semantics is an active research area

What is a good search result?

If you can not measure it, you can not improve it.   ---Lord Kelvin

- Some sample queries
  - 'TREC'
  - Ellen Voorhees home page
- Are these results good? bad? indifferent?
  - How do you know?
  - Does your neighbor agree?

What is a good search result?

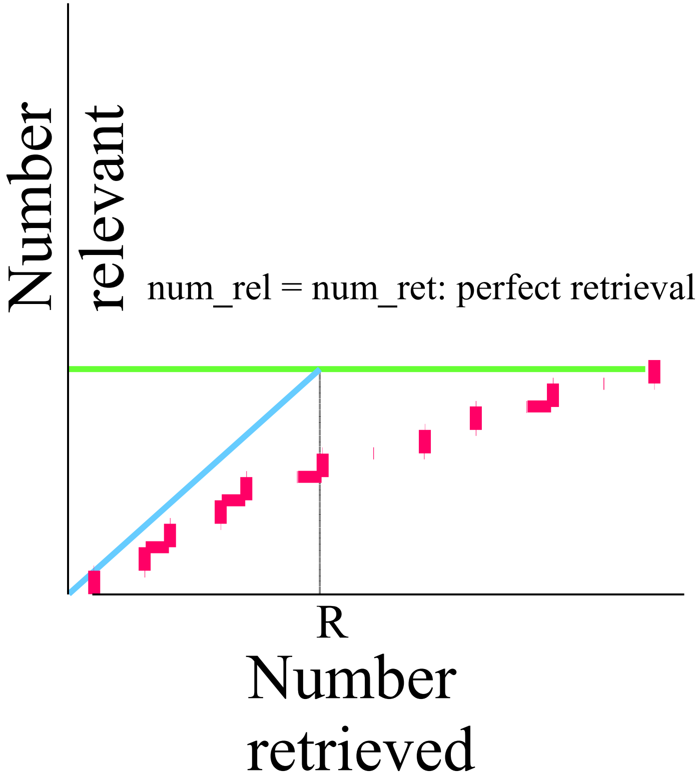
If you can not measure it, you can not improve it. ---Lord Kelvin

- Search is inherently a user activity
  - purpose/goals of the search
  - background knowledge of the searcher
  - corpus being searched
- Different users have different (conflicting) criteria for success

➢ There is no single Truth

Ranked Retrieval Evaluation

num_rel =R



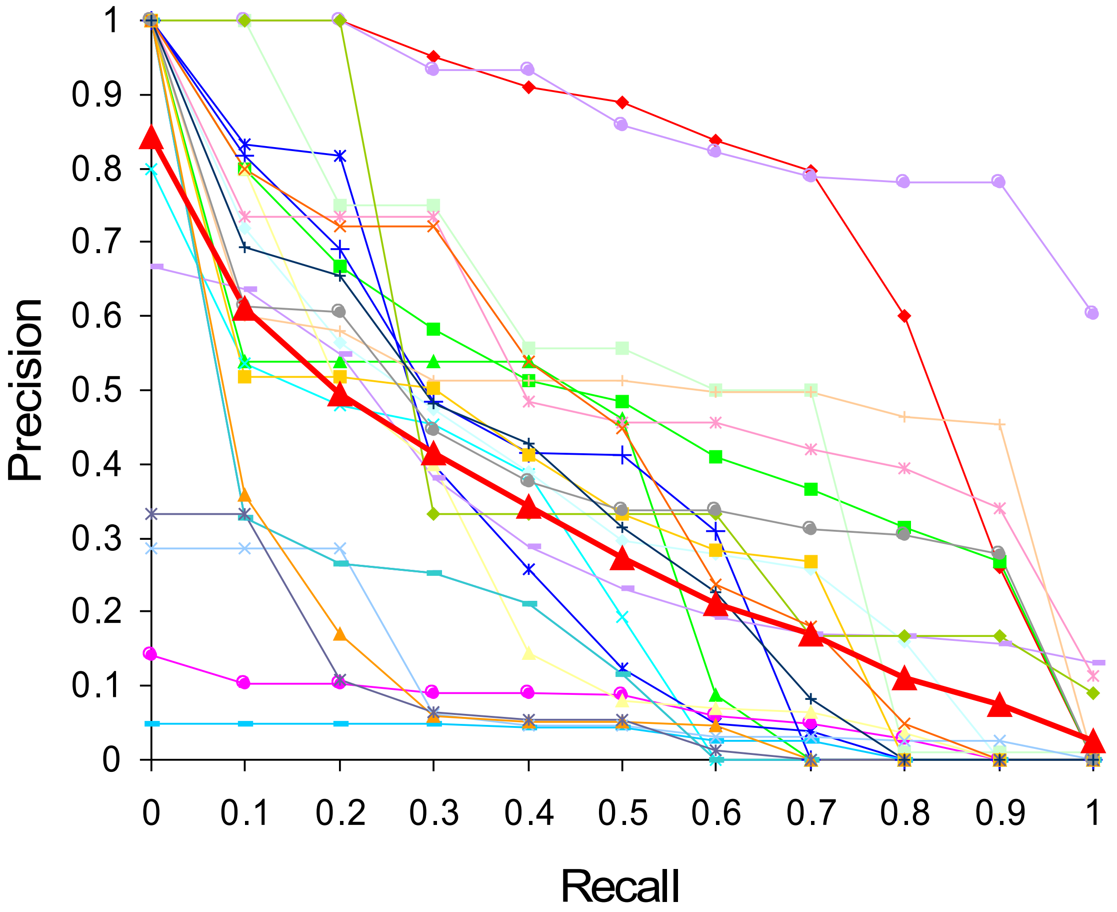
Retrieval Evaluation

- Quality of a single search result set can be summarized as a [recall, precision] pair

$$\text{recall} = \frac{\text{\# relevant retrieved}}{\text{\# relevant}} \qquad \text{precision} = \frac{\text{\# relevant retrieved}}{\text{\# retrieved}}$$

  - inversely related in practice
  - recall is hard to measure and users tend to vastly overestimate it
  - optimum of [1,1] is not achievable by humans

Interpolated R-P Curves for Individual Searches



Cranfield Tradition

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

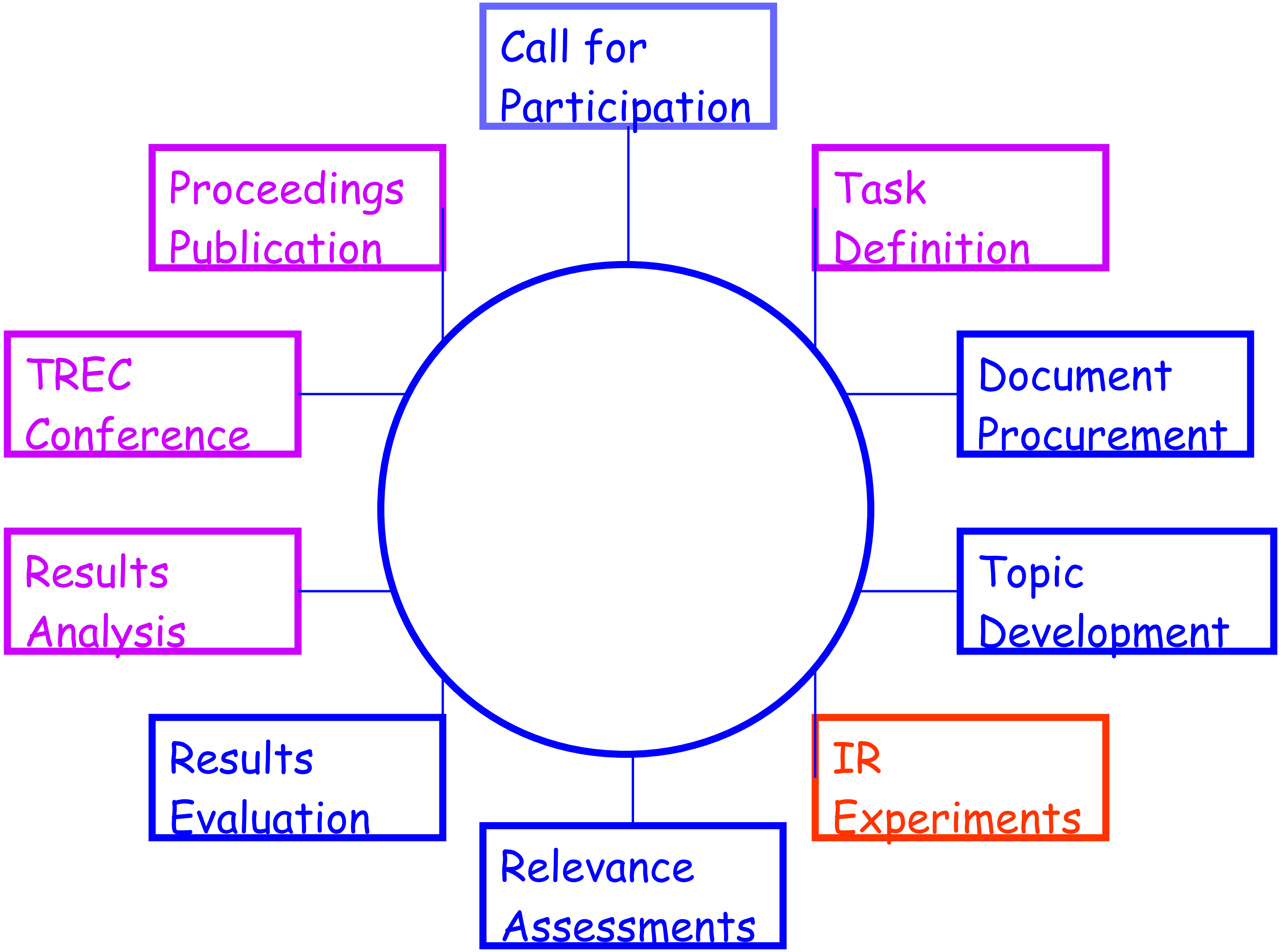
TREC

- A workshop series that provides the infrastructure for large-scale evaluation of (text) retrieval technology
  - realistic test collections
  - uniform, appropriate scoring procedures
  - a forum for the exchange of research ideas and for the discussion of research methodology

TREC Philosophy

- TREC is a modern example of the Cranfield tradition
  - system evaluation based on test collections
- Emphasis on advancing the state of the art from evaluation results
  - TREC's primary purpose is not competitive benchmarking
  - experimental workshop: sometimes experiments fail!

Yearly Conference Cycle



Call for Participation

Task Definition

Document Procurement

Topic Development

IR Experiments

Relevance Assessments

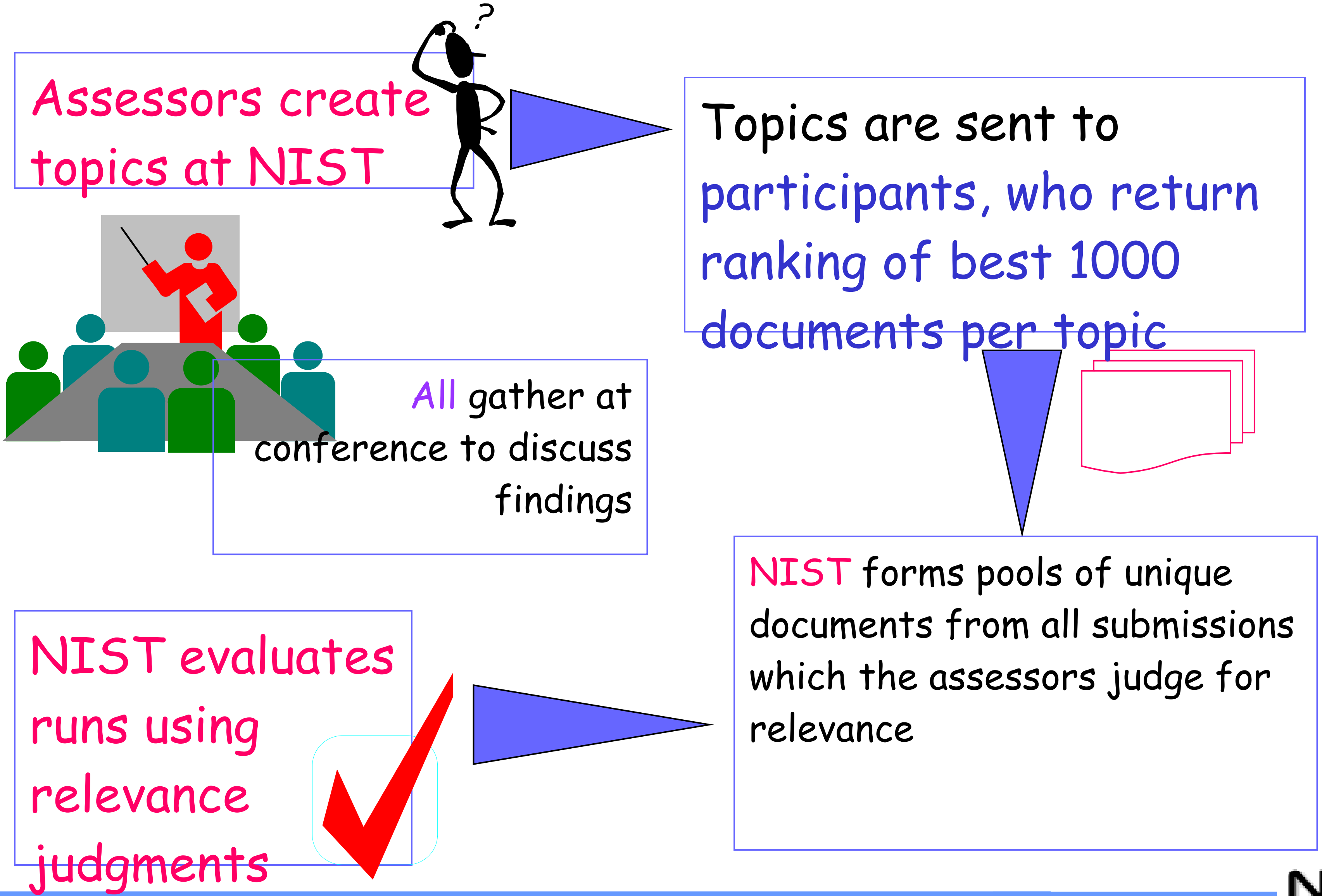
Results Evaluation

Results Analysis

TREC Conference

Proceedings Publication

NIST TREC Approach



Assessors create topics at NIST

Topics are sent to participants, who return ranking of best 1000 documents per topic

NIST forms pools of unique documents from all submissions which the assessors judge for relevance

NIST evaluates runs using relevance judgments

All gather at conference to discuss findings

Example Topics

·Document will discuss hydroponics: the science of growing plants in water or some substance other than soil.

·Commercial uses of Magnetic Levitation.

·What modern instances have there been of old fashioned piracy, the boarding or taking control of boats?

·Are there reliable and consistent predictors of mutual fund performance?

·Identify instances where a journalist has been put at risk (e.g., killed, arrested, taken hostage) in the performance of his work.

·Health studies primarily in the U.S. have caused reductions in tobacco sales here, but the economic impact has caused U.S. tobacco companies to look overseas for customers. What impact have the health and economic factors had overseas?

·Aside from the United States, which country offers the best living conditions and quality of life for a U.S. retiree?

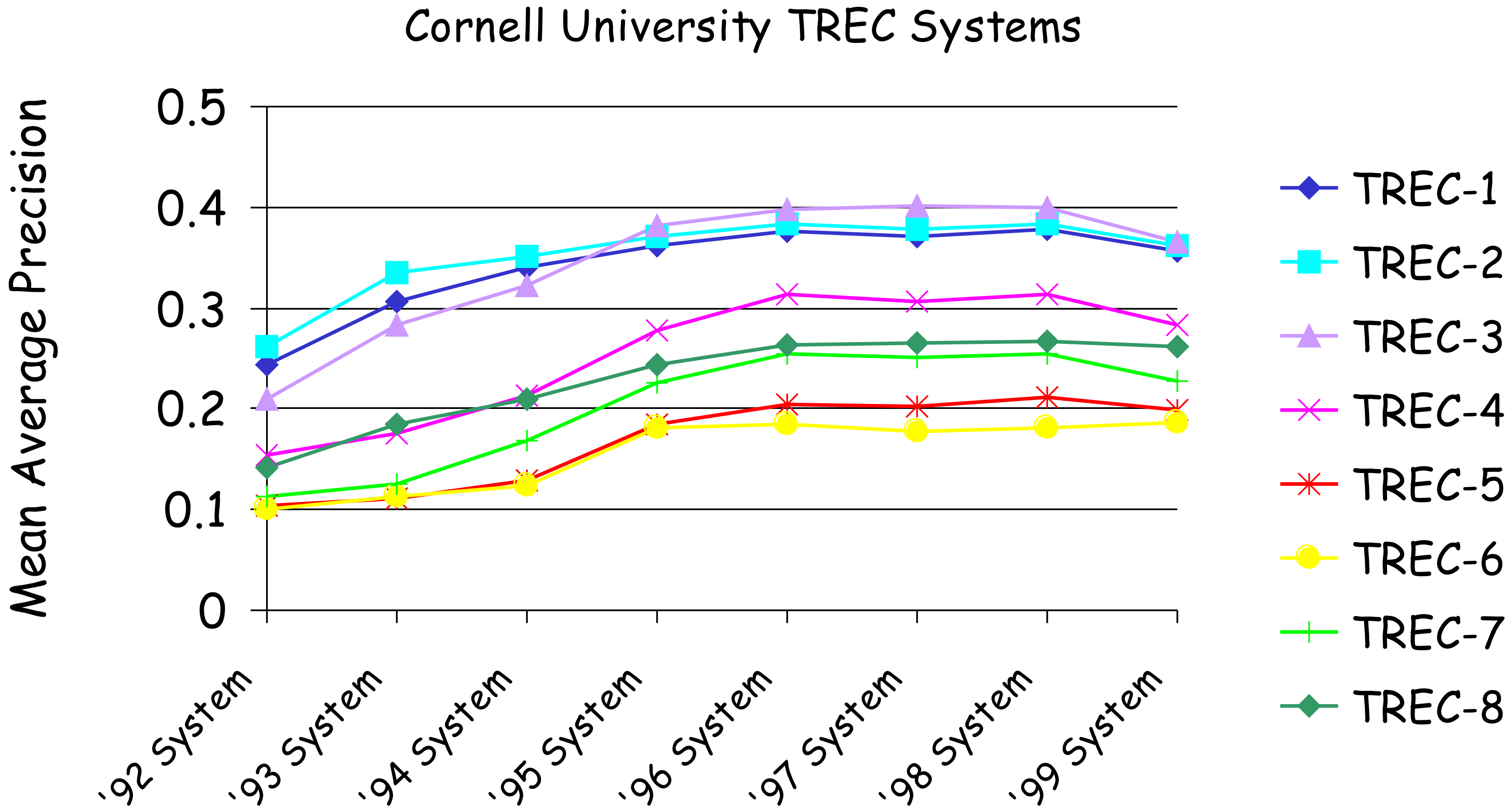
TREC 2009 Participants

- Applied Discovery
- Beijing Institute of Technology
- Beijing U. of Posts and Telecommunications
- Cairo Microsoft Innovation Center
- Carnegie Mellon University
- Chinese Academy of Sciences (2)
- Clearwell Systems, Inc.
- Clearly Gottlieb Steen & Hamilton, with Backstop LLC
- Dalian University of Technology
- Delft University of Technology
- EMC - CMA - R&D
- Equivio
- Fondazione Ugo Bordoni
- Fraunhofer SCAI
- Fudan University
- H5
- Heilongjiang Inst. of Technology
- Integreon
- International Inst. of Information Technology, Hyderabad
- Know-Center
- Lehigh University
- Logik Systems, Inc.
- Microsoft Research Asia
- Microsoft Research Cambridge
- Milwaukee School of Engineering
- Mugla University
- National Institute of Information and Communications Technology
- Northeastern University
- Open Text Corporation
- Peking University
- Pohang U. of Science & Technology
- Purdue University
- Queensland University of Technology
- RMIT University
- Sabir Research
- South China University of Technology
- SUNY Buffalo
- Tsinghua University
- Universidade do Porto
- University College Dublin
- University of Alaska, Fairbanks
- University of Amsterdam (2)
- University of Applied Science Geneva
- University of Arkansas, Little Rock
- University of California, Santa Cruz
- University of Delaware (2)
- University of Glasgow
- University of Illinois, Urbana-Champaign
- University of Iowa
- University of Lugano
- University of Maryland, College Park
- University of Massachusetts, Amherst
- The University of Melbourne
- University of Padova
- University of Paris
- University of Pittsburgh
- University of Twente
- University of Waterloo (2)
- Ursinus College
- Yahoo! Research
- York University (2)
- ZL Technologies, Inc.

Participation in TREC

About 300 distinct groups have participated in at least one TREC.

TREC Impacts



TREC Tracks

- Task that focuses on a particular subproblem of text retrieval
- Tracks invigorate TREC & keep TREC ahead of the state-of-the-art
  - specialized collections support research in new areas
  - first large-scale experiments debug what the task really is
  - provide evidence of technology's robustness

TREC Tracks

- Set of tracks in a particular TREC depends on:
  - interests of participants
  - appropriateness of task to TREC
  - needs of sponsors
  - resource constraints
- Need to submit proposal for new track in writing to NIST

The TREC Tracks

Personal documents		Blog Spam
Retrieval in a domain		Chemical IR Genomics
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		Interactive, HARD, Feedback
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	

Ad Hoc Technologies

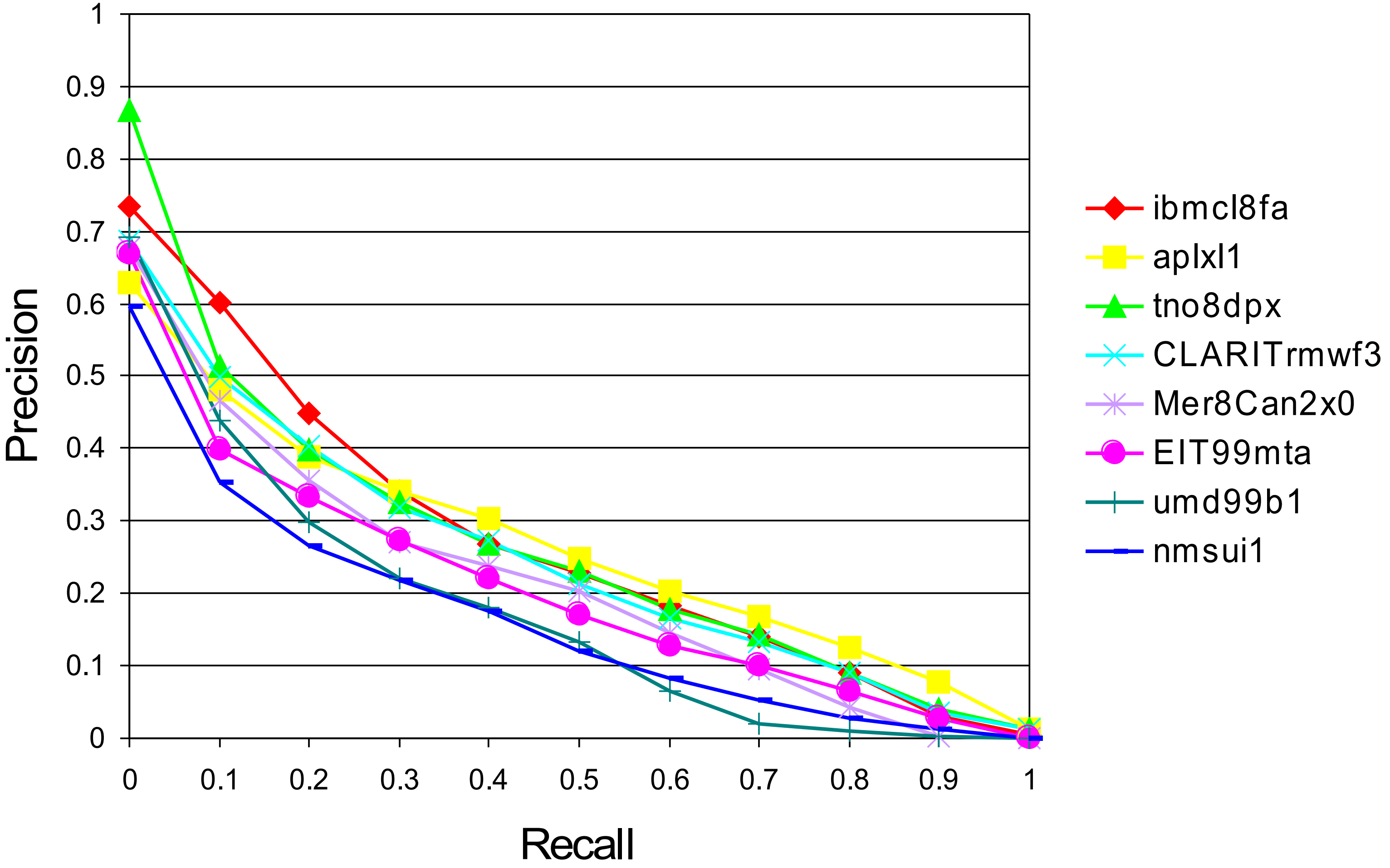
Streamed Text

Adaptive Filtering

T11SU measure; red line is scaled utility of retrieving no documents

Beyond English

X to EFGI Results



National Institute of Standards and Technology

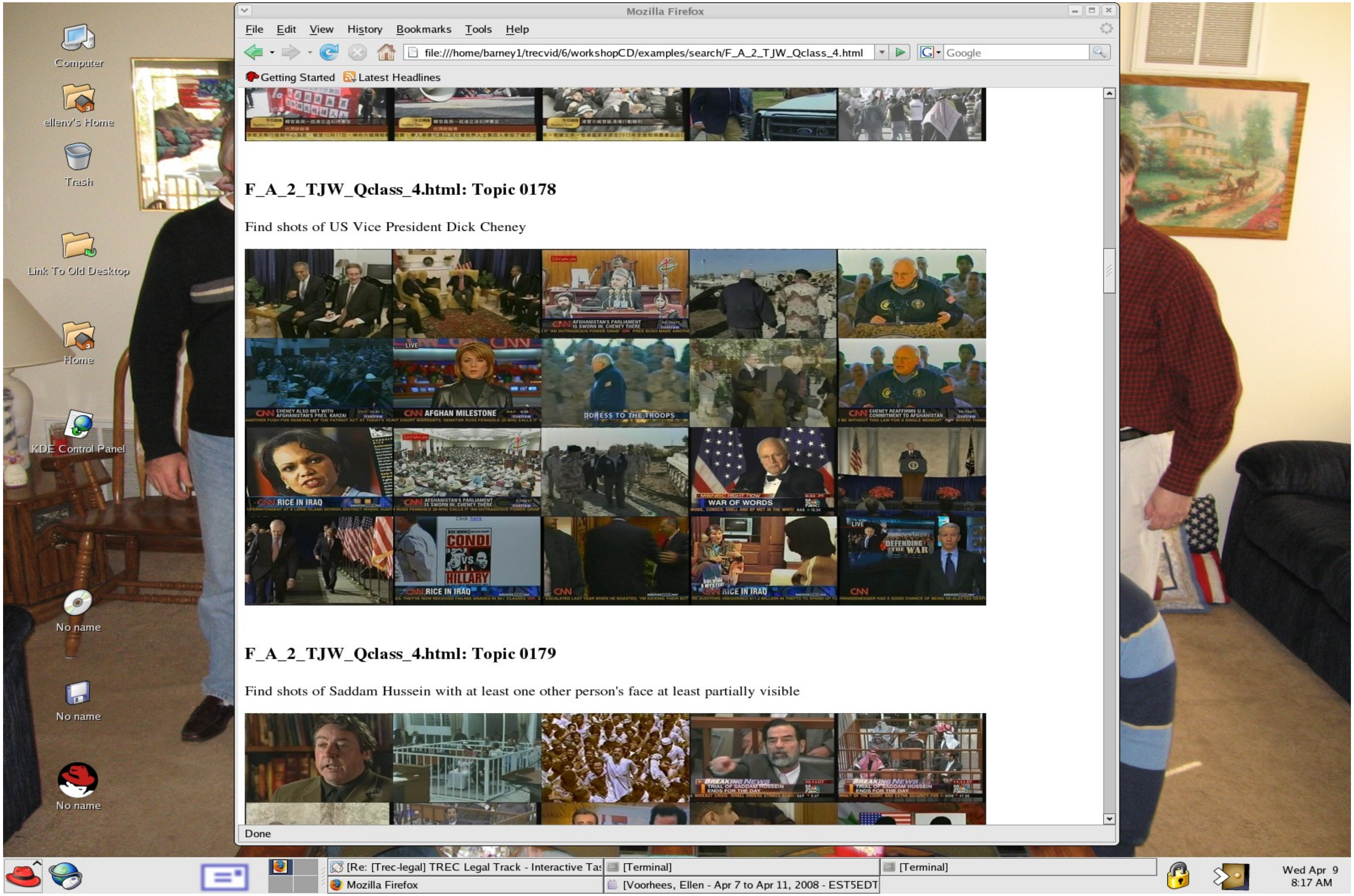
Beyond Text: Video

People in uniform and in formation

Soccer goalposts

Condoleezza Rice

Soldiers, police or guards escorting a prisoner



F_A_2_TJW_Qclass_4.html: Topic 0178

Find shots of US Vice President Dick Cheney

F_A_2_TJW_Qclass_4.html: Topic 0179

Find shots of Saddam Hussein with at least one other person's face at least partially visible

Answers, Not Documents

- TREC Question Answering track
  - common task for NLP & IR communities
  - reinvigorated research on QA systems
  - explosion of QA workshops, journal issues
- Original emphasis on factoid questions
  - research roadmap developed by community
  - slowly expanded types and difficulty of questions in test set

2007 Question Series Task

254 House of Chanel

254.1	FACT	Who founded the House of Chanel?
254.2	FACT	In what year was the company founded?
254.3	FACT	Who is the president of the House of Chanel?
254.4	FACT	Who took over the House of Chanel in 1983?
254.5	LIST	What women have worn Chanel clothing to award ceremonies?
254.6	LIST	What museums have displayed Chanel clothing?
254.7	FACT	What Chanel creation is the top-selling fragrance in the world?
254.8	Other	

70 series in test set with 6-10 questions per series

19 People
17 Organizations
19 Things
15 Events

360 total factoid questions
85 total list questions
70 total "other" questions

2007 Series Task Results

2009 Web Diversity Results

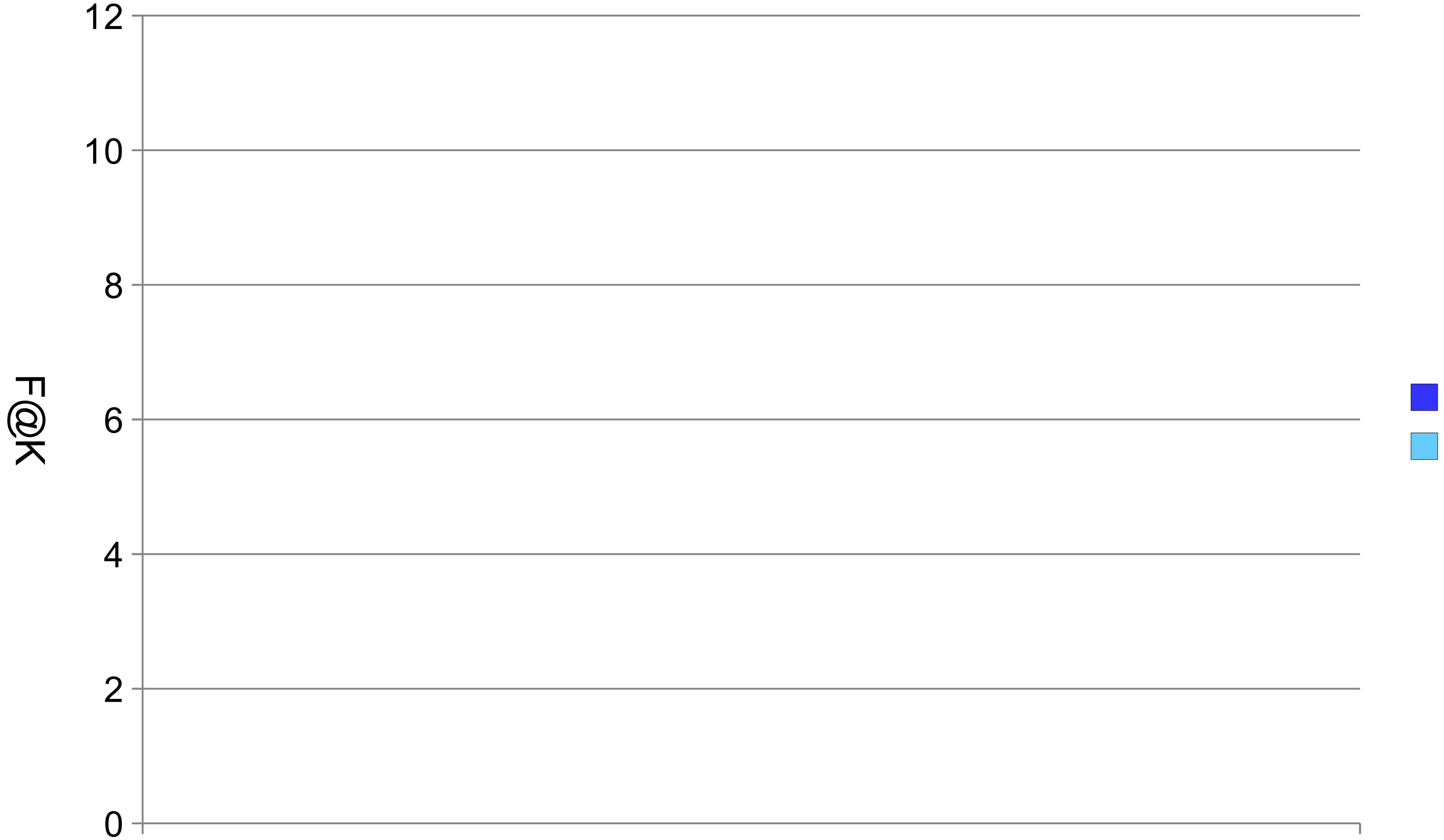
Best diversity run per group per category by α-NDCG

Legal Track

- Goal: evaluate search technology for discovery of electronically stored data
- That technology must support how legal discovery actually happens
  - set-based retrieval
  - large, heterogeneous information space
  - large "relevant" result sets
  - need for high recall
  - costs that are proportional to precision
  - parties want confidence that process worked for *this* case (average performance of less interest)

2009 Legal Batch Results

Average F1@K for Batch Task Runs and Baselines



BM25: A TREC Success Story

See Stephen Robertson, *How Okapi Came to TREC*. In **TREC: Experiment and Evaluation in Information Retrieval**, MIT Press, 2005, pp. 287—299.

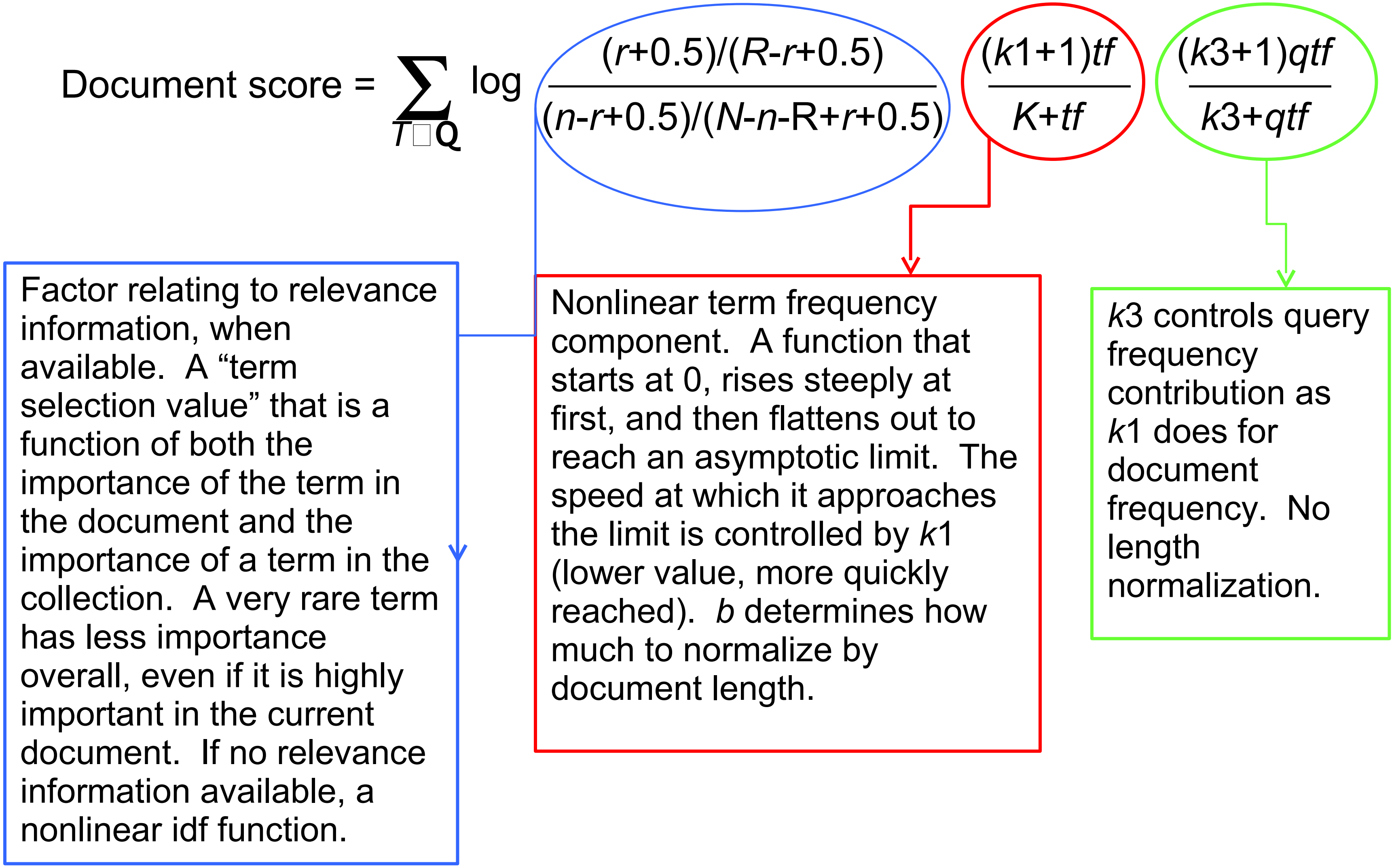
- BM25 is a (family of) function(s) for
  - assigning weights to individual terms in the query, *and*
  - combining term weights to score documents
- Developed by the Okapi group
  - first used in their TREC-2 (1993) runs
  - motivated by failure of TREC-1 system to handle documents of widely varying lengths
  - further refined in later experiments
- Subsequently adopted by many others

BM25

$$\text{Document score} = \sum_{T \in \mathbf{Q}} \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-\mathrm{R}+r+0.5)} \frac{(k1+1)tf}{K+tf} \frac{(k3+1)qtf}{k3+qtf}$$

Q	a query containing terms *T*
tf	the frequency of occurrence of the term within the current document
qtf	the frequency of occurrence of the term in the query statement
dl, *avdl*	the document length of the current document, the average document length
N	the number of documents in the collection
n	the number of documents containing the current term
R	the number of documents known to be relevant to the query; set to 0 if none known
r	the number of relevant documents containing the current term; set to 0 if none known
$K=k1((1-b)+b\times dl/avdl)$; and	
k1,b,k3	tuning parameters with $k1\geq 0$, $0\leq b\leq 1$, $k3\geq 0$

BM25



$$\text{Document score} = \sum_{T \in Q} \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-\mathrm{R}+r+0.5)} \; \frac{(k1+1)tf}{K+tf} \; \frac{(k3+1)qtf}{k3+qtf}$$

Factor relating to relevance information, when available. A “term selection value” that is a function of both the importance of the term in the document and the importance of a term in the collection. A very rare term has less importance overall, even if it is highly important in the current document. If no relevance information available, a nonlinear idf function.

Nonlinear term frequency component. A function that starts at 0, rises steeply at first, and then flattens out to reach an asymptotic limit. The speed at which it approaches the limit is controlled by *k*1 (lower value, more quickly reached). *b* determines how much to normalize by document length.

*k*3 controls query frequency contribution as *k*1 does for document frequency. No length normalization.

TREC Impacts

- Test collections
- Incubator for new research areas
- Common evaluation methodology and improved measures for text retrieval
- Open forum for exchange of research
- Technology transfer