

Evaluation

INFM 718X/LBSC 718X

Session 6

Douglas W. Oard

Evaluation Criteria

- Effectiveness
 - System-only, human+system
- Efficiency
 - Retrieval time, indexing time, index size
- Usability
 - Learnability, novice use, expert use

IR Effectiveness Evaluation

- User-centered strategy
 - Given several users, and at least 2 retrieval systems
 - Have each user try the same task on both systems
 - Measure which system works the “best”
- System-centered strategy
 - Given documents, queries, and relevance judgments
 - Try several variations on the retrieval system
 - Measure which ranks more good docs near the top

Good Measures of Effectiveness

- Capture some aspect of what the user wants
- Have predictive value for other situations
 - Different queries, different document collection
- Easily replicated by other researchers
- Easily compared
 - Optimally, expressed as a single number

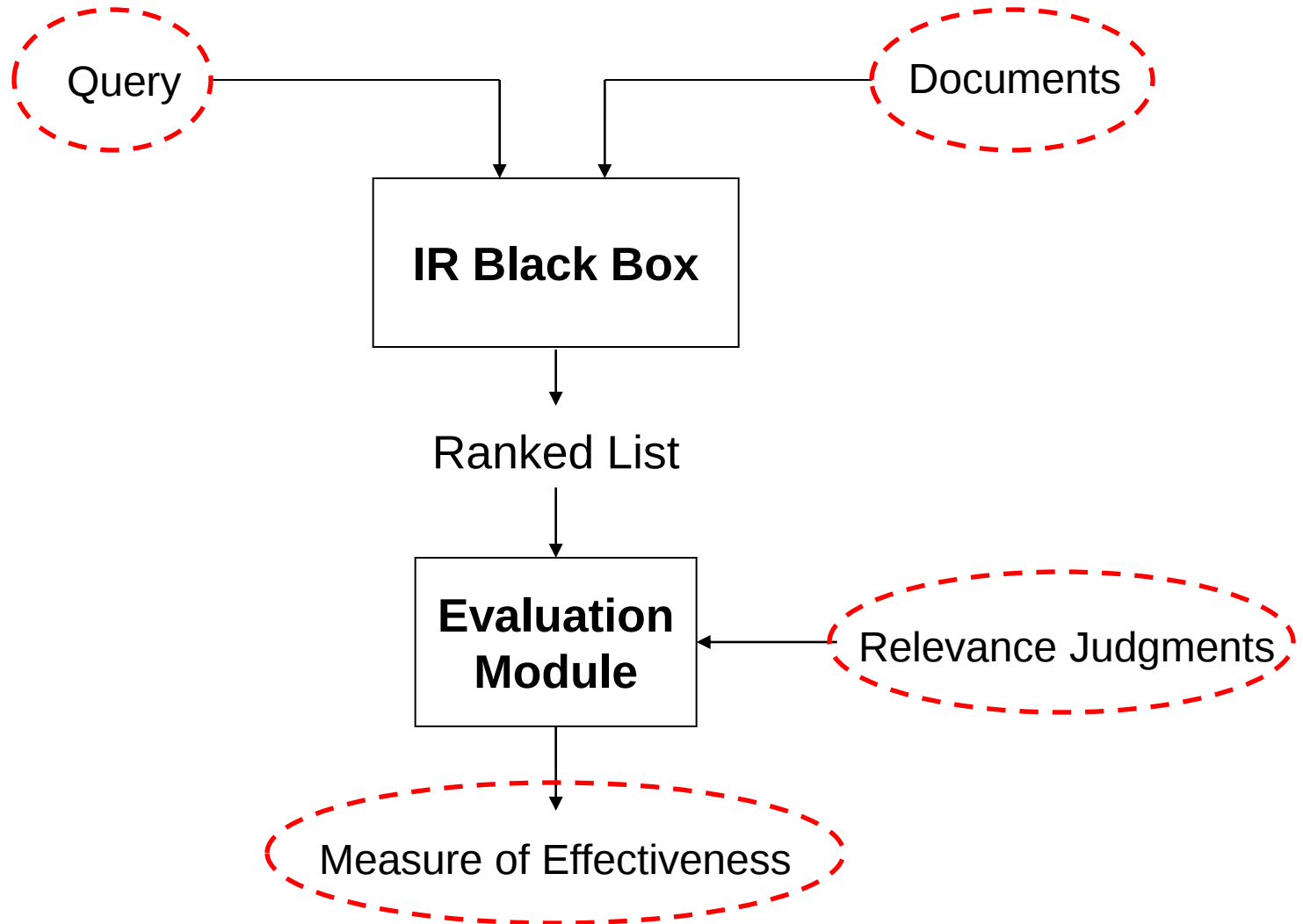
Comparing Alternative Approaches

- Achieve a meaningful improvement
 - An application-specific judgment call
- Achieve reliable improvement in unseen cases
 - Can be verified using statistical tests

Evolution of Evaluation

- Evaluation by **inspection** of examples
- Evaluation by **demonstration**
- Evaluation by **improvised** demonstration
- Evaluation on **data** using a figure of merit
- Evaluation on **test data**
- Evaluation on **common** test data
- Evaluation on common, **unseen** test data

Automatic Evaluation Model



These are the four things we need!

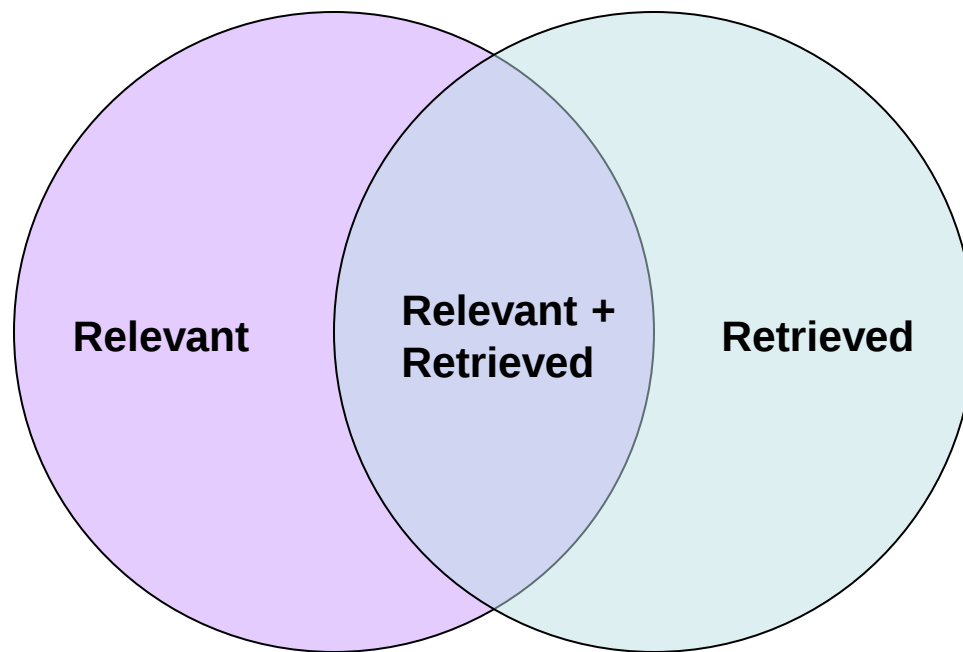
IR Test Collection Design

- Representative document collection
 - Size, sources, genre, topics, ...
- “Random” sample of representative queries
 - Built somehow from “formalized” topic statements
- Known binary relevance
 - For each topic-document pair (topic, not query!)
 - Assessed by humans, used only for evaluation
- Measure of effectiveness
 - Used to compare alternate systems

Defining “Relevance”

- **Relevance** relates a topic and a document
 - Duplicates are equally relevant by definition
 - Constant over time and across users
- **Pertinence** relates a task and a document
 - Accounts for quality, complexity, language, ...
- **Utility** relates a user and a document
 - Accounts for prior knowledge

Space of all documents



Not Relevant + Not Retrieved

Set-Based Effectiveness Measures

- Precision
 - How much of what was found is relevant?
 - Often of interest, particularly for interactive searching
- Recall
 - How much of what is relevant was found?
 - Particularly important for law, patents, and medicine
- Fallout
 - How much of what was irrelevant was rejected?
 - Useful when different size collections are compared

Effectiveness Measures

Doc \ Action	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Miss
Not relevant	False Alarm	Irrelevant Rejected

User-Oriented	Precision = $\frac{\text{Relevant Retrieved}}{\text{Retrieved}}$	System-Oriented
	Recall = $\frac{\text{Relevant Retrieved}}{\text{Relevant}} = 1 - \text{Miss}$	
	Fallout = $\frac{\text{Irrelevant Rejected}}{\text{Not Relevant}} = 1 - \text{FA}$	

Balanced F Measure (F_1)

- Harmonic mean of recall and precision

$$F_1 = \frac{1}{\frac{0.5}{P} + \frac{0.5}{R}}$$

Variation in Automatic Measures

- System
 - What we seek to measure
- Topic
 - Sample topic space, compute expected value
- Topic+System
 - Pair by topic and compute statistical significance
- Collection
 - Repeat the experiment using several collections

IIT CDIP v1.0 Collection

Scanned

OCR

Metadata

Philip Morris U.S.A. Inter-Office Correspondence
Benefits Department Richmond, Virginia

To: Distribution Date: May 30, 1997
From: Lisa Halle
Subject: CIGNA Well-Being Newsletter - Future Strategy

During our last CIGNA Action Plan meeting, the issue of whether to stop previewing articles and discontinue sending CIGNA Well-Being newsletter to our employees was a matter of discussion. I have done some research, and wanted to present you with my findings and preliminary recommendation for PM's strategy regarding future newsletters. I believe everyone's input is valuable, and would appreciate hearing from each of you on whether you concur with my recommendations.

Background Information

CIGNA Well-Being newsletters are sent on a quarterly basis. The process we have been using is to have one of the analysts preview both national and local articles slated for a particular newsletter, and then recommend to either skip or send that issue. Offensive local articles can be replaced with another of similar length at no cost to PM. If we opt to replace or modify a national article, it costs PM \$3,000 per issue.

Since 1996, we have opted to either skip or send issues as follows:

Date of Newsletter	Decision	Comments
Spring 1996	Send	Deleted advertisement for CIGNA Time-Life video featuring Dr. Eugene General. Keep prior to sending.

Date of Newsletter	Decision	Comments
Fall 1996	Send	A national article on heart attacks contained one minor reference to smoking which was deemed no worse than what the warning label states. Also, a breast cancer article included into a free Time-Life video offer on subjects such as breast cancer, breast lumps, mastectomies, pregnancy, menopause, and osteoporosis. Decision to send based on benefits that overwhelmingly positive reaction to Time-Life video series. No complaints received from member population based on newsletter content.
Winter 1996	Skip	National article entitled "A Breath of Fresh Air" listed smoking as one of the threats to the environment which can trigger an asthma attack, and went further to say "Do not allow smoking in your house or in any environment that you can control".
Spring 1997	Send	Contained nothing objectionable.
Summer 1997	Skip	National article contained objectionable secondhand smoke reference.

In summary we have opted to skip three (3) of the last six (6) newsletters. To my knowledge, these are the only ones we have ever skipped.

The process of reviewing articles and making a recommendation to send or skip a issue varies, depending on content. Typically, it is immediately clear if something is objectionable. Other times, it may require discussion with others and management. I would say between phone calls with CIGNA, previewing of captioned versions and full versions of local and national articles, discussion if necessary, editing receipt of newsletters, and fielding complaints, I have spent as little as four (4) hours, to as long as several days on this activity. The issue surrounding the Time-Life video series required the most time, as you may remember from discussions in staff meetings.

Recommendation

Philip Moxx's. U.S.A. x.dr~am~c.
cvrrespoaa.aa
Benffrts Departmext Rieh>pwna, Yfe&ia
Ta: Dishlbutfon Data aday 90,1997.
From: Lisa Fislla
Sbj.csr CIGNA WeWedng Newsbttsr -
Yntsre StratsU
During our last CIGNA Aatfoa Plan
meadng, tlu iasuo of wLetSae to iOop
per'Irw+ng
artieles aod discontinue mndia6 CIGNA
Well-Being aawslener to om employees
was a
msiter of disanision . I Imvm done
somme reaearc>>, and wanted to
pruedt you with my
Sadings and pcdiminary
recwmmeadatioa for PM's atratezy
leprding l4aas aewelattee* .
I believe .vayone'a input is valusble, and
would epproolate hoarlng fmaa aach of
you on
whetlne you concur with my
reeommendatioa

...

Title: CIGNA WELL-BEING
NEWSLETTER - FUTURE
STRATEGY

Organization Authors:
PMUSA, PHILIP MORRIS
USA

Person Authors: HALLE, L

Document Date: 19970530

Document Type: MEMO,
MEMORANDUM

Bates Number:
2078039376/9377

Page Count: 2

Collection: Philip Morris

“Complaint” and “Production Request”

...12. On January 1, 2002, Echinoderm announced record results for the prior year, primarily attributed to strong demand growth in overseas markets, particularly China, for its products. The announcement also touted the fact that Echinoderm was unique among U.S. tobacco companies in that it had seen no decline in domestic sales during the prior three years.

13. Unbeknownst to shareholders at the time of the January 1, 2002 announcement, defendants had failed to disclose the following facts which they knew at the time, or should have known:

- a. The Company's success in overseas markets resulted in large part from bribes paid to foreign government officials to gain access to their respective markets;
- b. The Company knew that this conduct was in violation of the Foreign Corrupt Practices Act and therefore was likely to result in enormous fines and penalties;
- c. The Company intentionally misrepresented that its success in overseas markets was due to superior marketing.
- d. Domestic demand for the Company's products was dependent on pervasive and ubiquitous advertising, including outdoor, transit, point of sale and counter top displays of the Company's products, in key markets. Such advertising violated the marketing and advertising restrictions to which the Company was subject as a party to the Attorneys General Master Settlement Agreement ("MSA").
- e. The Company knew that it could be ordered at any time to cease and desist from advertising practices that were not in compliance with the MSA and that the inability to continue such practices would likely have a material impact on domestic demand for its products. ...

All documents which describe, refer to, report on, or mention any “in-store,” “on-counter,” “point of sale,” or other retail marketing campaigns for cigarettes.

An Ad Hoc “Production Request”

<ProductionRequest>

<RequestNumber>**148**</RequestNumber>

<RequestText>**All documents concerning the Company's FMLA policies, practices and procedures.**</RequestText>

<BooleanQuery>

<FinalQuery>**(policy OR policies OR practice! or procedure! OR rule! OR guideline! OR standard! OR handbook! OR manual!) w/50 (FMLA OR leave OR "Family medical leave" OR absence)**</FinalQuery>

<NegotiationHistory>

<ProposalByDefendant>**(FMLA OR "federal medical leave act") AND (policies OR practices OR procedures)**</ProposalByDefendant>

<RejoinderByPlaintiff>**(FMLA OR "federal medical leave act") AND (leave w/10 polic!)**</RejoinderByPlaintiff>

<Consensus1>**(policy OR policies OR practice! or procedure! OR rule! OR guideline! OR standard! OR handbook! OR manual!) AND (FMLA OR leave OR "Family medical leave" OR absence)**</Consensus1>

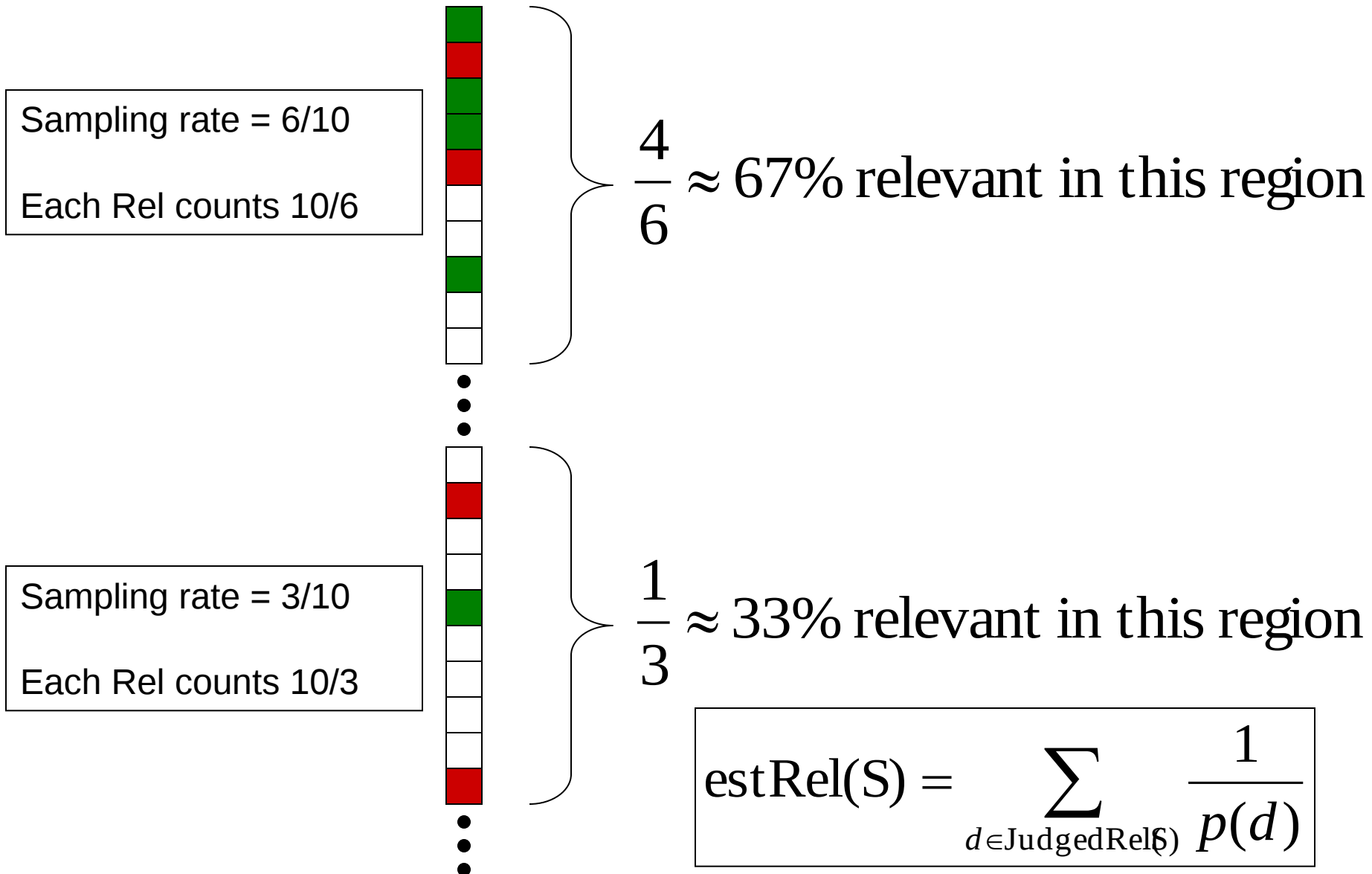
</NegotiationHistory>

</BooleanQuery>

<FinalB>**40863**</FinalB>

<RequestSource>**2008-H-7**</RequestSource>

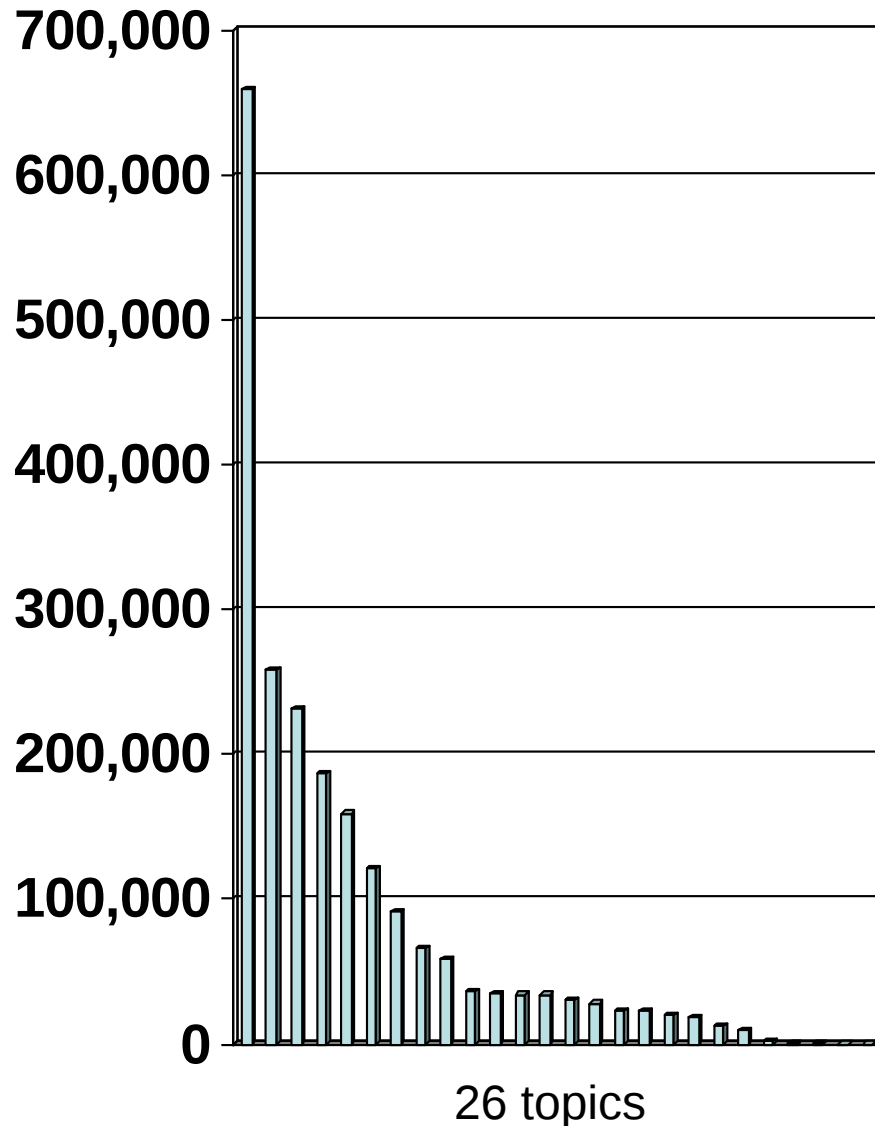
Estimating Retrieval Effectiveness



Relevance Assessment

- All volunteers
 - Mostly from law schools
- Web-based assessment system
 - Based on document images
- 500-1,000 documents per assessor
 - Sampling rate varies with (minimum) depth

2008 Est. Relevant Documents

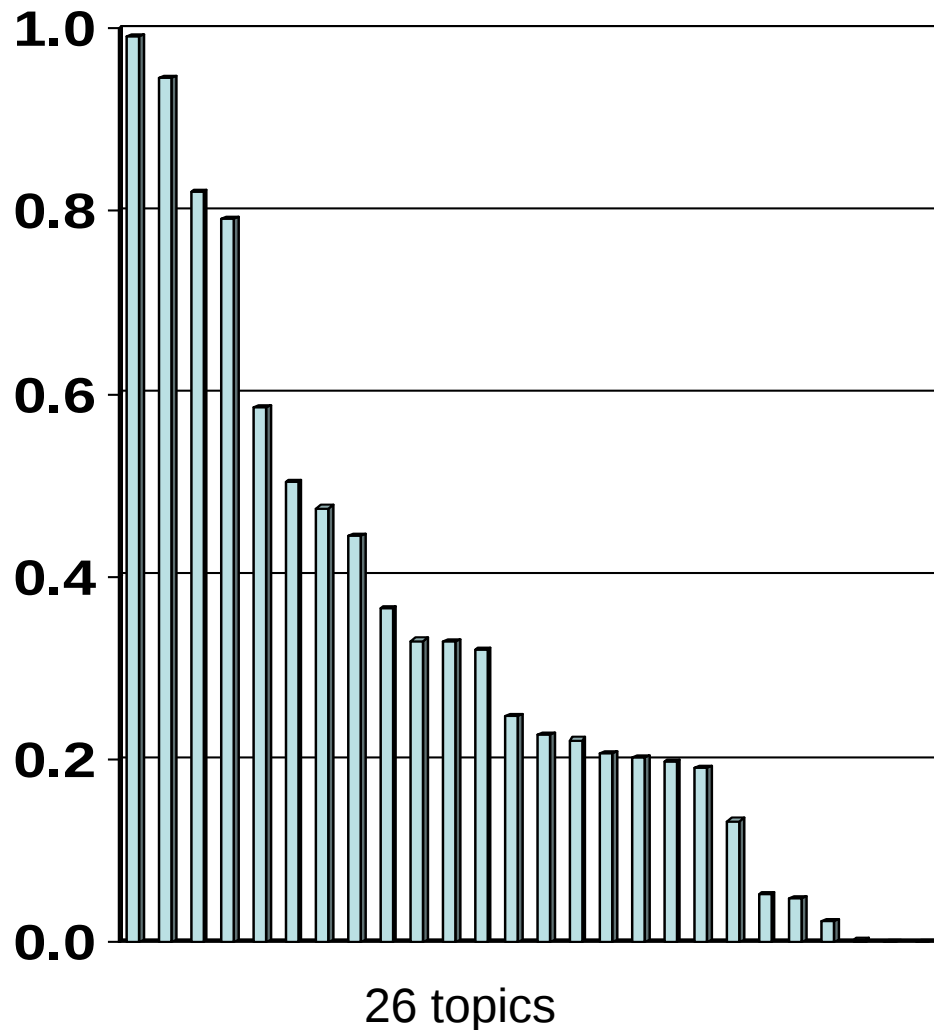


Mean estRel = 82,403 (26 topics)
• 5x 2007 mean estRel (16,904)

Max estRel=658,339, Topic 131
(rejection of trade goods)

Min estRel=110 Topic 137
(intellectual property rights)

2008 (cons.) Boolean Estimated Recall



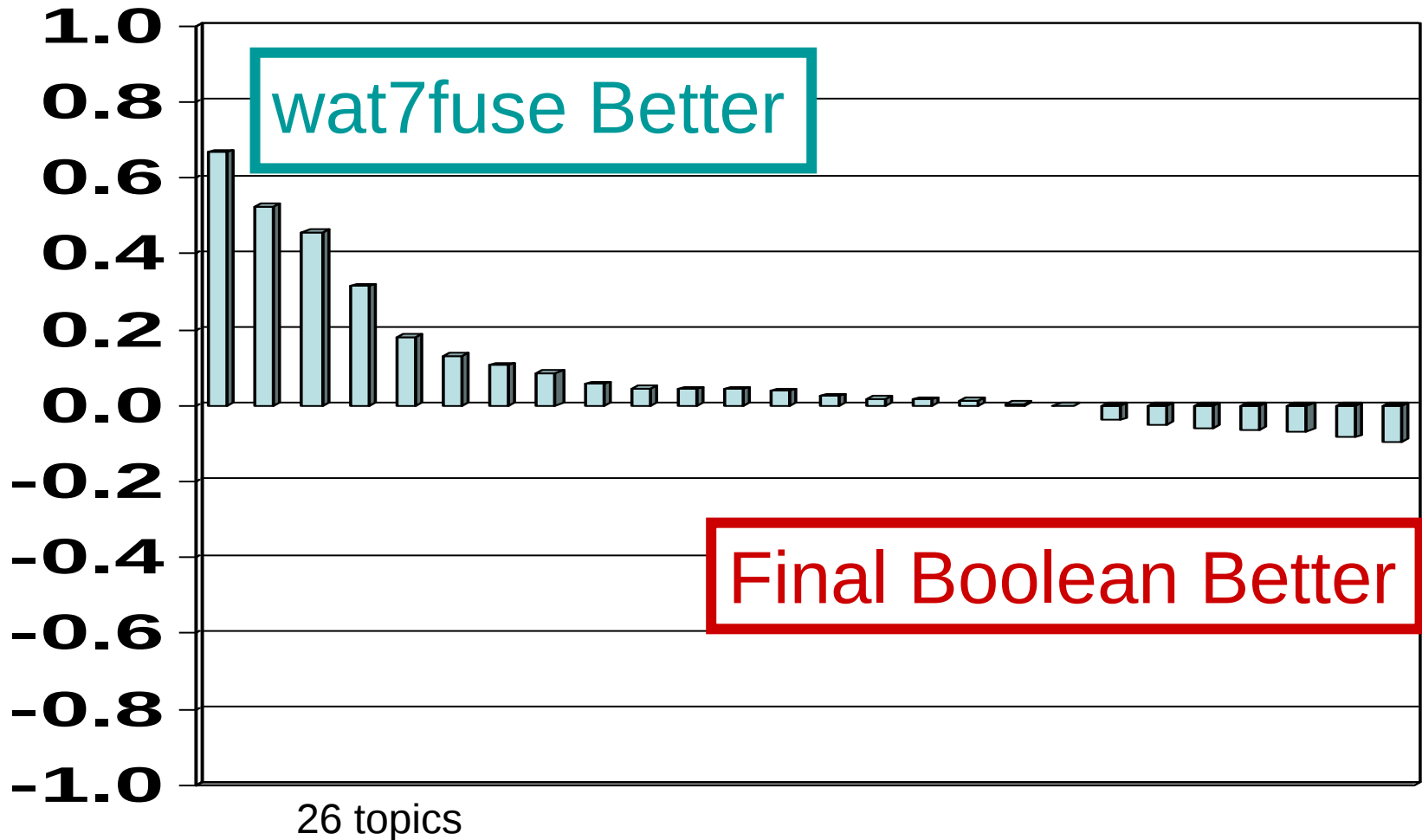
Mean estR=0.33 (26 topics)

- Missed 67% of relevant documents (on average)

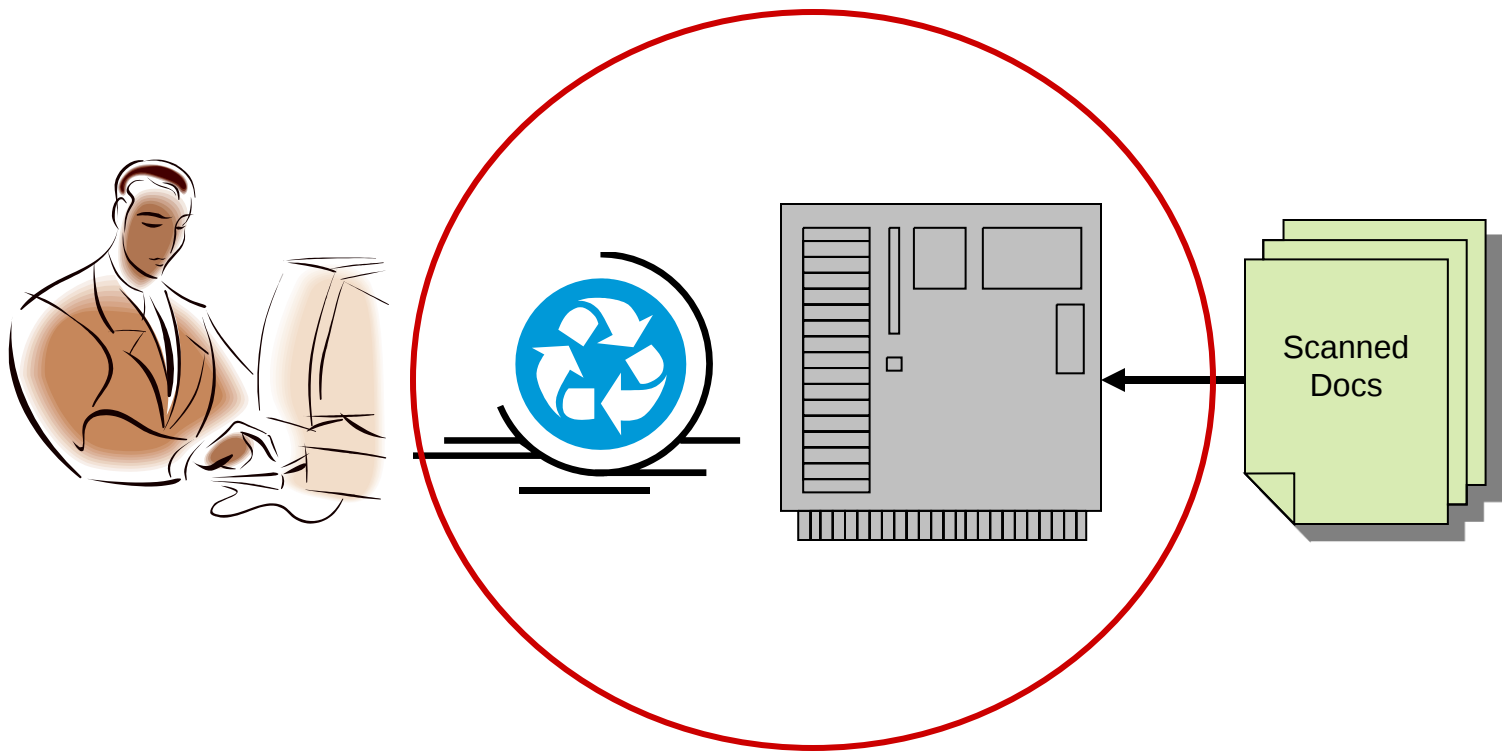
Max estR =0.99, Topic 127
(sanitation procedures)

Min estR=0.00, Topic 142
(contingent sales)

2008 Δ estR@B: wat7fuse vs. Boolean

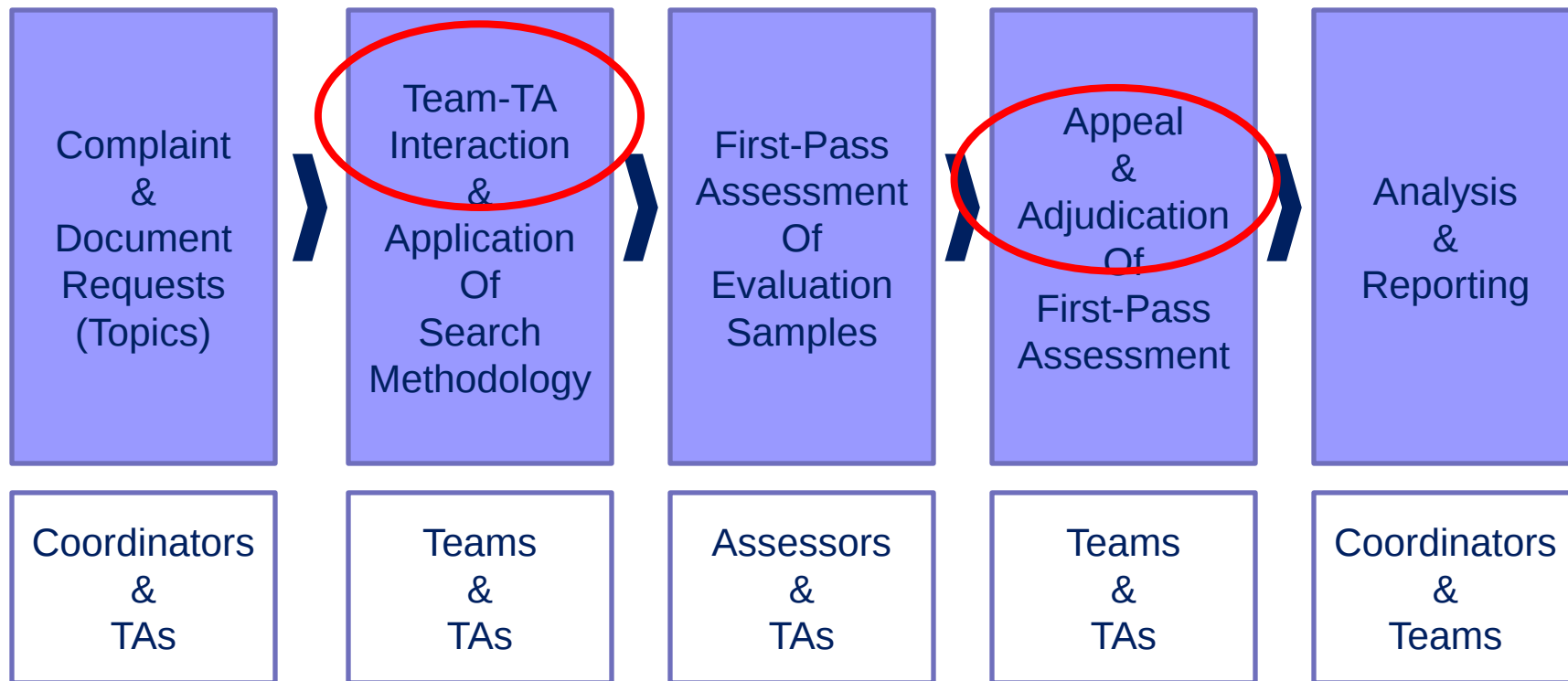


Evaluation Design



Interactive Task

Interactive Task: Key Steps





Interactive Task: Participation

■ 2008

- ☐ 4 Participating Teams (2 commercial, 2 academic)
- ☐ 3 Topics (and 3 TAs)
- ☐ Test Collection: MSA Tobacco Collection

■ 2009

- ☐ 11 Participating Teams (8 commercial, 3 academic)
- ☐ 7 Topics (and 7 TAs)
- ☐ Test Collection: Enron Collection

■ 2010

- ☐ 12 Participating Teams (6 commercial, 5 academic, 1 govt)
- ☐ 4 Topics (and 4 TAs)
- ☐ Test Collection: Enron Collection (new EDRM version)

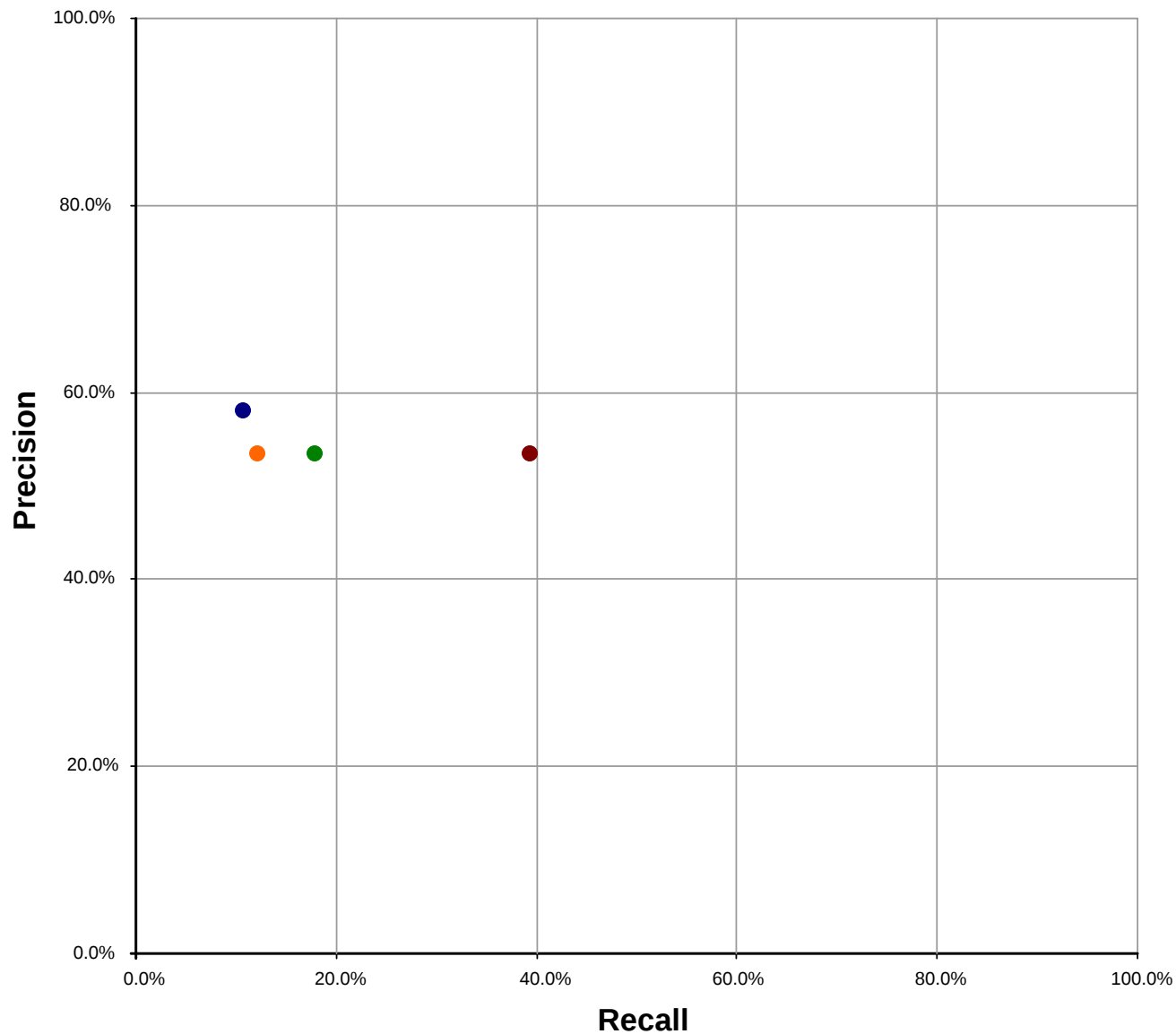
UB	CI	H5	Pitt	AdHoc	N	n	a	r
R	R	R	R	R	5,727	46	46	38
R	R	R	R	N	24	5	5	4
R	R	R	N	R	11,965	98	98	78
R	R	R	N	N	995	9	9	9
R	R	N	R	R	131	5	5	3
R	R	N	R	N	0	0	0	0
R	R	N	N	R	1,547	13	13	2
R	R	N	N	N	220	5	5	2
R	N	R	R	R	1,901	15	15	11
R	N	R	R	N	46	5	5	2
R	N	R	N	R	17,082	145	145	111
R	N	R	N	N	10,291	84	84	61
R	N	N	R	R	176	5	5	1
R	N	N	R	N	19	5	5	2
R	N	N	N	R	7,679	62	61	23
R	N	N	N	N	9,531	77	77	17
N	R	R	R	R	8,068	65	65	49
N	R	R	R	N	101	5	5	2
N	R	R	N	R	73,280	541	540	393
N	R	R	N	N	28,409	235	235	146
N	R	N	R	R	1,185	10	10	4
N	R	N	R	N	37	5	4	3
N	R	N	N	R	23,688	193	193	84
N	R	N	N	N	20,078	171	164	57
N	N	R	R	R	5,321	43	43	33
N	N	R	R	N	371	5	5	2
N	N	R	N	R	151,787	800	795	552
N	N	R	N	N	293,439	1,100	1,095	621
N	N	N	R	R	2,253	18	18	6
N	N	N	R	N	456	5	5	2
N	N	N	N	R	526,099	1,100	1,087	234
N	N	N	N	N	5,708,286	1,625	1,579	111
TOTAL					6,910,192	6,500	6,421	2,663

2008 Interactive Topics

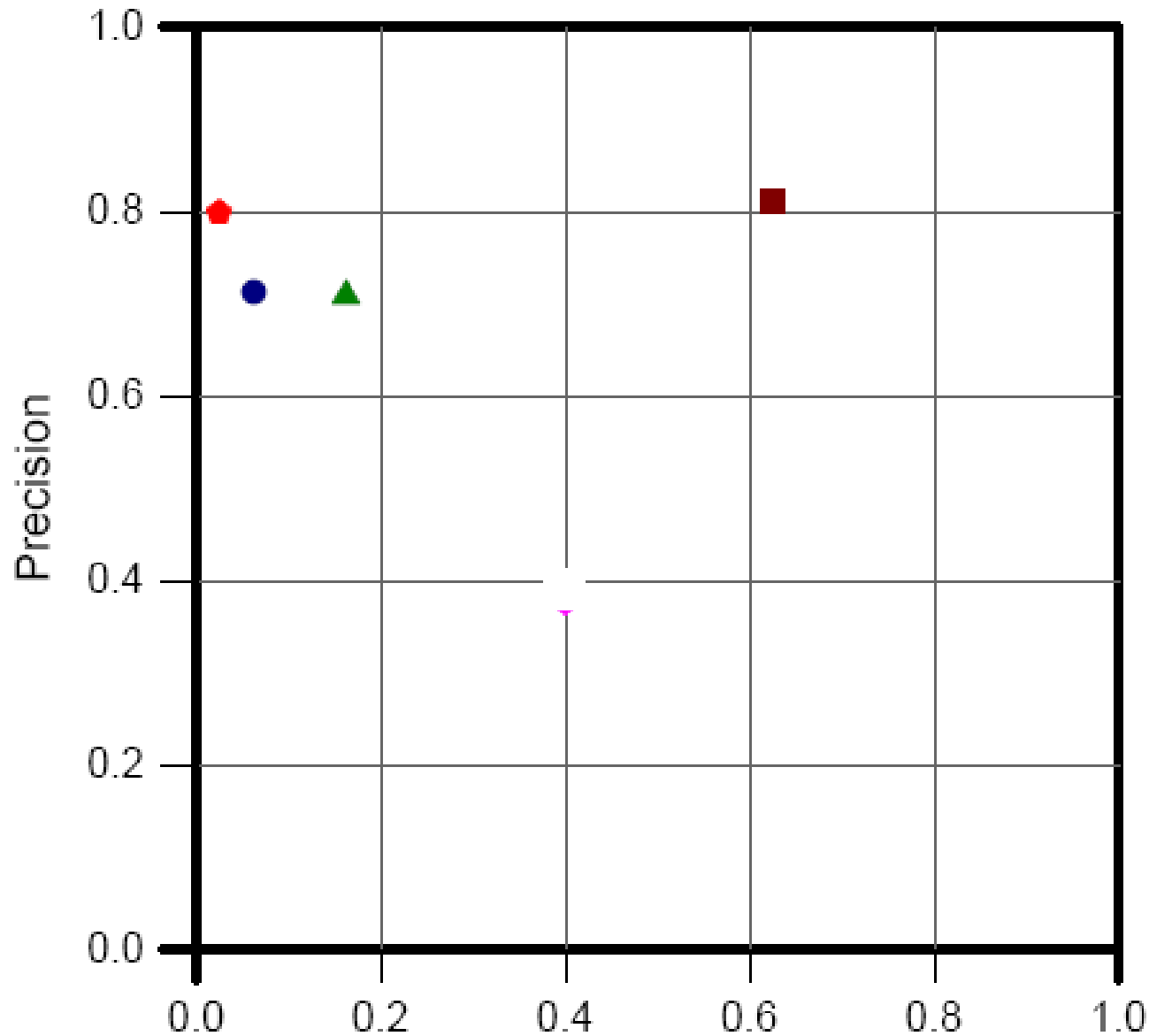
Topic	Samples	Est Nrel Pre- adjudication	Est Nrel: Post- Adjudication	Relevance Density
102	4,500		562,402 $\pm$ 73,000	8.1%
103	6,500	914,258 $\pm$ 72,000	786,862 $\pm$ 54,000	11.4%
104	2,500		45,614 $\pm$ 25,000	0.7%

Topic 103

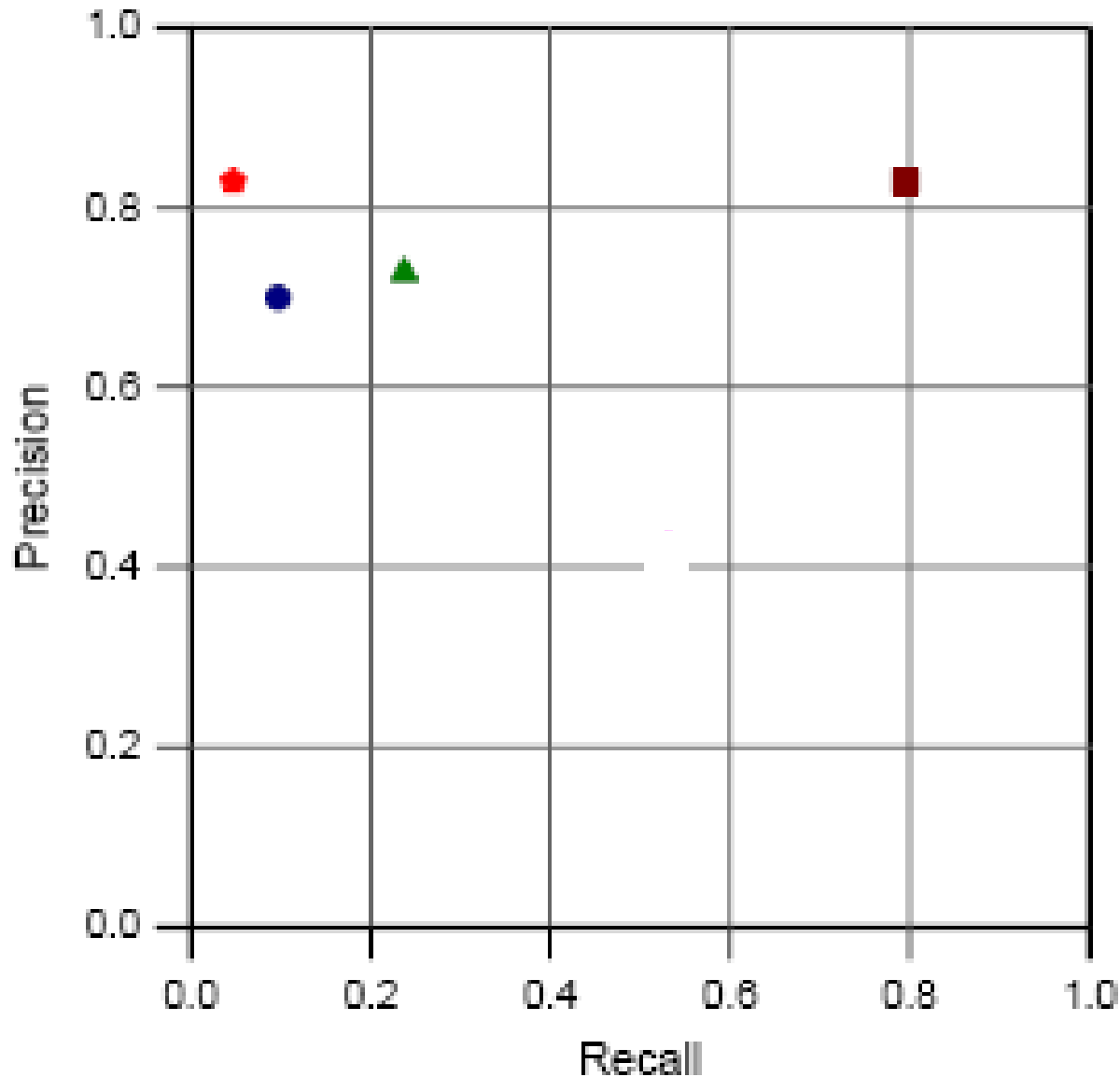
Pre-Adjudication Results



Post-Adjudication Results



Results on Good OCR

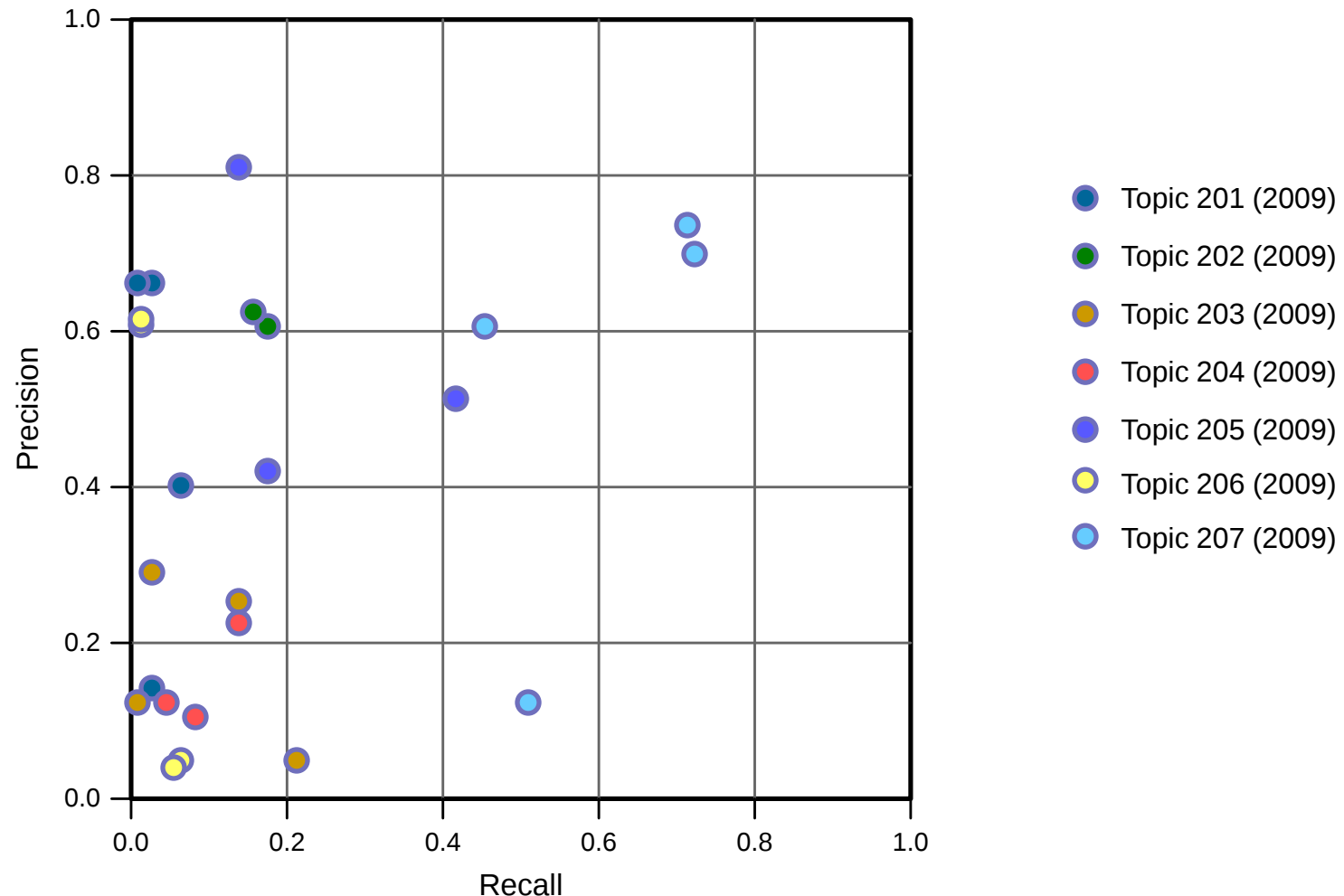


High OCR-accuracy documents only

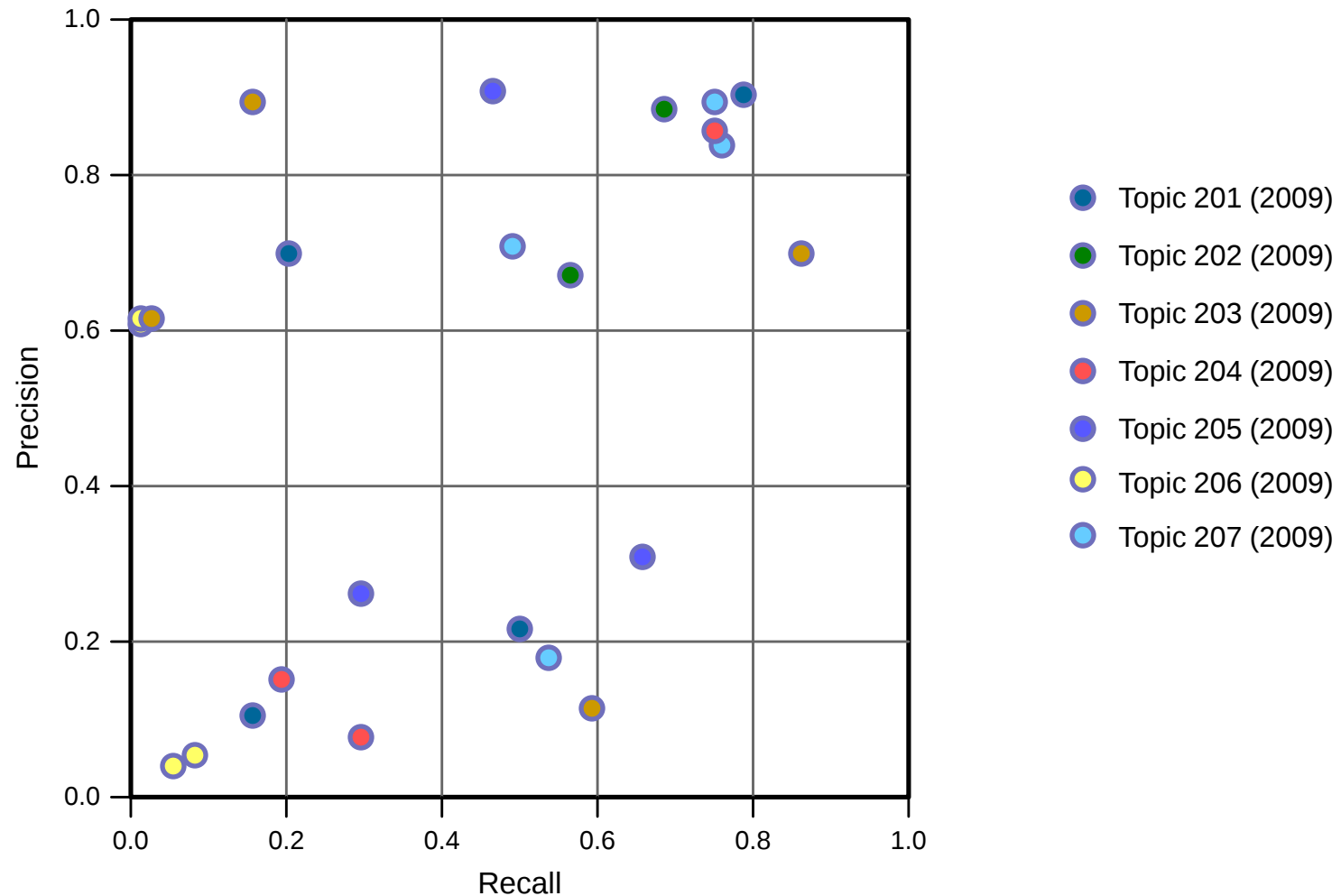
TREC Enron Email Test Collection Version 1

- **Enron Collection**
 - A collection of emails produced by Enron in response to requests from the Federal Energy Regulatory Commission (FERC)
 - First year used in the Legal Track
- **Size of Collection (post-deduplication)**
 - 569,034 messages
 - 847,791 documents
 - Over 3.8 million pages
- **Distribution Format**
 - Extracted Text (in EDRM XML interchange format)
 - Native .msg files

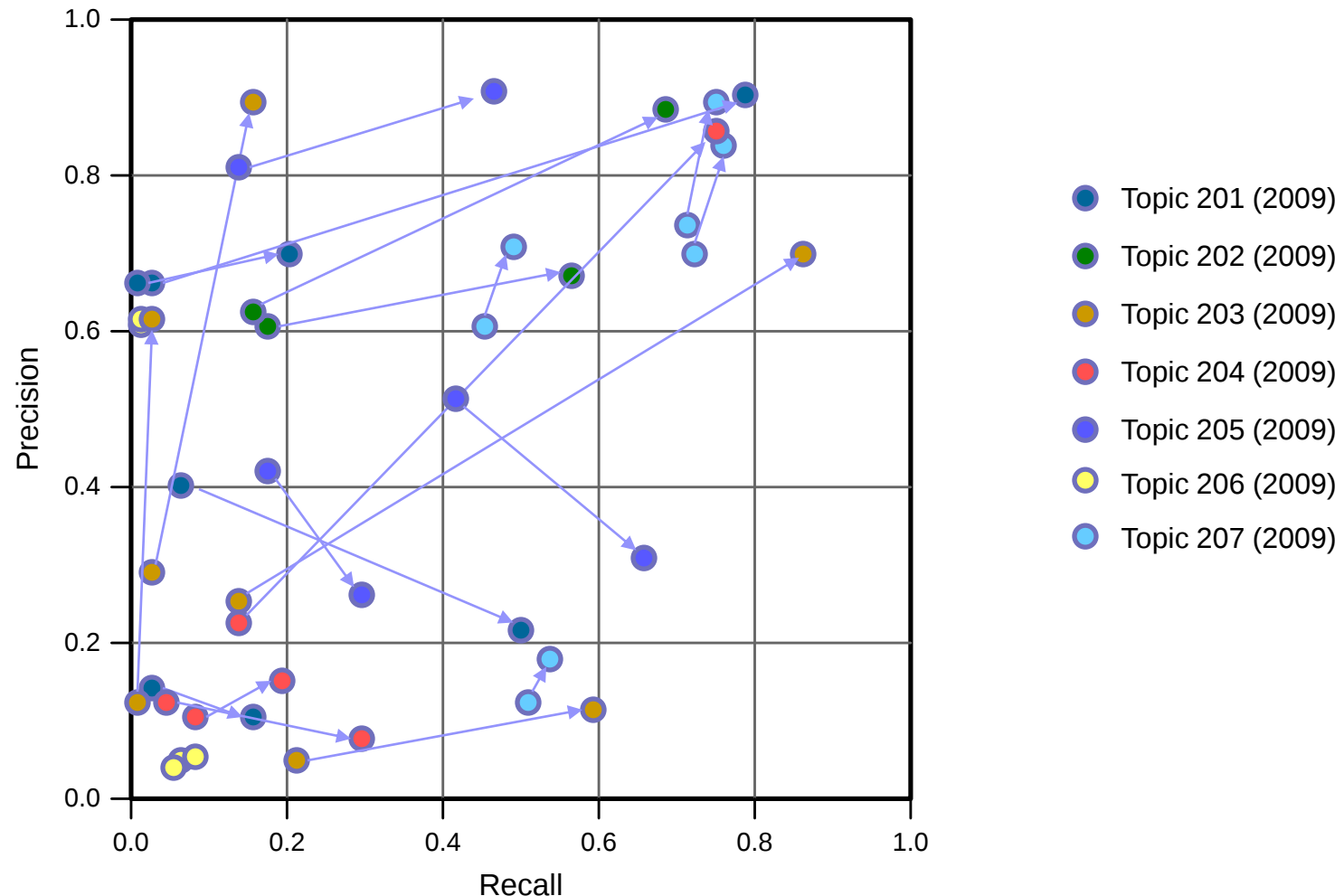
2009 Results (pre-adjudication)



2009 Results (post-adjudication)



2009 Results (pre- to post-adj)





EDRM Enron V2 Dataset

Email from ~150 Enron executives

1.3M records captured by FERC

Processed to several formats by ZL/EDRM

- EDRM XML (text+native) ~100GB
- PST ~100GB

Deduped, reformatted by U. Waterloo

- 455,449 messages + 230,143 attachments = 685,592 docs
- Text (1.2 GB compressed; 5.5GB uncompressed)
- Mapping from PST docs to EDRM document identifiers

Used for both Learning and Interactive tasks

- Participants submitted EDRM document identifiers



Topic 301 (2010)

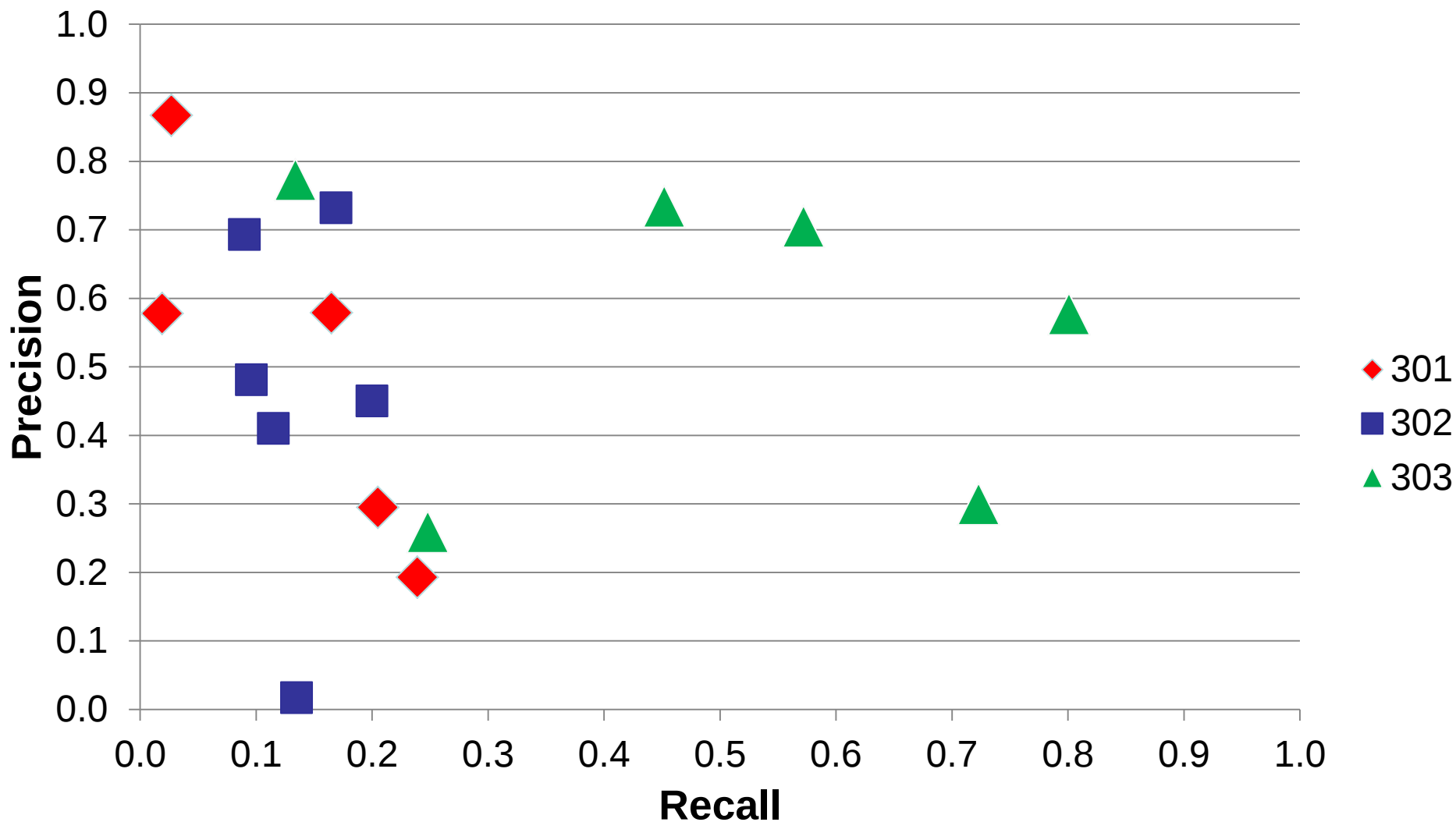
- Document Request

- ☐ All documents or communications that describe, discuss, refer to, report on, or relate to onshore or offshore oil and gas drilling or extraction activities, whether past, present or future, actual, anticipated, possible or potential, including, but not limited to, all business and other plans relating thereto, all anticipated revenues therefrom, and all risk calculations or risk management analyses in connection therewith.

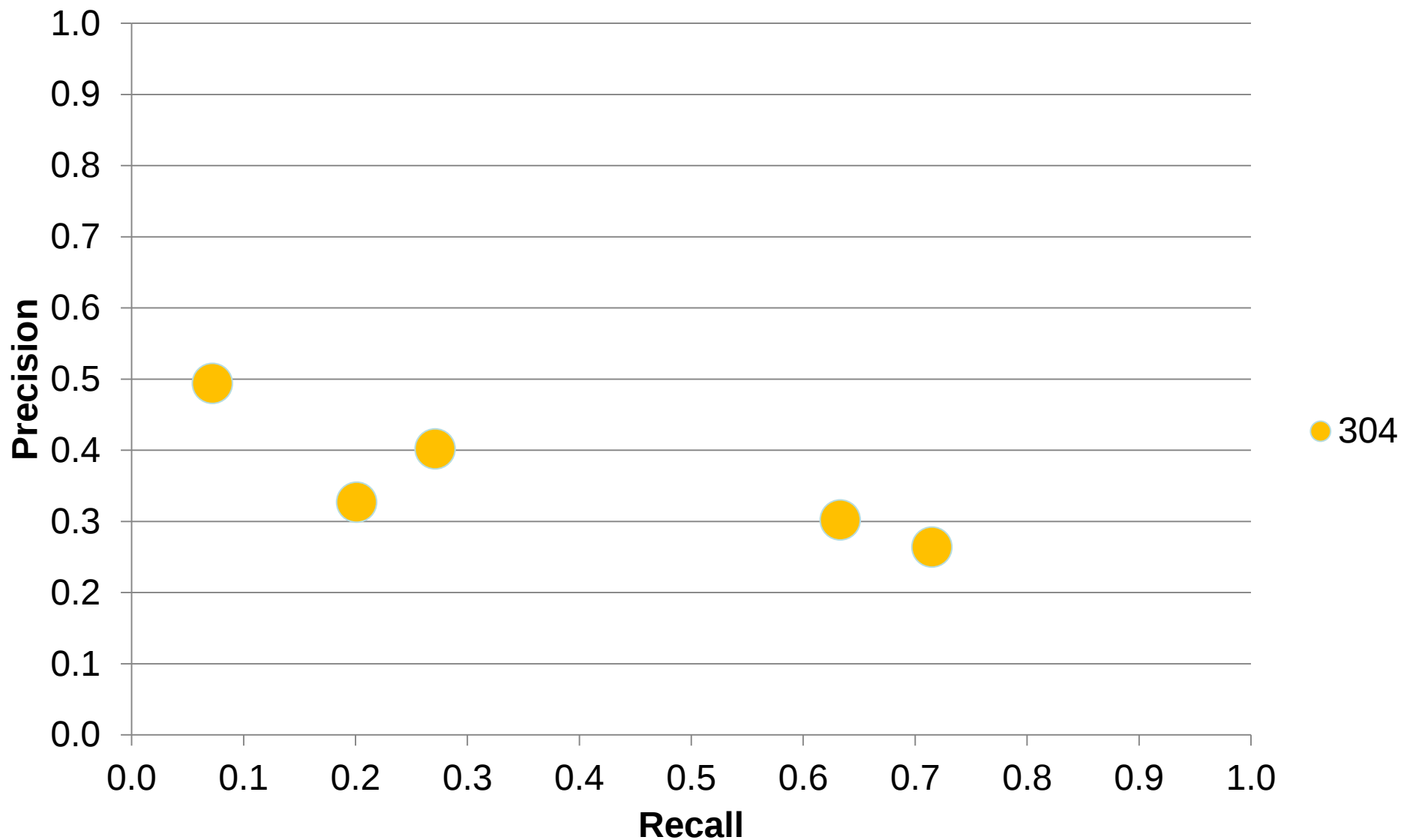
- Topic Authority

- ☐ Mira Edelman (Hughes Hubbard)

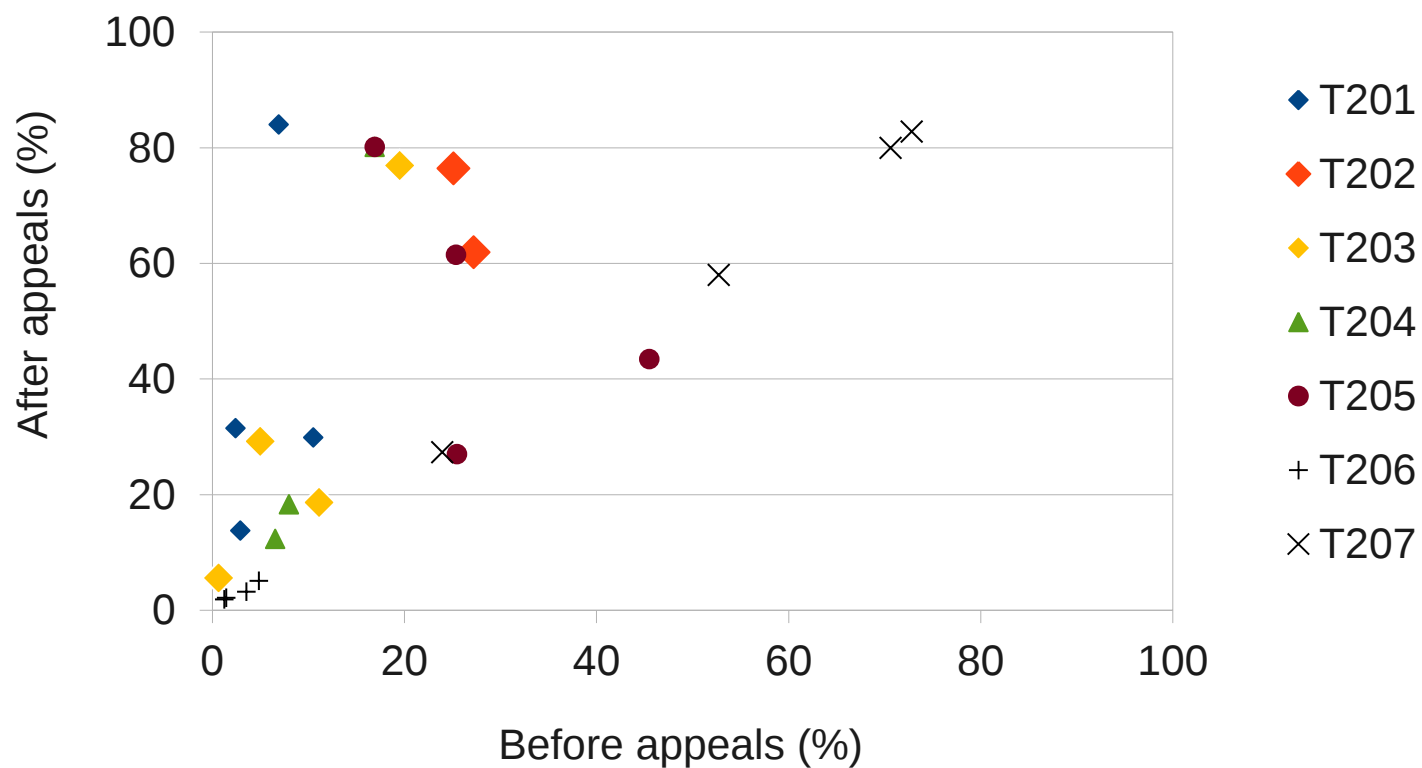
2010 Post-Adj Relevance Results



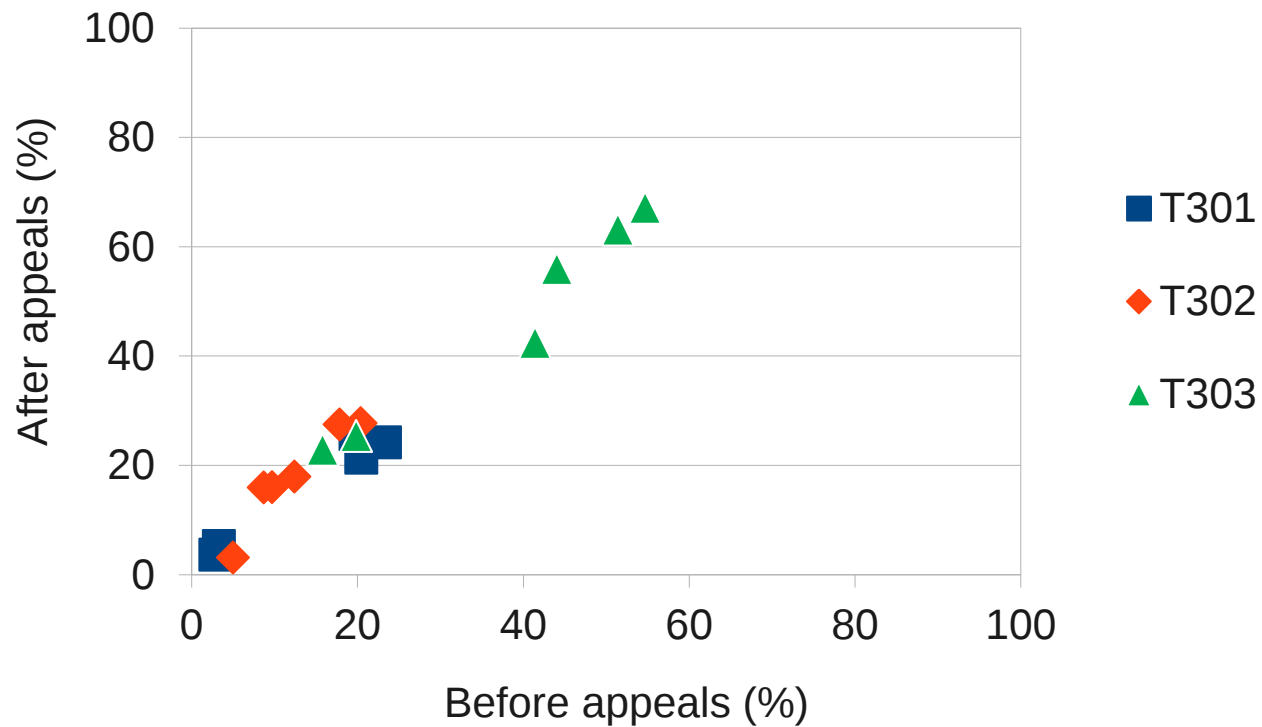
2010 Post-Adj Privilege Results



2009 Change in F1



2010 Change in F1



User Studies

- Goal is to account for interface issues
 - By studying the interface component
 - By studying the complete system
- Formative evaluation
 - Provide a basis for system development
- Summative evaluation
 - Designed to assess performance

Blair and Maron (1985)

- A classic study of retrieval effectiveness
 - Earlier studies used unrealistically small collections
- Studied an archive of documents for a lawsuit
 - 40,000 documents, ~350,000 pages of text
 - 40 different queries
 - Used IBM's STAIRS full-text system
- Approach:
 - Lawyers wanted at least 75% of all relevant documents
 - Precision and recall evaluated only after the lawyers were satisfied with the results

Blair and Maron's Results

- Mean precision: 79%
- Mean recall: 20% (!!)
- Why recall was low?
 - Users can't anticipate terms used in relevant documents

“accident” might be referred to as “event”, “incident”, “situation”, “problem,” ...
 - Differing technical terminology
 - Slang, misspellings
- Other findings:
 - Searches by both lawyers had similar performance
 - Lawyer's recall was not much different from paralegal's

Additional Effects in User Studies

- Learning
 - Vary topic presentation order
- Fatigue
 - Vary system presentation order
- Topic+User (Expertise)
 - Ask about prior knowledge of each topic

Batch vs. User Evaluations

- Do batch (black box) and user evaluations give the same results? If not, why?
- Two different tasks:
 - Instance recall (6 topics)

What countries import Cuban sugar?

What tropical storms, hurricanes, and typhoons have caused property damage or loss of life?

- Question answering (8 topics)

Which painting did Edvard Munch complete first, “Vampire” or “Puberty”?

Is Denmark larger or smaller in population than Norway?

Results

- Compared of two systems:
 - a baseline system
 - an improved system that was provably better in batch evaluations
- Results:

	Instance Recall		Question Answering	
	Batch MAP	User recall	Batch MAP	User accuracy
Baseline	0.2753	0.3230	0.2696	66%
Improved	0.3239	0.3728	0.3544	60%
Change	+18%	+15%	+32%	-6%
p-value (paired t-test)	0.24	0.27	0.06	0.41

Qualitative User Studies

- Observe user behavior
 - Instrumented software, eye trackers, etc.
 - Face and keyboard cameras
 - Think-aloud protocols
 - Interviews and focus groups
- Organize the data
 - For example, group it into overlapping categories
- Look for patterns and themes
- Develop a “grounded theory”