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

NIST Special Publication 500-274

Information Technology:

The Sixteenth Text Retrieval Conference

TREC 2007

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

December 2008

The National Institute of Standards and Technology was established in 1988 by Congress to "assist industry in the development of technology ... needed to improve product quality, to modernize manufacturing processes, to ensure product reliability ... and to facilitate rapid commercialization ... of products based on new scientific discoveries."

NIST, originally founded as the National Bureau of Standards in 1901, works to strengthen U.S. industry's competitiveness; advance science and engineering; and improve public health, safety, and the environment. One of the agency's basic functions is to develop, maintain, and retain custody of the national standards of measurement, and provide the means and methods for comparing standards used in science, engineering, manufacturing, commerce, industry, and education with the standards adopted or recognized by the Federal Government.

As an agency of the U.S. Commerce Department, NIST conducts basic and applied research in the physical sciences and engineering, and develops measurement techniques, test methods, standards, and related services. The Institute does generic and precompetitive work on new and advanced technologies. NIST's research facilities are located at Gaithersburg, MD 20899, and at Boulder, CO 80303. Major technical operating units and their principal activities are listed below. For more information visit the NIST Website at http://www.nist.gov, or contact the Publications and Program Inquiries Desk, 301-975-NIST.

Office of the Director
- Baldrige National Quality Program
- Public and Business Affairs
- Civil Rights and Diversity
- International and Academic Affairs

Technology Services
- Standards Services
- Measurement Services
- Information Services
- Weights and Measures

Advanced Technology Program
- Economic Assessment
- Information Technology and Electronics
- Chemistry and Life Sciences

Manufacturing Extension Partnership Program
- Center Operations
- Systems Operation
- Program Development

Electronics and Electrical Engineering Laboratory
- Semiconductor Electronics
- Optoelectronics[1]
- Quantum Electrical Metrology
- Electromagnetics

Materials Science and Engineering Laboratory
- Intelligent Processing of Materials
- Ceramics
- Materials Reliability[1]
- Polymers
- Metallurgy
- NIST Center for Neutron Research

NIST Center for Neutron Research

Nanoscale Science and Technology

Chemical Science and Technology Laboratory
- Biochemical Science
- Process Measurements
- Surface and Microanalysis Science
- Physical and Chemical Properties[2]
- Analytical Chemistry

Physics Laboratory
- Electron and Optical Physics
- Atomic Physics
- Optical Technology
- Ionizing Radiation
- Time and Frequency[1]
- Quantum Physics[1]

Manufacturing Engineering Laboratory
- Precision Engineering
- Manufacturing Metrology
- Intelligent Systems
- Fabrication Technology
- Manufacturing Systems Integration

Building and Fire Research Laboratory
- Materials and Construction Research
- Building Environment
- Fire Research

Information Technology Laboratory
- Mathematical and Computational Sciences[2]
- Advanced Network Technologies
- Computer Security
- Information Access
- Software Diagnostics and Conformance Testing
- Statistical Engineering

[1]At Boulder, CO 80303
[2]Some elements at Boulder, CO

NIST Special Publication 500-274

Information Technology:

The Sixteenth Text Retrieval Conference

TREC 2007

Ellen M. Voorhees
and
Lori P. Buckland,
Editors

Information Access Division
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

December 2008

U.S. Department of Commerce
Carlos M. Gutierrez, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Deputy Director

Reports on Information Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) stimulates U.S. economic growth and industrial competitiveness through technical leadership and collaborative research in critical infrastructure technology, including tests, test methods, reference data, and forward-looking standards, to advance the development and productive use of information technology. To overcome barriers to usability, scalability, interoperability, and security in information systems and networks, ITL programs focus on a broad range of networking, security, and advanced information technologies, as well as the mathematical, statistical, and computational sciences. This Special Publication 500-series reports on ITL's research in tests and test methods for information technology, and its collaborative activities with industry, government, and academic organizations.

National Institute of Standards and Technology Special Publication 500-274
Natl. Inst. Stand. Technol. Spec. Publ. 500-274, 163 pages (December 2008)

Foreword

This report constitutes the proceedings of the 2007 Text REtrieval Conference, TREC 2007, held in Gaithersburg, Maryland, November 6–9, 2007. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Intelligence Advanced Research Projects Activity (IARPA). Approximately 150 people attended the conference, including representatives from 18 countries. The conference was the sixteenth in an ongoing series of workshops to evaluate new technologies for text retrieval and related information-seeking tasks.

The workshop included plenary sessions, discussion groups, a poster session, and demonstrations. Because the participants in the workshop drew on their personal experiences, they sometimes cite specific vendors and commercial products. The inclusion or omission of a particular company or product implies neither endorsement nor criticism by NIST. Any opinions, findings, and conclusions or recommendations expressed in the individual papers are the authors' own and do not necessarily reflect those of the sponsors.

I gratefully acknowledge the tremendous work of the TREC program committee and the track coordinators.

Ellen Voorhees
September 12, 2008

TREC 2007 Program Committee

Ellen Voorhees, NIST, chair
James Allan, University of Massachusetts at Amherst
Chris Buckley, Sabir Research, Inc.
Gordon Cormack, University of Waterloo
Susan Dumais, Microsoft
Donna Harman, NIST
Bill Hersh, Oregon Health & Science University
David Lewis, David Lewis Consulting
John Prager, IBM
Steve Robertson, Microsoft
Mark Sanderson, University of Sheffield
Ian Soboroff, NIST
Richard Tong, Tarragon Consulting
Ross Wilkinson, CSIRO

TREC 2007 Proceedings

Overview Papers

Other Papers

(Contents of these papers are found on the TREC 2007 Proceedings CD.)

Appendix

(Contents of the Appendix are found on the TREC 2007 Proceedings CD.)

Papers: Alphabetical by Organization

(Contents of these papers are found on the TREC 2007 Proceedings CD.)

Erasmus MC
Cross Language Information Retrieval for Biomedical Literature

Exegy, Inc.
Exegy at TREC 2007 Million Query Track

Fitchburg State College
FSC at TREC

Fondazione Ugo Bordoni
FUB, IASI-CNR and University of Tor Vergata at TREC

Fudan University
FDU at TREC 2007: Opinion Retrieval of Blog Track
WIM at TREC 2007
FDUQA on TREC 2007 QA Track

Georgetown University Medical Center
Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics

IASI "Antonio Ruberti"
FSC at TREC

IBM Haifa Research Lab
Lucene and Juru at TREC 2007: 1-Million Queries Track

Indiana University
WIDIT in TREC 2007 Blog Track: Combining Lexicon-Based Methods to Detect Opinionated Blogs

Illinois Institute of Technology
IIT TREC 2007 Genomics Track: Using Concept-Based Semantics in Context for Genomics Literature Passage Retrieval

Kobe University
TREC 2007 Blog Track Experiments at Kobe University
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2007 Blog Track

Kyoto University
Passage Retrieval with Vector Space and Query-Level Aspect Models

Language Computer Corporation
Question Answering with LCC's CHAUCER-2 at TREC 2007

Long Island University
TREC 2007 Legal Track Interactive Task: A Report from the LIU Team

Lymba Corporation
Lymba's PowerAnswer 4 in TREC 2007

Michigan State University
Michigan State University at the 2007 TREC ciQA Task

Microsoft, USA
Overview of the TREC 2007 Enterprise Track

Microsoft Research Asia
Research on Enterprise Track of TREC 2007 at SJTU APEX Lab

MIT
CSAIL at TREC 2007 Question Answering

Mitsubishi
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

Mitsubishi and Southern Connecticut State University
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

Mitsubishi and University of Massachusetts, Amherst
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

MSR Cambridge, UK
Overview of the TREC 2007 Enterprise Track

National Archives and Records Administration
Overview of the TREC 2007 Legal Track

National Institute of Informatics
Opinion Retrieval Experiments Using Generative Models: Experiments for the TREC 2007 Blog Track

National Institute of Standards and Technology
Overview of TREC 2007
Overview of the TREC 2007 Blog Track
Overview of the TREC 2007 Enterprise Track
Overview of the TREC 2007 Question Answering Track

National Library of Medicine
Combining Resources to Find Answers to Biomedical Questions

Northeastern University
The Hedge Algorithm for Metasearch at TREC 2007
Million Query Track 2007 Overview

National Taiwan University
NTU at TREC 2007 Blog Track

Open Text Corporation
Experiments with the Negotiated Boolean Queries of the TREC 2007 Legal Discovery Track
The Open University at TREC 2007 Enterprise Track
Overview of the TREC 2007 Legal Track

Oregon Health & Science University
The OHSU Biomedical Question Answering System Framework
TREC 2007 Genomics Track Overview

Pfizer Corporation
TREC 2007 Genomics Track Overview

Queens College
Testing an Entity Ranking Function for English Factoid QA

RMIT University
TREC 2007 ciQA Track at RMIT and CSIRO
RMIT University at the TREC 2007 Enterprise Track

The Robert Gordon University
The Robert Gordon University at the Opinion Retrieval Task of the 2007 TREC Blog Track

Saarland University
The Alyssa System at TREC QA 2007: Do We Need Blog06?

Sabir Research, Inc.
Examining Overfitting in Relevance Feedback: Sabir Research at TREC 2007

Shanghai Jiao Tong University
Research on Enterprise Track of TREC 2007 at SJTU APEX Lab

Swiss Institute of Bioinformatics
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

Technical University Berlin
Feed Distillation Using AdaBoost and Topic Maps

TNO
Cross Language Information Retrieval for Biomedical Literature

Tokyo Institute of Technology
TREC 2007 Question Answering Experiments at Tokyo Institute of Technology

Tsinghua National Laboratory for Information Science and Technology
THUIR at TREC 2007: Enterprise Track

Tufts University
Relaxed Online SVMs in the TREC Spam Filtering Track

University of Alaska, Fairbanks
Collection Selection Based on Historical Performance for Efficient Processing

University at Albany SUNY
UAlbany's ILQUA at TREC 2007

University of Alicante
Using IR-n for Information Retrieval of Genomics Track

University of Arkansas at Little Rock
Topic Categorization for Relevancy and Opinion Detection
UALR at TREC-ENT 2007

University of Amsterdam
Query and Document Models for Enterprise Search
Bootstrapping Language Associated with Biomedical Entities
Access to Legal Documents: Exact Match, Best Match, and Combinations
Parsimonious Language Models for a Terabyte of Text
The University of Amsterdam at the TREC 2007 QA Track
Language Modeling Approaches to Blog Postand Feed Finding

University of Buffalo
Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics

University of Glasgow
University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier
Overview of the TREC 2007 Blog Track

University of Geneva
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

University Hospital of Geneva
Combining Resources to Find Answers to Biomedical Questions
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

University Hospital of Geneva and University of Geneva
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

University of Illinois at Chicago
TREC Genomics Track at UIC
UIC at TREC 2007 Blog Track

University of Illinois at Urbana-Chanpaign
Language Models for Genomics Information Retrieval: UIUC at TREC 2007 Genomics Track

The University of Iowa
Exploring the Legal Discovery and Enterprise Tracks at the University of Iowa

Universität Karlsruhe
Semantic Extensions of the Ephyra QA System for TREC 2007

University of Lethbridge
University of Lethbridge's Participation in TREC 2007 QA Track

University of Maryland, College Park
TREC 2007 ciQA Task: University of Maryland
Overview of the TREC 2007 Legal Track
Overview of the TREC 2007 Question Answering Track

University of Massachusetts, Amherst
UMass Complex Interactive Question Answering (ciQA) 2007: Human Performance as Question Answerers
UMass at TREC 2007 Blog Distillation Task
CIIR Experiments for TREC Legal 2007
Indri at TREC 2007: Million Query (1MQ) Track
Million Query Track 2007 Overview

The University of Melbourne
Entity-Based Relevance Feedback for Genomic List Answer Retrieval
Passage Retrieval with Vector Space and Query-Level Aspect Models

University of Missouri, Kansas City
Evaluation of Query Formulations in the Negotiated Query Refinement Process of Legal e-Discovery: UMKC at TREC 2007 Legal Track

University of North Carolina, Chapel Hill
Using Interactions to Improve Translation Dictionaries: UNC, Yahoo! and ciQA
Overview of the TREC 2007 Question Answering Track

University of Neuchatel
IR-Specific Searches at TREC 2007: Genomics & Blog Experiments

University of North Texas
Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics

University "Tor Vergata"
FSC at TREC

University of Rome "La Sapienza"
The Pronto QA System at TREC 2007: Harvesting Hyponyms, Using Nominalisation Patterns, and Computing Answer Cardinality

University of Sydney
The Pronto QA System at TREC 2007: Harvesting Hyponyms, Using Nominalisation Patterns, and Computing Answer Cardinality

University of Maryland, College Park
Overview of the TREC 2007 Legal Track

University of Twente
University of Twente at the TREC 2007 Enterprise Track: Modeling Relevance Propagation for the Expert search Task
Million Query Track 2007 Overview
Cross Language Information Retrieval for Biomedical Literature
Parsimonious Language Models for a Terabyte of Text

University of Washington Information School
University of Washington (UW) at Legal TREC Interactive 2007

University of Waterloo
TREC 2007 Spam Track Overview
University of Waterloo Participates in the TREC 2007 Spam Track
Complex Interactive Question Answering Enhanced with Wikipedia
Using Subjective Adjectives in Opinion Retrieval from Blogs
Enterprise Search: Identifying Relevant Sentences and Using Them for Query Expansion
MultiText Legal Experiments at TREC 2007

Wuhan University
CSIR at TREC 2007 Expert Search Task
WHU at Blog Track 2007

Yahoo! Research
CIIR Experiments for TREC Legal 2007
(University of Massachusetts, Amherst)

Yahoo! Research Barcelona
Using Interactions to Improve Translation Dictionaries: UNC, Yahoo! and ciQA

York University, Toronto
York University at TREC 2007: Enterprise Document Search
York University at TREC 2007: Genomics Track

Papers: Organized by Track

(Contents of these papers are found on the TREC 2007 Proceedings CD.)

Blog

Technical University Berlin
Feed Distillation Using AdaBoost and Topic Maps

University of Arkansas at Little Rock
Topic Categorization for Relevancy and Opinion Detection

University of Amsterdam
Language Modeling Approaches to Blog Post and Feed Finding

University of Buffalo
Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics

University of Glasgow
University of Glasgow at TREC 2007:
Experiments in Blog and Enterprise Tracks with Terrier

Overview of the TREC 2007 Blog Track

University of Illinois at Chicago
UIC at TREC 2007 Blog Track

University of Massachusetts, Amherst
UMass at TREC 2007 Blog Distillation Task

University of North Texas
Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics

University of Texas, Austin
University of Texas School of Information at TREC 2007

University "Tor Vergata"
FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track

University of Waterloo
Using Subjective Adjectives in Opinion Retrieval from Blogs

Wuhan University
WHU at Blog Track 2007

Enterprise

Chinese Academy of Sciences
Research on Enterprise Track of TREC 2007

CSIRO ICT Centre
Research on Enterprise Track of TREC 2007

CWI
Overview of the TREC 2007 Enterprise Track

Dalian University of Technology
DUTIR at TREC 2007 Enterprise Track

Fudan University
WIM at TREC 2007

Microsoft Research Asia
Research on Enterprise Track of TREC 2007 at SJTU APEX Lab

Microsoft, USA
Overview of the TREC 2007 Enterprise Track

MSR Cambridge, UK
Overview of the TREC 2007 Enterprise Track

National Institute of Standards and Technology
Overview of the TREC 2007 Enterprise Track

The Open University
The Open University at TREC 2007 Enterprise Track

RMIT University
RMIT University at the TREC 2007 Enterprise Track

Shanghai Jiao Tong University
Research on Enterprise Track of TREC 2007 at SJTU APEX Lab

Tsinghua National Laboratory for Information Science and Technology
THUIR at TREC 2007: Enterprise Track

University of Arkansas, Little Rock
UALR at TREC-ENT 2007

University of Amsterdam
Query and Document Models for Enterprise Search

University of Glasgow
University of Glasgow at TREC 2007:
Experiments in Blog and Enterprise Tracks with Terrier

The University of Iowa
Exploring the Legal Discovery and Enterprise Tracks at the University of Iowa

University of Twente
University of Twente at the TREC 2007 Enterprise Track: Modeling Relevance Propagation for the Expert Search Task

University of Waterloo
Enterprise Search: Identifying Relevant Sentences and Using Them for Query Expansion

Wuhan University
CSIR at TREC 2007 Expert Search Task

York University, Toronto
York University at TREC 2007: Enterprise Document Search

Genomics

Arizona State University
Passage Relevancy Through Semantic Relatedness

Concordia University
Interactive Retrieval Using Weights

Dalian University of Technology
DUTIR at TREC 2007 Genomics Track

Erasmus MC
Cross Language Information Retrieval for Biomedical Literature

European Bioinformatics Institute
Information Retrieval and Information Extraction in TREC Genomics 2007

Illinois Institute of Technology
IIT TREC 2007 Genomics Track: Using Concept-Based semantics in Context for Genomics Literature Passage Retrieval

Kyoto University
Passage Retrieval with Vector Space and Query-Level Aspect Models

National Library of Medicine
Combining Resources to Find Answers to Biomedical Questions

Oregon Health & Science University
TREC 2007 Genomics Track Overview

The OHSU Biomedical Question Answering System Framework

Pfizer Corporation
TREC 2007 Genomics Track Overview

Swiss Institute of Bioinformatics
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

TNO
Cross Language Information Retrieval for Biomedical Literature

University of Alicante
Using IR-n for Information Retrieval of Genomics Track

University of Amsterdam
Bootstrapping Language Associated with Biomedical Entities

University of Geneva
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

University Hospital of Geneva
Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics

Combining Resources to Find Answers to Biomedical Questions

University of Illinois at Chicago
TREC Genomics Track at UIC

University of Illinois at Urbana-Chanpaign
Language Models for Genomics Information Retrieval: UIUC at TREC 2007 Genomics Track

The University of Melbourne
Passage Retrieval with Vector Space and Query-Level Aspect Models

Entity-Based Relevance Feedback for Genomic List Answer Retrieval

University of Neuchatel
IR-Specific Searches at TREC 2007: Genomics & Blog Experiments

University of Twente
Cross Language Information Retrieval for Biomedical Literature

York University
York University at TREC 2007: Genomics Track

Legal

Carnegie Mellon University
Stuctured Queries for Legal Search

CogiTech
CIIR Experiments for TREC Legal 2007

Dartmouth College
Overview of the TREC 2007 Legal Track
Dartmouth College at TREC 2007 Legal Track

Million Query

IBM Haifa Research Lab
Lucene and Juru at TREC 2007: 1-Million Queries Track

Northeastern University
The Hedge Algorithm for Metasearch at TREC 2007

Million Query Track 2007 Overview

University of Alaska, Fairbanks
Collection Selection Based on Historical Performance for Efficient Processing

University of Amsterdam
Parsimonious Language Models for a Terabyte of Text

University of Massachusetts, Amherst
Million Query Track 2007 Overview

Indri at TREC 2007: Million Query (1MQ) Track

University of Twente
Parsimonious Language Models for a Terabyte of Text

Question Answering

Carnegie Mellon University
Semantic Extensions of the Ephyra QA System for TREC 2007

Concordia University
Concordia University at the TREC 2007 QA Track

CSIRO ICT Centre
TREC 2007 ciQA Track at RMIT and CSIRO

Drexel University
Drexel at TREC 2007: Question Answering

EffectiveSoft
Intellexer Question Answering

Fitchburg State College
FSC at TREC

Fudan University
FDUQA on TREC 2007 QA Track

IBM India Research Lab
ITD-IBMIRL System for Question Answering Using Pattern Matching, Semantic Type and Semantic Category Recognition

Indian Institute of Technology
ITD-IBMIRL System for Question Answering Using Pattern Matching, Semantic Type and Semantic Category Recognition

Language Computer Corporation
Question Answering with LCC's CHAUCER-2 at TREC 2007

Lymba Corporation
Lymba's PowerAnswer 4 in TREC 2007

Michigan State University
Michigan State University at the 2007 TREC ciQA Task

MIT
CSAIL at TREC 2007 Question Answering

National Institute of Standards and Technology
Overview of the TREC 2007 Question Answering Track

Queens College
Testing an Entity Ranking Function for English Factoid QA

RMIT University
TREC 2007 ciQA Track at RMIT and CSIRO

Saarland University
The Alyssa System at TREC QA 2007: Do We Need Blog06?

Tokyo Institute of Technology
TREC 2007 Question Answering Experiments at Tokyo Institute of Technology

University at Albany SUNY
UAlbany's ILQUA at TREC 2007

University of Amsterdam
The University of Amsterdam at the TREC 2007 QA Track

University of Glasgow
Persuasive, Authorative and Topical Answers for Complex Question Answering

Universität Karlsruhe
Semantic Extensions of the Ephyra QA System for TREC 2007

University of Lethbridge
University of Lethbridge's Participation in TREC 2007 QA Track

University of Maryland, College Park
TREC 2007 ciQA Task: University of Maryland
Overview of the TREC 2007 Question Answering Track

University of Massachusetts, Amherst
UMass Complex Interactive Question Answering (ciQA) 2007:
Human Performance as Question Answerers

University of North Carolina, Chapel Hill
Using Interactions to Improve Translation Dictionaries: UNC, Yahoo! and ciQA
Overview of the TREC 2007 Question Answering Track

University of Rome "La Sapienza"
The Pronto QA System at TREC 2007: Harvesting Hyponyms, Using Nominalisation Patterns, and Computing Answer Cardinality

University of Strathclyde
Persuasive, Authorative and Topical Answers for Complex Question Answering

University of Sydney
The Pronto QA System at TREC 2007: Harvesting Hyponyms, Using Nominalisation Patterns, and Computing Answer Cardinality

University of Waterloo
Complex Interactive Question Answering Enhanced with Wikipedia

Yahoo! Research Barcelona
Using Interactions to Improve Translation Dictionaries: UNC, Yahoo! and ciQA

Spam

Fudan University
WIM at TREC 2007

Mitsubishi
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

Mitsubishi and Southern Connecticut State University
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

Mitsubishi and University of Massachusetts, Amherst
Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007

Tufts University
Relaxed Online SVMs in the TREC Spam Filtering Track

University of Waterloo
TREC 2007 Spam Track Overview
University of Waterloo Participates in the TREC 2007 Spam Track

Abstract

This report constitutes the proceedings of the 2007 Text REtrieval Conference, TREC 2007, held in Gaithersburg, Maryland, November 6–9, 2007. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Intelligence Advanced Research Projects Activity (IARPA). TREC 2007 had 95 participating groups including participants from 18 countries.

TREC 2007 is the latest in a series of workshops designed to foster research in text retrieval and related technologies. This year's conference consisted of seven different tasks: search in support of legal discovery of electronic documents, search within and between blog postings, question answering, detecting spam in an email stream, enterprise search, search in the genomics domain, and strategies for building fair test collections for very large corpora.

The conference included paper sessions and discussion groups. The overview papers for the different "tracks" and for the conference as a whole are gathered in this bound version of the proceedings. The papers from the individual participants and the evaluation output for the runs submitted to TREC 2007 are contained on the disk included in the volume. The TREC 2007 proceedings web site (`http://trec.nist.gov/pubs.html`) also contains the complete proceedings, including system descriptions that detail the timing and storage requirements of the different runs.

Overview of TREC 2007

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

1 Introduction

The sixteenth Text REtrieval Conference, TREC 2007, was held at the National Institute of Standards and Technology (NIST) November 6–9, 2007. The conference was co-sponsored by NIST and the Intelligence Advanced Research Projects Activity (IARPA). TREC 2007 had 95 participating groups from 18 countries. Table 2 at the end of the paper lists the participating groups.

TREC 2007 is the latest in a series of workshops designed to foster research on technologies for information retrieval. The workshop series has four goals:

- to encourage retrieval research based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

TREC 2007 contained seven areas of focus called "tracks". Six of the tracks ran in previous TRECs and explored tasks in question answering, blog search, detecting spam in an email stream, enterprise search, search in support of legal discovery, and information access within the genomics domain. A new track called the million query track investigated techniques for building fair retrieval test collections for very large corpora.

This paper serves as an introduction to the research described in detail in the remainder of the proceedings. The next section provides a summary of the retrieval background knowledge that is assumed in the other papers. Section 3 presents a short description of each track—a more complete description of a track can be found in that track's overview paper in the proceedings. The final section looks toward future TREC conferences.

2 Information Retrieval

Information retrieval is concerned with locating information that will satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. There is increasing interest, however, in finding appropriate information regardless of the medium that happens to contain that information. Thus

"document" can be interpreted as any unit of information such as a blog post, an email message, or an invoice.

The prototypical retrieval task is a researcher doing a literature search in a library. In this environment the retrieval system knows the set of documents to be searched (the library's holdings), but cannot anticipate the particular topic that will be investigated. We call this an *ad hoc* retrieval task, reflecting the arbitrary subject of the search and its short duration. Other examples of ad hoc searches are web surfers using Internet search engines, lawyers performing patent searches or looking for precedent in case law, and analysts searching archived news reports for particular events. A retrieval system's response to an ad hoc search is generally an ordered list of documents sorted such that documents the system believes are more likely to satisfy the information need are ranked before documents it believes are less likely to satisfy the need. The tasks within the million query and legal tracks are examples of ad hoc search tasks. The feed task in the blog track is also an ad hoc search task, though in this case the documents to be ranked are entire blogs rather than blog postings.

In a *categorization* task, the system is responsible for assigning a document to one or more categories from among a given set of categories. Deciding whether a given mail message is spam is one example of a categorization task. The polarity task in the blog track, in which opinions were determined to be pro, con or both, is a second example.

Information retrieval has traditionally focused on returning entire documents in response to a query. This emphasis is both a reflection of retrieval systems' heritage as library reference systems and an acknowledgement of the difficulty of returning more specific responses. Nonetheless, TREC contains several tasks that do focus on more specific responses. In the question answering track, systems are expected to return precisely the answer; the system response to a query in the expert-finding task in the enterprise track is a set of people; and the task in the genomics track explores the trade-offs between different granularities of responses (whole documents, passages, and aspects).

2.1 Test collections

Text retrieval has a long history of using retrieval experiments on test collections to advance the state of the art [4, 8], and TREC continues this tradition. A test collection is an abstraction of an operational retrieval environment that provides a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. Test collections consist of three parts: a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics. We call the result of a retrieval system executing a task on a test collection a run.

2.1.1 Documents

The document set of a test collection should be a sample of the kinds of texts that will be encountered in the operational setting of interest. It is important that the document set reflect the diversity of subject matter, word choice, literary styles, document formats, etc. of the operational setting for the retrieval results to be representative of the performance in the real task. Frequently, this means the document set must be large. The initial TREC test collections contain 2 to 3 gigabytes of text and 500,000 to 1,000,000 documents. While the document sets used in various tracks throughout the years have been smaller and larger depending on the needs of the track and the availability of data, the general trend has been toward ever-larger document sets to enhance the realism of the evaluation tasks. Similarly, the initial TREC document sets consisted mostly of newspaper or newswire articles, but later document sets have included a much broader spectrum of

```
<num> Number:  951
<title> Mutual Funds
<desc> Description:  Blogs about mutual funds performance and trends.
<narr> Narrative:  Ratings from other known sources (Morningstar) or
relative to key performance indicators (KPI) such as inflation, currency
markets and domestic and international vertical market outlooks.  News
about mutual funds, mutual fund managers and investment companies.
Specific recommendations should have supporting evidence or facts linked
from known news or corporate sources.  (Not investment spam or pure,
uninformed conjecture.)
```

Figure 1: A sample TREC 2007 topic from the blog track feed task.

document types (such as recordings of speech, web pages, scientific documents, blog posts, email messages, and business documents). Each document is assigned an unique identifier called the DOCNO. For most document sets, high-level structures within a document are tagged using a mark-up language such as SGML or HTML. In keeping with the spirit of realism, the text is kept as close to the original as possible.

2.1.2 Topics

TREC distinguishes between a statement of information need (the topic) and the data structure that is actually given to a retrieval system (the query). The TREC test collections provide topics to allow a wide range of query construction methods to be tested and also to include a clear statement of what criteria make a document relevant. What is now considered the "standard" format of a TREC topic statement—a topic id, a title, a description, and a narrative—was established in TREC-5 (1996). But topic formats vary in support of the task. The spam track has no topic statement at all, for example, and the topic statements used in the legal track contain much more information as might be available from a negotiated request to produce. An example topic taken from this year's blog track feed task is shown in figure 1.

The different parts of the traditional topic statements allow researchers to investigate the effect of different query lengths on retrieval performance. The description ("desc") field is generally a one sentence description of the topic area, while the narrative ("narr") gives a concise description of what makes a document relevant. The "title" field has served different purposes in different years. In TRECs 1–3 the field is simply a name given to the topic. In later ad hoc collections (ad hoc topics 301 and following), the field consists of up to three words that best describe the topic. For some of the test collections where topics were suggested by queries taken from web search engine logs, the title field contains the original query (sometimes modified to correct spelling or similar errors).

Participants are free to use any method they wish to create queries from the topic statements. TREC distinguishes among two major categories of query construction techniques, automatic methods and manual methods. An automatic method is a means of deriving a query from the topic statement with no manual intervention whatsoever; a manual method is anything else. The definition of manual query construction methods is very broad, ranging from simple tweaks to an automatically derived query, through manual construction of an initial query, to multiple query reformulations based on the document sets retrieved. Since these methods require radically different amounts of (human) effort, care must be taken when comparing manual results to ensure that the runs are truly comparable.

TREC topics are generally constructed specifically for the task they are to be used in. When outside resources such as web search engine logs are used as a source of topics the sample selected for inclusion

in the test set is vetted to insure there is a reasonable match with the document set (i.e., neither too many nor too few relevant documents). Topics developed at NIST are created by the NIST *assessors*, the set of people hired to both create topics and make relevance judgments. Most of the NIST assessors are retired intelligence analysts. The assessors receive track-specific training by NIST staff for both topic development and relevance assessment.

2.1.3 Relevance judgments

The relevance judgments are what turns a set of documents and topics into a test collection. Given a set of relevance judgments, the ad hoc retrieval task is then to retrieve all of the relevant documents and none of the irrelevant documents. TREC usually uses binary relevance judgments—either a document is relevant to the topic or it is not. To define relevance for the assessors, the assessors are told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors are instructed to judge a document as relevant regardless of the number of other documents that contain the same information.

Relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [6]. Furthermore, a set of static, binary relevance judgments makes no provision for the fact that a real user's perception of relevance changes as he or she interacts with the retrieved documents. Despite the idiosyncratic nature of relevance, test collections are useful abstractions because the *comparative* effectiveness of different retrieval methods is stable in the face of changes to the relevance judgments [9].

The relevance judgments in early retrieval test collections were complete. That is, a relevance decision was made for every document in the collection for every topic. The size of the TREC document sets makes complete judgments infeasible. For example, with one million documents and assuming one judgment every 15 seconds (which is *very* fast), it would take approximately 4100 hours to judge a single topic. Thus by necessity TREC collections are created by judging only a subset of the document collection for each topic and then estimating the effectiveness of retrieval results from the judged sample.

The technique most often used in TREC for selecting the sample of documents for the human assessor to judge is pooling [7]. In pooling, the top results from a set of runs are combined to form the pool and only those documents in the pool are judged. Runs are subsequently evaluated assuming that all unpooled (and hence unjudged) documents are not relevant. In more detail, the TREC pooling process proceeds as follows. When participants submit their retrieval runs to NIST, they rank their runs in the order they prefer them to be judged. NIST chooses a number of runs to be merged into the pools, and selects that many runs from each participant respecting the preferred ordering. For each selected run, the top X (frequently $X = 100$) documents per topic are added to the topics' pools. Many documents are retrieved in the top X for more than one run, so the pools are generally much smaller than the theoretical maximum of $X \times$ *the-number-of-selected-runs* documents (usually about 1/3 the maximum size).

The critical factor in pooling is that unjudged documents are assumed to be not relevant when computing traditional evaluation scores. This treatment is a direct result of the original premise of pooling: that by taking top-ranked documents from sufficiently many, diverse retrieval runs, the pool will contain the vast majority of the relevant documents in the document set. If this is true, then the resulting relevance judgment sets will be "essentially complete", and the evaluation scores computed using the judgments will be very close to the scores that would have been computed had complete judgments been available.

Various studies have examined the validity of pooling's premise in practice. Harman [5] and Zobel [10] independently showed that early TREC collections in fact had unjudged documents that would have been

judged relevant had they been in the pools. But, importantly, the distribution of those "missing" relevant documents was highly skewed by topic (a topic that had lots of known relevant documents had more missing relevant), and uniform across runs. Zobel demonstrated that these "approximately complete" judgments produced by pooling were sufficient to fairly compare retrieval runs. Using the leave-out-uniques (LOU) test, he evaluated each run that contributed to the pools using both the official set of relevant documents published for that collection and the set of relevant documents produced by removing the relevant documents uniquely retrieved by the run being evaluated. For the TREC-5 ad hoc collection, he found that using the unique relevant documents increased a run's 11 point average precision score by an average of 0.5 %. The maximum increase for any run was 3.5 %. The average increase for the TREC-3 ad hoc collection was somewhat higher at 2.2 %.

As document sets continue to grow, the proportion of documents contained in standard-sized pools shrinks. At some point, pooling's premise must become invalid. The test collection created in the Robust and HARD tracks in TREC 2005 showed that this point is not at some absolute pool size, but rather when pools are shallow relative to the number of documents in the collection [2]. With shallow pools, the sheer number of documents of a certain type fill up the pools to the exclusion of other types of documents. This produces judgments sets that are biased against runs that retrieve the less popular document type, resulting in an invalid evaluation.

Several recent TREC tracks have investigated new ways of sampling from very large documents sets to obtain judgment sets that support fair evaluations. The primary goal of the terabyte track that was part of TRECs 2004–2006 was to investigate new pooling strategies to build reusable, fair collections at a reasonable cost despite collection size. The new million query track is a successor to the terabyte track in that it has the same goal, but a different approach. The goal in the million query track is to test the hypothesis that a test collection containing very many topics, each of which has a modest number of well-chosen documents judged for it, will be an adequate tool for comparing retrieval techniques. The legal track has used a different sampling strategy still to address the challenging problem of comparing recall-oriented (see below) searches of large document sets for both ranked and unranked result sets.

2.2 Evaluation

Retrieval runs on a test collection can be evaluated in a number of ways. In TREC, ad hoc tasks are evaluated using the `trec_eval` package written by Chris Buckley of Sabir Research [1]. This package reports about 85 different numbers for a run, including *recall* and *precision* at various cut-off levels plus single-valued summary measures that are derived from recall and precision. Precision is the proportion of retrieved documents that are relevant (number-retrieved-and-relevant/number-retrieved), while recall is the proportion of relevant documents that are retrieved (number-retrieved-and-relevant/number-relevant). A cut-off level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list. The `trec_eval` program reports the scores as averages over the set of topics where each topic is equally weighted. (An alternative is to weight each relevant document equally and thus give more weight to topics with more relevant documents. Evaluation of retrieval effectiveness historically weights topics equally since all users are assumed to be equally important.)

Precision reaches its maximal value of 1.0 when only relevant documents are retrieved, and recall reaches its maximal value (also 1.0) when all the relevant documents are retrieved. Note, however, that these theoretical maximum values are not obtainable as an average over a set of topics at a single cut-off level because different topics have different numbers of relevant documents. For example, a topic that has fewer than ten relevant documents will have a precision score at ten documents retrieved less than 1.0 regardless of how

the documents are ranked. Similarly, a topic with more than ten relevant documents must have a recall score at ten documents retrieved less than 1.0. For a single topic, recall and precision at a common cut-off level reflect the same information, namely the number of relevant documents retrieved. At varying cut-off levels, recall and precision tend to be inversely related since retrieving more documents will usually increase recall while degrading precision and vice versa.

Of all the numbers reported by `trec_eval`, the interpolated recall-precision curve and mean average precision (non-interpolated) are the most commonly used measures to describe TREC retrieval results. A recall-precision curve plots precision as a function of recall. Since the actual recall values obtained for a topic depend on the number of relevant documents, the average recall-precision curve for a set of topics must be interpolated to a set of standard recall values. The particular interpolation method used is given in Appendix A, which also defines many of the other evaluation measures reported by `trec_eval`. Recall-precision graphs show the behavior of a retrieval run over the entire recall spectrum.

Mean average precision (MAP) is the single-valued summary measure used when an entire graph is too cumbersome. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved (using zero as the precision for relevant documents that are not retrieved). The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the individual topics in the run. The average precision measure has a recall component in that it reflects the performance of a retrieval run across all relevant documents, and a precision component in that it weights documents retrieved earlier more heavily than documents retrieved later.

The measures described above are traditional retrieval evaluation measures that assume (relatively) complete judgments. As concerns about traditional pooling arose, new measures and new techniques for estimating existing measures given a particular judgment sampling strategy have been investigated. Bpref is a measure that explicitly ignores unjudged documents in the retrieved sets, and thus it can be used when judgments are known to be far from complete [3]. It is defined as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones. The sampling strategies used in the million query and legal tracks have corresponding methods for estimating the value of evaluation measures based on the sampled documents. The track overview paper gives the details of the evaluation methodology used in that track.

3 TREC 2007 Tracks

TREC's track structure began in TREC-3 (1994). The tracks serve several purposes. First, tracks act as incubators for new research areas: the first running of a track often defines what the problem *really* is, and a track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task. The tracks also demonstrate the robustness of core retrieval technology in that the same techniques are frequently appropriate for a variety of tasks. Finally, the tracks make TREC attractive to a broader community by providing tasks that match the research interests of more groups.

Table 1 lists the different tracks that were in each TREC, the number of groups that submitted runs to that track, and the total number of groups that participated in each TREC. The tasks within the tracks offered for a given TREC have diverged as TREC has progressed. This has helped fuel the growth in the number of participants, but has also created a smaller common base of experience among participants since each participant tends to submit runs to a smaller percentage of the tracks.

This section describes the tasks performed in the TREC 2007 tracks. See the track reports later in these proceedings for a more complete description of each track.

Table 1: Number of participants per track and total number of distinct participants in each TREC

	TREC															
Track	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07
Ad Hoc	18	24	26	23	28	31	42	41								
Routing	16	25	25	15	16	21										
Interactive			3	11	2	9	8	7	6	6	6					
Spanish			4	10	7											
Confusion				4	5											
Merging				3	3											
Filtering				4	7	10	12	14	15	19	21					
Chinese					9	12										
NLP					4	2										
Speech						13	10	10	3							
XLingual						13	9	13	16	10	9					
High Prec						5	4									
VLC							7	6								
Query							2	5	6							
QA								20	28	36	34	33	28	33	31	28
Web								17	23	30	23	27	18			
Video										12	19					
Novelty											13	14	14			
Genomics												29	33	41	30	25
HARD												14	16	16		
Robust												16	14	17		
Terabyte													17	19	21	
Enterprise														23	25	20
Spam														13	9	12
Legal															6	14
Blog															16	24
Million Q																11
Participants	22	31	33	36	38	51	56	66	69	87	93	93	103	117	107	95

3.1 The blog track

The blog track first started in TREC 2006. Its purpose is to explore information seeking behavior in the blogosphere, in particular to discover the similarities and differences between blog search and other types of search. The TREC 2007 track contained three tasks, an opinion retrieval task that was the main task in 2006; a subtask of the opinion task in which systems were to classify the kind of the opinion detected (the polarity task); and a blog distillation (also called a feed search) task.

The document set for all tasks was the blog corpus created for the 2006 track and distributed by the University of Glasgow (see `http://ir.dcs.gla.ac.uk/test_collections`). This corpus was collected over a period of 11 weeks from December 2005 through February 2006. It consists of a set of uniquely-identified XML feeds and the corresponding blog posts in HTML. For the opinion and polarity tasks, a "document" in the collection is a single blog post plus all of its associated comments as identified by a Permalink. The collection is a large sample of the blogosphere as it existed in early 2006 that retains all of the gathered material including spam, potentially offensive content, and some non-blogs such as RSS feeds. Specifically, the collection is 148GB of which 88.8GB is permalink documents, 38.6GB is feeds, and 28.8GB is homepages. There are approximately 3.2 million permalink documents.

In the opinion task, systems were to locate blog posts that expressed an opinion about a given target. Targets included people, organizations, locations, product brands, technology types, events, literary works,

etc. For example, three of the test set topics asked for opinions regarding Coretta Scott King, JSTOR, and Barilla brand pasta. Targets were drawn from a log of queries submitted to a commercial blog search engine. The query from the log was used as the title field of the topic statement; the NIST assessor who selected the query created the description and narrative parts of the topic statement to explain how he or she interpreted that query.

The systems' job in the opinion task was to retrieve posts expressing an opinion of the target without regard to the kind (polarity) of the opinion. Nonetheless, the relevance assessors did differentiate among different types of posts during the assessment phase as they had done in 2006. A post could remain unjudged if it was clear from the URL or header that the post contains offensive content. If the content was judged, it was marked with exactly one of: irrelevant (not on-topic), relevant but not opinionated (on-topic but no opinion expressed), relevant with negative opinion, relevant with mixed opinion, or relevant with positive opinion. These judgments supported the polarity subtask. For the polarity subtask, participants' systems labeled each document in the ranking submitted to the opinion task with the predicted judgment (positive, negative, mixed) of that document.

The goal in the blog distillation task was for systems to find blogs (not individual posts) with a principal, recurring interest in the subject matter of the topic. Such technology is needed, for example, when a user wishes to find blogs in an area of interest to follow regularly. The system response for the feed task was a ranked list of up to 100 feed ids (as opposed to permalink ids.) Topic creation and relevance judging for the feed task were performed collaboratively by the participants.

Twenty-four groups total participated in the blog track including 20 in the opinion task, 11 in the polarity subtask, and 9 in the feed task.

To address the question of specific opinion-finding features that are useful for good performance in the opinion task, participants were asked to submit both a topic-relevance-only baseline and an opinion-finding run. Results from this comparison were mixed, with some systems showing a marked increase in effectiveness over good baselines by using opinion-specific features, but others showing serious degradation. Nonetheless, as in the 2006 track the correlation between topic-relevance effectiveness and opinion-finding effectiveness remains very high, indicating that topic-relevance effectiveness is still a dominant factor in good opinion finding.

3.2 The enterprise track

TREC 2007 was the third year of the enterprise track, a track whose goal is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task. Enterprise data generally consists of diverse types such as published reports, intranet web sites, and email, and a goal is to have search systems deal seamlessly with the different data types.

Because of the track's focus on supporting a user of an organization's data, the data set and task abstraction are particularly important. The document set in the first two years of the track was a crawl of the World-Wide Web Consortium web site. This year the document set was instead a crawl of `www.csiro.au`, the web site of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), which is Australia's national science agency. CSIRO employs people known as science communicators who enhance CSIRO's public image and promote the capabilities of CSIRO by managing information and interacting with various constituencies. In the course of their work, science communicators can come upon an area of focus for which no good overview page exists. In such a case a communicator would like to find a set of key pages and people in that area as a first step in creating an overview page (or to stand as a substitute for such a page). This "missing page" problem was the motivation for the two tasks in the track.

In the document search task systems were to retrieve a set of key pages related to the target topic. As in previous years, a key page was defined as an authoritative page that is principally about the target topic. In the search-for-experts task systems returned a ranked list of email addresses representing individuals who are experts in the target topic. Unlike previous years, there was no a priori list of people made available to the systems. Instead, systems were required to mine the document set to find people and decide whether they are experts in a given field. Systems were required to return a list of up to 20 documents in support of the nomination of an expert.

The topics for the track were developed by current CSIRO science communicators, with the same set of topics used for both tasks. Communicators were given a CSIRO query log and asked to develop topics using queries taken from the log or something similar to those. In addition to the query, the communicators were also asked to supply examples of key pages for the area of the query, one or two CSIRO staff members who are experts in that area, and a short description of the information they would consider relevant to include in the overview page.

Systems were provided with the query and description as the official topic statement. Systems could also access the communicator-provided key page examples for relevance feedback experiments. The experts supplied by the science communicators were used as the relevance judgments for the expert search task. Document pools were judged by participants based on the full topic statements to produce the relevance judgments for the document task.

Twenty groups total participated in the enterprise track, with 16 groups participating in the document task and 16 in the expert search task. Comparison between feedback and non-feedback runs in the document task shows that successfully exploiting the example key pages was challenging: only a few teams submitted feedback runs that were more effective than their own non-feedback runs. The results from the expert-finding task suggest that systems are finding only people associated with a given topic rather than actual expertise. For example, systems suggested the science communicators as experts for some topics.

3.3 The genomics track

The goal of genomics track is to provide a forum for evaluation of information access systems in the genomics domain. It was the first TREC track devoted to retrieval within a specific domain, and thus a subgoal of the track is to explore how exploiting domain-specific information improves access. The task in the TREC 2007 track was similar to the passage retrieval task introduced in 2006. In this task systems retrieve excerpts from the documents that are then evaluated at several levels of granularity to explore a variety of facets. The task is motivated by the observation that the best response for a biomedical literature search is frequently a direct answer to the question, but with the answer placed in context and linking to original sources.

The document collection used for 2007 was the same as that used for 2006. This document collection is a set of full-text articles from several biomedical journals that were made available to the track by Highwire Press. The documents retain the full formatting information (in HTML) and include tables, figure captions, and the like. The test set contains about 160,000 documents from 49 journals and is about 12.3 GB of HTML. A passage is defined to be any contiguous span of text that does not include an HTML paragraph token (<p> or <\p>). Systems returned a ranked list of passages in response to a topic where passages were specified by byte offsets from the beginning of the document.

The format of the topic statements differed from that of 2006. The 2007 topics were questions asking for lists of specific entities such as drugs or mutations or symptoms. The questions were solicited from practicing biologists and represent actual information needs. The test set contained 36 questions.

Relevance judgments were made by domain experts. The judgment process involved several steps to enable system responses to be evaluated at different levels of granularity. Passages from different runs were pooled, using the maximum extent of a passage as the unit for pooling. (The maximum extent of a passage is the contiguous span between paragraph tags that contains that passage, assuming a virtual paragraph tag at the beginning and end of each document.) Judges decided whether a maximum span was relevant (contained an answer to the question), and, if so, marked the actual extent of the answer in the maximum span. In addition, the assessor listed the entities of the target type contained within the maximum span. A maximum span could contain multiple answer passages; the same entity could be covered by multiple answer passages and a single answer passage could contain multiple entities.

Using these relevance judgments, runs were then evaluated at the document, passage, and aspect (entity) levels. A document is considered relevant if it contains a relevant passage, and it is considered retrieved if any of its passages are retrieved. The document level evaluation was a traditional ad hoc retrieval task (when all subsequent retrievals of a document after the first were ignored). Passage- and aspect-level evaluation was based on the corresponding judgments. Aspect-level evaluation is a measure of the diversity of the retrieved set in that it rewards systems that are able to find more different aspects. Passage-level evaluation is a measure of how well systems are able to find the particular information within a document that answers the question.

The genomics track had 25 participants. Results from the track showed that effectiveness as measured at the three different granularities was highly correlated. As in the blog track, this suggests that basic recognition of topic relevance remains a dominating factor for effective performance in each of these tasks.

3.4 The legal track

The legal track was started in 2006 to focus specifically on the problem of e-discovery, the effective production of digital or digitized documents as evidence in litigation. Since the legal community is familiar with the idea of searching using Boolean expressions of keywords, Boolean search is used as a baseline in the track. The goal of the track is thus to evaluate the effectiveness of Boolean and other search technologies for the e-discovery problem.

The TREC 2007 track contained three tasks, the main task, an interactive task, and a relevance feedback task. The document set used for all tasks was the IIT Complex Document Information Processing collection, which was also the corpus used in the 2006 track. This collection consists of approximately seven million documents drawn from the Legacy Tobacco Document Library hosted by the University of California, San Francisco. These documents were made public during various legal cases involving US tobacco companies and contain a wide variety of document genres typical of large enterprise environments. A document in the collection consists of the optical character recognition (OCR) output of a scanned original plus metadata.

The main task was an ad hoc search task using as topics a set of hypothetical requests for production of documents. The production requests were developed for the track by lawyers and were designed to simulate the kinds of requests used in current practice. Each production request includes a broad complaint that lays out the background for several requests and one specific request for production of documents. The topic statement also includes a negotiated Boolean query for each specific request. Stephen Tomlinson of Open Text, a track coordinator, ran the negotiated Boolean queries to produce the task's reference run. Participants could use the negotiated Boolean query, the set of documents that matched the Boolean query, and the size of the retrieved set of the Boolean query (B) in any way (including ignoring them completely) for their submitted runs. For each topic systems returned a ranked list of up 25 000 documents (or up to B documents if B was larger than 25 000).

Because of the size of the document collection and the legal community's interest in being able to evaluate the effectiveness of the (unranked) Boolean run, special pools were built from the submitted runs to support Estimated-Recall-at-B as the evaluation measure. The pooling method sampled a total of approximately 500 documents from the set of submitted runs respecting the property that documents at ranks closer to one had a higher probability of being selected for inclusion in the pools. (See the track overview paper for more details.) Note that it is not currently known how reusable the resulting collection is (that is, whether the judgments can be usefully exploited to evaluate runs that did not contribute to the pools). The relevance assessments were made by legal professionals (mostly law students) who followed the legal community's typical work practices.

Iterative search methods generally offer increased effectiveness as compared to the single running of a static query, even if that query is the result of prior negotiation. The feedback and interactive search tasks were introduced into the legal track to explore the level of performance obtainable by iterative search methods in e-discovery and to investigate how best to evaluate those techniques. Both tasks used a subset of the topics from the TREC 2006 legal track.

The goal in the interactive task was for a user to find as many relevant documents as possible for a topic while actively engaging with the retrieval system. Twelve topics were available for this task, ranked in priority order. Participants in the interactive task could do as many of the twelve topics as desired, but were required to perform them in priority order. Submissions consisted of up to 100 documents per topic, which were scored using a utility measure (gaining one point for each relevant document retrieved and losing a half point for each nonrelevant retrieved).

For the relevance feedback task, systems re-ran the TREC 2006 topics exploiting the relevance judgments produced as a result of the TREC 2006 track. Documents that had been judged in 2006 were removed from the submissions ("residual collection" evaluation) and new pools were formed for 10 topics (a subset of the 12 topics used in the interactive task)[1]. The main evaluation measure used in the task was again Estimated-Recall-at-(residual)-B.

A total of 14 groups participated in the legal track: 12 in the main task, 3 in the interactive task, and 3 in the relevance feedback task. Results from the TREC 2007 tasks confirm results from the TREC 2006 track with respect to the Boolean run. Collectively the runs produced by track participants retrieve many relevant documents not retrieved by the negotiated Boolean queries of the reference run, but the average effectiveness of the reference Boolean run is at least as great as the average effectiveness of the other individual runs (with respect to Estimated-Recall-at-B). In other words, all of the runs, including the reference Boolean run, have significant room for improvement with respect to consistently obtaining high recall.

3.5 The million query track

The million query track was a new track in TREC 2007. One of the main goals of the track was to investigate a specific retrieval evaluation hypothesis: that a test collections built using many topics with few, shallow judgments may be a better evaluation tool than a test collection built from fewer topics with relatively thorough judgments. The track also provided an opportunity for participants to explore ad hoc retrieval on a large document set.

The retrieval task of the track was an ad hoc search task over the GOV2 document set. GOV2 is a collection of web pages from within the `.gov` domain spidered in early 2004. The collection contains about 25 million documents and is available from the University of Glasgow (see `http://ir.dcs.`

[1]The new judgments made for the 2007 tasks were created by a different assessor from the one who judged the topic in TREC 2006.

gla.ac.uk/test_collections). The topics for the track were taken from a web search engine log and consisted only of the equivalent of the standard TREC topic statement's title field (some of these topics later had standard topic statements developed for them during the assessing phase). The test set consisted of 10,000 queries, including the title field from some of the topics that had been used in previous years' terabyte tracks.

Relevance judging was performed by both NIST assessors and track participants. The judging procedure was as follows:

1. The assessment system presented the judge with 5 queries randomly selected from the test set.
2. The judge selected one of the queries; the others were returned to the query pool.
3. The judge wrote a description and narrative for this query, thus creating a standard TREC topic statement.
4. The system presented a GOV2 document to the judge and obtained a 3-way judgment (highly relevant, relevant, not relevant) for it.
5. The process continued until at least 40 documents were judged. The judge could continue past 40 documents if he or she wanted to.

The documents to be judged were selected by one of two different sampling methods, the minimal test collection method and the statistical evaluation method, each of which supports a particular evaluation strategy. The details of the sampling and corresponding evaluation methods are given in the track overview paper in these proceedings. The target was to have half the queries that were judged have 20 documents selected by both methods, a quarter of the queries have 40 documents selected by the minimal test collection sampling method, and the remaining queries have 40 documents selected by the statistical evaluation method. Approximately 1800 queries were judged, with a small set receiving judgments from multiple people.

The judgments gathered in this way allow evaluation using the appropriate measure(s) associated with the selection method. The use of the terabyte topics allows runs to be evaluated over those topics using `trec_eval` and the standard NIST-produced relevance judgments created in the terabyte track as a third evaluation strategy. The 24 runs submitted by 11 groups were each evaluated using the three evaluation strategies in turn. The three different strategies agreed with one another with respect to "big picture" results: all three strategies found the same three clusters of systems with similar effectiveness. More fine-grained comparisons differed across strategies, though, in that rankings of systems within clusters varied depending on the evaluation strategy used. The rankings produced by the two sampling-based evaluation methods were more similar to each other than either was to the ranking produced by evaluation over the terabyte topics.

3.6 The question answering (QA) track

The goal of the question answering track is to develop systems that return actual answers, as opposed to ranked lists of documents, in response to a question. The 2007 track contained two tasks, the main task that was a series task similar to the task used since 2004, and a complex interactive QA (ciQA) task introduced in 2006.

The questions in the main task were organized into a set of series. A series consisted of a number of "factoid" (questions with fact-based, short answers) and list questions that each related to a common, given target. The final question in a series was an explicit "Other" question, which systems were to answer by

retrieving information pertaining to the target that had not been covered by earlier questions in the series. Answers were required to be supported by a document from the corpus used in the track.

The 2007 main task differed from the task in earlier years in that the corpus consisted of both newswire documents (the AQUAINT-2 collection) and blog documents (the same corpus as was used in the blog track). Introducing blogs into the track created two significant new challenges for QA systems. First, since language use in blogs can be much more informal than in newswire, systems were required to handle language that is not well-formed. Second, blog data also contains discourse structures that are less formal and reliable than newswire, so systems had to do more vetting of candidate responses to determine if those responses were indeed answers.

Despite the introduction of the blog data, which was expected to increase the difficulty of the QA task, individual component scores for the best systems were greater in 2007, after having generally declined each year since TREC 2004. While it is possible that the questions in the 2007 test set are intrinsically easier than previous years, no procedural changes in the way questions were formed were instituted, so large changes in difficulty are not likely.

The ciQA task was introduced in TREC 2006 and is a blend of the TREC 2005 relationship QA task and the TREC 2005 HARD track. The goal of the task is to extend systems' abilities to answer more complex information needs than those covered in the main task and to provide a limited form of interaction with the user in a QA setting.

As in 2006, the questions used in the task contained two parts, a specific question derived from templates of relationship question types, and a narrative that provided more explanation for the specific question. The system response to a question was a ranked list of information "nuggets" supported by AQUAINT documents (the blog corpus was not used in the ciQA task), where each nugget provides evidence for the relationship in question.

The interaction was accomplished using the NIST assessor as the surrogate user and web forms to implement the interface. Unlike 2006, the forms were hosted at the individual participants' home site, so any type of web-based QA system could be used in the task. For each topic, the assessor was given a list of URLs, one URL per participating run. The lists of URLs for different topics were sorted differently, and assessors processed each list in the order given, to control for presentation order effects. Assessors clicked on a URL to begin an interaction and had a maximum of five minutes to finish the task for that pair of run/topic. Participants were responsible for instrumenting the application to capture the results of the interaction.

The protocol for the ciQA task had participants submit initial runs prior to the interaction, perform the interaction, and then submit final runs that (presumably) made use of the information gathered in the interaction. Retrieval results were scored using Pyramid nuggets F-score. In addition, an exit questionnaire gathered data on the assessors' perceptions of the interactions.

Results from the ciQA task showed that, unlike in TREC 2006, most runs were more effective than a sentence-retrieval baseline run. However, many interactions degraded effectiveness; that is, the final run score was less than the corresponding initial run's score. Analysis of the data collected from the exit questionnaire suggested a possible contributing factor for the decrease in effectiveness through interaction: NIST assessors are unusual users in that they already know a lot about the topic, yet the typical users assumed by many participating systems were naive users searching for basic information. Future instantiations of interactive tasks will need to take this mismatch into consideration.

A total of 28 groups participated in the QA track. The main task had 21 participants and the ciQA task had 7 participants.

3.7 The spam track

The spam track was first run in TREC 2005. The goal of the track is to evaluate how well systems are able to separate spam and ham (non-spam) when given an email sequence. The TREC 2007 track repeated the three 2006 tasks using new data. The tasks all involved classifying email messages as ham or spam, differing in the amount and frequency of the feedback the system received.

For each task the track used a test jig that implements a simple interface between the evaluation infrastructure on the one hand and a participant's classifier on the other. The jig takes an email stream, a set of ham/spam judgments, and a classifier, and runs the classifier on the stream reporting the evaluation results of that run based on the judgments. In the main on-line filtering task, the classifier receives the correct designation for a message as soon as it classifies the message (this represents ideal user feedback). In the delayed feedback extension to the task, the classifier might eventually receive the correct designation for a message, but the designation for a given message m may come after some number of intervening messages that must be classified before the feedback for m is received, or the feedback may never come at all. In the partial feedback extension to the task, feedback is provided only for messages sent to a subset of the users of a mail server, though the filter is expected to filter messages to all users. In the active learning task, the classifier must explicitly request the correct designation for a document, and may do so for only a given number N of messages.

The track used both a private email stream and a public email stream. Participants ran their own filters on the public corpora using the jig and submitted the evaluation output to NIST. For the private corpora, participants submitted their filters to NIST. NIST passed the filters onto the University of Waterloo after stripping all identification of which filters came from which participant. The University of Waterloo used the jig to run the filters on the private stream and returned the evaluation results to NIST, who then forwarded the evaluation results to the appropriate participant.

Twelve groups participated in the spam track. As in previous years of the track, the general effectiveness of the track's filters has improved relative to the then-current state-of-the-art. Comparison among the different types of training show that both delayed and partial feedback degrade filter effectiveness with respect to ideal feedback, but longer delay periods do not appear to cause more deterioration than shorter delay periods.

4 The Future

TREC 2007 contained a brainstorming session designed to get feedback as to what research areas individuals in the TREC community were personally interested in. In the spirit of true brainstorming, we asked for any ideas without initial filtering by feasibility concerns such as data availability or privacy issues. The session was lively with approximately 40 ideas suggested before discussion was stopped due to time constraints. Enough people expressed interest in three broad areas for those ideas to be further explored informally over a group lunch at the conference and discussion lists after the conference. The goal of the discussions was to formulate a proposal for a TREC track in the area to begin in TREC 2009. The three areas included:

informal text: a track to focus on data access tasks within social media contexts such as instant messaging systems or social tagging;

scientific literature: a track to focus on providing access to the scientific literature more broadly than within a single topic domain as in the genomics track; and

user interaction: a reprise of the TREC interactive track where the focus is on understanding how best to support humans in the search process.

There are five confirmed tracks for TREC 2008. The blog, enterprise, legal, and million query tracks will continue. A new track to examine the effectiveness of relevance feedback across different retrieval models and under different conditions (such as amount of relevance data) will begin. The question answering track will move to a new NIST evaluation conference called the Text Analysis Conference (TAC), see `http://www.nist.gov/tac`. The genomics and spam tracks are ending as TREC tracks, though tasks similar to those investigated in these tracks are expected to appear in other venues.

Acknowledgements

The track summaries in section 3 are based on the track overview papers authored by the track coordinators. My thanks to the coordinators who make the variety of different tasks addressed in TREC possible.

References

[1] Chris Buckley. trec_eval IR evaluation package. Available from `http://trec.nist.gov/trec_eval/`.

[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10:491–508, 2007.

[3] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004.

[4] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[5] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, October 1996. NIST Special Publication 500-236.

[6] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[7] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[8] Karen Sparck Jones. *Information Retrieval Experiment.* Butterworths, London, 1981.

[9] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[10] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998. ACM Press, New York.

Table 2: Organizations participating in TREC 2007

Arizona State University	RMIT University
Beijing U. of Posts & Telecommunications	The Robert Gordon University
Carnegie Mellon University	Saarland University
Carnegie Mellon University & U. Karlsruhe	Sabir Research, Inc
Chinese Academy of Sciences (2 groups)	Shanghai Jiao Tong University (2 groups)
Concordia University (2 groups)	South China University of Technology
CRM114 Team	St. Petersburg State U. & INRIA
CSIRO ICT Centre	SUNY Albany
Dalhousie University	SUNY Buffalo
Dalian University of Technology	Technical University Berlin
Dartmouth College	TNO, Twente University & EMC
Drexel University	Tokyo Institute of Technology
EffectiveSoft Ltd	Tsinghua University
European Bioinformatics Institute	Tufts University
Exegy Inc.	Twente University
Fitchburg State College	University of Alaska Fairbanks
Fondazione Ugo Bordoni, Italian National Research Council, & U. Roma 'Tor Vergata'	University of Alicante
Fudan University	University of Amsterdam (2 groups)
Heilongjiang Institute of Technology	University of Arkansas at Little Rock
IBM Cairo	University of Colorado School of Medicine
IBM Research Lab, Haifa	University of Glasgow
Indian Institute of Technology, Delhi	University and Hospitals of Geneva
Illinois Institute of Technology	University of Illinois at Chicago (2 groups)
Indiana University	University of Illinois at Urbana-Champaign
International Institute of Information Technology	University of Iowa (2 groups)
Jozef Stefan Institute	University of Lethbridge
Kobe University (2 groups)	University of Maryland, College Park
Kyoto University	University of Massachusetts
Language Computer Corporation	The University of Melbourne (2 groups)
Long Island University	University of Missouri at Kansas City
Lymba Corporation	University of Neuchatel
Massachusetts Institute of Technology	University of North Carolina
Michigan State University	Università di Roma 'La Sapienza'
The MITRE Corporation	University of Strathclyde
National Library of Medicine	University of Texas, Austin
National Taiwan University	University of Washington
National University of Defense Technology	University of Waterloo (2 groups)
Northeastern University	Ursinus College
Oregon Health & Science University	Weill Cornell Medical College
Open Text Corporation	Wuhan University
The Open University	York University
Peking University	Zhejiang University
Queens College, CUNY	

Overview of the TREC-2007 Blog Track

Craig Macdonald, Iadh Ounis
University of Glasgow
Glasgow, UK
{craigm,ounis}@dcs.gla.ac.uk

Ian Soboroff
NIST
Gaithersburg, MD, USA
ian.soboroff@nist.gov

1. INTRODUCTION

The goal of the Blog track is to explore the information seeking behaviour in the blogosphere. It aims to create the required infrastructure to facilitate research into the blogosphere and to study retrieval from blogs and other related applied tasks. The track was introduced in 2006 with a main opinion finding task and an open task, which allowed participants the opportunity to influence the determination of a suitable second task for 2007 on other aspects of blogs besides their opinionated nature. As a result, we have created the first blog test collection, namely the TREC Blog06 collection, for adhoc retrieval and opinion finding. Further background information on the Blog track can be found in the 2006 track overview [2].

TREC 2007 has continued using the Blog06 collection, and saw the addition of a new main task and a new subtask, namely a blog distillation (feed search) task and an opinion polarity subtask respectively, along with a second year of the opinion finding task. NIST developed the topics and relevance judgments for the opinion finding task, and its polarity subtask. For the blog distillation task, the participating groups created the topics and the associated relevance judgments. This second year of the track has seen an increased participation compared to 2006, with 20 groups submitting runs to the opinion finding task, 11 groups submitting runs to the polarity subtask, and 9 groups submitting runs to the blog distillation task. This paper provides an overview of each task, summarises the obtained results and draws conclusions for the future.

The remainder of this paper is structured as follows. Section 2 provides a short description of the used Blog06 collection. Section 3 describes the opinion finding task and its polarity subtask, providing an overview of the submitted runs, as well as a summary of the main used techniques by the participants. Section 4 describes the newly created blog distillation (feed search) task, and summarises the results of the runs and the main approaches deployed by the participating groups. We provide concluding remarks in Section 5.

2. THE BLOG06 TEST COLLECTION

All tasks in the TREC 2007 Blog track use the Blog06 collection, representing a large sample crawled from the blogosphere over an eleven week period from December 6, 2005 until February 21, 2006. The collection is 148GB in size, with three main components consisting of 38.6GB of XML feeds (i.e. the blog), 88.8GB of permalink documents (i.e. a single blog post and all its associated comments) and 28.8GB of HTML homepages (i.e. the main entry to the blog). In order to ensure that the Blog track experiments are conducted in a realistic and representative setting, the collection also includes spam, non-English documents, and some non-blogs documents such as RSS feeds.

The number of permalink documents in the collection is over 3.2 million, while the number of feeds is over 100,000 blogs. The permalink documents are used as a retrieval unit for the opinion finding task and its associated polarity subtask. For the blog distillation task, the feed documents are used as the retrieval unit. The collection has been distributed by the University of Glasgow since March 2006. Further information on the collection and how it was created can be found in [1].

3. OPINION FINDING TASK

Many blogs are created by their authors as a mechanism for self-expression. Extremely-accessible blog software has facilitated the act of blogging to a wide-ranging audience, their blogs reflecting their opinions, philosophies and emotions. The opinion finding task is an articulation of a user search task, where the information need seems to be of an opinion, or perspective-finding nature, rather than fact-finding. While no explicit scenario was associated with the opinion retrieval task, it aims to uncover the public sentiment towards a given entity (the "target"), and hence it can naturally be associated with settings such as tracking consumer-generated content, brand monitoring, and, more generally, media analysis. This is the second running of the opinion finding task in the Blog track. This year, it was the most popular task of the track, with 20 participating groups.

3.1 Topics and Relevance Judgments

Similar to TREC 2006, the opinion retrieval task involved locating blog posts that express an opinion about a given target [2]. The target can be a "traditional" named entity, e.g. a name of a person, location, or organisation, but also a concept (such as a type of technology), a product name, or an event. The task can be summarised as *What do people think about X, X* being a target. The topic of the post is not required to be the same as the target, but an opinion about the target had to be present in the post or one of the comments to the post, as identified by the permalink.

Topics used in the opinion finding task follow the familiar title, description, and narrative structure, as used in topics in other TREC test collections. 50 topics were again selected by NIST from a larger query log obtained from a commercial blog search engine. The topics were created by NIST using the same methodology as last year, namely selecting queries from the query log, and building topics around those queries [2]. An example of a TREC 2007 topic is included in Figure 1.

3.2 Pooling and Assessment Procedure

Participants could create queries manually or automatically from the 50 provided topics. They were allowed to submit up to six runs, including a compulsory automatic run using the title field of

```
<top>
  <num> Number: 930 </num>

  <title> ikea  </title>

  <desc> Description:
 Find opinions on Ikea or its products.
</desc>
  <narr> Narrative:
  Recommendations to shop at Ikea are
relevant opinions. Recommendations of
Ikea products are relevant opinions.
Pictures on an Ikea-related site that
are not related to the store or its
products are not relevant.
</narr>
</top>
```

Figure 1: Blog track 2007, opinion retrieval task, topic 930.

the topics, and another compulsory automatic run, using the title field of the topics, but with all opinion-finding features turned off. The latter was required to draw further conclusions on the extent to which a strong topic relevance baseline is required for an effective opinion retrieval system. It also helps to draw conclusions on the real effectiveness of the specifically used opinion finding approaches.

As mentioned in Section 2, for the purposes of the opinion finding task, the document retrieval unit in the collection is a single blog post plus all of its associated comments as identified by a permalink. However, participants were free to use any of the other Blog06 collection components for retrieval such as the XML feeds and/or the HTML homepages.

Overall, 20 groups participated in the opinion finding task, submitting 104 runs, including 98 automatic runs and 6 manual runs. The participants were asked to prioritise runs, in order to define which of their runs would be pooled. Like in TREC 2006, the guidelines of the Blog track encouraged participants to submit manual runs to improve the quality of the test collection. Each submitted run consisted of the top 1,000 opinionated documents (permalinks) for each topic. NIST formed the pools from the submitted runs using the three highest-priority runs per group, pooled to depth 80. In case of ties, the manual runs were preferred over the automatic runs, and among the automatic title-only tied runs, the compulsory ones were preferred.

NIST organised the relevance assessments for the opinion finding task, using the same assessment procedure defined in 2006 [2], with some further tightening up of the guidelines given to the assessors. In particular, the assessment procedure had two levels. The first level assesses whether a given blog post, i.e. a permalink, contains information about the target and is therefore relevant. The second level assesses the opinionated nature of the blog post, if it was deemed relevant in the first assessment level. Given a topic and a blog post, assessors were asked to judge the content of the blog posts. The following scale was used for the assessment:

0 *Not relevant.* The post and its comments were examined, and do not contain any information about the target, or refers to it only in passing.

1 *Relevant.* The post or its comments contain information about the target, but do not express an opinion towards it. To be assessed as "Relevant", the information given about the target should be substantial enough to be included in a report compiled about this entity.

Relevance Scale	Label	Nbr. of Documents	%
Not Relevant	0	42434	77.7%
Adhoc-Relevant	1	5187	9.5%
Negative Opinionated	2	1844	3.4%
Mixed Opinionated	3	2196	4.0%
Positive Opinionated	4	2960	5.4%
(Total)	-	54621	100%

Table 1: Relevance assessments of documents in the pool.

If the post or its comments are not only on target, but also contain an explicit expression of opinion or sentiment towards the target, showing some personal attitude of the writer(s), then the document had to be judged using the three labels below:

2 *Negatively opinionated.* Contains an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), and the opinion expressed is explicitly negative about, or against, the target.

3 *Mixed.* Same as (2), but contains both positive and negative opinions.

4 *Positively opinionated.* Same as (2), but the opinion expressed is explicitly positive about, or supporting, the target.

Posts that are opinionated, but for which the opinion expressed is ambiguous, mixed, or unclear, were judged simply as "mixed" (3 in the scale).

Table 1 shows a breakdown of the relevance assessment of the pooled documents, using the assessment procedure described above. About 78% of the pooled documents were judged as irrelevant. Moreover, there were roughly an equal percentage of negative and mixed opinionated documents, but slightly more positive opinionated documents, suggesting that overall, the bloggers had more positive opinions about the topics tackled by the TREC 2007 opinion finding topics set. Figure 2 shows the number of relevant positive and negative opinionated documents for each topic. Topic "northenvoice" (914) or topic "mashup camp" (925) have only relevant positive opinionated documents in the pool, whereas topic "censure" (943) or topic "challenger" (923) have more negative than positive opinionated documents in the pool, perhaps illustrating the nature of these tackled topics.

3.3 Overview of Results

Since the opinion finding task is an adhoc-like retrieval task, the primary measure for evaluating the retrieval performance of the participating groups is the mean average precision (MAP). Other metrics used for the opinion finding task are R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10).

Table 2 provides the average best, median and worst MAP measures for each topic, across all submitted 104 runs. While these are not "real" runs, they provide a summary of how well the spread of participating systems is performing. In particular, it is of interest to note that the retrieval performances of the participating groups in TREC 2007 are markedly higher than those reported in TREC 2006 on the same task. For example, the median MAP measure of the submitted runs for the opinion finding task has increased from 0.1059 in TREC 2006 [2] to 0.2416 in TREC 2007. Further investigation is required in order to conclude whether this is due to the TREC 2007 topics being easier than those used in TREC 2006, or if

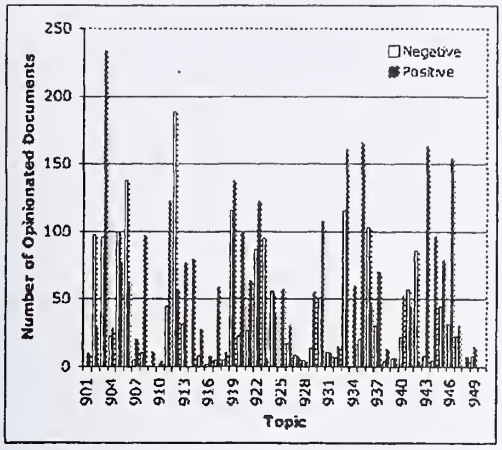


Figure 2: Number of positive and negative opinionated documents per topic in the pool.

	Opinion-finding MAP	Topic-relevance MAP
Best	0.5182	0.6382
Median	0.2416	0.3340
Worst	4.2e-05	0.0001

Table 2: Best, median and worst MAP measures for the 104 submitted runs to the opinion finding task.

the increase is due to the use of more effective retrieval approaches by the participants.

Table 3 shows the best-scoring opinion-finding title-only automatic run for each group in terms of MAP, and sorted in decreasing order. R-Prec, bPref and P@10 measures are also reported. Table 4 shows the best opinion-finding run from each group, in terms of MAP, regardless of the topic length used.

Each participating group was required to submit a compulsory automatic run, using only the title field of the topics, with all opinion finding features of the retrieval system turned off (i.e. a *topic-relevance baseline* run). The idea is to have a better understanding of the actual effectiveness of the opinion detection approaches deployed by the participating groups, allowing to draw conclusions as to whether the used opinion finding techniques actually help retrieving opinionated documents. Table 5 shows the best baseline run from each group, in terms of opinion-finding MAP. Comparing Tables 3 and 5, it is interesting to note that only one of the top five performing opinion finding runs was actually a topic-relevance baseline run. In particular, out of the 5 best opinion-finding performing runs in Table 3, only run uams07topic from the University of Amsterdam was a topic-relevance run.

In order to assess which opinion finding features and approaches deployed by the participating groups have actually worked, we compare the performance of the best performing opinion finding run of each group to its best submitted topic-relevance baseline. A relative increase in performance indicates that the used opinion finding features were useful. A relative decrease in performance indicates that the deployed opinion finding features did not help in retrieval. Table 6 shows the improvements of the best submitted compulsory automatic title-only runs over the baselines. Note that the best performing group on the opinion finding task, namely the UIC group, did not officially submit a baseline run, making it difficult to conclude on the success of their deployed opinion finding features. It

Evaluation Measure	ρ	τ
MAP	0.9778	0.8813
R-Prec	0.9677	0.8518
bPref	0.8118	0.9448
P@10	0.8032	0.9366

Table 8: Correlation of system rankings between opinion-finding performance measures and topic-relevance performance measures. Both Spearman's Correlation Coefficient (ρ) and Kendall's Tau (τ) are reported.

is interesting to note that the best opinion finding run by the University of Amsterdam has decreased the performance of the their strongly performing uams07topic topic-relevance baseline by over 57%. On the other hand, the opinion finding features used by the University of Glasgow, Indiana University, and the University of Arkansas at Little Rock seem to be helpful, improving their performance on the task by 15.8%, 14% and 13.9%, respectively, depsite their good performing baselines.

Given the two levels assessment procedure, it is possible to evaluate the submitted runs in a classical adhoc fashion, i.e. based on the relevance of their returned documents (judged 1 or above, as described in Section 3.2 above). Table 7 reports the best run from each group in terms of topic-relevance, regardless of the topic length.

Moreover, Table 8 reports the Spearman's ρ and Kendall's τ correlation coefficients between opinion finding and topic relevance measures. The overall rankings of systems on both opinion-finding and topic relevance measures are very similar, as stressed by the obtained high correlations. A similar finding was observed in TREC 2006 [2], suggesting again that good performances on the opinion finding task are strongly dominated by good performances on the underlying topic-relevance task. Figure 3(a) shows a scatter plot of opinion-finding MAP against topic-relevance MAP, which confirms that the correlation is very high.

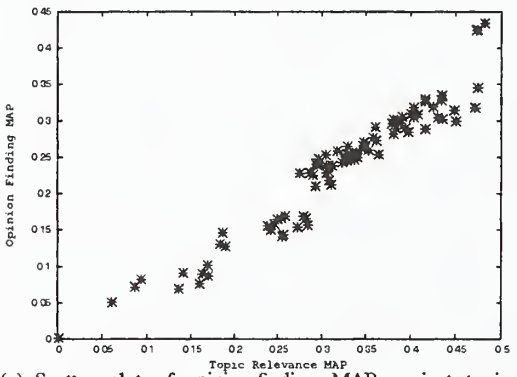
Finally, we report on the extent to which the 17,958 presumed splog feeds and their associated 509,137 spam posts, which were injected into the Blog06 collection during its creation have infiltrated the pool. Table 9 provides details on the number of presumed splog posts which infiltrated each element of the relevance scale. In total, 7,086 assumed splog documents were pooled, less than 1.5% of the splog posts in the collection. Moreover, there was a roughly equal number of relevant only and opinionated splog posts, though those that were opinionated were mostly positive. Figure 4 shows the average number of spam documents retrieved by all 104 submitted runs for each topic, in decreasing order.

Noticeably, unlike in last year's TREC 2006 topics set where the most spammed topics where about health, we note that topic 915 (namely "allianz") had by far the largest number of splog posts retrieved in the submitted runs (average 703 documents per run). Topic "grammys" (936) and topic "teri hatcher" also had a substantial number of splog posts retrieved (average 466 and 309 documents per run, respectively). These are widely popular topics, which might be prone to being spammed. Similar to TREC 2006 though, topics which retrieved far fewer spam documents, were concerning people not featuring in the tabloid news as often, such as topics 924 and 904: "mark driscoll" (23 documents) and "alterman" (9 documents), respectively.

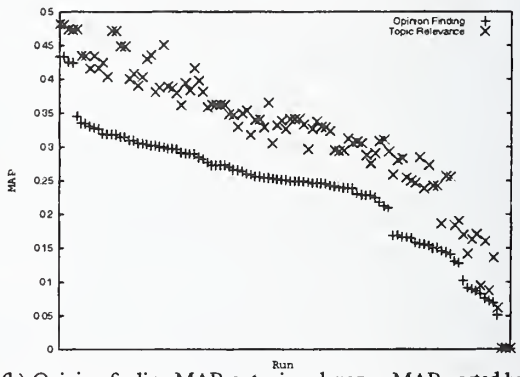
Next, we examined how the participating systems had been affected by spam documents. Table 10 shows the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), and for all the retrieved documents (Spam@all). The table also reports BadMAP, which is the Mean Average Precision when the pre-

Group	Run	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	**0.4341**	**0.4529**	**0.4724**	**0.690**
UAmsterdam (deRijke)	uams07topic	0.3453	0.3872	0.3953	0.562
UGlasgow (Ounis)	uogBOPFProxW	0.3264	0.3657	0.3497	0.552
DalianU (Yang)	DUTRun2	0.3190	0.3671	0.3686	0.600
FudanU (Wu)	FDUTOSVMSem	0.3143	0.3465	0.3499	0.460
CAS (Liu)	Relevant	0.3041	0.3600	0.3779	0.446
UArkansas Littlerock (Bayrak)	UALR07Blog1U	0.2911	0.3263	0.3134	0.580
IndianaU (Yang)	oqsnr2opt	0.2894	0.3572	0.3419	0.532
UNeuchatel (Savoy)	UniNEblog1	0.2770	0.3353	0.3074	0.492
FIU (Netlab team)	FlUbPL2	0.2728	0.3204	0.2925	0.454
UWaterloo (Olga)	UWopinion3	0.2631	0.3344	0.2980	0.496
Zhejiangu (Qiu)	EAGLE1	0.2561	0.3159	0.2867	0.428
CAS (NLPR-1ACAS)	NLPRPST	0.2542	0.3168	0.2945	0.462
BUPT (Weiran)	prisOpnBasic	0.2466	0.3018	0.2835	0.456
KobeU (Eguchi)	KobePrMIR01	0.246	0.3011	0.2744	0.440
NTU (Chen)	NTUAutoOp	0.2282	0.2614	0.2577	0.464
KobeU (Seki)	Ku	0.1689	0.2417	0.2190	0.254
RGU (Mukras)	rgu0	0.1686	0.2266	0.2163	0.288
UBuffalo (Ruiz)	UB2	0.1013	0.1297	0.1238	0.144
Wuhan (Lu)	NOOPWHU1	0.0011	0.0071	0.0072	0.008

Table 3: Opinion finding results: the automatic title-only run from each of 20 groups with the best MAP, sorted by MAP. The best in each column is highlighted.



(a) Scatter plot of opinion-finding MAP against topic-relevance MAP.



(b) Opinion finding MAP vs topic-relevance MAP, sorted by opinion-finding MAP.

Figure 3: Figures examining opinion-finding and topic-relevance MAP.

Relevance Scale	Nbr. of Splog Documents
Not Relevant	6357
Adhoc-Relevant	361
Negative Opinionated	78
Mixed Opinionated	98
Positive Opinionated	192
(Total)	7086

Table 9: Occurrences of presumed splog documents in the opinion finding task pool.

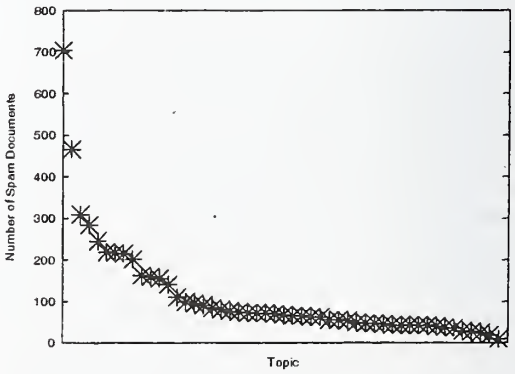


Figure 4: Distribution of number of spam documents retrieved per topic.

Group	Run	Automatic	Fields	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	yes	T	**0.4341**	**0.4529**	**0.4724**	**0.690**
UAmsterdam (deRijke)	uams07topic	yes	T	0.3453	0.3872	0.3953	0.562
IndianaU (Yang)	oqlr2fopt	yes	TDN	0.3350	0.3925	0.378	0.576
UGlasgow (Ounis)	uogBOPFProxW	yes	T	0.3264	0.3657	0.3497	0.552
DalianU (Yang)	DUTRun2	yes	T	0.3190	0.3671	0.3686	0.600
FudanU (Wu)	FDUTisdOpSVM	yes	T	0.3179	0.3467	0.3501	0.454
FIU (Netlab team)	FIUDDPH	yes	TD	0.3053	0.3498	0.3475	0.492
UNeuchatel (Savoy)	UniNEblog3	yes	TD	0.3049	0.3438	0.3266	0.516
CAS (Liu)	Relevant	yes	T	0.3041	0.3600	0.3779	0.446
UArkansas Littlerock (Bayrak)	UALR07BlogIU	yes	T	0.2911	0.3263	0.3134	0.580
UWaterloo (Olga)	UWopinion3	yes	T	0.2631	0.3344	0.298	0.496
CAS (NLPR-IACAS)	NLPRPTD2	yes	TD	0.2587	0.3088	0.2956	0.456
Zhejiangu (Qiu)	EAGLE1	yes	T	0.2561	0.3159	0.2867	0.428
BUPT (Weiran)	prisOpnBasic	yes	T	0.2466	0.3018	0.2835	0.456
KobeU (Eguchi)	KobePrMIR01	yes	T	0.2460	0.3011	0.2744	0.440
NTU (Chen)	NTUManualOp	no	T	0.2393	0.2659	0.2749	0.486
KobeU (Seki)	Ku	yes	T	0.1689	0.2417	0.219	0.254
RGU (Mukras)	rgu0	yes	T	0.1686	0.2266	0.2163	0.288
UBuffalo (Ruiz)	UB1	yes	TDN	0.1501	0.2001	0.1887	0.266
Wuhan (Lu)	NOOPWHU1	yes	T	0.0011	0.0071	0.0072	0.008

Table 4: Opinion finding results: best run from each of the 20 groups, regardless of the used topic length. The best in each column is highlighted.

sumed spam documents are treated as the relevant set. BadMAP shows when spam documents are retrieved at early ranks (a low BadMAP value is good, while a high BadMAP is bad as more spam documents are being retrieved at early ranks). From this table, we can see that some runs were less susceptible to spam documents than others. In particular, the run from UIC exhibited a perfect 0 BadMAP and the lowest Spam@10 and Spam@all measures, suggesting that this group has very successfully applied splog detection techniques (Indeed, UIC has experimented with a spam detection module in TREC 2007). In contrast, the run NTUAutoOp from NTU was affected much more by splog documents.

To see if runs that retrieved less spam documents were more likely to be high performing systems or low performing systems, we correlated the ranking of submitted runs by BadMAP, correlating this with opinion finding MAP. However, the correlation was low ($\rho = 0.01, \tau = 0.03$), showing that for this task, systems which did remove spam documents were not any more likely to have a higher opinion retrieval performance.

3.4 Polarity Subtask

The polarity subtask was introduced in TREC 2007 as a natural extension of the opinion task, and was intended to represent a text classification-related task, requiring participants to determine the polarity (or orientation) of the opinions in the retrieved documents, namely whether the opinions in a given document are positive, negative or mixed. Participants were encouraged to use last years 50 opinion task queries, with their associated relevance judgments for training. Indeed, during the assessment procedure in the TREC 2006 blog track, for each document in the pool, the NIST assessors have specified the polarity of the relevant documents as described in Section 3.2 above: relevant negative opinion (judged as 2 in qrels); relevant mixed positive and negative (judged as 3 in qrels); relevant positive opinion (judged as 4 in qrels).

Groups participating in the opinion task and wishing to submit runs to the polarity subtask were asked to provide a corresponding and separate file for a submitted run to the opinion task, which details the predicted polarity for each retrieved document for each query. Submitted runs included the same documents in the same order as for the opinion finding runs, but with an additional polarity predictive label. Overall, 11 groups submitted 38 runs to the polarity subtask, including 32 automatic runs and 6 manual runs.

	R-Acc
Best	0.2959
Median	0.1227
Worst	0.0004

Table 11: Best, median and worst R-accuracy measures for the 38 submitted runs to the polarity subtask.

The initial intention was to evaluate the submitted runs using a classification accuracy measure (i.e. set precision). However, a measure like classification accuracy is comparable between runs only when every run classifies every document in the test set. In the polarity subtask, each run only provides a classification for the documents in its associated ranked opinion finding run. This presents three problems: not every run classifies the same documents, the treatment of unclassified documents is undefined, and no standard cutoff in the ranking is apparent.

To provide scores that are suitably comparable between runs, we report a measure called "R-accuracy" (R-Acc). This is the fraction of retrieved documents above rank R that are classified correctly, where R is the number of opinion-containing documents for that topic. The proposed measure is analogous to R-precision where only the correctly-classified opinion documents are counted as relevant. We also report accuracy at fixed rank cutoffs (A@10 and A@1000) as a secondary metric. For all measures, unjudged retrieved documents have no correct classification. The assumption is that if a submitted run had known that the document was not opinionated then the run should not have retrieved it, i.e. by retrieving it the run assumes that the document was opinionated, and hence must have wrongly classified it. Table 11 provides the average best, median and worst R-Acc measures for each topic, across all submitted 38 runs.

Group	Run	MAP	R-prec	b-Bref	P@10
UAmsterdam (deRijke)	uams07topic	**0.3453**	**0.3872**	**0.3953**	**0.562**
FudanU (Wu)	FDUNOpRSVMT	0.3178	0.3447	0.3498	0.452
CAS (Liu)	Relevant	0.3041	0.3600	0.3779	0.446
DalianU (Yang)	DUTRun1	0.2890	0.3368	0.3249	0.502
UGlasgow (Ounis)	uogBOPFProx	0.2817	0.3366	0.3098	0.454
UNeuchatel (Savoy)	UniNEblog1	0.277	0.3353	0.3074	0.492
FIU (Netlab team)	FIUbPL2	0.2728	0.3204	0.2925	0.454
Zhejiangu (Qiu)	EAGLE1	0.2561	0.3159	0.2867	0.428
UArkansas Littlerock (Bayrak)	UALR07Base	0.2554	0.3145	0.2867	0.440
IndianaU (Yang)	oqsnr1Base	0.2537	0.323	0.3091	0.446
CAS (NLPR-IACAS)	NLPRPTONLY	0.2506	0.3166	0.2917	0.452
UWaterloo (Olga)	UWbasePhrase	0.2486	0.3087	0.2861	0.432
BUPT (Weiran)	prisOpnBasic	0.2466	0.3018	0.2835	0.456
NTU (Chen)	NTUAuto	0.2254	0.2795	0.2588	0.412
KobeU (Seki)	Ku	0.1689	0.2417	0.219	0.254
RGU (Mukras)	rgu0	0.1686	0.2266	0.2163	0.288
Wuhan (Lu)	NOOPWHU1	0.0011	0.0071	0.0072	0.008

Table 5: Opinion finding results: automatic title-only baseline runs from each of the group with the best MAP, sorted by MAP. In these runs, all opinion finding features are switched off. The best in each column is highlighted. Note that some groups did not submit the compulsory automatic title-only baseline run.

Table 12 shows the best-scoring title-only polarity detection run for each group in terms of R-accuracy, and sorted in decreasing order of R-accuracy, while Table 13 shows the same information, but regardless of the topic length. Noticeable from these tables is that the runs appear to be clustered into two groups, those above 11% polarity detection R-accuracy, and those below.

It is interesting to note that the Spearman's ρ and Kendall's τ correlation coefficients between the polarity detection R-accuracy results and their corresponding opinion-finding MAP results over the 38 submitted polarity runs are very high (ρ = 0.9345 and τ =0.8065). This can be explained by the fact that the systems which are more successful at retrieving opinionated documents ahead of relevant ones, will then have more documents for which they can make a correct classification. Systems which perform poorly at retrieving opinionated documents are by definition not going to have the chance to classify as many documents correctly, hence the strong correlation is expected.

3.5 Participant Approaches

There were a wide range of deployed techniques by the participating groups. In this section, we focus on those groups whose use of opinion finding features have markedly improved their topic-relevance baseline as shown in Table 6. Looking into the main features of the best submitted runs, we note the following:

Indexing All the participating groups only indexed the Permalink component of the Blog06 collection, but the group from the University of Waterloo, which used all three components of the collection namely, Permalinks, Feeds and Homepages.

Retrieval Similar to TREC 2006, most of the participating groups used a two-stage approach for document retrieval [2]. In the first stage, documents are ranked using a variety of document weighting models ranging from BM25 (e.g. University of Indiana and University of Waterloo) to Divergence From Randomness models (e.g. University of Glasgow and FIU (Netlab team)), through language modelling (e.g. University of Amsterdam). Many participants used off-the-shelf systems such as Indri or Terrier. In the second stage of the retrieval process, the retrieved documents are re-ranked taking into account opinion finding features, often through a combination of scores mechanism.

Opinion Finding Features From looking at the results, we observe that there were two main effective approaches for detecting opinionated documents, which both led to improvements over a topic-relevance baseline. The first approach, used for example by the University of Glasgow and FIU, consists in automatically building a weighted dictionary from the relevance assessments of the TREC 2006's opinion finding task. The weight of each term in the dictionary estimates its opinionated discriminability. The weighted dictionary is then submitted as a query to generate an opinionated score for each document of the collection. The second approach, tested for example by the University of Arkansas at Little Rock and the University of Waterloo, uses a pre-compiled list of subjective terms and indicators and re-ranks the documents based on the proximity of the query terms to the aforementioned pre-compiled list of terms.

In the following, we provide more details on methods used by the 5 best performing groups, whose approaches for detecting opinionated documents have worked well, compared to a topic-relevance baseline as shown in Table 6:

The **University of Glasgow (UoG)** experimented with two approaches for detecting opinionated documents, integrated into their Terrier search engine. The first purely statistical approach uses a compiled English word list collected from various available linguistic resources. UoG measured the opinionated discriminability of each term in the word list using an information theoretic divergence measure based on the relevance assessments of the TREC 2006's opinion finding task. They have then estimated the opinionated nature of each document in the collection with the PL2 Divergence from Randomness (DFR) weighting model, and using the weighted opinionated word list as a query. The same approach was used to detect polarity. Their second opinion detection approach uses OpinionFinder, a freely available toolkit, which identifies subjective sentences in text. For a given document, they adapted OpinionFinder to produce an opinion score for each document, based on the identified opinionated sentences. Using either of two opinion

Group	Best Baseline	Baseline MAP	Best Non-baseline	Non Baseline MAP	% Increase
UGlasgow (Ounis)	uogBOPFProx	0.2817	uogBOPFProxW	**0.3264**	**15.87%**
IndianaU (Yang)	oqsnr1Base	0.2537	oqsnr2opt	0.2894	14.07%
UArkansas Littlerock (Bayrak)	UALR07Base	0.2554	UALR07BlogIU	0.2911	13.98%
DalianU (Yang)	DUTRun1	0.289	DUTRun2	0.319	10.38%
UWaterloo (Olga)	UWbasePhrase	0.2486	UWopinion3	0.2631	5.83%
CAS (NLPR-IACAS)	NLPRPTONLY	0.2506	NLPRPST	0.2542	1.44%
NTU (Chen)	NTUAuto	0.2254	NTUAutoOp	0.2282	1.24%
FudanU (Wu)	FDUNOpRSVMT	0.3178	FDUTisdOpSVM	0.3179	0.03%
FIU (Netlab team)	FIUbPL2	0.2728	FIUdPL2	0.2728	0.00%
Wuhan (Lu)	NOOPWHU1	0.0011	OTWHU101	0.0011	0.00%
KobeU (Seki)	Ku	0.1689	KuKnn	0.1657	-1.89%
Zhejiangu (Qiu)	EAGLE1	0.2561	EAGLE2	0.2493	-2.66%
CAS (Liu)	Relevant	0.3041	DrapOpi	0.1659	-45.45%
RGU (Mukras)	rgu0	0.1686	rgu2	0.0892	-47.09%
UAmsterdam (deRijke)	uams07topic	**0.3453**	uams07mmqop	0.1459	-57.75%
BUPT (Weiran)	prisOpnBasic	0.2466	prisOpnC2	0.0821	-66.71%

Table 6: What worked. Improvements over the baselines, for automatic title-only runs. The best in each column is highlighted. Some groups did not submit title-only baseline runs (e.g. UIC group), and some did not submit any run with specific opinion finding features (e.g. UNeuchatel).

detection approaches, UoG used the opinionated scores of the documents as prior evidence, and integrated them with the relevance scores produced by the document weighting model used. All their six submitted runs used the PL2F field-based weighting model. One of their topic-relevance baselines included a DFR-based proximity model. They found that the use of the word list-based statistical opinion detection approach markedly improved their topic-relevance only baseline, leading to a substantial and marked improvement of 15.8% compared to the topic-relevance baseline (run uogBOPFProxW vs run uogBOPFProx). Interestingly, they also found that the opinion finding technique based on the Opinion-Finder tool was as effective as the statistical word list-based approach, although it was less efficient. They also reported that the use of proximity search is helpful.

The **University of Indiana (IndianaU)** focused on combining multiple sources of evidence to detect opinionated blog postings. Their approach to opinion blog retrieval consisted of first applying traditional retrieval methods to retrieve on-topic blogs and then boosting the ranks of opinionated blogs based on combined opinion scores generated by multiple assessment methods. Indiana's opinion assessment/detection method is comprised of High Frequency Module, which identifies opinion blogs based on the frequently used opinion terms, low frequency module, which leverages uncommon/rare term patterns (e.g., 'sooo good') for expressing opinions, IU Module, which makes use of 'I' and 'You' collocations (e.g. 'I believe') that qualify opinion sentences, Wilson's lexicon module, which makes use of Wilson's subjective lexicons, and opinion acronym module, which utilises the small set of opinion acronyms (e.g., 'imho') that are likely to be missed by preceding modules. Indiana's training data consisted of TREC 2006's opinion finding relevance data supplemented by the external IMDB movie review data, both of which were used to tune their opinion scoring and fusion module in an interactive system optimisation mechanism called the Dynamic Tuning Interface. All of the lexicon terms were scored with positive and negative values, which facilitated their participation in the polarity subtask. They found that their opinion finding approach improves upon the topic-relevance only baseline.

The **University of Arkansas at Little Rock (UArkansas)** used various opinion finding heuristics on top of a topic-relevance baseline. Their best performing opinion finding run re-ranked the documents returned by the baseline, by taking into account the proximity of words such as "I", "you", "me", "us", we" and opinion indicator words such as "like", "feel","think","hate" to the actual query words. They found that such a simple proximity-based approach could markedly improve the opinion finding retrieval effectiveness of their topic relevance baseline (about 14% improvement). UArkansas also experimented with a machine learning-based approach, which re-ranks the baseline results by associating a category to the queries. This approach while slightly improving upon the performance of the topic-relevance baseline, was comparatively less successful than the proximity-based approach.

The **Dalian University of Technology (DUT)** filtered out all non-English blog posts during indexing. They used an external resource, namely the Wikipedia, and a manually built sentiment lexicon resource to find opinions. In the polarity subtask, DUT used a method based on SVM, to assess the polarity of the retrieved blog posts. Judging by the results, DUT found that their used sentiment resources had improved their initial topic-relevance baseline MAP with about 11%.

The **University of Waterloo (UoW)** used a manually constructed list of 1336 subjective adjectives in document ranking. The top 1000 documents retrieved using BM25 were re-ranked based on the proximity of each query term instance to the subjective adjectives. Experiments were also conducted with different types of queries constructed from the topic titles: single terms and user-defined phrases, i.e. phrases enclosed in quotation marks by the user. Some improvements over the topic-relevance baseline were achieved (about 5.8% improvement) when the initial document set was retrieved using phrases, while the subjective adjective-based re-ranking was done using single terms. UoW concluded that subjective adjectives located close to any word from the query are useful indicators of the presence of opinions expressed about the query topic.

It is of interest to make some comments about the submitted official runs by some participating groups. The University of Illinois at Chicago (UIC) achieved the top scoring opinion finding run. How-

Group	Run	Fields	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	T	**0.4819**	0.5181	0.5484	**0.868**
UAmsterdam (deRijke)	uams07topic	T	0.4741	**0.523**	**0.5702**	0.762
FudanU (Wu)	FDUTisdOpSVM	T	0.4714	0.4889	0.5432	0.654
IndianaU (Yang)	oqlr2fopt	TDN	0.4347	0.4653	0.5022	0.822
CAS (Liu)	Relevant	T	0.4302	0.4949	0.5658	0.662
DalianU (Yang)	DUTRun2	T	0.4247	0.4750	0.5164	0.784
UGlasgow (Ounis)	uogBOPFProxW	T	0.4160	0.4436	0.4618	0.720
UNeuchatel (Savoy)	UniNEblog3	TD	0.4034	0.4296	0.4553	0.730
FIU (Netlab team)	FIUDDPH	TD	0.3907	0.4230	0.4692	0.714
UArkansas Littlerock (Bayrak)	UALR07Blog1U	T	0.3612	0.3975	0.4122	0.734
UWaterloo (Olga)	UWopinion3	T	0.3490	0.4040	0.4020	0.68
Zhejiangu (Qiu)	EAGLE2	T	0.3409	0.3809	0.3992	0.644
CAS (NLPR-IACAS)	NLPRTD	TD	0.3373	0.3804	0.3894	0.586
KobeU (Eguchi)	KobePrMIR01	T	0.3292	0.3655	0.3852	0.606
BUPT (Weiran)	prisOpnBasic	T	0.3267	0.3633	0.3735	0.684
NTU (Chen)	NTUManual	T	0.3051	0.3309	0.3631	0.582
RGU (Mukras)	rgu0	T	0.2798	0.3533	0.3651	0.560
KobeU (Seki)	Ku	T	0.2590	0.3357	0.3503	0.476
UBuffalo (Ruiz)	UB1	TDN	0.2421	0.2818	0.2956	0.484
Wuhan (Lu)	NOOPWHU1	T	0.0016	0.0111	0.0100	0.02

Table 7: Topic-relevance results: run from each of the 20 groups with the best topic-relevance MAP, sorted by MAP. The best in each column is highlighted.

ever, they did not submit the compulsory topic-relevance baseline. Therefore, it is difficult to assess the usefulness of their opinion finding features. Nevertheless, UIC's retrieval system contained two sub-systems. The opinion retrieval system (ORS), which was modified from the TREC 2006 version, and was used for the main task and a polarity classification system (PCS), which was newly designed for the polarity subtask. UIC experimented with a single query-independent SVM classifier and tested a spam detection module.

The runs submitted by the University of Amsterdam (UvA) raise a few interesting issues. While they had a strongly performing topic-relevance baseline run (see run uams07topic in Table 3), their used opinion finding features do not appear to be useful. UvA used the opinion finding task to compare the performance of an Indri implementation to their own mixture model. The mixture model combines different components of blog posts (e.g., headings, title, body) and assigns weights to these components based on tests on the TREC 2006 topics. Of both the baselines, the Indri system performed markedly better. To achieve better topical results, external (query) expansion on the AQUAINT-2 news corpus was performed. This expansion improves the performance of the Indri implementation, but hurts the mixture model. For opinion finding, UvA experimented with document priors in the mixture model based on either opinionated lexicons or the number of comments. The latter opinion finding features have not improved their opinion finding performance, markedly hurting their strongly performing uams07topic topic-relevance baseline run. In particular, run uams07topic is the 2nd top scoring title-only opinion finding run of the track, despite not using any opinion detection approach, suggesting that a strong retrieval baseline can do very well on the opinion finding task.

Interestingly, the Netlab team (FIU) used an approach that is very similar to the word list-based detection approach deployed by UoG, although developed separately. FIU used the DFR models, i.e. PL2 and the parameter free DPH, to assign both topic and opinion scores. A fully automatic and weighted dictionary was generated from TREC 2006's opinion finding relevance data. This dictionary was filtered and then submitted as a query to the Terrier search engine to get an initial query-independent opinion score of all retrieved documents. Ranking is done in two passages: a first topical-opinion ranking is obtained from the query-independent opinion score divided by the content rank, then the final topical-opinion ranking is established from the content score divided by the previous topical-opinion rank. Since FIU updated the final ranks but not the final topical-opinion scores in the re-ranking, trec_eval reported the same performance for all their official submitted runs. However, using the Terrier evaluation tool, which instead evaluates runs by ranks and not by scores, they show that FIU's opinion finding approach is actually effective. Indeed, their opinion finding run FIUIPL2 has about 17% improvement over their topic relevance baseline, an improvement in the same line as observed with UoG's wordlist-based approach, and expected given the similarities of the two groups's approaches.

3.6 Summary of Opinion Finding Task

The additional requirement that each participating group submits a compulsory topic-relevance baseline run allowed us to draw more conclusions on those opinion detection approaches that have worked and those that have not, providing additional insights for future work.

The overall opinion finding performance of the participating groups this year was markedly higher than the one observed for the TREC 2006 topics set. However, it is difficult to assess whether this increase in performance is due to the better deployed opinion finding systems and techniques or whether it is due to the difficulty level of the topics set. Answering this question requires running this year's systems on last year's topics.

Finally, similar to last year's conclusion, there appears to be no strong evidence that spam was a major hindrance to the retrieval performance of the participating groups.

4. BLOG DISTILLATION (FEED SEARCH) TASK

The blog distillation (feed search) task is a new task in the TREC 2007 Blog track, which was the result of the discussion that followed the introduction of the open task in TREC 2006. The task

Group	Run	Spam@10	Spam@all	BadMAP *10^-5
UIC (Zhang)	uic1c	**0.56**	**33.86**	**0.0**
UAmsterdam (deRijke)	uams07topic	0.92	104.14	2.8
UGlasgow (Ounis)	uogBOPFProxW	1.24	126.32	10.8
DalianU (Yang)	DUTRun2	0.74	55.66	3.0
FudanU (Wu)	FDUTOSVMSem	0.98	59.50	2.2
CAS (Liu)	Relevant	1.34	75.66	2.2
UArkansas Little Rock (Bayrak)	UALR07Blog1U	0.88	121.74	10.2
IndianaU (Yang)	oqsnr2opt	0.98	181.20	13.0
UNeuchatel (Savoy)	UniNEblog1	1.18	139.18	12.2
FIU (Netlab team)	FlUbPL2	1.42	131.98	11.8
UWaterloo (Olga)	UWopinion3	1.16	75.88	7.2
Zhejiangu (Qiu)	EAGLE1	1.24	121.74	9.6
CAS (NLPR-IACAS)	NLPRPST	1.22	124.68	8.4
BUPT (Weiran)	prisOpnBasic	1.32	80.22	7.2
KobeU (Eguchi)	KobePrMIR01	1.54	157.82	13.4
NTU (Chen)	NTUAutoOp	0.94	161.70	15.0
KobeU (Seki)	Ku	2.12	153.42	10.6
RGU (Mukras)	rgu0	1.30	86.30	5.6
UBuffalo (Ruiz)	UB2	4.92	86.44	4.2
Wuhan (Lu)	NOOPWHU1	1.56	101.96	4.8

Table 10: Spam measures for runs from Table 3, in the order given. Spam@10 is the mean number of spam posts in the top 10 ranked documents for each topic, Spam@all is the mean number of spam posts retrieved for each topic. BadMAP is the Mean Average Precision when the spam documents are treated as the relevant set. This shows when spam documents are retrieved at high ranks. For all measures, lower means the system was better at not retrieving spam documents. The best in each column is highlighted.

Group	Run	R-Acc	A@10	A@1000
UIC (Zhang)	uic75cpnm	**0.2295**	**0.3700**	**0.0493**
UAmsterdam (de Rijke)	uams07ipolt	0.1827	0.2640	0.0418
IndianaU (Yang)	oqsnr2optP	0.1799	0.2800	0.0401
DalianU (Yang)	DUTRun2P	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	0.1510	0.2380	0.0427
UGlasgow (Ounis)	uogBOPFPol	0.1460	0.2020	0.0397
NTU (Chen)	NTUAutoOpP	0.0967	0.1860	0.0296
CAS (Liu)	DrapStmSub	0.0818	0.1060	0.0243
BUPT (Weiran)	pUB21	0.0418	0.0340	0.0148
Wuhan (Lu)	OTPSWHU102	0.0032	0.0040	0.0010

Table 12: Best polarity run for each group, in terms of R-accuracy. Each polarity runs corresponds to an automatic title-only opinion finding run. The best in each column is highlighted. Not all groups submitted polarity runs corresponding to automatic title-only opinion finding runs.

focuses on an interesting feature of the blogs, namely the fact that feeds are aggregates of blog posts.

4.1 Motivations

Blog search users often wish to identify blogs (i.e. feeds) about a given topic, which they can subscribe to and read on a regular basis. This user task is most often manifested in two scenarios:

- Filtering: The user subscribes to a repeating search in their RSS reader.
- Distillation: The user searches for blogs with a recurring central interest, and then adds these to their RSS reader.

For TREC 2007, the latter scenario was investigated i.e. blog distillation, which is a feed search task. The blog distillation task can be summarised as *Find me a blog with a principle, recurring interest in X*. For a given target *X*, systems should suggest feeds that are principally devoted to X over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about *X* (i.e. a user may be interested in adding it to their RSS reader). This task is particularly interesting for the following reasons:

- A similar (yet-different) task has been investigated in the Enterprise track (Expert Search) in a smaller setting (around 1000 candidate experts on the W3C collection). For blog distillation, the Blog06 corpus contains around 100k blogs, and is a Web-like setting (with anchor text, linkage, spam, etc).
- A Topic distillation task was run in the Web track. In Topic distillation, site relevance was defined as (i) it is principally devoted to the topic, (ii) it provides credible information on the topic, and (iii) it is not part of a larger site also principally devoted to the topic.

While the definition of blog distillation as explained above is different, the idea is to provide the users with the key blogs about

Group	Run	Fields	R-Acc	A@10	A@1000
UIC (Zhang)	uic75cpnrm	T	**0.2295**	**0.3700**	**0.0493**
IndianaU (Yang)	oqlr2f2optP	TDN	0.1941	0.3080	0.0438
UAmsterdam (de Rijke)	uams07ipolt	T	0.1827	0.2640	0.0418
DalianU (Yang)	DUTRun2P	T	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	T	0.1510	0.2380	0.0427
UGlasgow (Ounis)	uogBOPFPol	T	0.1460	0.2020	0.0397
NTU (Chen)	NTUManualOpP	T	0.1161	0.2300	0.0348
CAS (Liu)	DrapStmSub	T	0.0818	0.1060	0.0243
BUPT (Weiran)	prisPolC2	T	0.0726	0.2020	0.0124
UBuffalo (Ruiz)	pUB11	TDN	0.0671	0.1000	0.0195
Wuhan (Lu)	OTPSWHU102	T	0.0032	0.0040	0.0010

Table 13: Best polarity run for each group, in terms of R-accuracy, regardless of the topic length. The best in each column is highlighted. Not all groups submitted polarity runs.

```
<top>
  <num> Number: 994 </num>

  <title> formula f1 </title>

  <desc> Description:
Blogs with interest in the formula
one (f1) motor racing, perhaps with
driver news, team news, or event
news.
</desc>

  <narr> Narrative:
  Relevant blogs will contain news
and analysis from the Formula f1
motor racing circuit. Blogs with
documents not in English are not
relevant.
</narr>
  </top>
```

Figure 5: Blog track 2007, blog distillation task, topic 994.

a given target. Note that point (iii) from the definition of the Web track Topic distillation task is not applicable in a blog setting.

4.2 Topics and Relevance Judgments

For the purposes of the blog distillation task, the retrieval document units are documents from the feeds component of the Blog06 collection. However, similar to the opinion finding task, the participating groups were free to use any other component of the Blog06 test collection in their submitted runs.

The topics for the blog distillation were created and judged by the participating groups. Each participating group has been asked to provide 6 or 7 topics along with some relevant feeds. A standard search system for documents on the Blog06 collection using the Terrier search engine [3] was provided by the University of Glasgow to help the participating groups in creating their blog distillation topics. The system displays the corresponding feed for each returned document (i.e. blog post), as well as all the documents for a given feed. Eight groups contributed each 5 to 7 topics. 45 topics were finally chosen by NIST from the proposed set of topics. A sample blog distillation topic is shown in Figure 5.

Overall, 9 groups submitted runs and agreed to help in their relevance judgments. Once runs were submitted, NIST formed pool and sent them to the University of Glasgow, where the community assessment system was hosted. The community judgments system interface was ported directly from the TREC Enterprise judgment system for expert search task developed by Soboroff et al. [4].

Participants were allowed to submit up to 4 runs, including a compulsory title-only run. Similar to the opinion finding task, the participants were asked to prioritise runs, in order to define which of their runs would be pooled. Each run has feeds ranked by their likelihood of having a principle (recurring) interest in the topic. Given the number of feeds in the collection (just over 100k feeds), each submitted run consisted of up to 100 feeds for each topic. A pool has then been formed by NIST from the 32 submitted runs, using the two highest-priority runs per group, pooled to depth 50.

For the assessment of the relevance of a feed, the assessors were asked to browse some of the documents of the feed, and then make a judgment on whether the feed has a recurring principle interest in the topic area. These guidelines are intentionally vague. A question that may arise is the number of documents (i.e. posts) that have to be read by the assessor for a given feed. Since there is no straightforward answer to this question, we decided to suggest that the assessors read enough documents of the feed such that they are certain that the feed has a more than passing interest in the topic area, and that they would be interested in subscribing to the feed in their RSS reader if they were interested in the topic area.

4.3 Overview of Results

The blog distillation task is another articulation of real user tasks in adhoc search behaviour on the blogosphere. Therefore, we use mean average precision (MAP) as the main metric for the evaluation of the retrieval performance of the submitted runs. In addition, we also report R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10).

All submitted runs were automatic. Table 14 provides the average best, median and worst MAP measures for each topic, across all submitted 32 runs. Figure 6 shows the distribution of the number of relevant feeds per topic in the pooled feeds, sorted in decreasing order. In particular, there appears to be a wide variance in the number of relevant feeds across the used 45 topics, with topics having as many as 153 relevant feeds (e.g. "christmas" (968) or "music" (978)), while other having as few as 5 relevant feeds (e.g. "Violence in Sudan" (964) or "machine learning" (982)).

Table 15 shows the best-scoring automatic title-only run from each participating group in terms of MAP, and sorted in decreasing order. Table 16 shows the best run from each group, regardless of the topic length used. Note that most of the 32 submitted runs were title-only runs. Indeed, there were 25 submitted runs using the title field only, 3 submitted runs used the title, description and narrative

	MAP
Best	0.4671
Median	0.2035
Worst	0.0006

Table 14: Best, median and worst MAP measures for the 32 submitted runs to the blog distillation task.

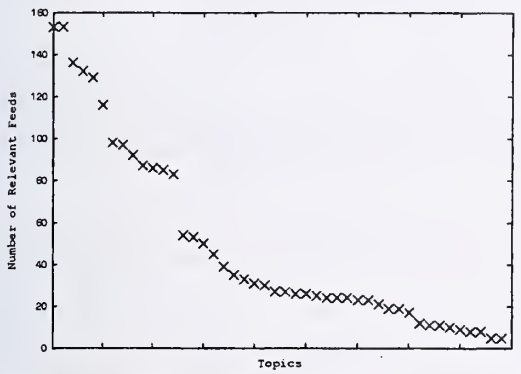


Figure 6: Distribution of number of relevant feeds per topic

Relevance Scale	Nbr. of Splogs
Not Relevant	2935
Relevant	255
(Total)	3190

Table 17: Occurrences of presumed splogs in the blog distillation task pool.

fields, 2 submitted runs used the title and description fields, and 2 submitted runs used the description field only. All the 10 best submitted blog distillation runs but one are title-only runs. Given the rather small number of submitted runs using long queries, it is difficult to draw conclusions as to whether the description and narrative fields of the topics might be helpful in the blog distillation task.

We examined whether the participating systems in the blog distillation task had been affected by spam, i.e. how many splog feeds have infiltrated the pool. Table 17 shows the breakdown of the feed distillation pool in terms of splog feeds. Moreover, Table 18 shows the extent to which the 17,958 presumed splogs have infiltrated the submitted runs. We use the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), in the retrieved documents (Spam@all), and finally BadMAP, which is the Mean Average Precision when the splog feeds are treated as the relevant set. Run UMaTiPCSwGR from UMass appears to be overall the least susceptible to splog feeds. On the contrary, run TDWHU200 was one of the most affected runs by splog feeds.

Similar to the analysis performed in Section 3.3, to see if runs that retrieved less splog feeds were more likely to be high performing systems or low performing systems, we correlated the ranking of submitted runs by BadMAP, correlating this with blog distillation MAP. For this task, a weak correlation was exhibited ($\rho = -0.193, \tau = -0.157$), showing some evidence that systems which did remove splogs were likely to have a higher retrieval performance.

4.4 Participant Approaches

There were a wide range of deployed indexing and retrieval approaches for the blog distillation task. The exploratory nature of most of the used techniques characterises the novelty of the task and its interesting underlying features. The main features of the submitted runs are summarised below:

Indexing Two types of indexes have been used. Three groups created an index using the Feeds component of the Blog06 collection, namely Carnegie Mellon University (CMU), the University of Texas, and the University of Wuhan. The rest of the groups only indexed the Permalinks component of the collection. Interestingly, CMU, the top performing group, experimented with both types of index, and concluded that an index based on the Feeds component of the Blog06 collection leads to a better retrieval performance on this task.

Retrieval Many groups approached the blog distillation task by connecting the task to other existing search tasks. For example, the University of Glasgow (UoG) explored the connection of blog distillation to the expert finding task of the Enterprise track, adapting their Voting Model paradigm to feed search. The University of Massachusetts looked at the blog distillation task as a resource selection problem in distributed search. Most of the groups that used an index based on Permalinks, have proposed various techniques to aggregate the scores of blog posts into a score for their composing feed. For the purposes of document retrieval, a range of document weighting models such as Language Modelling approaches and Divergence From Randomness models were used. Some groups have also experimented with classical information retrieval techniques, namely query expansion (e.g. CMU) and proximity search (e.g. UoG).

In the following, we provide a detailed description of the methods used by the top 3 performing groups in the blog distillation task:

Carnegie Mellon University (CMU) explored two indexing strategies, namely a large-document model (feed retrieval) and a small-document model (entry or blog post retrieval). Under the large-document model, feeds were treated as the unit of retrieval. Under the small-document model, the blog posts were treated as the unit of retrieval and aggregated to produce a final ranking of feeds. They found that the large-document approach outperformed the small-document approach on average. CMU also experimented with a query expansion method using the link structure and link text found within an external resource, namely the Wikipedia. CMU found that the used Wikipedia-based query expansion approach improves results under both the large- and small-document models.

The **University of Glasgow** (UoG) only indexed the Permalink component of the Blog06 collection. They investigated the connections between the blog distillation task and the expert search task. UoG adapted their Voting Model paradigm for Expert Search, by ranking feeds according to the number of on-topic posts each feed has (number of votes), and the extent to which the posts are about the topic area (strength of votes) - these two sources of evidence about the interests of each blogger were combined using the expCombMNZ voting technique. Posts are ranked using the PL2F Divergence From Randomness (DFR) field-based weighting model. They found that the additional use of a DFR-based term proximity model improves the topicality of the underlying ranking of blog posts, leading to a more accurate aggregated ranking of blog posts and a better feed search performance.

The **University of Massachusetts** (UMass) used language modelling approaches. UMass used the Permalink component of the Blog06 test collection for indexing. UMass looked at this task as a resource selection problem in distributed information retrieval,

Group	Run	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	**0.3695**	**0.4245**	**0.3861**	**0.5356**	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTiPCSwGR	0.2529	0.3334	0.2902	0.5111	**0.8093**
KobeU (Seki)	kudsn	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun1	0.2285	0.3105	0.2768	0.3711	0.5813
UTexas-Austin (Efron)	utblnrr	0.2197	0.3100	0.2649	0.4511	0.7245
UAmsterdam (deRijke)	uams07bdtblm	0.1605	0.2346	0.1820	0.3067	0.6320
WuhanU (Lu)	TDWHU200	0.0135	0.0419	0.0297	0.0578	0.1386

Table 15: Blog distillation results: the automatic title-only run from each of 8 groups with the best MAP, sorted by MAP. Note that 1 group (UBerlin) did not submit a title-only run. The best in each column is highlighted.

Group	Run	Fields	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	T	**0.3695**	**0.4245**	**0.3861**	**0.5356**	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	T	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTDPCSwGR	TD	0.2741	0.3356	0.3027	**0.5356**	**0.8407**
KobeU (Seki)	kudsn	T	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun4	TDN	0.2399	0.3126	0.2740	0.4378	0.7337
UTexas-Austin (Efron)	utblnrr	T	0.2197	0.3100	0.2649	0.4511	0.7245
UAmsterdam (deRijke)	uams07bdtblm	T	0.1605	0.2346	0.1820	0.3067	0.6320
UBerlin (Neubauer)	ADABoostM1	TDN	0.0176	0.0468	0.0330	0.0978	0.2881
WuhanU (Lu)	TDWHU200	T	0.0135	0.0419	0.0297	0.0578	0.1386

Table 16: Blog distillation results: one run from each of 9 groups with the best MAP, sorted by MAP. The best in each column is highlighted.

since each feed can be considered as a collection composed of blog posts. The most critical issue of resource selection is how a collection is represented. UMass applied two approaches for representation in this task. Further, since blogs which address many general and shallow topics are unlikely to be relevant in this task, UMass introduced an approach to penalise such blogs, and found that this improves the retrieval effectiveness.

Other approaches used by the participating groups included the investigation of blog specific approaches such as time-based priors and splog detection and filtering, or retrieval models variants to search from a feeds-based index. The University of Amsterdam (UvA) experimented with time-based priors. Their suggested idea is that more recent posts reflect better the current interest of a blogger. Results show that time-based priors, which order the feeds based on the score of the most relevant post from a feed, improve slightly over the baseline run. UvA also experimented with a relevant posts count, where for every feed the ratio of relevant posts to all posts in a feed is calculated and this score is combined with the feed relevance score from the baseline run. Results show that this has markedly decreased performance, suggesting that the combination parameters were not appropriate.

Kobe University (Seki et al.) experimented with splog detection, and filtering of non-English documents. Interestingly, their baseline is built by computing the similarity scores between a query and the posts included in the feed. They plotted a line for each blog site with the x-axis being the (normalised) post date and the y-axis being the computed similarity. The feeds are then ranked according to the descending order of the surface area under the plotted line. The intuition behind the proposed algorithm is that a relevant feed would frequently mention a given topic, and will constantly have a high similarity with the topic (query), resulting in a large surface area under the line of similarity scores. They found that filtering splogs and non-English documents improves their baseline.

Finally, the University of Texas' School of Information (UT) used a retrieval strategy based on a variant of the Kullback-Leibler (KL) divergence model. Given a query q the UT system derives a score for each feed f in the corpus by the negative KL-divergence between the query language model and the language model for f. The effectiveness of the proposed approach cannot be assessed without an experimental baseline.

4.5 Summary of Blog Distillation Task

The blog distillation task was a new task in TREC 2007. Overall, some of the deployed retrieval approaches achieved reasonable retrieval performances. One of the issues that might need to be further investigated in this task is whether it is beneficial to use the Feeds component of the Blog06 collection, instead of or in addition to the Permalinks component.

There was a wide variance in the distribution of relevant feeds in the used 45 topics, suggesting that the guidelines for the topic creation and assessments still require tightening for future iterations of this task. However, the task, as exemplified by the exploratory nature of the participants runs, promises much research in the future.

5. CONCLUSIONS

The TREC 2007 Blog track included two main tasks, namely the opinion finding and Blog distillation (aka feed search) tasks, which we believe are good articulations of real user tasks in adhoc search behaviour on the blogosphere. The used tasks address two interesting components of blogs: the feed itself and its constituent blog posts and their corresponding comments. As a consequence, a new topics set has been created for the opinion finding task, and a new test collection has been created for the Blog distillation task, therefore contributing to the creation of reusable resources for supporting research into blog search.

Much remains to be learned about opinion finding, even though the runs submitted this year show that some participants have been successful in proposing new opinion detection techniques, which show some marked improvements on the respective topic-relevance baseline. Indeed, this year's findings also consolidate the findings

Group	Run	Spam@10	Spam@all	BadMAP *10^-5
CMU (Callan)	CMUfeedW	2.8	22.5	48.2
UGlasgow (Ounis)	uogBDFeMNZP	2.2	22.4	28.0
UMass (Allen)	UMaTiPCSwGR	**0.6**	**3.1**	**3.1**
KobeU (Seki)	kudsn	1.5	9.2	10.0
DalianU (Yang)	DUTDRun1	3.6	21.6	56.2
UTexas-Austin (Efron)	utblnrr	2.0	15.5	23.7
UTexas-Austin (Efron)	utlc	2.1	13.44	19.6
UAmsterdam (deRijke)	uams07bdtblm	1.9	13.7	26.0
WuhanU (Lu)	TDWHU200	3.1	159.1	184.0

Table 18: Spam measures for runs from Table 15, in the order given. Spam@10 is the mean number of splog feeds in the top 10 ranked documents for each topic, Spam@all is the mean number of splog feeds retrieved for each topic. BadMAP is the Mean Average Precision when the splog feeds are treated as the relevant set. This shows when spam feeds are retrieved at high ranks. For all measures, lower means the system was better at not retrieving splogs.

of the previous Blog track 2006. In particular, a good performance in opinion finding is strongly dominated by its underlying topic-relevance baseline (i.e. opinion-finding MAP and topic-relevance MAP are very highly correlated). Indeed, a strongly performing topic-relevance baseline can still perform extremely well in opinion finding, as exemplified by the University of Amsterdam's submitted topic-relevance baseline. One possible methodology to have a better understanding of the deployed opinion detection techniques is to use a common and strong topic-relevance baseline for all participating groups.

For the polarity subtask, the overall performances of the participating groups are rather average, suggesting that the task of detecting the polarity of an opinion is still an open problem, which requires further research. We believe that polarity detection should be a more integral part of the opinion finding task, and not evaluated as in classification task-like manner. For future iterations of the opinion finding task, we believe that a better integration of the polarity component would involve creating a balanced number of topics, which explicitly specify whether they require positive or negative opinions to be retrieved. Evaluation can then be carried out in a more straightforward adhoc manner.

The Blog distillation task seems to have generated some very promising and interesting retrieval techniques. We plan to run the task again for 2008, in a similar fashion, but with clearer guidelines for the creation of the topics. This will provide further insights on the most effective techniques for this task.

Acknowledgments

The description of systems runs are based on paragraphs contributed by the participating groups. Thanks are also due to participants in the blog distillation task for creating and assessing topics. In particular all groups who submitted runs to the blog distillation task have participated in the relevance judging. Finally, we are grateful to Arjen de Vries for providing his assessment system, and to David Hannah for adapting it to the blog distillation task.

6. REFERENCES

[1] C. Macdonald and I. Ounis. The TREC Blog06 Collection : Creating and Analysing a Blog Test Collection *DCS Technical Report TR-2006-224*. Department of Computing Science, University of Glasgow. 2006. `http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf`

[2] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, I. Soboroff. Overview of TREC-2006 Blog track. In *Proceedings of TREC-2006*, Gaithersburg, USA, 2007.

[3] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of OSIR'2006 Workshop*, Seattle, USA, 2006

[4] I. Soboroff, A. de Vries, and N. Craswell. Overview of TREC-2006 Enterprise track. In *Proceedings of TREC-2006*, Gaithersburg, USA, 2006.

Overview of the TREC 2007 Enterprise Track

Peter Bailey
Microsoft, USA
pbailey@microsoft.com

Nick Craswell
MSR Cambridge, UK
nickcr@microsoft.com

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

1 Introduction

The goal of the enterprise track is to conduct experiments with enterprise data that reflect the experiences of users in real organizations. This year, the track has introduced a new corpus with the goal to be more representative of real-world enterprise search, by involving actual members of the organization in the topic development process, performing their real work tasks.

2 Collection

The CERC corpus (CSIRO Enterprise Research Collection, (`http://es.csiro.au/cerc/`)) represents the public-facing web of the Australian Commonwealth Scientific and Industrial Research Oganization (CSIRO). Here, we summarize the main characteristics of this corpus; a complete description of the collection is given in Bailey et al. (2007).

2.1 Data

The collection consists of all the `*.csiro.au` (public) websites as they appeared in March 2007. The resulting data set consists of 370 715 documents, with total size 4.2 gigabytes. The web crawler visited the outward-facing pages of CSIRO in a fashion similar to the crawl used in CSIRO's own search engine. In fact, the same crawler technology that CSIRO uses was used to gather the CSIRO documents (`http://www.funnelback.com/`). The corpus contains approximately 7.9 million hyperlinks, and 95% of pages have one or more outgoing links containing anchor text. One participant extracted email addresses of 3678 individuals, with 38% of documents containing at least one `mailto` field.

2.2 Users

A science communicator's role in CSIRO is to enhance CSIRO's public image and promote the capabilities of CSIRO by managing information and interacting with industry groups, government agencies, professional groups, media and the general public. Science Communicators read and create the outward-facing web pages of CSIRO (as opposed to internal documents). Therefore they were a natural choice when thinking of which users are a good match for our outward-facing crawl.

2.3 Tasks and Topics

The 2007 enterprise track defined two tasks: document search and expert search. Both search tasks are grounded in a 'missing overview page' scenario, where the science communicator has to construct a new overview page on the topic of interest, that enumerates the 'key pages' and a few 'key people' of interest. Given this scenario, the document search task models the problem of finding the set S of 'key pages', and the expert search task the problem of locating the 'key contacts' among CSIRO staff.

The primary method for involving Science Communicators was asking them to do topic development. A general email was sent to all science communicators, calling for them to create topics in their area. Examples of general queries from CSIRO's real public search site were given for inspiration. This yielded 25 usable topics from 9 science communicators from multiple CSIRO divisions. Being short of the standard 50 topics, we then approached one of these communicators who produced another 25 topics to complete the set.

Each topic description has a query and narrative, some examples of key reference URLs (on average 4 per topic) and a short list of key contacts (on average 3 per topic, varying from 1 to 11). The key reference URLs serve as a (admittedly somewhat poor) surrogate for click-log data. Note that both tasks have used the same set of topics.

2.4 Assessments

For document search we used community judging. NIST formed pools and sent them to CSIRO, where the assessment system was hosted. Track participants then judged the pools through the CSIRO system (adapted from the assessment system used in the Million Query track).

The guidelines instructed the assessors to read the query and narrative, and optionally carry out a Web search to learn more about the subject. The guidelines also emphasized that science communicators are web-savvy users – so judgments should take into account that navigational answers and relevant homepages are important results in exploratory search behaviour. Relevance judgments were made on a three-point scale:

2: Highly likely to be a 'key page'.
1: Possible as a candidate for a page in S, or otherwise informative to help build an overview page, but not highly likely.
0: Not a 'key page' as unlikely to be included in S, because, e.g., not relevant, off-topic, not an important page on the topic, on-topic but out-of-date, not the right kind of navigation point, or too informal or too narrow an audience.

After the workshop, we investigated to what extent the people making relevance judgements for the document search task have been exchangeable, comparing assessments made by participants ('bronze' judges) to sampled re-assessments for 33 topics by the topic authors ('gold' judges) and/or other science communicators familiar with the task ('silver' judges). The main finding from the study is that the bronze judges may not be able to substitute for topic and task experts, due to changes in the relative performance of assessed systems, and gold judges are preferred. The full details of this post-TREC study can be found in Bailey et al. (2008).

For expert search, we did no further judging, using the experts listed in the topic as our ground truth.

Table 1: Document search results for the automatic run with the highest MAP from each group.

Group	Run	MAP	NDCG	P@20
CAS	DocRun02	0.422	0.743	0.527
York	york07ed4	0.416	0.730	0.513
Waterloo	uwtbase	0.388	0.707	0.508
RMIT	RmitQ	0.388	0.698	0.471
SJTU	SJTUEntDS02	0.374	0.692	0.475
UvA	uams07bfb	0.369	0.675	0.445
Tsinghua	THUDSFULLSR	0.366	0.701	0.461
UALR	UALR07Ent1	0.357	0.662	0.428
Fudan	FDUBase	0.350	0.664	0.426
OU	ouTopicOnly	0.345	0.646	0.464
Glasgow	uogEDSF	0.337	0.675	0.413
DUT	DUTDST4	0.336	0.644	0.441
Iowa	uiowa07entD2	0.310	0.597	0.413
Hyberdad	QRYBASICRUN	0.246	0.487	0.408
CSIRO	CSIROdsQonly	0.194	0.352	0.378
St. Petersburg	insu2	0.028	0.185	0.041

3 Results

3.1 Document search

Systems return docids for document search. Participants submitted 43 automatic, 15 feedback and 5 manual runs. The pools for document search included the top 75 documents from two runs per participant.

Runs were evaluated on their capability to retrieve the key pages, using traditional retrieval measures including MAP and precision at fixed ranks; NDCG is reported to take into account the graded assessments.

Automatic runs may use the query and narrative fields of the topic, but each participating group had to submit at least one run using the query field only. Table 1 shows the best automatic run from each participating group based on mean average precision. Ordering on descending NDCG instead of MAP gives slightly different results; e.g., University of Waterloo's uwKLD run (using query expansion from pseudo-relevant documents) would come second and beat their best MAP-based uwtbase run, and the Open University's ouNarrAuto run (using the narrative for automatic query expansion) would give better results than the ouTopicOnly baseline. These observed differences seem to suggest that query expansion from documents or the topic narrative is more useful when trying to find the highly relevant documents than when just finding any type of relevant document.

Feedback runs can be thought of as simulating one type of click-based system. Using click logs, it is often possible to identify that we have seen this query before, and that one or two URLs were often clicked. In that case, it would be interesting to take those URLs as relevant and perform relevance feedback. Unfortunately, we do not have CSIRO click logs, but we can use the pages field of the topic, to simulate what would happen in such a case. Feedback runs should use the query and pages fields only (not the narrative field and no manual intervention).

There are at least two methods for evaluating relevance feedback in a way that allows a comparison between feedback and non-feedback runs. The predominant method in IR is to evaluate on the residual collection, that is, feedback documents are removed from all runs and the relevance judgments. In the web search engine community, another method known as

Table 2: Document search results for the automatic or feedback run with the highest MAP from each group, using residual ranking. Feedback runs are labeled with a '*'.

Group	Run	MAP	NDCG	P@20
Waterloo	uwRF*	0.395	0.691	0.479
York	york07ed4	0.386	0.677	0.472
UvA	uams07bfbex*	0.359	0.640	0.461
RMIT	RmitQ	0.357	0.633	0.423
CAS	DocRun02	0.353	0.666	0.457
UALR	UALR07Ent2*	0.344	0.623	0.423
SJTU	SJTUEntDS02	0.337	0.629	0.417
Fudan	FDUBase	0.320	0.591	0.382
Tsinghua	THUDSFULLSR	0.310	0.602	0.390
DUT	DUTDST2	0.298	0.577	0.386
OU	ouTopicOnly	0.296	0.582	0.401
Glasgow	uogEDSCLCDIS*	0.290	0.582	0.368
Iowa	uiowa07entD2	0.276	0.555	0.354
Hyberdad	QRYBASICRUN	0.202	0.413	0.353
CSIRO	CSIROdsQonly	0.127	0.282	0.305
St. Petersburg	insu2	0.024	0.146	0.033

promotion is used — the feedback documents are moved to the top of all rankings, or placed there if they have not been retrieved.

Table 2 summarizes the results using residual-collection evaluation. For these scores, the key pages from the topics have been removed from both the qrels and the run. This allows feedback and non-feedback runs to be compared directly, but the residual-collection scores in Table 2 are not comparable to the scores in Table 1. The overall best run is a feedback run, but the difference from the best automatic run is marginal (less than 1% in MAP). Not all groups submitted feedback runs, and for some groups that did, their feedback runs were worse than their non-feedback runs.

Table 3 reports again results for feedback runs, however this time using promotion evaluation. Here, the key pages are moved to or placed at the top of the ranking. This evaluation is another way to compare feedback and non-feedback runs to each other; by comparing the scores of baseline and feedback runs both with and without promotion, you can see if the feedback is generalizing beyond the feedback documents. The table lists only results for submitted feedback runs (so automatic runs are not included in this ranking). Only for Waterloo, UvA and Glasgow, using feedback information lead to their best results; the other teams submitted non-feedback runs that performed better than their feedback runs.

Manual runs involve humans in the loop at any stage, for example composing queries from the topics, manual term expansion, relevance feedback, or manual combination of results. Although DUT submitted a highly performing manual run (run DUTDST1, with MAP 0.402 and NDCG 0.725), it did not outperform the two best automatic runs (by CAS and York University), nor did it outperform the best feedback run (by University of Waterloo).

The remainder of this section reviews some highlights from the participant papers on their document search activities. Several teams experimented with web retrieval methods based on anchor text or determining a static ranking (e.g., by pagerank or URL length), but the results seem to indicate that the CSIRO data behaves differently from Web data and that these methods are less effective than expected. RMIT mentions the fact that most links originate from the non-content part of the CSIRO pages, i.e., layout structure such as menu bars; SJTU and Tsinghua

Table 3: Document search results for the feedback run with the highest MAP from each group, after promotion of the feedback documents.

Group	Run	MAP	NDCG	P@20
Waterloo	uwRF	0.500	0.787	0.585
UvA	uams07bfbex	0.470	0.750	0.555
UALR	UALR07Ent3	0.449	0.720	0.526
DUT	DUTDST3	0.424	0.696	0.523
Glasgow	uogEDSCLCDIS	0.411	0.714	0.482
Fudan	FDUFeedT	0.399	0.693	0.498
SJTU	SJTUEntDS04	0.387	0.706	0.501
Iowa	uiowa07entD4	0.370	0.672	0.474
CSIRO	CSIROdsQfb	0.256	0.435	0.436

Table 4: Expert ranking scores. The best run in each group according to MAP is shown.

Group	Run	MAP	P@5	P@20
Tsinghua	THUIRMPDD4	0.4632	0.2280	0.0910
SJTU	SJTUEntES03	0.4427	0.2360	0.0910
OU	ouExTitle	0.4337	0.2520	0.0950
CAS	ExpertRun02	0.3689	0.2040	0.0790
CSIRO	CSIROesQnarr	0.3655	0.2240	0.0770
Wuhan	WHU10	0.3399	0.1960	0.0710
Glasgow	uogEXFeMNZcP	0.3138	0.2200	0.0800
UvA	uams07exbl	0.3090	0.2080	0.0790
DUT	DUTEXP1	0.2630	0.1400	0.0580
Fudan	FDUn7e3	0.1788	0.1440	0.0610
Beijing	PRISRR	0.1571	0.0920	0.0440
Twente	qorwnewlinks	0.1481	0.1080	0.0540
Peking	zslrun	0.0944	0.0600	0.0220
Hyberbad	AUTORUN	0.0939	0.0560	0.0330
UALR	UALR07Exp1	0.0200	0.0160	0.0130

made independently the same observation and used the percentage of links to seperate layout from content and weight the latter stronger. Tsinghua reports an improvement using Pagerank and HITS, but the improved results are lower than the Lemur language modelling baseline without static weighting reported by RMIT. The participants who used the narrative, e.g. for query expansion, report improved effectiveness over their baseline systems.

3.2 Expert search

Expert finding systems participating in the 2007 enterprise track had to return email addresses to identify candidate experts. Since no canonical list of candidate experts could be made available, the track required participants to extract the email addresses of the 'key people' from the data. Participants submitted 45 automatic, 4 feedback and 6 manual runs.

The evaluation results, summarized in Table 4, measure the quality of the ranked list of people using traditional retrieval measures including MAP and precision at fixed ranks.

Tables 6 and 5 summarize the results of the feedback and manual runs. For expert search, the best runs are manual runs, but notice how many automatic runs have outperformed the

Table 5: Expert ranking scores of feedback runs.

Group	Run	MAP	P@5	P@20
CSIRO	CSIROesQpage	0.3660	0.2040	0.0670
Iowa	uiowa07entE1	0.2828	0.1640	0.0710
Twente	feedbackrun	0.2371	0.1480	0.0650

Table 6: Expert ranking scores of manual runs.

Group	Run	MAP	P@5	P@20
OU	ouExNarrRF	0.4787	0.2720	0.0990
OU	ouExNarr	0.4675	0.2680	0.0980
DUT	DUTEXP3	0.3404	0.1840	0.0680
DUT	DUTEXP2	0.3324	0.1920	0.0640
DUT	DUTEXP4	0.1876	0.1000	0.0440
UALR	UALR07Exp3	0.1840	0.1320	0.0360

other manual and the feedback runs.

We again highlight some findings from studying the participant papers. Most participants use some form of two-stage model. Several teams (e.g., SJTU, UvA) retrieved homepages of the identified candidate names to aid in the expertise assessment. Proximity between candidate mentions and query terms seems an important factor in SJTU, Glasgow and OU results. Both CAS and Twente experimented with query-specific graphs of expert-document pairs, but results are not yet conclusive. What we can however conclude from this year's experiments is that the lack of candidate list has complicated the task significantly when compared to previous years. Almost all participants have used template matching to identify candidates from email occurrences in the corpus, sometimes including sophisticated heuristics to circumvent anti-spam measures and to exclude general group email addresses from consideration. Several participants report however that they had missed about half of the candidates that were found relevant in the assessments (with correspondingly lower effectiveness).

To validate the outcome of the experiments, we asked one science communicator to look into the highly-ranked non-relevant responses, and classify those as follows:

E: Expert, but not key contact

K: Knowledgable, but not expert

N: Not knowledgable or expert

S: Science Communicator

U: Unknown status

None of these responses has been reconsidered as a 'key contact' missing from the topic definition. For three topics authored by this science communicator, we found that the systems identified five different science communicators (S) as the experts. Two of the ranked experts were deemed knowledgeable staff members but not experts (K), and four clearly not knowledgeable (N). The remaining twenty-eight highly-ranked non-relevant responses had unknown expertise (U).

We conclude from this minor investigation that the generic methods of expert identification are not taking into account the context of the situated task - science communicators created the topic set, and would not have nominated themselves as the key contact.

4 Summary

The third year of the enterprise track has introduced the CERC corpus (CSIRO Enterprise Research Collection). The data consists of a crawl of the public-facing web of the Australian Commonwealth Scientific and Industrial Research Oganization (CSIRO). The track involved CSIRO's science communicators in the topic development process, with the goal to model accurately the search activities of real members of the enterprise.

The newly introduced document search task is motivated by a 'missing overview page' scenario, where a search is conducted to find a set of 'key pages' related to the topic in question; for example, to assist the science communicator to create the missing overview page. The topics provided a small number of example 'key pages' to facilitate experiments with relevance feedback strategies.

The expert search task follows naturally from the missing page scenario, where the 'key contacts' among CSIRO staff should be identified. As opposed to previous years, the 2007 expert search task did not provide a pre-defined list of candidates, and fewer experts were expected per topic. The expertise judgments originate from the topic authors themselves, and encode inside knowledge. For example, highly-ranked non-relevant candidate experts for some topics turned out to be science communicators and other knowledgeable people that are not seen as experts.

References

P. Bailey, N. Craswell, I. Soboroff, and A.P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference*, Singapore, July 20–24 2008. To appear.

TREC 2007 Genomics Track Overview

William Hersh[1], Aaron Cohen[1], Lynn Ruslen[1], Phoebe Roberts[2]

[1]Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA
[2]Pfizer Corp., Cambridge, MA, USA

The TREC 2007 Genomics Track employed an entity-based question-answering task. Runs were required to nominate passages of text from a collection of full-text biomedical journal articles to answer the topic questions. Systems were assessed not only for the relevance of passages retrieved, but also how many aspects (entities) of the topic were covered and how many relevant documents were retrieved. We also classified the features of runs to explore which ones were associated with better performance, although the diversity of approaches and the quality of their reporting prevented definitive conclusions from being drawn.

For the TREC 2007 Genomics Track, we undertook a modification of the question answering extraction task used in the 2006 track [1]. We continued to task systems with extracting out relevant passages of text that answer topic questions. However for this year, instead of categorizing questions by generic topic type (GTT), we derived questions based on biologists' information needs where the answers were, in part, lists of named entities of a given type. Systems were required to return a passage of text, which provided one or more relevant list items within the context of supporting text.

Similar to 2006, systems were tasked to return passages of text. Relevance judges with expertise in biological research assigned the relevant passage "answers," or items belonging to a single named entity class, analogous to the assignment of MeSH aspects in 2006. After pooling the top nominated passages as in past years, judges selected relevant passages and then assigned one or more answer entities to each relevant passage. Passages had to contain one or more named entities of the given type with supporting text that answered the given question in order to be marked relevant. Judges created their own entity list for each topic, based on the passages they judged as relevant. Passages were given credit for each relevant and supported answer. This was required because it was assumed that the passage would not answer the list entity question unless it contains an entity of the type for which the judges were looking. The experts were instructed to perform their relevance judgments in this manner.

The evaluation measures for 2007 were a refinement of the measures used in 2006. We added a new character-based mean average precision (MAP) measure (called Passage2 MAP) to compare the accuracy of the extracted answers, modified from the original measure in 2006 (called Passage MAP). Passage2 MAP treated each individually retrieved character in published order as relevant or not, in a sort of "every character is a mini relevance-judged document" approach. This was done to increase the stability of the Passage MAP measure against arbitrary passage splitting techniques. We included the 2006 passage retrieval measure as well. The Aspect MAP measure remained the same, except that instead of using assigned MeSH aspects we used the answer entities assigned by the relevance judges. We continued to use Document MAP as is, i.e., a document that contained a passage judged relevant was deemed relevant.

Documents

We used the same full-text document corpus that we assembled for the TREC 2006 Genomics Track. The documents in this corpus came from the Highwire Press (www.highwire.org) electronic distribution of journals and were in HTML format. There were about 160,000 documents in the corpus from about 49 genomics-related journals. Highwire Press agreed to allow us to include their full text in HTML format, which preserved formatting, structure, table and figure legends, etc.. In 2006, we found some known issues with the document collection:

- The collection was not complete from the standpoint of each journal. That is, there were many journals where some articles appeared in the journal but did not make it into our collection. (Neither the article nor the MEDLINE record.) This was not an issue to us, since we viewed the corpus as a closed and fixed collection.
- Some of the PMIDs in the source data from Highwire Press were inconsistent with PubMed PMIDs (see next paragraph for an explanation).
- Some of the HTML files were empty or nearly empty (i.e., only contained a small amount of meaningless text). Some of this was due to errors in our processing, but some was also related to the incorrect PMID problem of Highwire. We froze the corpus for the test collection and, since these files were small, they were unlikely to have any relevant passages or even be retrieved by most systems.

Also discovered in 2006 were some errors between the PMIDs designated by Highwire and the actual PMIDs from NLM in MEDLINE. We identified 1,767 instances (about 1% of the 162K documents) where the Highwire file PMID was invalid, in the sense that it returned zero hits when searching for it on PubMed. Some invalid PMIDs are due to the fact that the corresponding documents represented errata and author responses to comments (e.g., author replies to letters). These were assigned PMIDs in publisher-supplied data, but NLM generally does not cite them separately in PubMed, and therefore deleted the PMIDs, although they remained in publisher data. There were documents already assigned a PMID submitted by Highwire that NLM, by policy, decided not to index at all, in which case, again, NLM deleted the PMID, but it was retained in Highwire data. We also found instances of invalid PMIDs in Highwire data for documents that were cited in PubMed but with a different PMID which is absent from Highwire data; such instances could be characterized as errors. In any case, we investigated the problem of invalid PMIDs and found that for all instances we checked, the problem was the original Highwire file having an invalid PMID. In other words, invalid PMIDs were in the Highwire data, not a result of our processing. For this reason, we decided not to delete these files from the collection. They represented, in our view, normal dirty data, whether due to errors or policy differences between NLM and publishers, and should be part of what real-world systems need to be able to handle.

Since the goal of the task was passage retrieval, we developed some additional data sources that aided researchers in managing and evaluating runs. As noted below, retrieved passages could contain any span of text that did not include any part of an HTML paragraph tag (i.e., one starting with <P or </P). We also used these delimiters to extract text that was assessed by the relevance judges. Because there was much confusion in the discussion about the different types of passages, we defined the following terms:

- Nominated passage - This is the passage that systems nominated in their runs and was scored in the passage retrieval evaluation.
- Maximum-length legal span - These were all the passages obtained by the delimited text of each document by the HTML paragraph tags. As noted below, nominated passages could not cross an HTML paragraph boundary. So these spans represented the longest possible passage that could be designated as relevant. As also noted below, we built pools of these spans for the relevance judges. The judges were given the entire span if any system nominated any part of the maximum-length legal span, even if no system nominated the entire span. However, the judges did not need to designate the entire span as relevant, and could select just a part of the span to be relevant.
- Relevant passage - These were the spans that the judges designated as definitely or possibly relevant, had to contain at least one answering entity of the given type, and had entities assign to them by the expert judges. A relevant passage must consist of all or part of a maximum-length legal span.

We note some other things about the maximum-length legal spans:

- The first and last spans were delimited at the beginning and end of the file respectively.
- Other HTML tags (e.g., <b>) could occur within the spans.
- "Empty" (zero character) spans were not included.

In order to facilitate our management of the data, and perhaps be of use to participants, we created a 215-megabyte file, legalspans.txt, which included all of the maximum-length legal spans for the collection. The first span for each document included all of the HTML prior to the first <p>, which contained the HTML header information and usually was not part of any relevant passage. This file identified all of the maximum-length legal spans in all of the documents, which consisted of all spans >0 bytes delimited by HTML paragraph tags. These spans were identified by the byte character offset and length in the HTML file. The index number of the first character of the file was 0.

These span definitions can be illustrated with the example in Table 1. The last line of the following data is sample text from an HTML file hypothetically named 12345.html (i.e., having PMID 12345). The numbers above the text represent the tens (top line) and ones (middle) digits for the file position in bytes.

The maximum-length legal spans in this example are from bytes 0-4, 8-29, and 39-50. Our legalspans.txt file would include the following data in PMID, offset, and length order:

```
12345 0  5
12345 8  22
12345 39 12
```

Let us consider the span 8-29 further. This is a maximum-length legal span because there is an HTML paragraph tag on either side of it. If a system nominates a passage that exceeds these boundaries, it will be disqualified for further analysis or judgment. But anything within the maximum-length legal span, e.g. 8-19, 18-19, or 18-28, could be nominated or relevant passages.

Table 1 - Example text for span definitions.

```
000000000011111111112222222222333333333344444444445
012345678901234567890123456789012345678901234567890
Aaa. <p> Bbbbb <b>cc</b> ddd. <p><p><p> Eee ff ggg.
```

We note that it would be possible for there to be more than one relevant passage in a maximum-length legal span. While this will be unlikely, our character-based scoring approach (see below) would handle it fine. However, this was a problem for the judges as the judging interface did not support an easy way to split a judged maximum-length span into multiple relevant passages. In this case judges were instructed to include all of the relevant text within a span in the relevant passage, even if that required the inclusion of some text that the judge thought not relevant. This was most likely to be an issue in spans originating in the references section of the original documents, where two references with informative titles are separated by one or more non-relevant references.

Topics

There were 36 official topics for the track in 2007, which were in the form of questions asking for lists of specific entities. The definitions for these entity types were based on controlled terminologies from different sources, with the source of the terms depending on the entity type. We gathered new information needs from working biologists. This was done by modifying the questionnaire used in 2004 to survey biologists about recent information needs. In addition to asking about information needs, biologists were asked if their desired answer was a list of a certain type of entity, such as genes, proteins, diseases, mutations, etc., and if so, to designate that entity type. Fifty information needs statements were selected after screening them against the corpus to ensure that relevant paragraphs with named entities were present, of which 36 were used as official topics and 14 used as sample topics. Table 2 lists the 36 topics and Table 3 shows the entities and the number of topics in which they occurred.

An example of our topic development process is as follows. Suppose that the information need was:
What is the genetic component of alcoholism?
This is transformed into a list question of the form:
What [GENES] are genetically linked to alcoholism?
Answers to this question are passages that relate one or more entities of type GENE to alcoholism. For example, a valid and relevant answer to the above question would be: *The DRD4 VNTR polymorphism moderates craving after alcohol consumption.* (from PMID 11950104 for those who want to know) And the GENE entity supported by this statement would be DRD4.

Table 2 - TREC 2007 Genomics Track official topics.

<200>What serum [PROTEINS] change expression in association with high disease activity in lupus?
<201>What [MUTATIONS] in the Raf gene are associated with cancer?
<202>What [DRUGS] are associated with lysosomal abnormalities in the nervous system?
<203>What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
<204>What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?
<205>What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
<206>What [TOXICITIES] are associated with zoledronic acid?
<207>What [TOXICITIES] are associated with etidronate?
<208>What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
<209>What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?
<210>What [MOLECULAR FUNCTIONS] are attributed to glycan modification?
<211>What [ANTIBODIES] have been used to detect protein PSD-95?
<212>What [GENES] are involved in insect segmentation?
<213>What [GENES] are involved in Drosophila neuroblast development?
<214>What [GENES] are involved axon guidance in C.elegans?
<215>What [PROTEINS] are involved in actin polymerization in smooth muscle?
<216>What [GENES] regulate puberty in humans?
<217>What [PROTEINS] in rats perform functions different from those of their human homologs?
<218>What [GENES] are implicated in regulating alcohol preference?
<219>In what [DISEASES] of brain development do centrosomal genes play a role?
<220>What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?
<221>Which [PATHWAYS] are mediated by CD44?
<222>What [MOLECULAR FUNCTIONS] is LITAF involved in?
<223>Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?
<224>What [GENES] are involved in the melanogenesis of human lung cancers?
<225>What [BIOLOGICAL SUBSTANCES] induce clpQ expression?
<226>What [PROTEINS] make up the murine signal recognition particle?
<227>What [GENES] are induced by LPS in diabetic mice?
<228>What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?
<229>What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?
<230>What [PATHWAYS] are involved in Ewing's sarcoma?
<231>What [TUMOR TYPES] are found in zebrafish?
<232>What [DRUGS] inhibit HIV type 1 infection?
<233>What viral [GENES] affect membrane fusion during HIV infection?
<234>What [GENES] make up the NFkappaB signaling pathway?
<235>Which [GENES] involved in NFkappaB signaling regulate iNOS?

Table 3 - TREC 2007 Genomics Track entities, definitions, sources of term, and topics with each entity.

Entity Type	Definition	Potential Source of Terms	Topics With Entity Type
ANTIBODIES	Immunoglobulin molecules having a specific amino acid sequence by virtue of which they interact only with the antigen (or a very similar shape) that induced their synthesis in cells of the lymphoid series (especially plasma cells).	MeSH	1
BIOLOGICAL SUBSTANCES	Chemical compounds that are produced by a living organism.	MeSH	3
CELL OR TISSUE TYPES	A distinct morphological or functional form of cell, or the name of a collection of interconnected cells that perform a similar function within an organism.	MeSH	2
DISEASES	A definite pathologic process with a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown.	MeSH	1
DRUGS	A pharmaceutical preparation intended for human or veterinary use.	MEDLINEplus	2
GENES	Specific sequences of nucleotides along a molecule of DNA (or, in the case of some viruses, RNA) which represent functional units of heredity.	iHoP, Harvester	11
MOLECULAR FUNCTIONS	Elemental activities, such as catalysis or binding, describing the actions of a gene product or bioactive substance at the molecular level.	GO	2
MUTATIONS	Any detectable and heritable change in the genetic material that causes a change in the genotype and which is transmitted to daughter cells and to succeeding generations	MeSH	1
PATHWAYS	A series of biochemical reactions occurring within a cell to modify a chemical substance or transduce an extracellular signal.	BioCarta, KEGG	2
PROTEINS	Linear polypeptides that are synthesized on ribosomes and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits.	MeSH	5
STRAINS	A genetic subtype or variant of a virus or bacterium.	Ad hoc	2
SIGNS OR SYMPTOMS	A sensation or subjective change in health function experienced by a patient, or an objective indication of some medical fact or quality that is detected by a physician during a physical examination of a patient.	MeSH	1
TOXICITIES	A measure of the degree and the manner in which which something is toxic or poisonous to a living organism.	MeSH	2
TUMOR TYPES	An abnormal growth of tissue, originating from a specific tissue of origin or cell type, and having defined characteristic properties, such as a recognized histology.	MeSH	1

Submissions

Submitted runs could contain up to 1000 passages per topic in ranked order that were predicted to be relevant to answering the topic question. Passages had to be identified by the PMID, the start offset into the text file in characters, and the length of the passage in characters.

Passages were required to be contiguous and not longer than one paragraph. This was operationalized by prohibiting any passage from containing HTML markup tags, i.e., those starting with <P or </P. Any passage that included those tags was ignored in the relevance judgment process but not omitted from the scoring process. (In other words, they were not including in the pooling and judgment for creating the gold standard, but they could be scored and may include some relevant characters.) Each participating group was be allowed to submit up to three official runs, each of which was used for building the judgement pools. Each passage also needed to be assigned a corresponding rank number and value, which was used to order nominated passages for rank-based performance computations. Rank values could be integers or floating point numbers, such as confidence values.

Each submitted run had to be submitted in a separate file, with each line defining one nominated passage using the following format based loosely on trec_eval. Each line in the file had to contain the following data elements, separated by white space (spaces or a tab characters):

- Topic ID - from 200 to 235.
- Doc ID - name of the HTML file minus the .html extension. This is the PMID that has been designated by Highwire, even though we now know that this may not be the true PMID assigned by the NLM (i.e., used in MEDLINE). But this is the official identifier for the document.
- Rank number - rank of the passage for the topic, starting with 1 for the top-ranked passage and preceding down to as high as 1000.
- Rank value - system-assigned score for the rank of the passage, an internal number that should descend in value from passages ranked higher.
- Passage start - the byte offset in the Doc ID file where the passage begins, where the first character of the file is offset 0.
- Passage length - the length of the passage in bytes, in 8-bit ASCII, not Unicode.
- Run tag - a tag assigned by the submitting group that should be distinct from all the group's other runs (and ideally any other group's runs, so it should probably have the group name, e.g., OHSUbaseline).

Here is an example of the submission file format:

```
200 12474524 1 1.0   1572 27  tag1
200 12513833 2 0.373 1698 54  tag1
200 12517948 3 0.222 99   159 tag1
201 12531694 1 0.907 232  38  tag1
201 12545156 2 0.456 789  201 tag1
```

A Perl script that checked runs to insure that the submission file was in the proper format was available (check_genomics.pl). Runs also needed to include a "dummy" passage for any topic for which no passages were retrieved. It was recommended that the dummy passage use "0" as a docid, "0" as the passage start, and "1" as the passage length. This worked for the Perl script and

did not correspond to a document in the collection.

Runs were also classified based on amount of intervention in converting topics to queries. We adopted the "usual" TREC rules (detailed at http://trec.nist.gov/act_part/guidelines/trec8_guides.html) for categorizing runs:

- Automatic - no human modification of topics into queries for your system whatsoever
- Manual - human modification of queries entered into your system (or any other system) but no modification based on results obtained (i.e., you cannot look at the output from your runs to modify the queries)
- Interactive - human interaction with the system, including modification of the queries or the system after viewing the output from your system or any other system (i.e., you look at the output from the topics and corpus and adjust your system to produce different output)

Relevance Judgments

The expert judging for this evaluation used the pooling method, with passages corresponding to the same topic ID pooled together. The judges were presented with the text of the maximum-length legal span containing each pooled passage, with pool composed of the top ranked 1000 passages for each topic. They then evaluated the text of the maximum-length legal span for relevance, and identified the portion of this text that contains an answer. This could be all of the text of the maximum legal span, or any contiguous substring. If a maximum legal span contained more than one relevant passage, judges were instructed to select the minimum contiguous passage that contained all relevant passages, even if the passages were separated by irrelevant text. Maximum legal spans comprised of the journal article bibliography frequently generated multiple relevant sub-passages that needed to all be included in the singe designated passage.

Judges were recruited from the institutions of track participants and other academic or research centers. They were required to have significant domain knowledge, typically in the form of a PhD in a life science. They were trained using a 12-page manual and a one-hour videoconference, with the option of testing out of the videoconference by successfully judging a mini-topic based on a practice topic from 2006 made up of an equal mix of definitely, possibly, and not relevant maximum-length legal spans. The self-training option had the unexpected benefit of highlighting and correcting potential problems with the judging tool or ambiguous guidelines before judging began in earnest. The training manual is on the track Web site at: http://ir.ohsu.edu/genomics/2007judgeguidelines.pdf

In summary, judges were given the following instructions:

1. Review the topic question and identify key concepts.
2. Identify relevant paragraphs and select minimum complete and correct excerpts.
3. Develop controlled vocabulary for entities based on the relevant passages and code entities for each relevant passage based on this vocabulary.

Judgments were made using database files created and accessed via the OpenOffice Base application. As shown in Figure 1, judges were presented passages as a form view of individual

records in the database with the topic, question, and text of the full-text legal passage. If part or all of the passage was relevant, the judges then:

1. Selected the level of relevance ("Definitely Relevant" or "Possibly Relevant").
2. Copied the relevant portion of the passage from the passage plain text field into the answer text box.
3. Selected entities (ENTITY1, ENTITY2, etc.) they had added using the Add Entities form (not shown).

A gold standard was created by extracting out the relevance passages and entities from the database file for each topic. Selected relevant text was transformed into file character offset and length using a text alignment algorithm. A summary of the gold standard developed from the results of the judging process is shown in Table 4. Topics ranged from a low of 1 relevant passage to a high of 377. Individual topics had a range of 1 to 300 relevant entities, with an average ranging between 1.0 to 3.5 entities assigned per relevant passage.

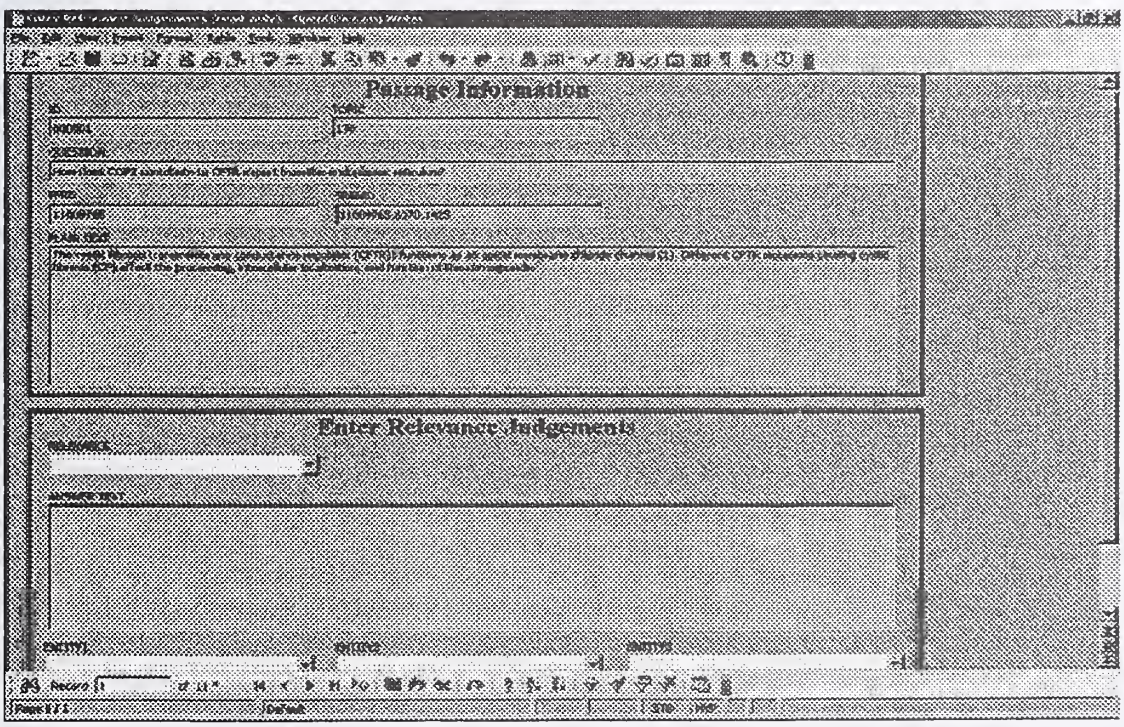


Figure 1 - Passage judgment form.

Table 4 - Relevant passages, relevant documents, mean and standard deviation (SD) of relevant passage length, number of aspects, and mean number of aspects per relevant passage.

Topic	Relevant Passages	Relevant Documents	Mean Relevant Passage Length	SD of Relevant Passage Length	Aspects	Mean Aspects Per Relevant Passage
200	320	193	2380.58	5387.02	300	2.15
201	37	12	1701.86	2894.64	7	1.16
202	53	43	522.77	293.60	28	1.45
203	321	147	2163.60	4237.72	245	1.91
204	164	74	1989.90	4670.61	36	1.79
205	93	65	788.67	1277.35	17	1.23
206	38	19	363.79	362.85	24	1.87
207	15	12	357.60	671.28	8	1.07
208	22	16	615.36	317.50	13	1.23
209	78	11	1239.63	720.81	15	1.50
210	71	57	669.79	623.70	21	1.10
211	57	42	191.68	217.10	29	1.14
212	358	133	1165.97	969.94	142	2.16
213	377	185	456.94	594.39	165	1.88
214	209	98	414.91	1095.21	54	1.42
215	137	73	750.96	580.54	80	1.66
216	42	34	1058.12	3141.51	13	1.12
217	38	34	1491.18	1019.48	34	1.03
218	163	74	632.23	635.55	80	1.28
219	22	16	623.64	503.66	43	3.41
220	16	6	425.75	218.10	6	1.75
221	183	87	1373.32	1705.58	108	1.44
222	57	42	1249.51	914.23	72	2.18
223	18	8	269.72	138.24	12	1.17
224	3	3	1009.33	666.59	1	1.00
225	1	1	745.00	0.00	1	1.00
226	152	57	753.82	1648.91	18	2.25
227	281	172	1307.02	863.14	183	2.25
228	15	14	632.20	413.79	13	1.87
229	150	57	528.81	978.41	34	1.79
230	82	29	1186.65	933.99	25	1.30
231	16	13	472.00	406.56	7	1.06
232	93	57	388.57	907.63	49	1.12
233	19	16	1186.68	1070.54	1	1.00
234	609	483	1777.02	3124.85	577	3.24
235	182	107	1963.25	1737.40	141	2.54
Mean	124.8	69.2	968.0	1276.2	72.3	1.63

Evaluation Measures

For this year's track, there were three levels of retrieval performance measured: passage retrieval, aspect retrieval, and document retrieval. Each of these provides insight into the overall performance for a user trying to answer the given topic questions. Each was measured by some variant of MAP. We again measured the three types of performance separately. There was not any summary metric to grade overall performance. A Python program to calculate these

measures (http://ir.ohsu.edu/genomics/trecgen2007_score.py) with the appropriate gold standard data files is available.

Passage-level retrieval performance - character-based MAP

The original passage retrieval measure for the 2006 track was found to be problematic in that non-content manipulations of passages had substantial effects on Passage MAP, with one group claiming that breaking passages in half with no other changes doubled their (otherwise low) score. To this end, we defined an alternative measure (Passage2 MAP) that calculated MAP as if each character in each passage were a ranked document. In essence, the output of passages was concatenated, with each character being from a relevant passage or not. We used Passage2 MAP as the primary passage retrieval evaluation measure in 2007.

The original Passage MAP measure was also calculated. This measure computed individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track [2]. For each nominated passage, a fraction of characters overlaps with those deemed relevant by the judges in the gold standard. At each relevant retrieved passage, precision was computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, remaining relevant passages that were not retrieved at all were added into the calculation as well, with precision set to 0 for relevant passages not retrieved. Then the mean of these average precisions over all topics was calculated to compute the mean average passage precision.

Aspect-level retrieval performance - aspect-based MAP

Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics. For 2007, the aspects were the different named entities of the given type for each question. To compute this, for each submitted run, the ranked passages were transformed to two types of values, either:

- the aspects of the gold standard passage that the submitted passage overlaps with, or
- not relevant

This resulted in an ordered list, for each run and each topic, of aspects and not-relevant. Because we were uncertain of the utility for a user of a repeated aspect (e.g., same aspect occurring again further down the list), we discarded them from the output to be analyzed and only kept the first appearance of an aspect. For these remaining aspects of a topic, we calculated Aspect MAP similar to how it was calculated for documents.

Document-level retrieval performance - document-based MAP

For the purposes of this measure, any PMID that had a passage associated with a topic ID in the set of gold standard passages was considered a relevant document for that topic. All other documents were considered nonrelevant for that topic. System run outputs were similarly collapsed, with the documents appearing in the same order as the first time the corresponding PMID appears in a nominated passage for that topic. For a given system run, average precision

was measured at each point of correct (relevant) recall for a topic, with Document MAP being the mean of the average precision values across topics.

Results

A total of 66 runs were submitted by 27 groups. Of the submitted runs, 49 were classified as automatic, 8 as manual, and 9 as interactive. Appendix 1 lists the type and description of each submitted run. Table 5 lists the performance statistics for all of the runs and for the runs subdivided by categories. Appendix 2 shows the overall scores for each run, sorted by each measure.

We also measured correlation of the four measures (Passage2 MAP, Passage MAP, Aspect MAP, and Document MAP) for each run. As is seen in Table 6, the new Passage2 MAP measure was highly correlated with Aspect MAP and Document MAP ($R^2 > 0.8$), with the older Passage MAP measure less correlated.

Table 5 - Descriptive statistics for all runs and subdivided by categories.

All	**Passage2 MAP**	**Passage MAP**	**Aspect MAP**	**Document MAP**
Min	0.0008	0.0029	0.0197	0.0329
Median	0.0377	0.0565	0.1311	0.1897
Mean	0.0398	0.0560	0.1326	0.1862
Max	0.1148	0.0976	0.2631	0.3286
Automatic				
Min	0.0008	0.0029	0.0197	0.0329
Median	0.0391	0.0587	0.1272	0.1954
Mean	0.0421	0.0582	0.1286	0.1891
Max	0.1097	0.0976	0.2494	0.3105
Manual				
Min	0.0032	0.0177	0.0204	0.0541
Median	0.0149	0.0276	0.1136	0.1696
Mean	0.0169	0.0328	0.0964	0.1526
Max	0.0458	0.0654	0.1503	0.2309
Interactive				
Min	0.0268	0.0394	0.1411	0.0892
Median	0.0384	0.0620	0.1865	0.1940
Mean	0.0475	0.0648	0.1868	0.2007
Max	0.1148	0.0968	0.2631	0.3286

Table 6 - MAP measure correlation matrix using Pearson correlation coefficient (all values significantly different from 0 with a significance level $p < .05$).

MAP	Passage2	Passage	Aspect	Document
Passage2	1	0.656	0.845	0.812
Passage	0.656	1	0.591	0.830
Aspect	0.845	0.591	1	0.775
Document	0.812	0.830	0.775	1

We attempted to analyze the automatic runs to discern whether there was any association between individual methods used (as reported in conference notebook papers and not final proceedings papers) and overall performance as measured by Passage2 MAP. The task was challenging since groups approached entity-based question answering with a myriad of methods. Submissions employed multiple approaches for query expansion, various levels of passage retrieval granularity, varying IR models with many different scoring schemes, and several methods of post-processing. In all, these runs exercised over 70 different features, any of which could have impacted Passage2 MAP separately or in combination. With so many features and a limited number of runs (43) having a corresponding notebook paper describing methods, data sparseness was an issue. We therefore distilled the features into high-level categories, or meta-features shown in Table 7.

If retrieval was done in two steps, e.g., to pare down results for secondary concept-based retrieval, and each step uses a different level of granularity for passage retrieval, we chose the granularity level of the second one in order to focus on features of the core strategy rather than a filtering step designed to reduce computer processing burdens. This only affected runs from ASU and Tsinghua. Each run was represented as a vector of meta-features deemed either present (1) or absent (0). The decision was binary since there is no uniform way to say something was partially done, such as in the case of fusion runs, or to weigh the impact of a paring step for concept-based retrieval. If fusion was done, the union of features used by the individual component runs was chosen since they presumably all contributed to the ultimate result. All meta-features were given the same weight. A hierarchical clustering algorithm using a centroid similarity metric grouped runs based on their meta-features, as shown in Figure 2. Runs were clustered as a "group" when their correlation was > 70%. Clustering using Dice's coefficient similarity measure produced similar results.

Originally, we had also clustered by statistical rank group. This simply revealed that many different paths lead to roughly the same performance, and was less informative as far as whether individual meta-features had an overall positive or negative impact. Although not used for clustering, the rank group is included in the heat map to indicate how a run performed. Given that the MAP measures were highly correlated (see Table 6), only Passage2 MAP rank is shown for clarity.

Table 7 - Meta-features of runs.

Meta-Feature Name	Description
SynExp	query expansion with synonyms
OrthExp	query expansion with orthographic variants using any source or method
ParGranularity	passage retrieval by paragraph
SentGranularity	passage retrieval by sentence
BlckGranularity	passage retrieval by block, including blocks of words or sentences greater than a single sentence yet smaller than a paragraph
ConcptIR	concept-based retrieval - a general retrieval strategy attempting to align concepts and, for some runs, relationships between a topic and a passage; uses external knowledge sources such as UMLS as a source of "concepts"; and finds concepts in the results as an inherent part of the retrieval process rather than a post-processing step to "trim" a passage
TermIR	term-based retrieval – a general retrieval strategy focusing on terms rather than concepts
FusionIR	fusion - combining results from 2 or more systems regardless of fusion operator used
TfIdfIR	passage retrieval using a vector space model with any variant of TF-IDF
OkapiIR	passage retrieval using a vector space model with any variant of Okapi
DfrIR	passage retrieval using a vector space model with any variant of divergence from randomness (DFR)
LatentSemIR	passage retrieval using a vector space model with any variant of latent semantic analysis
LmIR	passage retrieval using any language model
Feedback	feedback using pseudo-relevance feedback or a custom method
FilterPostProc	filter post-processing - removing passages for any reason
TrimPostProc	passage trimming - post-processing of passages by removing sentences from the ends regardless of method

Overview of the TREC 2007 Legal Track

Stephen Tomlinson, stomlins@opentext.com
Open Text Corporation, Ottawa, Ontario, Canada

Douglas W. Oard, oard@umd.edu
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA

Jason R. Baron, jason.baron@nara.gov
National Archives and Records Administration
Office of the General Counsel, Suite 3110, College Park, MD 20740, USA

Paul Thompson, Paul.Thompson@dartmouth.edu
Institute for Security Technology Studies
Dartmouth College, Hanover, NH 03755, USA

Abstract

TREC 2007 was the second year of the Legal Track, which focuses on evaluation of search technology for discovery of electronically stored information in litigation and regulatory settings. The track included three tasks: Ad Hoc (i.e., single-pass automatic) search, Relevance Feedback (two-pass search in a controlled setting with some relevant and nonrelevant documents manually marked after the first pass) and Interactive (in which real users could iteratively refine their queries and/or engage in multi-pass relevance feedback). This paper describes the design of the three tasks and analyzes the results.

1 Introduction

The use of information retrieval techniques in law has traditionally focused on providing access to legislation, regulations, and judicial decisions. Searching business records for information pertinent to a case (or "discovery") has also been important, but searching records in electronic form was until recently the exception rather than the norm. The goal of the Legal Track at the Text Retrieval Conference (TREC) is to assess the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records, an issue of increasing importance given the vast amount of information stored in electronic form to which access is increasingly desired in the context of current litigation. Ideally, the results of a study of how well comparative search methodologies perform when tasked to execute types of queries that arise in real litigation will serve to better educate the legal community on the feasibility of automated retrieval as well as its limitations. The TREC Legal Track was held for the first time in 2006, when 6 research teams participated in an ad hoc retrieval task. This year, 13 research teams participated in the track, which consisted of three tasks: 1) Ad Hoc, 2) Interactive, and 3) Relevance Feedback.

The results of the Legal Track are especially timely and important given recent changes in the U.S. Federal Rules of Civil Procedure that went into effect on December 1, 2006. The amended rules introduce a new category of evidence, namely, "Electronically Stored Information" (ESI) in "any medium," intended to stand on an equal footing with existing rules covering the production of "documents." Against the backdrop of the Federal Rules changes, the status quo in the legal profession, even in large and complex litigation, is

continued reliance on free-text Boolean searching to satisfy document (and now ESI) production demands [6]. An important aspect of e-discovery and thus of the TREC Legal Track is an emphasis on recall over precision. In light of the fact that a large percentage of requests for production of documents (and now ESI) routinely state that "all" such evidence is to be produced, it becomes incumbent on responding parties to attempt to maximize the number of responsive documents found as the result of a search.

The key goal of the TREC Legal Track is to apply objective benchmark criteria for comparing search technologies, using topics that approximate how real lawyers would go about propounding discovery in civil litigation, and a large, representative (unstructured and heterogeneous) document collection. Given the reality of the use of Boolean search in present day litigation, comparing the efficacy of Boolean search using negotiated queries with alternative methods is of considerable interest. The Legal Track has shown that alternative methods do identify many relevant documents that were missed by a reference implementation of a Boolean search, though no single alternative method has yet been shown to consistently outperform Boolean search without increasing the number of documents to review.

The remainder of this paper is organized as follows. Section 2 describes the Ad Hoc task, Section 3 describes the Interactive and Relevance Feedback tasks, Section 4 lists the individual topic results, Section 5 summarizes the workshop discussions and analysis conducted after the conference, and Section 6 concludes the paper.

2 Ad Hoc Task

In the Ad Hoc task, the participants were given requests to produce documents, herein called "topics", and a set of documents to search. The following sections provide more details, but an overview of the differences from the previous year is as follows:

- At the time of topic release, the B value (the number of documents matching the final negotiated Boolean query) was provided for each topic in 2007, along with an alphabetical list (by document-id) of the documents matching the Boolean query (the "refL07B" run) for optional use by participants.
- A new evaluation measure, Estimated Recall@B, where B is the number of documents matching the Boolean query, was established as the principal measure for the track (although other measures are also reported). The legal community is interested in knowing whether additional relevant documents (those missed by a Boolean query) can be found for the same number of retrieved documents.
- A new sampling method (herein called "L07") was used to produce estimates of the main measure for each topic for all submitted runs. All runs submitted to the Ad Hoc task were pooled this year, and all pooled runs were treated equally by the sampling procedure.
- The new topics were vetted to ensure that the B value for any topic was in the 100 to 25,000 range. (In 2006, B ranged from 1 to 128,195.)
- Participating teams were allowed to submit up to 25,000 documents for each topic (up from 5,000 in 2006).
- To facilitate cross-site comparisons, a "standard condition" run which just used the (typically one-sentence) request text field was requested from all groups. Additional runs which used other topic fields were also welcome, and encouraged.
- Three different Boolean queries were provided for each topic (defendant, plaintiff and final). In 2006, the plaintiff and final queries had (usually) been the same.

2.1 Document Collection

The 2007 Legal Track used the same collection as the 2006 Legal Track, the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0 (referred to here as "IIT CDIP 1.0") which is based on documents released under the tobacco "Master Settlement Agreement" (MSA). The MSA settled a range of lawsuits by the Attorneys General of several US states against seven US tobacco organizations (five tobacco companies and two research institutes). One part of this agreement required those organizations to make public all documents produced in discovery proceedings in the lawsuits by the states, as well as all documents produced in a number of other smoking and health-related lawsuits. Notable among the provisions is that the organizations were required to provide to the National Association of Attorneys General (NAAG) a copy of metadata and the scanned documents from the websites, and are forbidden from objecting to any subsequent distribution of this material.

The University of California San Francisco (UCSF) Library, with support from the American Legacy Foundation, has created a permanent repository, the Legacy Tobacco Documents Library (LTDL), for tobacco documents [10]. The IIT CDIP 1.0 collection is based on a snapshot, generated between November 2005 and January 2006, of the MSA subcollection of the LTDL. The snapshot consisted of 1.5 TB of scanned document images, as well as metadata records and Optical Character Recognition (OCR) produced from the images by UCSF. The IIT CDIP project subsequently reformatted the metadata and OCR, combined the metadata with a slightly different version obtained from UCSF in July 2005, and discarded some documents with formatting problems, to produce the IIT CDIP 1.0 collection [8]. The IIT CDIP 1.0 collection consists of 6,910,192 document records in the form of XML elements.

IIT CDIP 1.0 has had strengths and weaknesses as a collection for the Legal Track. Among the strengths are the wide range of document genres (including letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, and email) and the large number of documents. Among the weaknesses are that the documents themselves were released as a result of tobacco-related discovery requests, and thus may exhibit a skewed topic distribution when compared with the larger collections from which they were initially selected. See the 2006 TREC Legal Track overview paper for additional details about the IIT CDIP 1.0 collection [3].

2.2 Topics

Topic development in 2007 continued to be modeled on U.S. civil discovery practice. In the litigation context, a "complaint" is filed in court, outlining the theory of the case, including factual assertions and causes of action representing the legal theories of the case. In a regulatory context, often formal letters of inquiry serve a similar purpose by outlining the scope of the proposed investigation. In both situations, soon thereafter one or more parties create and transmit formal "requests for the production of documents" to adversary parties, based on the issues raised in the complaint or letter of inquiry. See the TREC 2006 Legal Track overview for additional background [3].

A survey of case law issued subsequent to the adoption of the new Federal Rules of Civil Procedure in December 2006 suggests that increasing attention is being paid by judges and lawyers to the idea of adversaries in litigation negotiating some form of "search protocol," including coming to consensus on what keywords will be used to search for relevant documents. In one reported case, a judge suggested to the parties that they reach consensus on what form of Boolean queries should be used [13]. In another case, a judge urged the parties to reflect upon recent scholarship discussing the use of "concept searches" to supplement traditional "keyword" searching [7, 9]. Although it remains unclear whether and to what extent lawyers are fully incorporating Boolean and other operators (e.g., proximity operators) in their proposed searches, as an example of best practices the TREC 2007 Legal Track chose to highlight the importance of negotiating Boolean queries by including for each newly created topic a three-stage Boolean query negotiation, consisting of (i) an initial Boolean query[1] as proposed by the receiving party on a discovery request, usually reading the request narrowly; (ii) a "counter"-proposal by the propounding party, usually including a broader set

[1]Although often referred to as "Boolean," these queries contain additional operators (e.g., proximity and truncation operators) that are commonly found in the query languages of commercial search systems that employ Boolean logic.

of terms; and (iii) a final "negotiated" query, representing what was deemed the consensus arrangement as agreed to by the parties without resort to further judicial intervention.

For the TREC 2007 Legal Track, four new hypothetical complaints were created by members of the Sedona Conference® Working Group on Electronic Document Production, a group of lawyers who play a leading role in the development of professional practices for e-discovery. These complaints described: (1) a wrongful death and product liability action based on the use of a certain type of radioactive phosphate resulting in contaminated candy and drinking water; (2) a patent infringement action on a device going by the name "Suck out the Bad, Blow in the Good," designed to ventilate smoke; (3) a shareholder class action suit alleging securities fraud and false advertising in connection with a fictional "Smoke Longer, Feel Younger" campaign relying on "60s-era folk music;" and (4) a fictional Justice Department antitrust investigation looking in to a planned merger and acquisition of a casualty and property insurance company by a tobacco company. As in 2006, in using fictional names and jurisdictions, the track coordinators attempted to ensure that no third party would mistake the academic nature of the TREC Legal Track for an actual lawsuit involving real-world companies or individuals, and any would-be link or association with either past or present real litigation was entirely unintentional.

For each of these four complaints, a set of topics (formally, "requests to produce") were initially created by the creator of the complaint, and revised by the track coordinators. The final topic set contained 50 topics, numbered from 52 to 101. An XML formatted version of the topics (fullL07_v1.xml) was created for (potentially automated) use by the participants.

2.3 Participation

12 research teams participated in this year's Ad Hoc task. The teams experimented with a wide variety of techniques including the following:

- Carnegie Mellon University: structured queries, Indri operators, Dirichlet smoothing, Okapi BM25, boolean constraints, wildcards.
- Dartmouth College: Combination of Expert Opinion (CEO) algorithm, Lemur/Indri, Lucene.
- Fudan University: Indri 2.3, Yatata, word distribution model, corpus pre-processing methods, query expansion, query shrink.
- Open Text Corporation: negotiated boolean queries, defendant boolean, plaintiff boolean, word proximity distances, vector query runs, blind feedback, fusion.
- Sabir Research, Inc.: SMART 16.0, statistical vector space model, ltu.Lnu weighting, Rocchio feedback weighting.
- University of Amsterdam: query formulations, run combinations, LUCENE engine version 1.9, vector-space retrieval model, parsimonious language modeling techniques.
- The University of Iowa (Eichmann): analysis of OCR, 3-4 ngram analysis, translation of boolean query, pseudo-relevance feedback on persons (authors, recipients and mentions) and production boxes.
- The University of Iowa (Srinivasan): Lucene library, Okapi reranking, metadata, wildcard expansion, blind feedback, query reduction.
- University of Massachusetts, Amherst: Indri, term dependence, Markov Random Field (MRF) model, pseudo-relevance feedback, Latent Concept Expansion (LCE), phrase dictionaries, synonym classes, proximity operators.
- University of Missouri, Kansas City: query formulations, vector space model, language model, Lucene, query expansion model, conceptual relevance framework.

- University of Waterloo: Wumpus search engine, cover density ranking, Okapi BM25 ranking, boolean terms, character 4-grams, pseudo-relevance feedback, logistic regression, fusion, CombMNZ combination method, proximity-ranked boolean queries, relaxed boolean.
- Ursinus College: document normalization, log normalization, power normalization, cosine normalization, enhanced OCR error detection, generalized vector space retrieval, query pruning.

The teams submitted a total of 68 experimental runs by the Aug 5, 2007 deadline (each team could submit a maximum of 8 runs). Please consult the individual team papers in the TREC proceedings for the details of the experiments conducted. Also, please check the track web site [1] for the slides of many of the participant presentations at the conference, along with links to the aforementioned individual team members' papers in the TREC proceedings.

2.4 Evaluation

2.4.1 Background on Estimating Precision and Recall

The most straightforward way to produce an unbiased estimate of the number of relevant documents retrieved would be to use simple random sampling (i.e., sampling in which all samples have an equal chance of being selected). Unfortunately, for our purpose, the individual estimates would usually be too inaccurate. For example, suppose the target collection has 7 million documents, and for a particular topic 700 of these are relevant. Suppose further that we have the resources to judge 1,000 documents. If we pick those 1,000 documents from a simple random sample of the collection, most likely 0 of the documents will be judged relevant, producing an estimate of 0 relevant documents, which is far too low. If 1 of the documents were to be judged relevant, then we would produce an estimate of 7,000 relevant documents, which is far too high.

TREC evaluations have typically dealt with this issue by using an extreme variant of stratified sampling. The primary stratum, known as the pool, is typically the set of documents ranked in the top-100 for a topic by the participating systems. Traditionally, all of the documents in the pool are judged. Contrary to the usual approach to stratified sampling, typically none of the unpooled documents are judged (these documents are just assumed non-relevant). For the older TREC collections of about 500,000 documents, [15] found that the results for comparing retrieval systems are reasonably reliable, even though that study also found that probably only 50%-70% of relevant documents for a topic were assessed, on average.

Traditional pooling can be too shallow for larger collections. As the judging pools have become relatively shallower, either from TREC collections becoming larger and/or the judging depth being reduced, concerns have been expressed with the reliability of results. For example, [4] recently reported bias issues with depth-55 judging for the 1 million-document AQUAINT corpus, and [12] estimated that fewer than 20% of the relevant documents were judged on average for the 7 million-document TREC 2006 Legal Track test collection. The TREC 2006 Terabyte Track [5] experimented with taking simple random samples of 200 documents from (up to) depth-1252 pools, and estimated the average precision score for each run based on this deeper pooling by using the "inferred average precision" (infAP) measure suggested by [14]. They found that infAP scores were highly correlated with Mean Average Precision (MAP) scores based on traditional depth-50 pooling.

2.4.2 The L07 Method

The L07 method for estimating recall and precision was based on how the recall and precision components are estimated in the infAP calculation. What distinguishes the L07 method is support for much deeper pooling by sampling higher-ranked documents with higher probability. For legal discovery, recall is of central concern. It was found last year by [12] that marginal precision exceeded 4% on average even at depth 9,000 for standard vector-based retrieval approaches. Hence we used depth-25000 pooling this year to get better coverage of the relevant documents. A simple random sample of a depth-25000 pool, however, would be unlikely to produce accurate estimates for recall at less deep cutoff levels. Hence we sampled higher-ranked

documents with higher probability in such a way that recall estimates at all cutoff levels up to max(25000,B) should be of similar accuracy. (Details are provided in the following sections.)

The L07 method was developed independently from the similar "statAP" method evaluated by Northeastern University in the TREC 2007 Million Query Track [2]. (The common ancestor was the infAP method, which also came from Northeastern.) Both methods associate a probability with each document judgment. The differences are in how the probabilities are assigned (which should not matter on average) and in the measures being estimated (we are estimating the recall and precision of a set, whereas statAP is estimating the "average precision" measure which factors in the ranks of the relevant documents). The L07 formulas are provided below, but please consult the Northeastern work for a more thorough discussion of the theoretical underpinnings of measure estimation than we aim to provide here.

2.4.3 Ad Hoc Task Pooling

As stated earlier, a total of 68 runs were submitted by the 12 research teams for the Ad Hoc task by the Aug 5, 2007 deadline. Each run included as many as 25,000 documents (sorted in a putative best-first order) for each of the 50 topics. All submitted runs, plus a 69th run described below, were pooled to depth 25,000 for each topic and then each pool was sampled. The pool sizes before sampling ranged from 195,688 (for topic 76) to 476,252 (for topic 84). (The pool sizes for all of the topics are listed in Section 4.)

The initial plan (given in the Ad Hoc task guidelines) was to assign judging probability p(d) = min(C / hiRank(d), 1) to each submitted document d, where hiRank(d) is the highest (i.e., best) rank at which any submitted run retrieved document d, and C is chosen so that the sum of all p(d) (for all submitted documents d) was the number of documents that could be judged (typically 500). It was hoped that C would be at least 10 for all topics, so that we would have the accuracy of at least 10 simple random sample points for estimates at all depths. After the runs came in, it turned out the C values would range from only 1.6 to 3.3 if judging only 500 documents, substantially limiting the accuracy of the estimates of all measures.

Running some experiments, it turned out for specific depths we could get greater accuracy. For example, if all resources went to a simple random sampling for estimating precision at depth-B, we could get the accuracy of at least 17 sample points for each topic. If instead all resources were directed to just depth-25000, we could get at least 26 sample points for each topic. Of course, if we targeted just one deep measure, we wouldn't have a lot of top-documents for training future systems or for contrasting our measure with traditional rank-based IR measures. Experiments also found that if we just did traditional depth-k pooling, we could only go to at least depth-12 for each topic. But if all resources went to top-12 documents, we wouldn't have the ability to estimate deeper measures.

The sampling process that we ultimately adopted was a hybrid of all of the above. The final p(d) formula for the probability of judging each submitted document d was as follows:

```
If (hiRank(d) <= 5) { p(d) = 1.0; }
Else if (hiRank(d) <= B) { p(d) = min(1.0, ((5/B)+(C/hiRank(d)))); }
Else { p(d) = min(1.0, ((5/25000)+(C/hiRank(d)))); }
```

This formula causes the the first judging bin of 500 documents to contain the top-5 documents from each run, and it causes measures at depths B and 25000 to have the accuracy of approximately 5+C simple random sample points. Measures at other depths will have the accuracy of approximately (at least) C simple random sample points. If C is set to the largest multiple of 0.01 which produces a bin of at most 500 documents, C ranges from 0.34 (topic 82) to 2.42 (topic 76). So by just dropping C by approximately 1 compared to the original plan, we gained more top document judging and at least 5-sample accuracy for depth-B and depth-25000.

To allow for the possibility that some assessors could judge more than 500 documents, the above process was adapted to have a first bin of approximately 500 documents and 5 additional bins of approximately 100 documents each, using the following approach. The C values were set so that the p(d) values would sum to 1,000, and an initial draw of approximately 1000 documents was done. Then the C values were set so that the p(d) values would sum to 900, and approximately 900 documents were drawn from the initial draw of 1000 (using the ratio of the probabilities); the approximately 100 documents that were not drawn became

"bin 6". This process was repeated to create "bin 5", "bin 4", "bin 3" and "bin 2". The approximately 500 documents drawn in the last step became "bin 1".

When the judgments were received from the assessors (as described in the next section), the final p(d) values were based on how many bins the assessor had completed (e.g., if 3 bins had been completed, then the p(d) values from choosing C so that the p(d) sum to 700 were used). If there had been partial judging of deeper bins, the judged documents from these bins were also kept, but with their p(d) reset to 1.0. Note that if the 1st bin was not completed, the topic had to be discarded. For each completed topic, the final number of assessed documents and corresponding C values are listed in Section 4.

Two "runs" deserve special mention. First, the reference Boolean run (refL07B), which would have been the 69th run, was not included in the pooling because it had been created by simply resorting one of the pooled runs (otL07fb) alphabetically by docno. Instead, a 69th run called randomL07 was created, which for each topic had 100 randomly chosen documents that were not retrieved by any of the other 68 runs for the topic. We only included 100 random documents per topic, not 25000, to reduce the number of judgments taken away from submitted runs. After the draw, it turned out that the 1st bin of 500 documents to be judged contained between 5 and 15 random documents (average 9.38).

2.5 Relevance Judgments

For the TREC 2007 Legal Track, the track coordinators primarily sought out second-year and third-year law students who would be willing to volunteer as assessors in order to fulfill a law school requirement or expectation to perform some form of pro bono service to the larger community. Based on a nationwide solicitation in mid-August 2007, we received an enthusiastic response from students at a variety of U.S. law schools. All 50 new Ad Hoc task topics for the second year were assigned to assessors, but judgments for 7 topics were not available in time for use in the evaluation.[2] Most of the assessors (42) were law students from a wide variety of institutions: Loyola-L.A. (23 volunteers), University of Indiana-Indianapolis (5), George Washington (3), Case Western Reserve (3), Loyola-New Orleans (2), Boston University (2), University of Dayton (2), University of Maryland (1), and University of Texas (1). Additionally, one Justice Department attorney and one archivist on staff in NARA's Office of the General Counsel participated.

This year, the assessors used a Web-based platform developed by NIST that was hosted at the University of Maryland to view scanned documents and to record their relevance judgments. Assessors found the interface easy to navigate, with the only reported problem being a technical one involving an inability to read or advance screens properly (due to use of a Web browser other than Firefox, the only one supported). Each assessor was given a set of approximately 500 documents to assess, which was labeled "Bin 1." Additional bins 2 through 6, each consisting of 100 documents, were available for optional additional assessment, depending on willingness and time. (It turned out that 8 of the assessors completed at least 1 of the optional bins, and 5 assessors completed all 5 optional bins.) In total, 24,404 judgments were produced for the 43 topics. The assessment phase extended from August 17, 2007 through September 24, 2007.

As in 2006, we provided the assessors with an updated "How To Guide" that explained that the project was modeled on the ways in which lawyers make and respond to real requests for documents, including in electronic form. Assessors were told to assume that they had been requested by a senior partner, or hired by a law firm or another company, to review a set of documents for "relevance." No special, comprehensive knowledge of the matters discussed in each complaint was expected (e.g., no need to be an expert in federal election law, product liability, etc.). The heart of the exercise was to look for relevant and nonrelevant documents within a topic. Relevance, consistent with all known legal definitions from Wigmore to Wikipedia, was to be defined broadly. Special rules were to be applied for any document of over 300 pages. The same process was used for assessment for the interactive and relevance feedback tasks (which had different topics, as described below). See the TREC 2006 Legal Track overview for additional background (including a discussion of inter-assessor agreement which was measured in 2006 but not in 2007) [3].

On the whole, there was less confusion reported by assessors as to the definitional scope of the assigned

[2]The assessments for one additional topic were completed after the deadline, and are available for research use, but results are reported in this paper for 43 topics.

topics in 2007 than in 2006, although some questions did arise. For example, for topic 75 ("All documents that memorialize any statement or suggestion from an elected federal public official that further research is necessary to improve indoor environmental air quality"), the assessor questioned whether "memorialize" would be broad enough to include a mere reference to a Superfund bill, without a quotation as such from the official. We responded that a quotation or allusion to an actual statement made by an official was necessary for the document to be responsive. On the same topic, the assessor wondered if a quote from an appointed federal official (e.g., from the EPA) would qualify, in light of the fact that the negotiated Boolean contained the term "public official" without further qualification. We responded that the topic, not the Boolean string, controlled interpretation, and that the topic contained the additional condition "elected," hence a mere quote from an EPA official without more would not be responsive.

In the case of topic 62, involving press releases concerning water contamination related to irrigation, the assessor reported afterwards that in performing the evaluation "it was sometimes difficult to determine what constituted a press release." Another post-assessment comment stated that because "assessments for responsiveness were done in different sessions, the triggers for responsiveness may not have been consistent," i.e., "sometimes a single word" convinced this assessor that the document was relevant, "while at other intervals I read on to see whether [a finding of relevance] would make more sense in the narrower context of the complaint."

The assessor of topic 80 found it difficult to determine if certain types of radio and magazine advertising were sufficiently clear so as to say that the document made "a connection between folk songs and music and the sale of cigarettes," as the topic required. In the words of the assessor: "While it was easy to identify a connection when a music magazine contained a cigarette ad or when a cigarette magazine contained a music article, other magazines were less obvious. An outdoor magazine[] that contains an interview with a musician as well as a cigarette ad, for example. Or a general interest[] magazine that contains a cigarette ad near its music section." In wondering "how close" the connection had to be, the assessor went on to conclude that "Ultimately, unless the cigarette ad was on the same page as the music section, or in the middle of it, I had to say there was no connection."

One assessor found an error in Complaint C, noting a one-time stray reference to a "Defendant Jones" (at Second Claim for Relief preceding paragraph 46), where all other references in the complaint were to "Defendant Smaug." This circumstance led to a lively debate among track coordinators as to whether the complaint should be left as is, amended for assessors still engaged, or alternatively discarded (we decided to leave it as is given the *de minimis* nature of the error). However, some form of sensitivity analysis might be profitably applied to see if eliminating the anomalous reference changed any run results.

The track coordinators asked assessors to record how much time they spent on their task. Based on 23 survey returns, assessors averaged 25 hours in accomplishing their review of the 500 documents in Bin 1, for an average of 20 documents per assessor per hour. (In 2006 the review rate averaged to 25 documents per hour.) Based on the 2007 returns, the total time devoted works out to approximately 1400 total hours spent on this year's Legal Track tasks (based on 28,141 total judgments divided by 20, including the 24,404 judgments for the 43 Ad Hoc topics, 3,238 judgments for the 10 Interactive/RF topics, and a bin of 499 judgments received after the official results went out). If second-year and third-year law student time were billed at the same rate as summer associates at law firms ($150/hour), those 1400 hours roughly translate to $200,000 in pro bono effort for performing combined relevance assessments during the Ad Hoc and Interactive/RF tasks in 2007.

Not only did the greater cadre this year of law students perform conscientiously during the compressed period of mid-August through mid-September for completing assessments, they appeared to enjoy and benefit from the exercise. Comments from post-assessment surveys included students saying: (i) "On a personal level, the documents were quite interesting. If I had had the time, I gladly would have done another bin of 500, but the semester is starting to get very busy." (ii) "I know more about the effects of cigarettes and smoking than I could have ever thought possible . . ." (iii) "I would love to help out in the future. I found my topic very interesting and enjoyed assessing documents." (iv) "I thought I was getting the short end of the stick because the U.S. Beet Sugar Association had to be the lamest topic of all time. But the documents were really interesting and I learned a lot about the sordid political wrangling over sugar." (v) "I thought

that the project was worthwhile from a purely practical standpoint, in that learning how to review massive amounts of information as efficiently as possible is a skill that all lawyers need to work on."

2.6 Results

Each reviewed document was judged relevant, judged non-relevant, or left as "gray." (Our "gray" category includes all documents that were presented to the assessor, but for which a judgment could not be determined. Among the most common reasons for this were documents that were too long to review (more than 300 pages, according to our "How To Guide") or for which there was a technical problem with displaying the scanned document image.)

A qrelsL07.normal file was created in the common trec_eval qrels format. Its 4th column was a 1 (judged relevant), 0 (judged non-relevant), -1 (gray) or -2 (gray). (In the assessor system, -1 was "unsure" (the default setting for all documents) and -2 was "unjudged" (the intended label for gray documents).)

A qrelsL07.probs file was also created, which was the same as qrelsL07.normal except that there was a 5th column which listed the p(d) for the document (i.e., the probability of that document being selected for assessment from the pool of all submitted documents). qrelsL07.probs can be used with the experimental l07_eval utility to estimate precision and recall to depth 25,000 for runs which contributed to the pool.

2.6.1 Estimating the Number of Relevant Documents for a Topic

To estimate the number of relevant, non-relevant and gray documents in the pool for each topic, the following procedure was used:

Let D be the set of documents in the target collection. For the Legal Track, $|D|$=6,910,912.

Let S be a subset of D.

Define $JudgedRel(S)$ to be the set of documents in S which were judged relevant.

Define $JudgedNonrel(S)$ to be the set of documents in S which were judged non-relevant.

Define $Gray(S)$ to be the set of documents in S which were reviewed but not judged relevant nor non-relevant.

Define $estRel(S)$ to be the estimated number of relevant documents in S:

$$estRel(S) = \min\left(\left(\sum_{d \in JudgedRel(S)} \frac{1}{p(d)}\right),\ (|S| - |JudgedNonrel(S)|)\right) \tag{1}$$

Note that $estRel(S)$ is 0 if $|JudgedRel(S)| = 0$.

Define $estNonrel(S)$ to be the estimated number of non-relevant documents in S:

$$estNonrel(S) = \min\left(\left(\sum_{d \in JudgedNonrel(S)} \frac{1}{p(d)}\right),\ (|S| - |JudgedRel(S)|)\right) \tag{2}$$

Note that $estNonrel(S)$ is 0 if $|JudgedNonrel(S)| = 0$.

Define $estGray(S)$ to be the estimated number of gray documents in S:

$$estGray(S) = \min\left(\left(\sum_{d \in Gray(S)} \frac{1}{p(d)}\right),\ (|S| - (|JudgedRel(S)| + |JudgedNonrel(S)|))\right) \tag{3}$$

Note that $estGray(S)$ is 0 if $|Gray(S)| = 0$.

Applying the above formulas, the estimated number of relevant documents in the pool, on average per topic, was 16,904(!). The number varied considerably by topic, from 18 (for topic 63) to 77,467 (for topic 71). (The estimates for all of the topics are listed in Section 4.) Obviously, traditional top-ranked pooling would not have been sufficient to cover the high numbers of relevant documents. On average (per topic), the estimated number of non-relevant documents in the pool was 298,678, and the estimated number of gray documents in the pool was 4,303.

2.6.2 Estimating Recall

The L07 approach to estimating recall is similar to how infAP estimates its recall component (i.e., it's based on the run's judged relevant documents found to depth k compared to the total judged relevant documents). In our case, we have to weight the judged relevant documents to account for the sampling probability.

For a particular ranked retrieved set S:

Let $S(k)$ be the set of top-k ranked documents of S. (Note that $|S(k)| = \min(k, |S|)$.)

Define $estRecall@k$ to be the estimated recall of S at depth k:

$$estRecall@k = \frac{estRel(S(k))}{estRel(D)} \tag{4}$$

The mean estimated recall of the reference Boolean run (refL07B) was just 22%. Hence the final negotiated boolean query was missing about 78% of the relevant documents (on average across all topics). Note that the estimated recall of refL07B varied considerably per topic, from 0% (topic 77) to 100% (topic 84). Figure 1 illustrates the breakdown of the estimated relevant documents into those matching the Boolean query and those only found by at least one of the ranked systems.

Section 4 lists the final Boolean query's estimated recall for each topic; it also lists several relevant documents (underlined) which did not match the final Boolean query. For example, it shows that for topic 74 ("All scientific studies expressly referencing health effects tied to indoor air quality") the final negotiated Boolean query of "`(scien! OR stud! OR research) AND ("air quality" w/15 health)`" missed matching relevant document vkm92d00. This document did not contain required Boolean terms such as "air" or "health", but it was judged relevant presumably because it referred to the "largest study ever" on whether "secondary smoke causes cancer" and also to a study of the "carcinogenic effects" of the "gas released from volatile organic compounds in shower water"[3].

Despite the low recall of the final Boolean query, none of the 68 submitted runs had a higher mean estimated Recall@B than the reference Boolean run (as Table 1 shows). This is a surprising result, since the refL07B run had been available to the participants since early July and thus could have been used by participating systems as one source of evidence. At least one participant (Open Text, which contributed the reference Boolean run) reported that combining other techniques with the Boolean run did not increase average recall at depth B. We anticipate that more participants will attempt to make use of the reference Boolean run next year.

One factor that may be limiting average recall across topics is that our Boolean queries targeted a B range between 100 and 25,000 to keep the size of submitted runs manageable. We should perhaps review whether our Boolean queries might be higher precision (and hence lower recall) than the Boolean queries used in practice.

Table 1 also lists mean estimated Recall@25000. The highest score (47%) was by a run which used both the final Boolean and request text fields (wat1fuse).

Section 4 lists the median scores for each topic in both Recall@B and Recall@25000. Although the final Boolean query had a higher recall than the median Recall@B for 31 of the 43 topics (and 4 tied), the median Recall@25000 was higher than the Boolean query's recall for 33 of the 43 topics (and 1 tied). The median run typically could not match the precision of the Boolean query to depth B, but by retrieving deeper it typically would find more relevant documents. (Table 1 shows that the average B value was 5004, while most of the runs retrieved the allowed maximum of 25,000 per topic.)

2.6.3 Estimating Precision

The L07 approach to estimating precision is similar to how infAP estimates its precision component (i.e., it's based on the precision of the run's judged documents to depth k). In our case, we have to weight the judged documents to account for the sampling probability. We also multiply by a factor to ensure that unretrieved documents are not inferred to be relevant.

[3]In the OCR output used by the participants, this latter phrase actually appeared as "Ras released f rom volatile organtc compounds in .houer water".

Run	Fields	Avg. Ret.	Est. R@B	Est. R25000	Est. P5	Est. P@B	Est. P25000	Est. Gray@B	S1J	GS10J	(raw) R-Prec
refL07B	bM	5004	**0.216**			**0.292**		0.042			
otL07fb	bM	5004	**0.216**	0.216	0.507	**0.292**	0.056	0.042	24/43	0.863	0.201
otL07fb2x	bM	6053	0.209	0.242	0.486	0.282	0.065	0.031	21/43	0.837	0.193
CMUL07ibs	bM	25000	0.208	0.392	0.452	0.267	0.137	0.022	24/43	0.832	0.151
wat5nofeed	brM	24999	0.196	0.447	0.537	0.263	0.159	0.016	24/43	0.864	0.204
CMUL07irs	bM	25000	0.194	0.395	0.372	0.242	0.149	0.013	23/43	0.813	0.133
otL07frw	brM	25000	0.193	0.428	**0.550**	0.278	0.150	0.015	26/43	**0.883**	**0.224**
CMUL07ibt	bM	25000	0.187	0.391	0.495	0.261	0.136	0.024	23/43	0.847	0.175
otL07pb	pM	18555	0.186	0.327	0.424	0.235	0.119	0.025	17/43	0.792	0.147
wat1fuse	brM	25000	0.186	**0.469**	0.529	0.271	0.155	0.012	21/43	0.837	0.197
CMUL07ibp	bM	25000	0.183	0.392	0.472	0.252	0.131	0.026	**26/43**	0.859	0.178
wat7bool	bM	7059	0.172	0.198	0.420	0.250	0.064	0.047	18/43	0.761	0.140
CMUL07o3	bdpM	25000	0.170	0.400	0.491	0.236	0.146	0.009	23/43	0.867	0.190
otL07fbe	bM	25000	0.169	0.369	0.492	0.273	0.160	0.015	22/43	0.792	0.178
IowaSL0704	bdpr	20071	0.167	0.382	0.461	0.232	0.123	0.015	21/43	0.760	0.173
UMass15	r	25000	0.165	0.354	0.465	0.229	0.122	0.006	23/43	0.789	0.172
IowaSL0703	bdpr	25000	0.164	0.403	0.451	0.216	0.139	0.009	19/43	0.808	0.181
otL07fv	bM	25000	0.163	0.357	0.467	0.225	0.129	0.005	20/43	0.856	0.176
UMass13	br	25000	0.162	0.322	0.442	0.195	0.118	0.007	20/43	0.861	0.175
IowaSL0705	dpr	7396	0.161	0.260	0.502	0.243	0.066	0.017	22/43	0.799	0.184
UMKC4	d	24419	0.157	0.412	0.391	0.253	0.145	0.014	16/43	0.749	0.144
IowaSL0706	br	7396	0.153	0.247	0.428	0.252	0.067	0.012	21/43	0.792	0.176
otL07rvl	rM	25000	0.153	0.420	0.471	0.235	0.151	0.010	15/43	0.850	0.187
CMUL07o1	bM	25000	0.152	0.361	0.467	0.204	0.133	0.009	24/43	0.848	0.168
UMass12	b	24028	0.150	0.285	0.350	0.197	0.117	0.005	21/43	0.804	0.125
SabL07arbn	bdpr	25000	0.149	0.321	0.393	0.203	0.117	0.009	19/43	0.790	0.132
UMass14	r	25000	0.147	0.362	0.464	0.208	0.113	0.010	25/43	0.813	0.167
wat8gram	bM	25000	0.142	0.389	0.419	0.256	0.143	0.009	21/43	0.781	0.137
IowaSL0707	brM	5004	0.142	0.142	0.409	0.237	0.053	0.013	20/43	0.781	0.173
wat6qap	bM	17179	0.142	0.239	0.385	0.194	0.089	0.031	13/43	0.733	0.113
UMKC6	b	25000	0.137	0.400	0.371	0.258	0.161	0.020	20/43	0.794	0.149
UMass11	r	25000	0.137	0.325	0.377	0.191	0.104	0.007	14/43	0.764	0.145
UMKC1	d	24419	0.135	0.426	0.436	0.241	0.153	0.019	20/43	0.752	0.124
IowaSL0702	bdpr	25000	0.134	0.363	0.429	0.205	0.132	0.007	19/43	0.816	0.183
SabL07ab1	bdpr	25000	0.132	0.316	0.371	0.204	0.123	0.013	18/43	0.747	0.119
CMUL07irt	bM	25000	0.132	0.294	0.467	0.189	0.115	0.013	22/43	0.829	0.158
UMass10	rM	23649	0.131	0.306	0.489	0.207	0.117	0.016	23/43	0.800	0.136
UMKC5	r	25000	0.126	0.411	0.370	0.260	**0.172**	0.012	18/43	0.750	0.148
wat3desc	rM	24999	0.124	0.394	0.426	0.235	0.143	0.010	20/43	0.775	0.160
CMUL07std	rM	25000	0.123	0.314	0.451	0.191	0.115	0.006	22/43	0.833	0.163
fdwim7xj	rM	25000	0.113	0.354	0.408	0.180	0.114	0.017	22/43	0.781	0.142
ursinus1	r	25000	0.113	0.329	0.340	0.195	0.125	0.016	16/43	0.751	0.131
ursinus2	r	25000	0.112	0.314	0.307	0.154	0.117	0.008	14/43	0.685	0.099
ursinus6	r	25000	0.110	0.298	0.242	0.153	0.108	0.009	10/43	0.628	0.089
IowaSL07Ref	r	25000	0.108	0.343	0.366	0.148	0.120	0.009	15/43	0.754	0.130
UMKC3	b	25000	0.107	0.391	0.444	0.243	0.161	0.031	17/43	0.790	0.131
fdwim7rs	r	25000	0.106	0.319	0.431	0.210	0.126	0.013	21/43	0.790	0.142
UIowa07LegE2	b	16708	0.106	0.268	0.283	0.224	0.114	0.021	11/43	0.651	0.109
UIowa07LegE0	r	24997	0.103	0.318	0.312	0.156	0.120	0.009	19/43	0.736	0.118
fdwim7ss	cir	25000	0.102	0.309	0.409	0.170	0.115	0.014	17/43	0.788	0.129
fdwim7sl	cir	25000	0.101	0.288	0.422	0.164	0.120	0.010	20/43	0.783	0.120
UMKC2	r	25000	0.100	0.409	0.416	0.226	0.171	0.016	17/43	0.770	0.125
ursinus7	bM	25000	0.099	0.283	0.265	0.161	0.109	0.010	12/43	0.601	0.086
SabL07ar2	r	25000	0.098	0.295	0.364	0.178	0.105	0.016	16/43	0.789	0.102
SabL07ar1	r	25000	0.097	0.288	0.369	0.174	0.105	0.015	15/43	0.784	0.101
ursinus4	r	25000	0.096	0.315	0.332	0.233	0.139	**0.131**	14/43	0.714	0.066
Dartmouth1	r	25000	0.083	0.275	0.285	0.137	0.102	0.010	15/43	0.682	0.095
wat2nobool	brM	25000	0.082	0.320	0.327	0.217	0.131	0.018	12/43	0.713	0.071
ursinus8	bM	25000	0.071	0.191	0.093	0.101	0.085	0.018	4/43	0.416	0.032
fdwim7ts	r	25000	0.070	0.177	0.163	0.109	0.067	0.012	6/43	0.592	0.044
ursinus3	r	25000	0.063	0.213	0.072	0.084	0.078	0.008	3/43	0.401	0.024
ursinus5	r	25000	0.063	0.220	0.072	0.081	0.083	0.009	3/43	0.396	0.023
wat4feed	brM	25000	0.061	0.224	0.261	0.151	0.092	0.025	9/43	0.600	0.063
catchup0701p	r	24016	0.061	0.171	0.126	0.130	0.098	0.023	5/43	0.528	0.041
UIowa07LegE1	r	24996	0.031	0.083	0.206	0.067	0.057	0.004	11/43	0.651	0.036
UIowa07LegE3	b	24999	0.028	0.110	0.194	0.090	0.070	0.012	9/43	0.550	0.035
otL07db	dM	368	0.027	0.027	0.301	0.026	0.006	0.003	15/43	0.576	0.074
UIowa07LegE5	b	24992	0.003	0.019	0.105	0.018	0.026	0.017	6/43	0.324	0.011
UIowa07LegE4	b	9879	0.002	0.004	0.023	0.012	0.009	0.010	1/43	0.101	0.002
randomL07		100	0.000	0.000	0.000	0.000	0.000	0.001	0/43	0.038	0.001

Table 1: Mean scores for submitted Ad Hoc task runs.

Run	Depths 1-5000	Depths 5001-10000	Depths 10001-15000	Depths 15001-20000	Depths 20001-25000
Ad Hoc - high	0.238	0.182	0.183	0.168	0.199
Ad Hoc - median	0.178	0.127	0.110	0.101	0.099
RF - high	0.218	0.197	0.146	0.197	0.150
RF - median	0.126	0.118	0.069	0.055	0.052

Table 2: High and Median Estimated Marginal Precision Rates

Define $estPrec@k$ to be the estimated precision of S at depth k:

$$estPrec@k = \frac{estRel(S(k))}{estRel(S(k)) + estNonrel(S(k))} \times \frac{|S(k)|}{k} \tag{5}$$

Note: we define $estPrec@k$ as 0 if both $estRel(S(k))$ and $estNonrel(S(k))$ are 0.

The mean estimated precision of the reference Boolean run (refL07B) was just 29%. Again, this varied by topic, from 0% (topic 77) to 97% (topic 69) (Section 4 lists the precision scores for all of the topics). As Table 1 shows, none of the 68 submitted runs had a higher mean estimated Precision@B than the reference Boolean run. The submitted run with the highest mean estimated Precision@25000 (17%) just used the request text field (UMKC5).

2.6.4 Marginal Precision Rates to Depth 25,000

Table 2 shows how precision falls with retrieval depth for the Ad Hoc task runs. The table includes the highest and median estimated marginal precision rates of the Ad Hoc task runs for depths 1-5000, 5001-10000, 10001-15000, 15001-20000 and 20001-25000. The median run was still maintaining 10% precision at the deepest stratum (depths 20001-25000), and some runs were close to 20% precision in this stratum. It appears for this test collection that depth 25,000 was not deep enough to cover all of the relevant documents that a run could potentially find.

We note however that the median precision in the deepest stratum exceeded 10% for only 6 of the 43 topics. These included topics 69, 74 and 71, which also had among the highest number of estimated relevant documents (as per the listing in Section 4). 13 of the 43 topics had more than 25,000 estimated relevant documents (hence 100% recall was not possible for these topics at depth 25,000). Perhaps it would be better for reusability to discard these 13 topics, but additional analysis will be needed before we can draw firm conclusions.

We hope to investigate reusability issues with the collection as (near) future work. But generally speaking, for runs that did not contribute to the pools, if the 25,000 documents retrieved for a topic are mostly a subset of the (approximately) 300,000 documents that were pooled for the topic, or if the unpooled documents contain few relevant documents, then the estimated measures from the 107_eval utility should still be comparable to the pooled systems' scores, particularly at deeper depths (e.g., at depth 25,000).

2.6.5 Estimated Gray Percentages

Table 1 also lists the estimated percentage of gray documents at depth B. The ursinus4 run retrieved a lot more gray documents (13%) than the other runs; we therefore suspect this run's approach favored longer documents. Boolean runs retrieved 4-5% gray, perhaps because the Boolean constraints matched some long documents. Most other runs retrieved less than 2% gray. These systematic differences suggest that it may be productive to reconsider the techniques being used for dealing with long documents, both in system design and in our assessment process.

2.6.6 Random Run Results

It was hoped that the randomL07 run would give us at least a rough indication of the number of relevant documents that may have been outside of the pooled results of participating systems. A total of 446 of the randomL07 documents were judged (over 43 topics), and 3 of these were judged relevant (0.7%). Typically, about 6.5 million documents of the almost 7 million documents in the collection were not submitted by any participating system and thus not included in our pooling process. If 0.7% of those are relevant, that would suggest another 50,000 relevant documents (per topic). However, when we reviewed the 3 judged relevant documents from the randomL07 run (zsm67e00 for topic 58, cdf53e00 for topic 70 and dkb74d00 for topic 71), none of them appeared to actually be relevant to us (though the official qrels have not been altered). So perhaps the overall precision of the unpooled documents is actually much less than 0.7% (though even 0.1% would represent more than 5,000 relevant documents per topic).

There were 6 randomL07 documents that were considered "gray" (i.e., not judged relevant nor non-relevant). None of these 6 documents were very long. For one of them, the PDF document would not display (bev71d00 for topic 84). 2 of the 6 were non-English documents (tpv77e00 and xyu37e00 for topic 52). The other 3 had relatively little text (lkf03d00 and xqx21a00 for topic 69, and qge12c00 for topic 89). However, these documents do not seem to be typical of the gray documents retrieved by system runs, which generally were very long documents.

2.6.7 Table Glossary

Table 1 lists the mean scores for each of the 68 submitted runs for the Ad Hoc task and the 2 additional reference runs. The following glossary explains the codes used in that table.

"Fields": The topic fields used by the run: 'b' Boolean query (final negotiated), 'c' complaint, 'd' defendant Boolean (initial proposal), 'i' instructions and definitions, 'p' plaintiff Boolean (rejoinder query), 'r' request text, 'M' manual processing was involved, 'F' feedback run (2006 relevance assessments were used, applicable to RF task only).

"Avg. Ret.": The Average Number of Documents Retrieved per Topic.

"R@B" and "R@25000": Estimated Recall at Depths B and 25000.

"P5", "P@B" and "P25000": Estimated Precision at Depths 5, B and 25000.

"Gray@B": Estimated percentage of Gray documents at Depth B.

"S1J": Success of the First Judged Document.

"GS10J": Generalized Success@10 on Judged Documents (1.08^{1-r} where r is the rank of the first relevant document, only counting judged documents, or zero if no relevant document is retrieved). GS10J is a robustness measure which exposes the downside of blind feedback techniques [11]. Intuitively, it is a predictor of the percentage of topics for which a relevant document is retrieved in the first 10 rows.

"R-Prec": R-Precision (raw Precision at Depth R, where R is the raw number of known relevant documents). Estimation is not used for this measure. It is provided so that we can see if the results differ with a traditional IR measure.

For the reference Boolean run (refL07B), only measures at depth B are shown so that the ordering of Boolean results does not matter.

3 Interactive and Relevance Feedback Tasks

In 2006, most teams applied existing information retrieval systems to obtain what might best be characterized as baseline results. Moreover, in 2006 the relevance assessment pools were (with two exceptions) drawn from near the top of submitted ranked retrieval runs. Both factors tend to reduce the utility of the available relevance judgments for the 2006 topic set somewhat. We therefore created two opportunities for teams to contribute runs that would permit us to enrich the 2006 relevance judgments: a Relevance Feedback task, and an Interactive task. The same document collection was used in 2006 and 2007, so participation in these tasks did not require indexing a new collection.

The objective in the Relevance Feedback task was to automatically discover previously unknown relevant documents by augmenting the evidence available from the topic description with evidence available from the relevance assessments that were created in 2006. Teams could use positive and/or negative judgments in conjunction with the metadata for and/or full text from the judged documents to refine their models. This task provides a simple and well controlled model for assessing the utility of a two-pass search process.

Interactive searchers have an even broader range of strategies available for enhancing their results, including iteratively improving their query formulation (based on their examination of search results) and/or performing more than one iteration of relevance feedback. There is, therefore, significant scope for research on processes by which specific search technologies can best be employed. Participants in the Interactive task could use any combination of: systems of their own design, the Legacy Tobacco Document Library system (LTDL, a Web-based system provided by the University of California, San Francisco), or the Tobacco Documents Online system (TDO, the same Web-based system that was used for relevance assessment in the TREC 2006 Legal Track). Standardized questionnaires were used to collect information about the search process from the perspective of individual participants, and research teams could employ additional methods (e.g., observation or log file analysis) to collect complementary information at their option.

3.1 Topics

Twelve topics out of the first year's 39 completed topics were selected for the Interactive task. These topics were chosen by the track coordinators based on a variety of factors, including (i) not being too closely tied to a "tobacco-related" topic, so as to mitigate whatever inherent bias exists; (ii) the absolute number of relevant documents found for each topic in year one, with topics returning under a threshold of 50 documents being considered lesser priority; (iii) relatively high kappa scores from year 1 on inter- assessor agreement, and (iv) their inherent interest. The top three interactive topics ended up, in priority order, involving the subjects of pigeon deaths (topic 45), memory loss (topic 51), and the placement of tobacco products in G-rated movies (topic 7). A total of 10 of the 12 interactive topics were completely assessed by volunteers. These same 10 topics were used for the Relevance Feedback task in 2007.

3.2 Evaluation

Eight Relevance Feedback runs were submitted by 3 research teams. Participating teams were allowed to submit up to max(25000,B)+1000 documents per topic. "Residual evaluation" was used for the Relevance Feedback task. Hence, before pooling, any documents that were judged last year (of which there were at most 1000 per topic) were removed from the Relevance Feedback runs. For topics with B_r>25000, which was the case for two of the Relevance Feedback topics, depth-B_r pooling was used; other topics were pooled to depth 25,000. The resulting ranked lists were therefore truncated at max(25000,B_r), where B_r is the number of documents matching the reference Boolean query (refL06B) after last year's judged documents are removed. Because B_r>25000 for two topics, R@B_r scores can exceed R@25000 in the Relevance Feedback task, which is not possible for the Ad Hoc task.

The pools were then enriched before judgment with three additional runs:

- A special "oldrel07" run was added which included 25 relevant documents (or all if less than 25 were available) randomly chosen from last year's judgments for each topic.
- A special "oldnon07" run was added which included 25 non-relevant documents randomly chosen from last year's judgments for each topic.
- A special "randomRF07" run was added which included 100 randomly chosen documents that were not otherwise pooled (or judged last year).

Finally, all documents from the Interactive task runs were included in this year's pools (even if they were judged in 2006).

The p(d) formula for the Relevance Feedback task was the same as for the Ad Hoc task except that: (1) all documents from the Interactive task and from the oldrel07 and oldnon07 runs were assigned probability

1.0 so that they would all be presented to the assessor, (2) topics with B_r>25000 were sampled to depth B_r, with p(d) of min(1.0, ((5/25000)+(C/hiRank(d))) for documents with hiRank(d) between 6 and 25000 inclusive, and (3) the sum of the p(d) could be just 250 instead of 500 for some of the topics (because fewer documents needed to be judged to maintain the same accuracy (C value) as in the Ad Hoc task).

For the Interactive task, teams could submit as many as 100 documents per topic, but submission of nonrelevant documents was penalized. A simple utility function (the number of submitted relevant documents minus half the number of submitted nonrelevant documents) was chosen as the principal evaluation measure for the Interactive task in order to encourage participants to submit fewer than 100 documents when that many relevant documents could not be found.

3.3 Interactive Task Results

Table 3 shows the results for each Interactive task team. Eight teams from three sites participated:

Long Island University (LIU). Nine participants worked in groups of three, with one group assigned to each of the highest priority topics. All three groups searched only the LTDL. A tenth participant reviewed each group's top 100 retrieved results and only those considered relevant were submitted. The estimated total search time (across all three searchers in a team) was 39 hours for topic 51, 39 hours for topic 45, and 25 hours for topic 7. Results were reported as both Bates numbers and URL's; the DOCNO was extracted from the URL for pooling and scoring. The reported results for LIU were corrected after they were initially distributed to remove 14 LTDL documents that are not contained in the TREC Legal Track test collection.

Sabir Research (Sabir). One participant worked alone to search the eight highest priority topics. Multiple relevance feedback iterations were performed based on judgments from 2006, plus judgments for an additional 10 previously unjudged documents that were added at each iteration. The limited multi-pass interaction in this process was intended as a contrastive condition to the one-pass relevance feedback runs from the same site; comparison with results from other sites may be less informative because manual query refinement was not performed in this case. The process required an average of about 16 minutes per topic. Results were reported as both Bates numbers and URL; the DOCNO was therefore extracted from the URL.

University of Washington (UW). Sixteen participants worked in six teams, each consisting of two or three participants, and the results for each team were submitted and scored separately. The average time per topic was not reported. Results were submitted as Bates numbers and were automatically mapped to DOCNO values based on exact string match. This process resulted in frequent failures (44% of all values reported as Bates numbers proved to be unmappable); inspection of the values revealed that some were very similar to valid Bates numbers that were present in the collection (e.g., with a dash in place of a slash or with a brief prefix indicating the source), but that others were not. After relevance judgments were completed, the mapping script was modified to accommodate some patterns that were detected by inspection and the UW runs were rescored. The rescored values are shown in italics in Table 3 where they differ from the initially computed (completely assessed) results.

The evaluation design that we chose can in some sense be thought of as comparing the opinion of one person (the relevance assessor) with the opinion of one or more other people (the experiment participants). From the LIU results, we can see that a moderately good level of agreement can be achieved, with agreement on 96 (81%) of the 118 positive judgments made by the participants. Perhaps the most interesting conclusion that we can draw from comparing the results from the seven LIU and UW teams that tried a fully interactive process is that searchers exhibited substantial variation. For example, among the six UW teams (which were given consistent instructions) we see a variation of at least a factor of two in the number of relevant documents found (in the opinion of the relevance assessor) for each of the three topics. Of course, we must caveat this result by noting that the process used for recording Bates numbers by UW participants exhibited substantial variation both by team and by topic, so variations in the effectiveness of the mapping process are a possible confounding factor.

Team	Score	45S	45R	45N	51S	51R	51N	7S	7R	7N
LIU	**85.0**	13.5	18	9	20.0	24	8	**51.5**	54	5
UW6	*66.5*	**28.0**	**35**	14	*15.0*	*19*	*8*	23.5	38	29
UW2	*51.5*	*24.5*	32	*15*	15.5	31	31	*11.5*	*24*	25
UW1	48.0	17.0	28	22	**24.0**	**35**	22	7.0	10	6
UW3	*43.0*	*16.0*	*17*	*2*	9.5	23	27	17.5	30	25
UW5	*39.0*	*12.5*	21	*17*	*15.0*	23	*16*	11.5	23	23
UW4	*36.0*	*18.0*	*28*	*20*	*5.5*	*17*	*23*	*12.5*	20	*15*
Sabir	-10.5	-7.0	11	36	-0.5	0	1	3.0	3	12
oldrel07	14.5	5.5	12	13	12.0	16	8	-3.0	6	18
refL06B	12.0	6.5	71	129	7.0	73	132	-1.5	16	35
randomRF07	-23.5	-10.0	0	20	-5.0	1	12	-8.5	0	17
oldnon07	-34.5	-11.5	0	23	-10.5	1	23	-12.5	0	25

Table 3: Mean scores for all Interactive task teams (S=Score, R=relevant, N=Not relevant).

Table 3 also lists the scores of the reference Boolean run (refL06B). Most of the Interactive teams scored higher than the Boolean run on all 3 topics. It should be noted that, unlike the participants, the Boolean run was not limited to 100 retrieved documents, and its results were sampled unevenly (it includes the documents selected for the residual RF evaluation along with the documents from the Interactive, oldrel07 and oldnon07 runs).

3.4 Relevance Feedback Task Results

Table 4 shows the results for the 10 Relevance Feedback runs. Three research teams participated, and two additional reference reference runs were also scored (refL06B and randomRF07). The following summarizes each team's submissions:

Carnegie Mellon University (CMU07). The CMU team treated the Relevance Feedback task as a supervised learning problem and retrieved documents using Indri queries that approximated Support Vector Machines (SVMs) learned from training documents. Both keyword features and simple structured "term.field" features were investigated. Named-Entity tags, the LingPipe sentence breaker, and metadata provided in the collection with each document were used to generate the field information. The CMU07RSVMNP run included negative and positive weight terms, while the CMURFBSVME run was formulated using only terms with positive weights.

Open Text Corporation (ot). The two runs submitted by Open Text performed the Relevance Feedback task, but without actually using the available positive or negative assessments. Run otRF07fb ranked the documents in the reference Boolean run (refL06B). Run otRF07fv performed ranked retrieval using the query terms from the same Boolean query.

Sabir Research (sab07). Sabir's runs provided a baseline for multi-pass relevance feedback runs that were submitted for the Interactive task.

Mean scores over just 10 topics may not be very reliable, so little should be read from the result that no participating system outperformed the reference Boolean run on the mean Est. $R@B_r$ measure or that every run (other than the random run) outperformed the reference Boolean run on the mean Est. $P@B_r$ measure. An encouraging result from this pilot study is that the median of the 5 feedback runs outscored the reference Boolean run in Est. $R@B_r$ for 7 of the 10 topics, in contrast with the Ad Hoc task in which the median run outscored the reference Boolean run in just 8 of the 43 topics. Of course, these results were obtained on different topics, by different numbers of teams, and 10 topics remains a small sample however you slice it (the track hopes to conduct a larger study in the upcoming year). Some useful insights may come from failure analysis of the topics for which the feedback runs were still outscored by the reference Boolean run.

Run	Fields	Avg. Ret.	Est. $R@B_r$	Est. R25000	Est. P5	Est. $P@B_r$	Est. P25000	Est. $Gray@B_r$	S1J	GS10J	(raw) R-Prec
refL06B	bM	20185	**0.386**			0.106		**0.051**			
otRF07fb	bM	20185	**0.386**	0.367	0.380	0.106	0.044	**0.051**	3/10	0.735	0.100
CMU07RSVMNP	F-bM	25159	0.380	0.598	**0.570**	0.170	**0.155**	0.007	5/10	**0.870**	0.182
CMU07RBase	bM	25100	0.353	0.578	0.500	0.115	0.102	0.024	**7/10**	0.843	0.117
CMU07RFBSVME	F-bM	25116	0.334	0.578	**0.570**	0.118	0.072	0.034	6/10	0.855	0.173
sab07legrf2	F-bdpr	36625	0.333	0.515	0.500	0.173	0.099	0.012	4/10	0.788	0.199
sab07legrf3	F-bdpr	36625	0.321	**0.616**	0.520	**0.194**	0.117	0.012	4/10	0.814	**0.202**
sab07legrf1	F-brM	25106	0.278	0.475	0.480	0.132	0.072	0.013	5/10	0.802	0.197
otRF07fv	bM	36625	0.248	0.444	0.355	0.156	0.064	0.007	3/10	0.612	0.065
randomRF07		100	0.002	0.002	0.040	0.004	0.000	0.000	0/10	0.159	0.004

Table 4: Mean scores for submitted Relevance Feedback task runs.

4 Individual Topic Results

This section provides summary information for each of the assessed topics of the Ad Hoc and Relevance Feedback tasks.

The 44 assessed Ad Hoc topics are listed first (including one topic (#62) whose judgments arrived after the official results had been released). The information provided is as follows:

Topic : The topic numbers range from 52 to 101. In parentheses are the year of the topic set (2007), the label of the complaint (A, B, C or D), and the request number inside the complaint. (The complaints run several pages and are available on the track web site [1].)

Request Text : The one-sentence statement of the request.

Initial Proposal by Defendant, Rejoinder by Plaintiff and Final Negotiated Boolean Query :

The following syntax was used for the Boolean queries:

- `AND, OR, NOT, ()`: As usual.
- `x BUT NOT y`: Same meaning as `(x AND (NOT (y)))`
- `x`: Match this word exactly (case-insensitive).
- `x!`: Truncation – matches all strings that begin with substring `x`.
- `!x`: Truncation – matches all strings that end with substring `x`.
- `x?y`: Single-character wildcard – matches all strings that begin with substring `x`, end with substring `y`, and have exactly one-character in between `x` and `y`.
- `x*y`: Muliple-character wildcard – matches all strings that begin with substring `x`, end with substring `y`, and have 0 or more characters between `x` and `y`.
- `"x", "x y", "x y z"`, etc.: Phrase – match this string or sequence of words exactly (case-insensitive).
- `"y x!", "x! y"`, etc.: If `!` is used internal to a phrase, then do the truncated match on the words with `!`, and exact match on the others. (The `*` and `?` wildcard operators may also be used inside a phrase.)
- `w/k`: Proximity – `x w/k y` means match `"x a b ... c y"` or `"y a b ... c x"` if `"a b ... c"` contains `k` or fewer words.
- `x w/k1 y w/k2 z`: Chained proximity – a match requires the same occurrence of `y` to satisfy `x w/k1 y` and `y w/k2 z`.

Sampling and Est. Rel. : The number of pooled documents is given (i.e., all distinct documents from all of the submitted runs for the topic), followed by the number presented to the assessor, the number the assessor judged relevant, the number the assessor judged non-relevant, and the number the assessor left as "gray" (defined earlier). The "C" value of the p(d) formula is given (the measures at depths B and 25000 have approximately the accuracy of 5+C simple random sample points, as described earlier). "Est. Rel." is the estimated number of relevant documents in the pool for the topic (based on the sampling results).

Final Boolean Result Size (B), Est. Recall and Est. Precision : "B" is the number of documents matching the final negotiated Boolean query (which always was between 100 and 25000 in 2007). "Est. Recall" is the estimated recall of the final negotiated Boolean query result set. "Est. Precision" is the estimated precision of the final negotiated Boolean query result set.

Participant High Recall@B and Median Recall@B : The highest estimated recall at depth B of the participant runs is listed, followed in parentheses by the run's identifier; if more than one run tied for the highest score, just one of them is listed (chosen randomly) and the number of other tied runs is stated. The median estimated recall at depth B is based on the median score of 70 runs (the 68 participant runs along with the refL07B and randomL07 runs).

Participant High Recall@25000 and Median Recall@25000 : Same as previous line except that the measures are at depth 25000 instead of depth B.

5 Deepest Sampled Relevant Documents : The identifiers of the 5 deepest sampled documents that were judged relevant are listed in descending depth order. The identifier is underlined if the document was not retrieved by the final negotiated Boolean query. Each identifier is followed by its weight in the estimation formulas (i.e., the estimated number of relevant documents it represents) which is 1/p(d) (i.e., the reciprocal of the probability of being selected for judging). In parentheses is the identifier of the run which retrieved the document at the highest rank, followed by that rank (which can range from 1 to 25000); if multiple runs retrieved the document at that rank, just one of them is listed. For example, for topic 52, the entry of "hdz83f00-93.4 (otL07fbe-515)" indicates that the hdz83f00 document was judged relevant and, because it is not underlined, it matched the final Boolean query. It counts as 93.4 estimated relevant documents in the estimation formulas (because it was selected for judging with probability 1/93.4). The otL07fbe run retrieved this document at rank 515; any other run that retrieved the document did so at the same or deeper rank (i.e., if the pooling had been to less than depth 515, the document would not have been in the pool). Note that the document content and metadata can be found online by appending the document identifier to the url "http://legacy.library.ucsf.edu/tid/" (e.g., `http://legacy.library.ucsf.edu/tid/hdz83f00`). One can do failure analysis for the final Boolean query for most topics by reviewing the underlined document identifiers.

Topic 52 (2007-A-1)

Request Text: Please produce any and all documents that discuss the use or introduction of high-phosphate fertilizers (HPF) for the specific purpose of boosting crop yield in commercial agriculture.

Initial Proposal by Defendant: `"high-phosphate fertilizer!" AND (boost! w/5 "crop yield") AND (commercial w/5 agricultur!)`

Rejoinder by Plaintiff: `(phosphat! OR hpf OR phosphorus OR fertiliz!) AND (yield! OR output OR produc! OR crop OR crops)`

Final Negotiated Boolean Query: `(("high-phosphat! fertiliz!" OR hpf) OR ((phosphat! OR phosphorus) w/15 (fertiliz! OR soil))) AND (boost! OR increas! OR rais! OR augment! OR affect! OR effect! OR multipl! OR doubl! OR tripl! OR high! OR greater) AND (yield! OR output OR produc! OR crop OR crops)`

Sampling: 361264 pooled, 1000 assessed, 55 judged relevant, 941 non-relevant, 4 gray, "C"=4.68, **Est. Rel.:** 257.4

Final Boolean Result Size (B): 3078, **Est. Recall:** 97.6%, **Est. Precision:** 10.6%

Participant High Recall@B: 100.0% (wat5nofeed), **Median Recall@B:** 48.2%

Participant High Recall@25000: 100.0% (wat1fuse and 10 others), **Median Recall@25000:** 73.1%

5 Deepest Sampled Relevant Documents: hdz83f00-93.4 (otL07fbe-515), dud53d00-21.8 (UMass15-106), huw23d00-18.8 (UMKC2-91), zge78d00-17.0 (SabL07ar1-82), ehe58c00-13.4 (wat5nofeed-64)

Topic 53 (2007-A-2)

Request Text: Please produce any and all documents concerning the effect of Maleic hydrazide (MH) on the tumorigenicity in hamsters.

Initial Proposal by Defendant: `"maleic hydrazide" AND tumorigenicity AND hamster!`

Rejoinder by Plaintiff: `("maleic hydr?zide" OR MH OR pesticid! OR "weed killer" OR herbicid! OR ((growth OR sprout!) w/3 (inhibitor! OR retardant)) OR "potassium salt" OR De-cut OR "Drexel MH" OR Gro-taro OR C4N2H4O2) AND (hamster! OR mice OR mouse OR rat OR rats OR rodent! OR subject! OR animal!)`

Final Negotiated Boolean Query: `("maleic hydr?zide" OR (MH AND (pesticid! OR "weed killer" OR herbicid! OR (growth OR sprout!) w/3 (inhibitor! OR retardant))) OR "potassium salt" OR De-cut OR "Drexel MH" OR Gro-taro OR C4N2H4O2) AND (tumor! OR oncogenic OR oncology! OR pathology! OR pathogen!) AND (hamster! OR mice OR mouse OR rat OR rats OR rodent!)`

Sampling: 309106 pooled, 499 assessed, 140 judged relevant, 359 non-relevant, 0 gray, "C"=2.26, **Est. Rel.**: 31632.5

Final Boolean Result Size (B): 4066, **Est. Recall**: 8.3%, **Est. Precision**: 60.7%

Participant High Recall@B: 12.0% (UMKC6), **Median Recall@B**: 3.0%

Participant High Recall@25000: 44.7% (ursinus4), **Median Recall@25000**: 16.7%

5 Deepest Sampled Relevant Documents: piq79c00-3407.2 (otL07rvl-24173), cdn65e00-3009.0 (UIowa07LegE5-17078), bin58d00-2556.7 (otL07fbe-11824), urb90d00-2284.6 (wat6qap-9507), pcp77c00-2194.5 (UMKC2-8839)

Topic 55 (2007-A-4)

Request Text: Please produce any and all documents concerning the known radioactivity of apatite rock.

Initial Proposal by Defendant: `apatite w/15 radioactiv!`

Rejoinder by Plaintiff: `(Apatite OR "CA5(PO4)3OH" OR "CA5(PO4)3F" OR "CA5(PO4)3Cl" OR Fluorapatite OR Chlorapatite OR Hydroxylapatite) OR ((rock OR geolo!) AND (radioactiv! OR unstable OR instabil! OR radiat! OR radium OR polonium OR lead))`

Final Negotiated Boolean Query: `(Radioactiv! OR unstable OR instabil! OR radiat! OR radium OR polonium OR lead) AND (apatite OR "CA5(PO4)3OH" OR "CA5(PO4)3F" OR "CA5(PO4)3Cl" OR Fluorapatite OR Chlorapatite OR Hydroxylapatite)`

Sampling: 380213 pooled, 496 assessed, 46 judged relevant, 440 non-relevant, 10 gray, "C"=1.27, **Est. Rel.**: 5564.7

Final Boolean Result Size (B): 580, **Est. Recall**: 1.7%, **Est. Precision**: 22.5%

Participant High Recall@B: 6.4% (wat4feed), **Median Recall@B**: 0.6%

Participant High Recall@25000: 52.2% (fdwim7ss), **Median Recall@25000**: 3.8%

5 Deepest Sampled Relevant Documents: vwc40e00-1871.6 (wat4feed-3799), dmm74e00-1507.2 (fdwim7ss-2740), jwr99d00-1289.2 (fdwim7ss-2206), qsj72f00-92.3 (UMKC1-574), dtn01e00-91.4 (wat4feed-548)

Topic 56 (2007-A-5)

Request Text: Please produce any and all documents concerning soil water management as it pertains to commercial irrigation.

Initial Proposal by Defendant: `("soil water" w/10 manage!) AND "commercial irrigation"`

Rejoinder by Plaintiff: `(Soil! OR sewage OR sewer! OR septic OR drain! OR dirt OR field! OR groundwater OR (ground w/3 water)) AND (manage! OR control!) AND irrigat!`

Final Negotiated Boolean Query: `(((Soil! OR sewage OR sewer! OR septic OR drain! OR dirt OR field! OR groundwater OR (ground w/3 water)) AND (manage! OR "control system")) AND irrigat!)`

Sampling: 319017 pooled, 499 assessed, 112 judged relevant, 361 non-relevant, 26 gray, "C"=2.02, **Est. Rel.**: 2461.0

Final Boolean Result Size (B): 3288, **Est. Recall**: 46.8%, **Est. Precision**: 42.0%

Participant High Recall@B: 64.5% (UMKC5), **Median Recall@B**: 26.0%

Participant High Recall@25000: 87.3% (otL07frw), **Median Recall@25000**: 56.3%

5 Deepest Sampled Relevant Documents: nxw38d00-445.2 (ursinus8-2785), lmz34c00-421.3 (fdwim7ts-2369), amx11c00-291.1 (UMKC6-1055), rin34d00-283.6 (wat3desc-1007), cyp41a00-255.1 (UMKC2-842)

Topic 57 (2007-A-6)

Request Text: Please produce any and all documents that discuss methods for decreasing sugar loss in sugar-beet crops.

Initial Proposal by Defendant: `"sugar beet" AND "sugar loss"`

Rejoinder by Plaintiff: `sugar! AND (beet OR beets OR crop OR crops) AND (lost OR loss OR losses OR decreas! OR wane! OR reduc! OR prevent!)`

Final Negotiated Boolean Query: `(sugar-beet OR sugarbeet OR beet OR beets OR crop OR crops) w/75 (lost OR loss OR losses OR decreas! OR wane! OR reduc! OR prevent!) w/75 sugar!`

Sampling: 307648 pooled, 1000 assessed, 340 judged relevant, 643 non-relevant, 17 gray, "C"=4.67, **Est. Rel.**: 49048.1

Final Boolean Result Size (B): 3006, **Est. Recall**: 3.2%, **Est. Precision**: 64.4%

Participant High Recall@B: 5.7% (ursinus6), **Median Recall@B**: 2.9%

Participant High Recall@25000: 39.0% (wat4feed), **Median Recall@25000**: 13.4%

5 Deepest Sampled Relevant Documents: vet80a00-2564.3 (wat4feed-24583), urp52d00-2502.5 (wat4feed-23396), vxr81a00-2436.5 (fdwim7ss-22194), rrx98e00-2423.5 (catchup0701p-21963), fgo02a00-2354.2 (wat4feed-20777)

Topic 58 (2007-A-7)

Request Text: Please produce any and all documents that discuss health problems caused by HPF, including, but not limited to immune disorders, toxic myopathy, chronic fatigue syndrome, liver dysfunctions, irregular heart-beat, reactive depression, and memory loss.

Initial Proposal by Defendant: `phosphat! AND ("immune disorder!" OR "toxic myopathy" OR "chronic fatigue`

syndrome" OR "liver dysfunction!" OR "irregular heart-beat" OR "reactive depression" OR "memory loss") AND (cause OR relate OR associate! OR derive! OR correlate!)

Rejoinder by Plaintiff: (HPF OR phosphat! OR phosphorus OR fertiliz!) AND (illness! OR health OR disorder! OR toxic! OR "chronic fatigue" OR dysfunction! OR irregular OR memor! OR immun! OR myopath! OR liver! OR kidney! OR heart! OR depress! OR loss OR lost))

Final Negotiated Boolean Query: Phosphat! w/75 (caus! OR relat! OR assoc! OR derive! OR correlat!) w/75 (health OR disorder! OR toxic! OR "chronic fatigue" OR dysfunction! OR irregular OR memor! OR immun! OR myopath! OR liver! OR kidney! OR heart! OR depress! OR loss OR lost)

Sampling: 346836 pooled, 495 assessed, 41 judged relevant, 454 non-relevant, 0 gray, "C"=1.37, **Est. Rel.:** 1150.6

Final Boolean Result Size (B): 8183, **Est. Recall:** 94.0%, **Est. Precision:** 11.8%

Participant High Recall@B: 94.0% (wat7bool and 1 other), **Median Recall@B:** 7.0%

Participant High Recall@25000: 94.9% (otL07fb2x), **Median Recall@25000:** 9.4%

5 Deepest Sampled Relevant Documents: rmw20d00-571.8 (otL07fb-1204), mqo61a00-284.1 (wat6qap-471), riy94d00-70.5 (wat8gram-101), bcx01d00-66.5 (wat8gram-95), zyd58c00-23.7 (wat4feed-33)

Topic 59 (2007-A-8)

Request Text: Please produce any and all studies, reports, discussions or analyses of the limestone quicklime wastewater treatment method that discusses this treatment's effectiveness in minimizing water contamination.

Initial Proposal by Defendant: (limestone OR quicklime) AND "wastewater treatment"

Rejoinder by Plaintiff: (Limestone OR quicklime) AND (wastewater OR waste! OR water! OR sewage OR sewer! OR dispos! OR irrigate! OR well OR wells OR treat! OR purify! OR purification OR reduc! OR septic OR clean! OR steril! OR minim!)

Final Negotiated Boolean Query: (Limestone OR quicklime) w/75 (wastewater OR waste! OR water! OR sewage OR sewer! OR dispos! OR irrigate! OR well OR wells) w/75 (treat! OR purify! OR purification OR reduc! OR septic OR clean! OR steril! OR minim!)

Sampling: 329585 pooled, 499 assessed, 15 judged relevant, 482 non-relevant, 2 gray, "C"=1.09, **Est. Rel.:** 111.8

Final Boolean Result Size (B): 240, **Est. Recall:** 0.9%, **Est. Precision:** 0.8%

Participant High Recall@B: 60.8% (UMKC4), **Median Recall@B:** 7.2%

Participant High Recall@25000: 100.0% (CMUL07ibp and 12 others), **Median Recall@25000:** 63.7%

5 Deepest Sampled Relevant Documents: yuz34f00-38.1 (UMKC4-202), uql43d00-29.6 (CMUL07ibp-84), aqo33d00-13.3 (CMUL07std-20), bti91f00-11.2 (CMUL07ol-16), eex53e00-9.6 (SabL07ar1-13)

Topic 60 (2007-A-9)

Request Text: Please produce any and all documents that discuss phosphate precipitation as a method of water purification.

Initial Proposal by Defendant: (phosphate w/3 precipitation) AND (water w/3 purification)

Rejoinder by Plaintiff: phosphat! AND (precip! OR septic OR method!) AND purif!

Final Negotiated Boolean Query: (phosphat! w/75 (precip! OR septic OR method!)) AND ((water! OR waste!) w/75 purif!)

Sampling: 279129 pooled, 700 assessed, 10 judged relevant, 669 non-relevant, 21 gray, "C"=2.49, **Est. Rel.:** 83.2

Final Boolean Result Size (B): 1496, **Est. Recall:** 7.2%, **Est. Precision:** 0.5%

Participant High Recall@B: 71.8% (UMass11 and 1 other), **Median Recall@B:** 7.6%

Participant High Recall@25000: 100.0% (CMUL07ibp and 16 others), **Median Recall@25000:** 90.6%

5 Deepest Sampled Relevant Documents: ake51c00-53.7 (UMass10-163), tcg42d00-18.1 (ursinus3-48), rbc63d00-4.4 (ursinus1-11), oma59c00-1.0 (UMass13-3), bmc55c00-1.0 (otL07fbe-3)

Topic 61 (2007-A-10)

Request Text: Please produce any and all waste treatment schedules that discuss phosphate concentrations in water.

Initial Proposal by Defendant: ("waste treatment" w/3 schedule!) AND (phosphate w/5 water)

Rejoinder by Plaintiff: schedul! AND (phosphat! OR phosphor!) AND (water OR waste! OR runoff OR irrigat! OR drain! OR sewage OR sewer OR liquid!)

Final Negotiated Boolean Query: Treat! w/150 schedul! w/150 (phosphat! OR phosphor!) w/150 (water OR waste! OR runoff OR irrigat! OR drain! OR sewage OR sewer OR liquid!)

Sampling: 252532 pooled, 507 assessed, 64 judged relevant, 440 non-relevant, 3 gray, "C"=1.11, **Est. Rel.:** 372.1

Final Boolean Result Size (B): 296, **Est. Recall:** 43.9%, **Est. Precision:** 50.1%

Participant High Recall@B: 57.0% (otL07frw), **Median Recall@B:** 8.0%

Participant High Recall@25000: 98.9% (otL07fv), **Median Recall@25000:** 78.7%

5 Deepest Sampled Relevant Documents: bzz83e00-37.8 (wat6qap-116), una63e00-34.4 (otL07fb2x-91), nwe73e00-34.1 (otL07fb2x-89), txe73e00-29.4 (wat3desc-65), hpo02a00-27.0 (fdwim7sl-55)

Topic 62 (2007-A-11) *(This topic's assessments were completed too late for inclusion in the official set of 43 topics.)*

Request Text: Please produce any and all documents relating to any and all press releases concerning water contamination related to irrigation.

Initial Proposal by Defendant: "press release!" AND "water contamination" AND irrigation

Rejoinder by Plaintiff: ("press release" OR "press releases" OR news! OR report! OR public! OR announce! OR noti!) AND (contamina! OR toxic! OR pollut!) AND (irrigat! OR water)

Final Negotiated Boolean Query: ("press release" OR "press releases" OR news! OR report! OR public! OR

announce! OR noti!) w/150 water w/100 (contamina! OR toxic! OR pollut!) w/150 irrigat!

Sampling: 355008 pooled, 499 assessed, 25 judged relevant, 466 non-relevant, 8 gray, "C"=0.50, **Est. Rel.:** 271.0

Final Boolean Result Size (B): 354, **Est. Recall:** 1.1%, **Est. Precision:** 1.9%

Participant High Recall@B: 47.5% (otL07fbe), **Median Recall@B:** 1.1%

Participant High Recall@25000: 67.3% (otL07fbe), **Median Recall@25000:** 30.3%

5 Deepest Sampled Relevant Documents: ahq25c00-64.0 (otL07fbe-334), aov39e00-59.6 (otL07fbe-189), bei86c00-57.8 (CMUL07o1-157), cul10a00-35.2 (UMKC6-35), ufm82c00-13.1 (CMUL07ibp-8)

Topic 63 (2007-A-12)

Request Text: Please produce any and all documents that specifically discuss an exclusivity clause in a sugar contract.

Initial Proposal by Defendant: "sugar contract" AND "exclusivity clause"

Rejoinder by Plaintiff: Sugar AND (contract! OR agreement! OR deal! OR exclusiv!)

Final Negotiated Boolean Query: (Sugar w/20 (contract! OR agreement! OR deal!)) AND exclusiv!

Sampling: 341624 pooled, 506 assessed, 11 judged relevant, 479 non-relevant, 16 gray, "C"=0.79, **Est. Rel.:** 18.6

Final Boolean Result Size (B): 294, **Est. Recall:** 26.8%, **Est. Precision:** 2.4%

Participant High Recall@B: 89.3% (UMass13), **Median Recall@B:** 16.1%

Participant High Recall@25000: 100.0% (otL07fv and 7 others), **Median Recall@25000:** 83.9%

5 Deepest Sampled Relevant Documents: hko97e00-8.6 (otL07pb-8), whv93d00-1.0 (wat1fuse-5), avt49c00-1.0 (otL07fv-4), gko97e00-1.0 (otL07pb-3), axb56e00-1.0 (UIowa07LegE1-3)

Topic 64 (2007-A-13)

Request Text: Please produce any and all documents that specifically discuss any and all deceptive implied health claims related to sugar.

Initial Proposal by Defendant: deceptive AND (health w/5 claim!) w/75 sugar

Rejoinder by Plaintiff: (Decept! OR deceive OR false OR inaccurate OR mislead! OR misinform! OR misguid! OR untrue OR claim! OR state! OR declar! OR inform!) AND sugar

Final Negotiated Boolean Query: (Decept! OR deceive OR false OR inaccurate OR mislead! OR misinform! OR misguid! OR untrue) w/15 (claim! OR state OR statement! OR declare! OR inform!) w/75 sugar

Sampling: 335933 pooled, 594 assessed, 16 judged relevant, 558 non-relevant, 20 gray, "C"=1.64, **Est. Rel.:** 159.1

Final Boolean Result Size (B): 131, **Est. Recall:** 6.9%, **Est. Precision:** 8.3%

Participant High Recall@B: 41.0% (wat4feed), **Median Recall@B:** 0.6%

Participant High Recall@25000: 99.4% (wat1fuse), **Median Recall@25000:** 3.8%

5 Deepest Sampled Relevant Documents: fzw07d00-79.8 (wat4feed-133), ubd22d00-18.2 (wat4feed-97), bbb63e00-14.1 (wat4feed-50), bhe58c00-13.1 (wat4feed-43), hvh77e00-11.9 (wat4feed-36)

Topic 65 (2007-A-14)

Request Text: Please produce any and all documents that explicitly discuss candy packaging, the labeling of candy, or which provide examples of candy packages or wrappers.

Initial Proposal by Defendant: candy w/5 (packag! OR label! OR wrapper!)

Rejoinder by Plaintiff: Candy AND (pack! OR label! OR wrap! OR adverti! OR box OR ingredient! OR contain!)

Final Negotiated Boolean Query: Candy w/15 (pack! OR label! OR wrap! OR adverti! OR box OR ingredient! OR contain!)

Sampling: 338958 pooled, 500 assessed, 58 judged relevant, 425 non-relevant, 17 gray, "C"=1.04, **Est. Rel.:** 609.7

Final Boolean Result Size (B): 8700, **Est. Recall:** 67.2%, **Est. Precision:** 7.8%

Participant High Recall@B: 97.0% (otL07rvl), **Median Recall@B:** 65.7%

Participant High Recall@25000: 99.2% (wat1fuse), **Median Recall@25000:** 68.8%

5 Deepest Sampled Relevant Documents: abh6aa00-159.8 (CMUL07ibp-183), aue3aa00-143.0 (SabL07ab1-162), czp15d00-75.4 (wat5nofeed-82), gtt92d00-41.3 (otL07fbe-44), rim00e00-35.8 (UMKC6-38)

Topic 66 (2007-A-15)

Request Text: Please produce any and all documents concerning the formation of the U.S. Beet Sugar Association.

Initial Proposal by Defendant: ((U.S. OR US) w/3 "Beet Sugar Association") AND (impact OR policy OR policies OR legislation)

Rejoinder by Plaintiff: ("beet sugar" w/30 association!) AND (form OR formed OR start! OR create! OR founded OR began OR first OR initiat! OR initial OR begin! OR conceive!)

Final Negotiated Boolean Query: "beet sugar" AND association! AND (form OR formed OR start! OR create! OR founded OR began OR first OR initiat! OR initial OR begin! OR conceive!)

Sampling: 415787 pooled, 488 assessed, 25 judged relevant, 455 non-relevant, 8 gray, "C"=1.01, **Est. Rel.:** 454.6

Final Boolean Result Size (B): 162, **Est. Recall:** 9.0%, **Est. Precision:** 27.1%

Participant High Recall@B: 9.0% (otL07fb2x and 1 other), **Median Recall@B:** 1.8%

Participant High Recall@25000: 99.1% (UMass11), **Median Recall@25000:** 6.7%

5 Deepest Sampled Relevant Documents: hls25e00-400.7 (UMass10-440), jgu96d00-21.4 (CMUL07ibt-64), ewn59e00-6.4 (wat6qap-8), ogi57d00-5.0 (wat7bool-6), apk48c00-1.0 (otL07pb-5)

Topic 67 (2007-A-16)

Request Text: Please produce any and all documents that explicitly refer to "The Sugar Program," and/or discuss the

formation, contemplation or existence of a sugar cartel, or that discuss the sugar lobby in the context of Sugar Acts passed by Congress.

Initial Proposal by Defendant: `"Sugar Program" AND "sugar cartel"`

Rejoinder by Plaintiff: `"Sugar Program" OR (Sugar AND (lobby! OR Congress OR "sugar acts" OR law! OR polic! OR legis! OR regulat! OR ordinance! OR control! OR cartel! OR combine OR syndicate OR trust OR conspir!))`

Final Negotiated Boolean Query: `"Sugar Program" OR ((sugar OR sucrose) AND (cartel OR combine)) OR ((Sugar OR sucrose) w/15 (lobby! OR Congress OR "sugar acts" OR law! OR polic! OR legis! OR regulat! OR ordinance! OR control!))`

Sampling: 383624 pooled, 493 assessed, 75 judged relevant, 418 non-relevant, 0 gray, "C"=1.01, **Est. Rel.:** 41189.6

Final Boolean Result Size (B): 13241, **Est. Recall:** 1.4%, **Est. Precision:** 5.7%

Participant High Recall@B: 13.0% (UMass12), **Median Recall@B:** 3.8%

Participant High Recall@25000: 26.9% (ursinus8), **Median Recall@25000:** 8.0%

5 Deepest Sampled Relevant Documents: lgo18d00-4085.5 (ursinus8-22561), idj24d00-3706.9 (UIowa07LegE3-14476), gog85a00-3670.2 (UMKC3-13938), izo81d00-3627.2 (SabL07ar2-13343), clc07e00-2190.5 (UIowa07LegE0-12801)

Topic 69 (2007-B-1)

Request Text: All documents referring or relating to indoor smoke ventilation.

Initial Proposal by Defendant: `"indoor smoke ventilation"`

Rejoinder by Plaintiff: `(indoor OR inside) AND ("smoke ventilation" OR filtration)`

Final Negotiated Boolean Query: `(indoor OR inside) w/25 ("smoke ventilation" OR filtration)`

Sampling: 226267 pooled, 495 assessed, 280 judged relevant, 110 non-relevant, 105 gray, "C"=1.55, **Est. Rel.:** 37457.5

Final Boolean Result Size (B): 5123, **Est. Recall:** 8.3%, **Est. Precision:** 96.9%

Participant High Recall@B: 13.7% (wat2nobool), **Median Recall@B:** 11.5%

Participant High Recall@25000: 61.0% (UMKC3), **Median Recall@25000:** 39.6%

5 Deepest Sampled Relevant Documents: khk42d00-3732.5 (catchup0701p-22822), vyh91f00-2931.7 (SabL07ar1-10985), bfm91f00-2212.0 (UMKC3-6149), mjf08d00-2122.2 (otL07fbe-5715), art52c00-2042.9 (CMUL07ibs-5354)

Topic 70 (2007-B-2)

Request Text: All documents that make reference to the smell of baked goods, including but not limited to baked cookies.

Initial Proposal by Defendant: `smell w/3 ("baked goods" OR "baked cookie!")`

Rejoinder by Plaintiff: `(smell OR aroma) AND baked OR pie! OR bread! OR cake OR food!`

Final Negotiated Boolean Query: `(smell OR aroma) w/15 ("baked good!" OR "baked cookie!" OR pie! OR bread! OR cake! OR foodstuff!)`

Sampling: 352690 pooled, 499 assessed, 43 judged relevant, 452 non-relevant, 4 gray, "C"=1.26, **Est. Rel.:** 19828.1

Final Boolean Result Size (B): 1381, **Est. Recall:** 0.1%, **Est. Precision:** 1.7%

Participant High Recall@B: 1.1% (UIowa07LegE2), **Median Recall@B:** 0.1%

Participant High Recall@25000: 35.4% (ursinus4), **Median Recall@25000:** 1.6%

5 Deepest Sampled Relevant Documents: ajk32d00-3551.3 (ursinus4-15444), eth06c00-2631.9 (UIowa07LegE1-7002), bue80c00-2388.1 (otL07fbe-5760), fyl99d00-2202.2 (UIowa07LegE1-4959), eli23e00-1957.4 (ursinus1-4053)

Topic 71 (2007-B-3)

Request Text: All documents discussing the condition of bromhidrosis (a/k/a body odor).

Initial Proposal by Defendant: `bromhidrosis`

Rejoinder by Plaintiff: `bromhidrosis OR ((body OR human OR person) AND odor!))`

Final Negotiated Boolean Query: `bromhidrosis OR ((body OR human OR person) w/3 odor!)`

Sampling: 356441 pooled, 697 assessed, 308 judged relevant, 376 non-relevant, 13 gray, "C"=2.28, **Est. Rel.:** 77466.9

Final Boolean Result Size (B): 4527, **Est. Recall:** 4.3%, **Est. Precision:** 69.5%

Participant High Recall@B: 5.8% (CMUL07ibt and 2 others), **Median Recall@B:** 3.4%

Participant High Recall@25000: 23.1% (otL07pb), **Median Recall@25000:** 11.2%

5 Deepest Sampled Relevant Documents: myo01d00-3186.1 (wat4feed-20024), rsi55d00-3112.8 (ursinus5-18803), cqr42e00-3084.7 (fdwim7xj-18361), tfp94a00-3049.7 (IowaSL07Ref-17826), dns25e00-2710.8 (UMKC5-13499)

Topic 72 (2007-B-4)

Request Text: All documents referring to the scientific or chemical process(es) which result in onions have the effect of making persons cry.

Initial Proposal by Defendant: `("scientific process!" OR "chemical process!") AND onion AND cry`

Rejoinder by Plaintiff: `((scien! OR research! OR chemical) AND onion!) AND (cries OR cry! OR tear!)`

Final Negotiated Boolean Query: `((scien! OR research! OR chemical) w/25 onion!) AND (cries OR cry! OR tear!)`

Sampling: 400625 pooled, 499 assessed, 11 judged relevant, 477 non-relevant, 11 gray, "C"=0.60, **Est. Rel.:** 97.7

Final Boolean Result Size (B): 119, **Est. Recall:** 77.9%, **Est. Precision:** 43.0%

Participant High Recall@B: 77.9% (otL07fb), **Median Recall@B:** 3.1%

Participant High Recall@25000: 100.0% (UMKC5 and 6 others), **Median Recall@25000:** 50.5%

5 Deepest Sampled Relevant Documents: tmj97c00-20.7 (otL07fb-94), hes71f00-20.6 (otL07fbe-91), brq10e00-18.8

(wat7bool-54), cmj97c00-12.6 (otL07frw-16), vlj97c00-12.2 (CMUL07ibp-15)

Topic 73 (2007-B-5)

Request Text: Any advertisements or draft advertisements that target women seen in the kitchen or cooking.

Initial Proposal by Defendant: `advertisement AND "target! women"`

Rejoinder by Plaintiff: `(("ad campaign" OR advertis!) AND (woman OR women OR girl OR female) AND (kitchen OR cook!)`

Final Negotiated Boolean Query: `(("ad campaign" OR advertis!) w/25 (woman OR women OR girl OR female)) AND (kitchen OR cook!)`

Sampling: 370452 pooled, 498 assessed, 72 judged relevant, 426 non-relevant, 0 gray, "C"=0.74, **Est. Rel.**: 31894.5

Final Boolean Result Size (B): 4085, **Est. Recall:** 4.9%, **Est. Precision:** 41.5%

Participant High Recall@B: 8.7% (UMKC5), **Median Recall@B:** 1.0%

Participant High Recall@25000: 51.6% (UMKC2), **Median Recall@25000:** 4.9%

5 Deepest Sampled Relevant Documents: tdc58e00-4345.7 (UMKC2-24574), qyf46e00-4279.2 (UIowa07LegE5-21967), nth79d00-4259.8 (CMUL07irs-21294), myc81a00-4067.3 (UIowa07LegE5-16135), djv94c00-3504.2 (wat6qap-8668)

Topic 74 (2007-B-6)

Request Text: All scientific studies expressly referencing health effects tied to indoor air quality.

Initial Proposal by Defendant: `"health effect!" w/10 "air quality"`

Rejoinder by Plaintiff: `(scien! OR stnd! OR research) AND ("air qnality" OR health)`

Final Negotiated Boolean Query: `(scien! OR stnd! OR research) AND ("air qnality" w/15 health)`

Sampling: 225883 pooled, 499 assessed, 299 judged relevant, 196 non-relevant, 4 gray, "C"=1.50, **Est. Rel.**: 62406.2

Final Boolean Result Size (B): 20516, **Est. Recall:** 22.2%, **Est. Precision:** 77.2%

Participant High Recall@B: 32.8% (wat3desc), **Median Recall@B:** 24.3%

Participant High Recall@25000: 40.0% (UMKC4), **Median Recall@25000:** 29.2%

5 Deepest Sampled Relevant Documents: vkm92d00-3123.0 (Dartmouth1-19609), gew24e00-3095.9 (SabL07ab1-18917), ljq87c00-3070.3 (otL07fbe-18296), mhf57d00-2942.7 (UIowa07LegE3-15606), bof25d00-2912.1 (otL07pb-15048)

Topic 75 (2007-B-7)

Request Text: All documents that memorialize any statement or suggestion from an elected federal public official that further research is necessary to improve indoor environmental air quality.

Initial Proposal by Defendant: `"public official" AND research w/10 ("indoor air quality" OR "indoor environment! air quality")`

Rejoinder by Plaintiff: `(("public official" OR senator OR representative OR congressman OR congresswoman OR president OR vice-president OR VP) AND ((research OR scienc! OR stnd!) AND (indoor AND (environment! w/5 "air quality") AND (statement OR "public debate" OR suggestion OR remark!)`

Final Negotiated Boolean Query: `("public official" OR senator OR representative OR congressman OR congresswoman OR president OR vice-president OR VP) AND ((research OR scienc! OR stud!) w/25 indoor w/25 environment! w/5 "air quality") AND (statement OR "pnblic debate" OR snggestion OR remark!)`

Sampling: 263784 pooled, 499 assessed, 23 judged relevant, 469 non-relevant, 7 gray, "C"=0.91, **Est. Rel.**: 228.1

Final Boolean Result Size (B): 788, **Est. Recall:** 13.5%, **Est. Precision:** 5.3%

Participant High Recall@B: 45.6% (SabL07ab1 and 1 other), **Median Recall@B:** 4.0%

Participant High Recall@25000: 100.0% (fdwim7ss and 11 others), **Median Recall@25000:** 86.1%

5 Deepest Sampled Relevant Documents: bxc52c00-89.4 (UMKC4-188), zku96d00-60.8 (SabL07ab1-90), kiu34e00-30.2 (SabL07ab1-34), twh67c00-28.7 (otL07fb2x-32), iif12f00-1.0 (CMUL07o1-5)

Topic 76 (2007-B-8)

Request Text: All documents that make reference to any public meeting or conference held in Washington, D.C. on the subject of indoor air quality.

Initial Proposal by Defendant: `("pnblic meeting" OR conference) AND "Washington, D.C." AND "indoor air quality"`

Rejoinder by Plaintiff: `((meeting OR conference OR event OR symposium) AND (Washington! OR "D.C." OR "District of Columbia") AND (indoor AND "air qnality")`

Final Negotiated Boolean Query: `(meeting OR conference OR event OR symposium) AND (Washington! OR "D.C." OR "District of Columbia") AND (indoor w/5 "air quality")`

Sampling: 195688 pooled, 1000 assessed, 254 judged relevant, 730 non-relevant, 16 gray, "C"=6.49, **Est. Rel.**: 4408.9

Final Boolean Result Size (B): 22518, **Est. Recall:** 50.8%, **Est. Precision:** 10.9%

Participant High Recall@B: 77.8% (CMUL07std), **Median Recall@B:** 50.8%

Participant High Recall@25000: 78.4% (ursinus2), **Median Recall@25000:** 52.1%

5 Deepest Sampled Relevant Documents: ijz83e00-415.2 (UIowa07LegE3-2968), qwf00a00-403.1 (UIowa07LegE1-2873), agq30c00-396.2 (UIowa07LegE3-2819), bwf69d00-345.7 (fdwim7ts-2430), jxp57d00-304.8 (SabL07ar1-2122)

Topic 77 (2007-B-9)

Request Text: All documents that refer or relate to the effect of smoke on bystanders, excluding studies on cigarette smoking or tobacco smoke.

Initial Proposal by Defendant: `(effect AND smoke) w/5 bystander BUT NOT (tobacco OR cigarette)`

Rejoinder by Plaintiff: `(smok! OR effect) AND (bystander! OR "third party" OR passive!)`

Final Negotiated Boolean Query: `(smok! AND bystander!) BUT NOT (tobacco OR cigarette)`

Sampling: 345347 pooled, 499 assessed, 11 judged relevant, 477 non-relevant, 11 gray, "C"=0.83, **Est. Rel.:** 11233.8

Final Boolean Result Size (B): 154, **Est. Recall:** 0.0%, **Est. Precision:** 0.0%

Participant High Recall@B: 0.2% (wat4feed), **Median Recall@B:** 0.0%

Participant High Recall@25000: 53.1% (wat4feed), **Median Recall@25000:** 0.0%

5 Deepest Sampled Relevant Documents: ivm59c00-4116.8 (otL07rvl-19345), zul52d00-3883.9 (wat4feed-14441), usn97c00-1805.7 (wat4feed-2346), gof44c00-1153.1 (IowaSL0706-1244), rrp87d00-250.6 (wat4feed-219)

Topic 78 (2007-B-10)

Request Text: All documents referencing patents on odors, excluding tobacco or cigarette related patents.

Initial Proposal by Defendant: `(patent! w/15 odor!) BUT NOT (tobacco OR cigarette)`

Rejoinder by Plaintiff: `patent! AND odor!`

Final Negotiated Boolean Query: `(patent! AND odor!) BUT NOT (tobacco OR cigarette!)`

Sampling: 288711 pooled, 499 assessed, 24 judged relevant, 440 non-relevant, 35 gray, "C"=1.30, **Est. Rel.:** 835.4

Final Boolean Result Size (B): 1611, **Est. Recall:** 25.3%, **Est. Precision:** 11.7%

Participant High Recall@B: 61.3% (otL07pb), **Median Recall@B:** 7.7%

Participant High Recall@25000: 99.9% (CMUL07ibt and 7 others), **Median Recall@25000:** 38.4%

5 Deepest Sampled Relevant Documents: anc13c00-232.2 (wat8gram-1081), agu15d00-191.2 (IowaSL0704-611), xph02a00-130.5 (UIowa07LegE2-285), wau60c00-52.2 (otL07fv-81), gnb97c00-45.0 (otL07pb-68)

Topic 79 (2007-C-1)

Request Text: All documents making a connection between the music and songs of Peter, Paul, and Mary, Joan Baez, or Bob Dylan, and the sale of cigarettes.

Initial Proposal by Defendant: `(music OR songs) AND (((peter w/2 paul AND (paul w/2 mary)) OR "bob Dylan" OR "joan baez") AND sale`

Rejoinder by Plaintiff: `((peter AND paul AND mary) OR dylan OR baez) AND (sale! OR sell! OR advertis! OR promot! OR market!)`

Final Negotiated Boolean Query: `(((peter w/3 paul) AND (paul w/3 mary)) OR (simon w/3 garfunkel) OR "Bob Dylan" OR "Joan Baez") AND (sale! OR sell! OR advertis! OR promot! OR market!)`

Sampling: 409225 pooled, 491 assessed, 35 judged relevant, 448 non-relevant, 8 gray, "C"=1.03, **Est. Rel.:** 1486.6

Final Boolean Result Size (B): 317, **Est. Recall:** 6.5%, **Est. Precision:** 50.2%

Participant High Recall@B: 9.5% (otL07frw), **Median Recall@B:** 3.3%

Participant High Recall@25000: 100.0% (otL07fbe), **Median Recall@25000:** 10.9%

5 Deepest Sampled Relevant Documents: thr11a00-766.2 (otL07fbe-932), ecp25d00-478.5 (IowaSL0706-545), bea05d00-51.9 (otL07fbe-294), goq26e00-48.3 (SabL07arbn-208), dhv36c00-28.4 (fdwim7xj-53)

Topic 80 (2007-C-2)

Request Text: All documents making a connection between folk songs and music and the sale of cigarettes.

Initial Proposal by Defendant: `"folk songs" AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Rejoinder by Plaintiff: `folk AND (sale OR sell! OR promot! OR advertis! OR market!)`

Final Negotiated Boolean Query: `("folk songs" OR "folk music" OR "folk artists") AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Sampling: 364619 pooled, 999 assessed, 391 judged relevant, 602 non-relevant, 6 gray, "C"=4.40, **Est. Rel.:** 38649.9

Final Boolean Result Size (B): 331, **Est. Recall:** 0.6%, **Est. Precision:** 81.9%

Participant High Recall@B: 0.8% (otL07rvl and 1 other), **Median Recall@B:** 0.6%

Participant High Recall@25000: 39.9% (otL07rvl and 1 other), **Median Recall@25000:** 15.8%

5 Deepest Sampled Relevant Documents: vxf58c00-2519.3 (UIowa07LegE0-22342), oxb28d00-2496.5 (ursinus5-21938), cwq61c00-2392.0 (IowaSL0702-20178), iyb70f00-2362.3 (UMKC6-19703), bcm43f00-2074.0 (wat4feed-15594)

Topic 82 (2007-C-4)

Request Text: All documents discussing the color of the paper used to make cigarettes in connection with increasing sales.

Initial Proposal by Defendant: `(color! w/2 paper) AND (increas! w/3 sales)`

Rejoinder by Plaintiff: `(color! OR shade! OR pastel! OR tint!) AND paper AND (sale! OR sell!)`

Final Negotiated Boolean Query: `((color! OR shade! OR pastel! OR tint!) w/5 paper) AND (increas! w/15 (sale! OR sell!))`

Sampling: 418281 pooled, 491 assessed, 176 judged relevant, 310 non-relevant, 5 gray, "C"=0.34, **Est. Rel.:** 75558.9

Final Boolean Result Size (B): 888, **Est. Recall:** 0.8%, **Est. Precision:** 61.2%

Participant High Recall@B: 1.1% (otL07pb), **Median Recall@B:** 0.4%

Participant High Recall@25000: 33.1% (otL07fbe), **Median Recall@25000:** 4.7%

5 Deepest Sampled Relevant Documents: nlq90a00-4664.5 (UMass10-23634), bim31f00-4627.1 (otL07fbe-21093), qwf74f00-4596.9 (otL07fbe-19384), dmv71d00-4396.0 (CMUL07irs-12373), hko20f00-4366.2 (otL07fbe-11712)

Topic 83 (2007-C-5)

Request Text: All documents discussing using psychedelic colors to increase sales of cigarettes.

Initial Proposal by Defendant: `"psychedelic color!" AND (increas! w/3 sales)`

Rejoinder by Plaintiff: `psyched*lic AND (sale! OR sell! OR promot! OR advertis! OR market`

Final Negotiated Boolean Query: `psyched*lic AND color! AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Sampling: 385402 pooled, 496 assessed, 44 judged relevant, 452 non-relevant, 0 gray, "C"=0.52, **Est. Rel.:** 13987.5

Final Boolean Result Size (B): 281, **Est. Recall:** 0.8%, **Est. Precision:** 32.1%

Participant High Recall@B: 1.2% (otL07frw), **Median Recall@B:** 0.4%

Participant High Recall@25000: 33.3% (wat4feed), **Median Recall@25000:** 1.6%

5 Deepest Sampled Relevant Documents: bco01e00-4467.9 (wat4feed-21831), aqm49d00-4228.5 (UMass10-14250), xfk10d00-3935.5 (fdwim7sl-9612), ait55f00-967.7 (ursinus4-624), ect68c00-45.9 (SabL07ab1-131)

Topic 84 (2007-C-6)

Request Text: All documents referencing "Bonnie and Clyde" or a James Bond film or the films of Stanley Kubrick in connection with sales of cigarettes.

Initial Proposal by Defendant: `(Bonnie w/2 Clyde) OR "James Bond film" OR "Stanley Kubrick"`

Rejoinder by Plaintiff: `((Bonnie w/3 Clyde) OR ("James Bond" OR "Agent 007" OR Goldfinger OR "Dr. No" OR "Dr No" OR "From Russia With Love" OR Thunderball OR "Sean Connery" OR "Roger Moore") OR ("Stanley Knbrick" OR (Kubrick w/3 (movie! OR film)) OR "2001: A Space Odyssey" OR HAL OR Lolita OR "James Mason" OR "Dr Strangelove" OR "Dr. Strangelove" OR "Peter Sellers")) AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Final Negotiated Boolean Query: `((Bonnie w/3 Clyde) OR ("James Bond" OR "Agent 007" OR Goldfinger OR "Dr No" OR "Dr. No" OR "From Russia With Love" OR Thunderball) OR ("Stanley Kubrick" OR (Knbrick w/3 (movie! OR film)) OR "2001: A Space Odyssey" OR Lolita OR "Dr Strangelove" OR "Dr. Strangelove")) AND (sale! OR sell! OR promot! OR advertis! OR market!)`

Sampling: 476252 pooled, 498 assessed, 63 judged relevant, 431 non-relevant, 4 gray, "C"=0.73, **Est. Rel.:** 450.1

Final Boolean Result Size (B): 2493, **Est. Recall:** 100.0%, **Est. Precision:** 18.6%

Participant High Recall@B: 100.0% (CMUL07ibp and 3 others), **Median Recall@B:** 38.3%

Participant High Recall@25000: 100.0% (CMUL07ibp and 7 others), **Median Recall@25000:** 85.7%

5 Deepest Sampled Relevant Documents: mrj70e00-137.8 (CMUL07ibs-139), qcm09c00-71.6 (otL07fb-61), gqd62d00-41.4 (wat5nofeed-33), mel03f00-41.4 (ursinus7-33), gsd19d00-29.6 (otL07db-23)

Topic 85 (2007-C-7)

Request Text: All documents discussing or referencing generally accepted accounting principles in connection with the decision to record as sales products shipped to distributors on a sale-or-return basis, and the implementation thereof.

Initial Proposal by Defendant: `("gaap" OR "generally accepted accounting principle!") AND (revenne! OR records OR recording OR account!)) AND (sale w/5 retnrn)`

Rejoinder by Plaintiff: `("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial accounting standards board" OR "sab" OR "Staff acconnting bulletin" OR "sas" OR (statement w/2 "auditing standards")) AND ((sale! OR allowance OR reserve! OR right! OR entitle! OR could) w/5 return!)`

Final Negotiated Boolean Query: `("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial acconnting standards board" OR "sab" OR "Staff accounting bulletin" OR "sas" OR (statement w/2 "anditing standards")) AND (revenue! OR recording OR records OR acconnt!) AND (sale! OR allowance OR reserve!) AND ((right! OR entitle! OR could) w/5 retnrn!)`

Sampling: 361317 pooled, 497 assessed, 96 judged relevant, 392 non-relevant, 9 gray, "C"=0.80, **Est. Rel.:** 3890.7

Final Boolean Result Size (B): 1305, **Est. Recall:** 13.8%, **Est. Precision:** 44.3%

Participant High Recall@B: 31.6% (IowaSL0705 and 1 other), **Median Recall@B:** 9.8%

Participant High Recall@25000: 77.5% (ursinus7), **Median Recall@25000:** 42.8%

5 Deepest Sampled Relevant Documents: lma14c00-221.5 (otL07fbe-1170), yip12d00-214.3 (ursinus3-957), ynz51d00-206.1 (SabL07ar1-783), yde76d00-203.7 (otL07fb-742), xsh35f00-201.7 (UIowa07LegE2-710)

Topic 86 (2007-C-8)

Request Text: All documents discussing or referencing both generally accepted accounting principles and the defendants' decision to restate its financial results.

Initial Proposal by Defendant: `restate AND ("gaap" OR "generally accepted accounting principle!" OR "financial accounting standards board" OR "staff accounting bulletin" OR (statement w/2 "auditing standards")) AND (revenue! OR records OR recording)`

Rejoinder by Plaintiff: `(restate! OR revise!) AND (gaap OR "generally accepted accounting principle!" OR fasb OR financial OR sab OR accounting OR sas OR anditing`

Final Negotiated Boolean Query: `(restate! OR revise!) AND ("gaap" OR "generally accepted accounting principle!" OR "fasb" OR "financial accounting standards board" OR "sab" OR "staff accounting bnlletin" OR "sas" OR (statement w/2 "auditing standards")) AND (revenue! OR records OR recording)`

Sampling: 370179 pooled, 499 assessed, 21 judged relevant, 465 non-relevant, 13 gray, "C"=1.21, **Est. Rel.:** 8830.1

Final Boolean Result Size (B): 6446, **Est. Recall:** 4.8%, **Est. Precision:** 8.2%

Participant High Recall@B: 24.2% (UIowa07LegE0), **Median Recall@B:** 4.8%

Participant High Recall@25000: 51.2% (IowaSL07Ref), **Median Recall@25000:** 6.9%

5 Deepest Sampled Relevant Documents: gnl85a00-3408.1 (otL07fbe-12953), vgm85c00-981.3 (wat4feed-4971), ohk68c00-830.7 (wat3desc-2826), jbm58c00-755.8 (UIowa07LegE0-2210), rsc58c00-607.5 (catchup0701p-1390)

Topic 87 (2007-C-9)

Request Text: All documents discussing Securities and Exchange Commission 10b-5 reports or reporting requirements.

Initial Proposal by Defendant: `"10b-5 Report!" AND (Securities w/3 "exchange commission")`

Rejoinder by Plaintiff: `10! AND (SEC OR (Securities w/3 "Exchange Commission"))`

Final Negotiated Boolean Query: `10b-5 AND (SEC OR (securities w/3 "exchange commission"))`

Sampling: 351913 pooled, 496 assessed, 41 judged relevant, 444 non-relevant, 11 gray, "C"=1.36, **Est. Rel.**: 875.3

Final Boolean Result Size (B): 138, **Est. Recall**: 3.8%, **Est. Precision**: 40.1%

Participant High Recall@B: 6.5% (UMKC5 and 1 other), **Median Recall@B**: 1.8%

Participant High Recall@25000: 100.0% (UMKC2 and 6 others), **Median Recall@25000**: 11.0%

5 Deepest Sampled Relevant Documents: bgb65a00-758.0 (UMKC5-1215), hse51c00-21.5 (UMass12-132), ckq92f00-18.0 (fdwim7ts-70), xbn93c00-13.1 (UMKC1-34), cqp92e00-7.5 (wat8gram-14)

Topic 89 (2007-D-1)

Request Text: Submit all documents listing monthly and/or annual sales for companies in the property and casualty insurance business in the United States between 1980 and the present.

Initial Proposal by Defendant: `(("monthly sales" OR "annual sales") AND ("property insurance" OR "casualty insurance")) AND (United States OR U.S.) AND (198* OR 199*)`

Rejoinder by Plaintiff: `(month! OR annual!) AND (sales OR sell! OR revenue) AND insurance AND (198* OR 199*)`

Final Negotiated Boolean Query: `(((month! OR annual!) w/15 (sales OR sell! OR revenue)) AND ((property OR casualty) AND insurance)) BUT NOT (England OR "Great Britain" OR U.K. OR UK)`

Sampling: 345011 pooled, 496 assessed, 78 judged relevant, 392 non-relevant, 26 gray, "C"=1.16, **Est. Rel.**: 6083.6

Final Boolean Result Size (B): 3636, **Est. Recall**: 9.9%, **Est. Precision**: 16.8%

Participant High Recall@B: 21.0% (wat3desc), **Median Recall@B**: 8.9%

Participant High Recall@25000: 91.9% (otL07fbe), **Median Recall@25000**: 17.1%

5 Deepest Sampled Relevant Documents: mwv44d00-3850.5 (otL07fbe-19428), qxd35a00-561.1 (SabL07ab1-2850), qhc48c00-495.2 (IowaSL0707-1800), jyl13d00-259.1 (otL07fbe-467), ltq35f00-203.1 (SabL07ab1-327)

Topic 90 (2007-D-2)

Request Text: Submit all documents listing monthly and/or annual sales for companies in the property and casualty insurance business in England for all available years.

Initial Proposal by Defendant: `(("monthly sales" OR "annual sales") AND ("property insurance" OR "casualty insurance")) AND England`

Rejoinder by Plaintiff: `(month! OR annual!) AND (sales OR sell! OR revenue) AND Insurance AND (England OR Brit! OR U.K. OR UK)`

Final Negotiated Boolean Query: `(((month! OR annual!) w/15 sales) AND ((property OR casualty) AND insurance)) AND (England OR "Great Britain" OR U.K. OR UK)`

Sampling: 330155 pooled, 492 assessed, 34 judged relevant, 458 non-relevant, 0 gray, "C"=1.30, **Est. Rel.**: 1066.1

Final Boolean Result Size (B): 2665, **Est. Recall**: 10.3%, **Est. Precision**: 9.0%

Participant High Recall@B: 63.9% (otL07fbe), **Median Recall@B**: 14.6%

Participant High Recall@25000: 99.3% (otL07fbe), **Median Recall@25000**: 33.7%

5 Deepest Sampled Relevant Documents: fds95c00-393.5 (otL07fbe-1954), umo55a00-159.9 (CMUL07irt-297), wyu71f00-153.4 (wat4feed-280), cmr03f00-91.2 (otL07pb-143), hre33f00-85.8 (otL07fbe-133)

Topic 92 (2007-D-4)

Request Text: Submit all documents relating to competition or market share in the property and casualty insurance industry, including, but not limited to, market studies, forecasts and surveys.

Initial Proposal by Defendant: `("market stud!" OR forecast! OR survey!) AND "market share" AND ("property insurance" OR "casualty insurance")`

Rejoinder by Plaintiff: `(competition OR market OR share) AND insurance`

Final Negotiated Boolean Query: `("market stud!" OR forecast! OR survey!) AND (competition OR share) AND (property OR casualty) AND insurance`

Sampling: 313137 pooled, 498 assessed, 117 judged relevant, 369 non-relevant, 12 gray, "C"=1.63, **Est. Rel.**: 18070.2

Final Boolean Result Size (B): 9401, **Est. Recall**: 12.8%, **Est. Precision**: 21.0%

Participant High Recall@B: 18.6% (ursinus6), **Median Recall@B**: 8.6%

Participant High Recall@25000: 36.0% (UMass15), **Median Recall@25000**: 13.5%

5 Deepest Sampled Relevant Documents: ahe53c00-3532.2 (UMass13-19613), ggr75d00-3162.8 (wat8gram-14031), pkr48d00-2733.5 (CMUL07irt-9829), cxu05d00-1413.5 (SabL07ab1-9283), ams51d00-1309.8 (ursinus6-7038)

Topic 94 (2007-D-6)

Request Text: Submit all documents relating to insurance price lists, pricing plans, pricing policies, pricing forecasts, pricing strategies, pricing analyses, and pricing decisions.

Initial Proposal by Defendant: `("price lists" OR "pricing plans" OR "pricing policies" OR "pricing forecasts" OR "pricing strategies" OR "pricing analyses" OR "pricing decisions") AND ("property insurance" OR "casualty`

insnrance")

Rejoinder by Plaintiff: ((price OR pricing) AND (list! OR plan! OR polic! OR Forecast! OR strateg! OR analys! OR decision!)) AND insurance

Final Negotiated Boolean Query: ((price OR pricing) w/15 (list! OR plan! OR polic! OR forecast! OR strateg! OR analys! OR decision!)) AND insurance

Sampling: 279484 pooled, 500 assessed, 104 judged relevant, 391 non-relevant, 5 gray, "C"=1.36, **Est. Rel.:** 40068.5

Final Boolean Result Size (B): 12080, **Est. Recall:** 6.7%, **Est. Precision:** 23.7%

Participant High Recall@B: 21.4% (CMUL07irs), **Median Recall@B:** 4.5%

Participant High Recall@25000: 45.5% (CMUL07irs), **Median Recall@25000:** 7.5%

5 Deepest Sampled Relevant Documents: nhw23f00-3901.8 (UMKC3-24159), fek48d00-3813.1 (CMUL07irs-21845), bsv84a00-3521.7 (wat6qap-16200), clg85a00-3281.1 (wat6qap-12980), ayy84a00-1831.4 (UIowa07LegE2-10294)

Topic 95 (2007-D-7)

Request Text: Submit all documents discussing or relating to the historical, current, or future financial impact of tobacco usage on the property and casualty insurance industry.

Initial Proposal by Defendant: ("historical" OR "current" OR "futnre") AND "financial impact" AND nsage AND ("property insnrance" OR "casnalty insurance")

Rejoinder by Plaintiff: (financial OR (increas! w/3 cost!) OR (smoking w/5 (illness OR sick! OR death!)) AND insnrance

Final Negotiated Boolean Query: ("financial impact" OR (increas! w/3 cost!) OR ("smoking-related" w/5 (illness OR sick! OR death!)) AND insnrance

Sampling: 315430 pooled, 499 assessed, 120 judged relevant, 379 non-relevant, 0 gray, "C"=1.53, **Est. Rel.:** 34111.6

Final Boolean Result Size (B): 16324, **Est. Recall:** 18.5%, **Est. Precision:** 33.5%

Participant High Recall@B: 27.1% (CMUL07ibt), **Median Recall@B:** 8.7%

Participant High Recall@25000: 36.6% (CMUL07ibt), **Median Recall@25000:** 12.9%

5 Deepest Sampled Relevant Documents: qns51a00-3690.8 (otL07fbe-21566), gvn44a00-3435.7 (CMUL07irs-16802), pem65a00-2424.3 (UMKC1-14407), yns76d00-2418.1 (CMUL07ibp-14266), yxb15a00-2363.2 (UMass10-13092)

Topic 96 (2007-D-8)

Request Text: Submit all documents that discuss entry conditions into the property and casualty insurance industry.

Initial Proposal by Defendant: "entry condition!" AND ("property insnrance" OR "casualty insurance"

Rejoinder by Plaintiff: (entry AND (barrier! OR condition!)) AND ((property OR casualty) w/10 insurance)

Final Negotiated Boolean Query: (entry w/10 (barrier! OR condition!)) AND ((property OR casualty) w/10 insnrance)

Sampling: 279511 pooled, 499 assessed, 140 judged relevant, 349 non-relevant, 10 gray, "C"=1.62, **Est. Rel.:** 43945.8

Final Boolean Result Size (B): 103, **Est. Recall:** 0.1%, **Est. Precision:** 38.3%

Participant High Recall@B: 0.2% (IowaSL0706), **Median Recall@B:** 0.1%

Participant High Recall@25000: 39.4% (UMKC1), **Median Recall@25000:** 14.7%

5 Deepest Sampled Relevant Documents: bts05f00-3598.7 (UMKC1-20802), ypn31e00-3594.6 (otL07fbe-20717), wgp97d00-3427.9 (UMKC5-17662), gyd20e00-3425.8 (UIowa07LegE5-17627), wpc45c00-3297.4 (UIowa07LegE3-15687)

Topic 97 (2007-D-9)

Request Text: Submit all documents that relate to any plans of, interest in, or efforts undertaken for any acquisition, divestiture, joint venture, alliance, or merger of any kind within or related to the property and casualty insurance industry.

Initial Proposal by Defendant: (plan OR interest OR effort) AND (acquisition OR divestitnre OR "joint venture" OR alliance OR merger) AND ("property insnrance" OR "casualty insurance")

Rejoinder by Plaintiff: (acquisition OR divestiture OR venture OR alliance OR merger) AND insurance

Final Negotiated Boolean Query: (acquisition OR divestiture OR "joint ventnre" OR alliance OR merger) AND ((property OR casnalty) AND insurance)

Sampling: 256752 pooled, 499 assessed, 90 judged relevant, 404 non-relevant, 5 gray, "C"=2.36, **Est. Rel.:** 9032.0

Final Boolean Result Size (B): 13296, **Est. Recall:** 29.4%, **Est. Precision:** 18.1%

Participant High Recall@B: 33.0% (CMUL07ibs), **Median Recall@B:** 10.1%

Participant High Recall@25000: 71.7% (otL07pb), **Median Recall@25000:** 11.6%

5 Deepest Sampled Relevant Documents: bib83c00-2696.3 (otL07pb-13811), cht55f00-1758.8 (CMUL07ibs-12259), rnr35f00-1451.4 (ursinus4-7542), czr93f00-1100.7 (otL07pb-4432), oht84f00-779.0 (UIowa07LegE3-2600)

Topic 98 (2007-D-10)

Request Text: Submit all documents that describe the policies and procedures relating to the retention and destruction of documents (hard copy or electronic) for any company in the property and casualty insurance industry.

Initial Proposal by Defendant: record w/2 (schedule OR retention OR destruction) AND ("property insurance" OR "casnalty insurance")

Rejoinder by Plaintiff: (schedule OR retention OR destr!) AND insurance

Final Negotiated Boolean Query: (record w/5 (schedule OR retention OR destr!)) AND ((property OR casualty) AND insurance)

Sampling: 256036 pooled, 499 assessed, 100 judged relevant, 385 non-relevant, 14 gray, "C"=1.33, **Est. Rel.:** 26640.9

Final Boolean Result Size (B): 682, **Est. Recall:** 0.5%, **Est. Precision:** 19.2%

Participant High Recall@B: 2.3% (wat4feed), **Median Recall@B:** 0.4%

Participant High Recall@25000: 39.1% (fdwim7sl), **Median Recall@25000:** 15.4%

5 Deepest Sampled Relevant Documents: lbh20f00-3725.4 (otL07pb-19437), yav99c00-3613.9 (catchup0701p-17338), pvl48d00-3389.9 (ursinus3-14001), aql40f00-2878.1 (UIowa07LegE5-9020), qfb94a00-2747.0 (UMass12-8108)

Topic 99 (2007-D-11)

Request Text: Submit all documents describing natural disasters leading to claims handled by the property and casualty insurance industry.

Initial Proposal by Defendant: `"natural disaster" AND ("property insurance" OR "casualty insurance")`

Rejoinder by Plaintiff: `("natural disaster!" OR devastation OR catastroph!) AND insurance`

Final Negotiated Boolean Query: `("natural disaster!" OR earthqnake! OR fire! OR flood!) AND ((property OR casualty) AND insnrance)`

Sampling: 286700 pooled, 498 assessed, 84 judged relevant, 414 non-relevant, 0 gray, "C"=2.36, **Est. Rel.:** 7484.0

Final Boolean Result Size (B): 19716, **Est. Recall:** 20.6%, **Est. Precision:** 8.2%

Participant High Recall@B: 52.1% (CMUL07o3), **Median Recall@B:** 31.1%

Participant High Recall@25000: 55.0% (CMUL07o1), **Median Recall@25000:** 33.8%

5 Deepest Sampled Relevant Documents: pzl46e00-3320.6 (UIowa07LegE3-23332), fex21c00-1283.7 (fdwim7sl-4492), frz84a00-695.5 (otL07fbe-1993), oqc77e00-613.0 (UMKC5-1713), yhc77e00-440.1 (UMKC5-1169)

Topic 100 (2007-D-12)

Request Text: Submit all documents representing or referencing a formal statement by a CEO of a tobacco company describing a company merger or acquisition policy or practice.

Initial Proposal by Defendant: `"formal statement" AND (CEO OR C.E.O. OR "Chief Executive Officer") AND (merger OR acquisition)`

Rejoinder by Plaintiff: `(CEO OR C.E.O. OR chief OR head) AND (merger OR acquisition)`

Final Negotiated Boolean Query: `(CEO OR C.E.O. OR "Chief Executive Officer") AND (merger OR acqnisition)`

Sampling: 310268 pooled, 497 assessed, 93 judged relevant, 404 non-relevant, 0 gray, "C"=1.31, **Est. Rel.:** 8710.0

Final Boolean Result Size (B): 11480, **Est. Recall:** 43.9%, **Est. Precision:** 31.6%

Participant High Recall@B: 45.0% (wat5nofeed), **Median Recall@B:** 19.7%

Participant High Recall@25000: 61.2% (wat1fuse), **Median Recall@25000:** 24.0%

5 Deepest Sampled Relevant Documents: qjh54d00-3378.7 (UIowa07LegE1-13650), bek21f00-719.2 (otL07pb-1372), hzl04e00-651.3 (ursinus4-1191), oqa08c00-528.8 (IowaSL0707-900), rrk60d00-518.3 (wat7bool-877)

Topic 101 (2007-D-13)

Request Text: Submit all documents specifically referencing a lawsuit filed against a named property and casualty insurance company (as either sole or joint defendant).

Initial Proposal by Defendant: `(lawsnit OR "complaint filed") AND ("property insurance" OR "casualty insurance")`

Rejoinder by Plaintiff: `(lawsnit OR complaint OR pleading) AND insurance`

Final Negotiated Boolean Query: `(lawsuit OR "complaint filed") AND ((property OR casualty)) AND insurance`

Sampling: 204551 pooled, 1000 assessed, 184 judged relevant, 785 non-relevant, 31 gray, "C"=6.68, **Est. Rel.:** 8950.9

Final Boolean Result Size (B): 6008, **Est. Recall:** 22.0%, **Est. Precision:** 27.4%

Participant High Recall@B: 30.8% (wat8gram), **Median Recall@B:** 10.0%

Participant High Recall@25000: 81.5% (wat3desc), **Median Recall@25000:** 25.3%

5 Deepest Sampled Relevant Documents: vhv39e00-1403.1 (wat4feed-13029), ihj31e00-492.2 (IowaSL0707-5569), rvl21f00-484.4 (wat3desc-5421), kon90d00-480.0 (wat8gram-5340), dsq44a00-478.4 (UMKC5-5310)

The 10 assessed Relevance Feedback topics are listed next. The summary information differs from that given earlier for the Ad Hoc topics as follows:

Topic : The 10 topics were selected from 2006, whose numbers ranged from 6 to 51. There were 5 complaints in 2006 (labelled A, B, C, D and E).

Initial Proposal by Defendant and Final Negotiated Boolean Query :

The "Rejoinder by Plaintiff" is not listed because it was usually the same as the Final Negotiated Boolean Query in 2006.

Sampling and Est. Rel. : The number of pooled documents includes not just the the residual output of the relevance feedback runs but also all of the documents submitted by the Interactive runs and the special oldrel07 and oldnon07 runs. The number presented to the assessor, the number the assessor judged relevant, the number the assessor judged non-relevant, and the number the assessor left as

"gray" are based on the full pool and hence includes rejudging of some documents judged in 2006 (particularly from the oldrel07 and oldnon07 runs). But the "Est. Rel." is just the estimated number of *residual* relevant documents in the pool for the topic (i.e., the number of relevant documents after those judged in 2006 are removed). Hence, unlike for the Ad Hoc topics, "Est. Rel." can be (and sometimes is) lower than "judged relevant".

Final Boolean Result Size B_r, Est. Recall and Est. Precision : "B_r" is the number of documents matching the final negotiated Boolean query after the documents judged in 2006 are removed; for 2 topics, B_r exceeded 25,000. "Est. Recall" and "Est. Precision" are for the residual Boolean result set.

Feedback High Recall@B_r and Median Recall@B_r : Only the 5 runs which used feedback (i.e., the runs which made use of the 2006 judgments) are considered for this listing.

Feedback High Recall@25000 and Median Recall@25000 : Again, only the 5 runs which used feedback are considered for this listing.

5 Deepest Sampled Relevant Documents : Only residual relevant documents are listed. The ranks following the run identifiers are residual ranks. For Interactive runs, all documents were assigned rank 1.

Topic 7 (2006-A-2)

Request Text: All documents discussing, referencing, or relating to company guidelines, strategies, or internal approval for placement of tobacco products in movies that are mentioned as G-rated.

Initial Proposal by Defendant: `(guidelines OR strategies OR "internal approval") AND placement AND "G-rated movie"`

Final Negotiated Boolean Query: `((guide! OR strateg! OR approv!) AND (place! or promot!)) AND (("G-rated" OR "G rated" OR family) W/5 (movie! OR film! OR picture!))`

Sampling: 119645 pooled, 500 assessed, 170 judged relevant, 318 non-relevant, 12 gray, "C"=2.58, **Est. Rel.**: 1280.2

Final Boolean Result Size (B_r): 425, **Est. Recall**: 0.9%, **Est. Precision**: 4.8%

Feedback High Recall@B_r: 16.7% (sab07legrf1), **Median Recall@B_r**: 10.3%

Feedback High Recall@25000: 95.0% (CMU07RFBSVME), **Median Recall@25000**: 88.4%

5 Deepest Sampled Relevant Documents: mak24d00-640.7 (CMU07RSVMNP-1896), lpd41a00-194.5 (CMU07RSVMNP-522), gvc91c00-55.7 (CMU07RFBSVME-416), czc42e00-50.0 (sab07legrf1-313), vwj35c00-44.2 (CMU07RFBSVME-238)

Topic 8 (2006-A-3)

Request Text: All documents discussing, referencing or relating to company guidelines, strategies, or internal approval for placement of tobacco products in live theater productions.

Initial Proposal by Defendant: `(guidelines OR strategies OR "internal approval") AND placement AND ("live theater" OR "live theatre")`

Final Negotiated Boolean Query: `((guide! OR strateg! OR approv!) AND (place! or promot!) AND (live W/5 (theatre OR theater OR audience"))`

Sampling: 119011 pooled, 244 assessed, 100 judged relevant, 143 non-relevant, 1 gray, "C"=2.12, **Est. Rel.**: 10983.9

Final Boolean Result Size (B_r): 623, **Est. Recall**: 1.7%, **Est. Precision**: 37.1%

Feedback High Recall@B_r: 4.2% (sab07legrf3), **Median Recall@B_r**: 2.5%

Feedback High Recall@25000: 50.0% (sab07legrf3), **Median Recall@25000**: 17.4%

5 Deepest Sampled Relevant Documents: cyy57d00-1942.2 (sab07legrf3-6733), jwf04e00-1695.8 (sab07legrf2-5440), eif71c00-1166.4 (CMU07RSVMNP-3225), ahw04c00-1129.4 (sab07legrf3-3093), cqs81e00-1015.9 (CMU07RBase-2703)

Topic 13 (2006-A-9)

Request Text: All documents to or from employees of a tobacco company or tobacco organization referring to the marketing, placement, or sale of chocolate candies in the form of cigarettes.

Initial Proposal by Defendant: `(marketing OR placement OR sale) AND "chocolate cigarettes" AND candy`

Final Negotiated Boolean Query: `(cand! OR chocolate) w/10 cigarette!`

Sampling: 123630 pooled, 250 assessed, 28 judged relevant, 212 non-relevant, 10 gray, "C"=3.01, **Est. Rel.**: 288.8

Final Boolean Result Size (B_r): 20240, **Est. Recall**: 99.3%, **Est. Precision**: 1.4%

Feedback High Recall@B_r: 100.0% (CMU07RFBSVME and 1 other), **Median Recall@B_r**: 76.7%

Feedback High Recall@25000: 100.0% (CMU07RFBSVME and 1 other), **Median Recall@25000**: 76.7%

5 Deepest Sampled Relevant Documents: egg72c00-153.4 (CMU07RFBSVME-480), exg09c00-67.3 (CMU07RFBSVME-206), xej32e00-35.2 (otRF07fb-107), akq55a00-12.3 (sab07legrf3-37), oyz23a00-4.3 (CMU07RFBSVME-13)

Topic 26 (2006-C-2)

Request Text: All documents discussing or referencing retail prices of tobacco products in the city of San Diego.

Initial Proposal by Defendant: `"retail prices" AND tobacco AND California`

Final Negotiated Boolean Query: `((retail OR net) w/2 pric!) AND ("San Diego" or ("S.D." w/3 Calif!))`

Sampling: 104952 pooled, 250 assessed, 95 judged relevant, 152 non-relevant, 3 gray, "C"=2.11, **Est. Rel.:** 15466.3

Final Boolean Result Size (B_r): 2301, **Est. Recall:** 3.1%, **Est. Precision:** 20.9%

Feedback High Recall@B_r: 11.2% (sab07legrf3), **Median Recall@B_r:** 3.9%

Feedback High Recall@25000: 71.0% (sab07legrf3), **Median Recall@25000:** 47.6%

5 Deepest Sampled Relevant Documents: dkm65e00-2785.0 (sab07legrf2-13265), mqp25d00-2650.1 (sab07legrf3-11898), ubo68c00-1820.0 (CMU07RFBSVME-6038), gcd65e00-1670.5 (CMU07RBase-5293), lpk24d00-1055.2 (sab07legrf2-2822)

Topic 27 (2006-C-3)

Request Text: All documents discussing or relating to the placement of product logos at events held in the State of California.

Initial Proposal by Defendant: `"product placement" AND "logos" AND California`

Final Negotiated Boolean Query: `("product placement" OR advertis! OR market! OR promot!) AND (logo! OR symbol OR mascot OR marqne OR mark) AND (California OR cal. OR calif. OR "CA")`

Sampling: 335971 pooled, 249 assessed, 120 judged relevant, 129 non-relevant, 0 gray, "C"=1.96, **Est. Rel.:** 23229.7

Final Boolean Result Size (B_r): 127525, **Est. Recall:** 36.7%, **Est. Precision:** 6.9%

Feedback High Recall@B_r: 64.0% (sab07legrf2), **Median Recall@B_r:** 49.1%

Feedback High Recall@25000: 59.1% (CMU07RSVMNP), **Median Recall@25000:** 20.0%

5 Deepest Sampled Relevant Documents: ofg48e00-3543.2 (CMU07RSVMNP-23835), uzn25d00-3455.1 (CMU07RBase-21917), ber19c00-3292.7 (sab07legrf2-18900), urt66d00-3156.6 (CMU07RSVMNP-16781), dah15d00-2441.8 (CMU07RSVMNP-9354)

Topic 30 (2006-C-6)

Request Text: All documents discussing or referencing the California Cartwright Act.

Initial Proposal by Defendant: `"California Cartwright Act"`

Final Negotiated Boolean Query: `California w/3 (antitrust OR monopol! OR anticompetitive OR restraint OR "unfair competition" OR "Cartwright")`

Sampling: 125617 pooled, 250 assessed, 24 judged relevant, 226 non-relevant, 0 gray, "C"=2.34, **Est. Rel.:** 7.0

Final Boolean Result Size (B_r): 202, **Est. Recall:** 28.6%, **Est. Precision:** 1.8%

Feedback High Recall@B_r: 57.1% (CMU07RFBSVME and 1 other), **Median Recall@B_r:** 42.9%

Feedback High Recall@25000: 100.0% (CMU07RFBSVME and 2 others), **Median Recall@25000:** 100.0%

5 Deepest Sampled Relevant Documents: idc90e00-1.0 (sab07legrf1-5), zce78c00-1.0 (CMU07RSVMNP-5), phh71c00-1.0 (CMU07RSVMNP-3), dzk44c00-1.0 (CMU07RBase-3), pbx64d00-1.0 (CMU07RBase-1)

Topic 34 (2006-D-1)

Request Text: All documents discussing or referencing payments to foreign government officials, including but not limited to expressly mentioning "bribery" and/or "payoffs."

Initial Proposal by Defendant: `(bribery OR payoffs) AND payments AND "foreign government officials"`

Final Negotiated Boolean Query: `(payment! OR transfer! OR wire! OR fund! OR kickback! OR payola OR grease OR bribery OR payoff!) AND (foreign w/5 (official! OR ministr! OR delegat! OR representative!))`

Sampling: 122598 pooled, 248 assessed, 105 judged relevant, 140 non-relevant, 3 gray, "C"=2.41, **Est. Rel.:** 20113.0

Final Boolean Result Size (B_r): 2380, **Est. Recall:** 1.8%, **Est. Precision:** 16.5%

Feedback High Recall@B_r: 6.5% (CMU07RSVMNP), **Median Recall@B_r:** 2.0%

Feedback High Recall@25000: 54.7% (CMU07RSVMNP), **Median Recall@25000:** 16.3%

5 Deepest Sampled Relevant Documents: rut15e00-2950.9 (CMU07RSVMNP-17353), yph30a00-2910.5 (CMU07RBase-16784), oyf37c00-2719.0 (sab07legrf1-14364), uva40f00-2686.7 (sab07legrf3-13995), nnj14e00-2587.3 (CMU07RSVMNP-12922)

Topic 37 (2006-D-4)

Request Text: All documents relating to defendants' tobacco advertising, marketing or promotion plans in China's capital.

Initial Proposal by Defendant: `(advertising OR marketing OR "promotion plans") AND (China OR Beijing)`

Final Negotiated Boolean Query: `(advertis! OR market! OR promot! OR encourag! OR incentiv!) AND (China OR Beijing OR Peking)`

Sampling: 149493 pooled, 250 assessed, 74 judged relevant, 175 non-relevant, 1 gray, "C"=2.85, **Est. Rel.:** 7086.3

Final Boolean Result Size (B_r): 38723, **Est. Recall:** 59.0%, **Est. Precision:** 13.6%

Feedback High Recall@B_r: 50.6% (sab07legrf1), **Median Recall@B_r:** 49.2%

Feedback High Recall@25000: 50.6% (sab07legrf1), **Median Recall@25000:** 49.2%

5 Deepest Sampled Relevant Documents: rgd60a00-2384.7 (CMU07RSVMNP-12994), aer19e00-1178.8 (sab07legrf2-4396), jmy90d00-1100.0 (otRF07fb-4019), awk95c00-1018.2 (CMU07RBase-3644), ziv19e00-519.1 (sab07legrf1-1651)

Topic 45 (2006-E-4)

Request Text: All documents that refer or relate to pigeon deaths during the course of animal studies.

Initial Proposal by Defendant: `"animal studies" AND "pigeon deaths"`

Final Negotiated Boolean Query: `(research OR stud! OR "in vivo") AND pigeon AND (death! OR dead OR die! OR dying)`

Sampling: 112239 pooled, 498 assessed, 91 judged relevant, 400 non-relevant, 7 gray, "C"=4.97, **Est. Rel.**: 83.2

Final Boolean Result Size (B_r): 2507, **Est. Recall**: 70.0%, **Est. Precision**: 2.4%

Feedback High Recall@B_r: 97.6% (sab07legrf2), **Median Recall@B_r**: 94.5%

Feedback High Recall@25000: 100.0% (CMU07RSVMNP and 2 others), **Median Recall@25000**: 100.0%

5 Deepest Sampled Relevant Documents: qwl16d00-16.0 (CMU07RSVMNP-82), jbr10a00-15.4 (otRF07fb-79), ivw00a00-8.3 (CMU07RSVMNP-42), wzn20a00-7.9 (otRF07fb-40), phg81d00-3.6 (otRF07fb-18)

Topic 51 (2006-E-10)

Request Text: All documents referencing or regarding lawsuits involving claims related to memory loss.

Initial Proposal by Defendant: `(lawsuits OR "tort claims") AND "memory loss"`

Final Negotiated Boolean Query: `((memory w/2 loss) OR amnesia OR Alzheimer! OR dementia) AND (lawsuit! OR litig! OR case OR (tort w/2 claim!) OR complaint OR allegation!)`

Sampling: 108434 pooled, 499 assessed, 83 judged relevant, 400 non-relevant, 16 gray, "C"=4.01, **Est. Rel.**: 65.7

Final Boolean Result Size (B_r): 6927, **Est. Recall**: 84.8%, **Est. Precision**: 0.8%

Feedback High Recall@B_r: 84.8% (CMU07RFBSVME), **Median Recall@B_r**: 23.9%

Feedback High Recall@25000: 84.8% (CMU07RFBSVME), **Median Recall@25000**: 46.7%

5 Deepest Sampled Relevant Documents: qcm10d00-7.7 (CMU07RBase-31), war78e00-1.0 (randomRF07-4), ptn85d00-1.0 (uw07T2-1), bee61e00-1.0 (uw07T2-1), sfw87e00-1.0 (liu07-1)

---

5 Workshop Discussion

There were several opportunities for interaction among the participants from the research and legal communities during the conference, culminating in the Thursday, November 8 workshop which discussed future plans for the track (which will continue for a 3rd year in 2008).

At the workshop, two smaller document sets were considered for 2008. One option was a collection of State Department cables from the 1970's, which would be a cleaner collection to work with (e.g., no OCR issues), but there were concerns about whether the legal community would accept results based on it. The other option was an Enron collection, which would feature email resembling modern e-discovery scenarios, but it was considered a difficult collection to work with (e.g., attachments in proprietary formats). It was decided to continue in 2008 with the same IIT CDIP collection as the past couple years, particularly since there were a lot of new participants in 2007 who would like the chance to fully focus on research issues in 2008 rather than deal with the details of using a new collection.

There were concerns raised at the workshop about the appropriateness of the Recall@B measure which was used as the primary measure in 2007; e.g., the real goal of discovery is to produce the set of relevant documents, not just to maximize success at a particular given size B. (We followup on the choice of measure in the next section.)

More focus on relevance feedback was suggested at the workshop, both to encourage more use of metadata (e.g., author, Bates number) and to enrich the relevance judgments for past topics to further improve their re-usability. Deeper and denser assessing was also suggested, even if it meant fewer new topics.

A proposal also discussed among track coordinators before and during the workshop concerned whether in future years the Legal Track should introduce and evaluate the concept of "highly relevant" documents, as a third category for purposes of assessment along with not relevant and relevant. The problem of isolating a true set of "hot" or "material" documents for use in later phases of discovery (e.g., depositions) and at trial, amongst a large universe of merely potentially tangentially relevant documents, remains a key concern for the legal profession. This issue will be explored further in Year 3 of the track.

We look forward to continuing the discussion in 2008!

5.1 Post-Workshop Analysis

After the conference, we analyzed the Ad Hoc runs from the perspective of trying to produce a set as close as possible to the desired set of R relevant documents. In particular, we looked at the F1 measure which

combines recall and precision into one measure (F1 = 2*Prec*Recall / (Prec+Recall)).

The reference Boolean run averaged an F1 of 0.14 over the 43 topics, whereas if we cutoff the Ad Hoc runs at their top-ranked R retrieved (where R is the estimated number of relevant documents, rounding up fractional estimated R values to the next integer) several Ad Hoc runs scored higher (to a high of 0.22 (otL07frw); the median was also 0.14). Hence, if the Ad Hoc runs can pick a good cutoff value, they apparently can produce a closer set to the optimal set of R documents than the reference Boolean run's set of B documents, taking both recall and precision into account.

This result suggests that we should enhance the Ad Hoc task in 2008 to require each system to additionally specify a cutoff value K for each topic. (Unfortunately, in 2007, we did not ask the Ad Hoc systems to specify a cutoff value before R was known.) A measure which balances recall and precision (such as F1) could then be used to evaluate whether automated approaches can produce a set of K documents closer to the optimal set of R relevant documents than the reference Boolean query result set (for which K=B). We would still ask the systems to submit their top-25,000 ranked documents (or whatever the agreed limit is in 2008) to enrich the pools and enable post hoc analysis of different choices of K.

6 Conclusion

In its second year, the TREC Legal Track made several advances. The Ad Hoc task developed a much deeper sampling approach to more accurately estimate recall and precision and evaluated a wider variety of automated search techniques thanks to a doubling in participation. A separate Interactive task was created for studying the effectiveness of "expert" searchers. A new Relevance Feedback task was created to study automated ways of making use of judgments from an initial sample. Baseline results for each task were established and several resources are now available to support further study going forward.

For the Ad Hoc task, 50 new topic statements, i.e., requests for documents with associated negotiated Boolean queries, were created. 12 research teams used a wide variety of (mostly automated) techniques to search the IIT CDIP collection (a complex collection of almost 7 million scanned documents) and submit a total of 68 result sets of 25,000 top-ranked documents for each topic. These submissions were pooled, producing a set of approximately 300,000 documents per topic. A new sampling scheme was used to select between 500 and 1000 documents from the pool for each topic. Volunteers from the legal community assessed 43 of the 50 result samples in time for reporting the results at the conference. Based on the samples, we can estimate that there were on average almost 17,000 relevant documents per topic, and that this number varied considerably by topic (from a low of 18 to a high of more than 77,000).

The deep sampling allows us to estimate the recall and precision of the final negotiated Boolean query more accurately than before. On average (over the 43 topics), the reference Boolean query found just 22% of the relevant documents that are estimated to exist. Its precision averaged 29%. Again, these numbers varied considerably by topic (the recall ranged from 0% to 100%, while the precision ranged from 0% to 97%). It is quite striking that, on average per topic, 78% of the relevant documents were only found by participant research techniques and not by the reference Boolean query.

Surprisingly, when recall was estimated at depth B, where B is the number of documents matched by the reference Boolean query, no system participating in the Ad Hoc task submitted results that improved over the reference Boolean run (on average), despite the systems' collective success at finding relevant documents missed by the Boolean query. However, post hoc analysis using the F1 measure (which balances recall and precision) found that the Ad Hoc systems potentially can produce a set of results closer to the optimal set of R relevant documents than the reference Boolean result set. Unfortunately, this latter possibility was not properly evaluated in 2007 because the systems were not asked to specify a cutoff value before R was known (where R is the estimated number of relevant documents). We should consider refining the methodology in 2008 to require each system to specify a cutoff value K for each topic for targeting a measure (such as F1) which balances the demand for recall with the cost of reviewing unresponsive documents.

A new Interactive task was created in 2007 to followup on an interesting result from 2006, which was that the sole expert searcher achieved a higher mean R-precision than any of the automated runs in 2006 (albeit based on the shallower sampling used in 2006, which underestimated R considerably). In 2007, 8

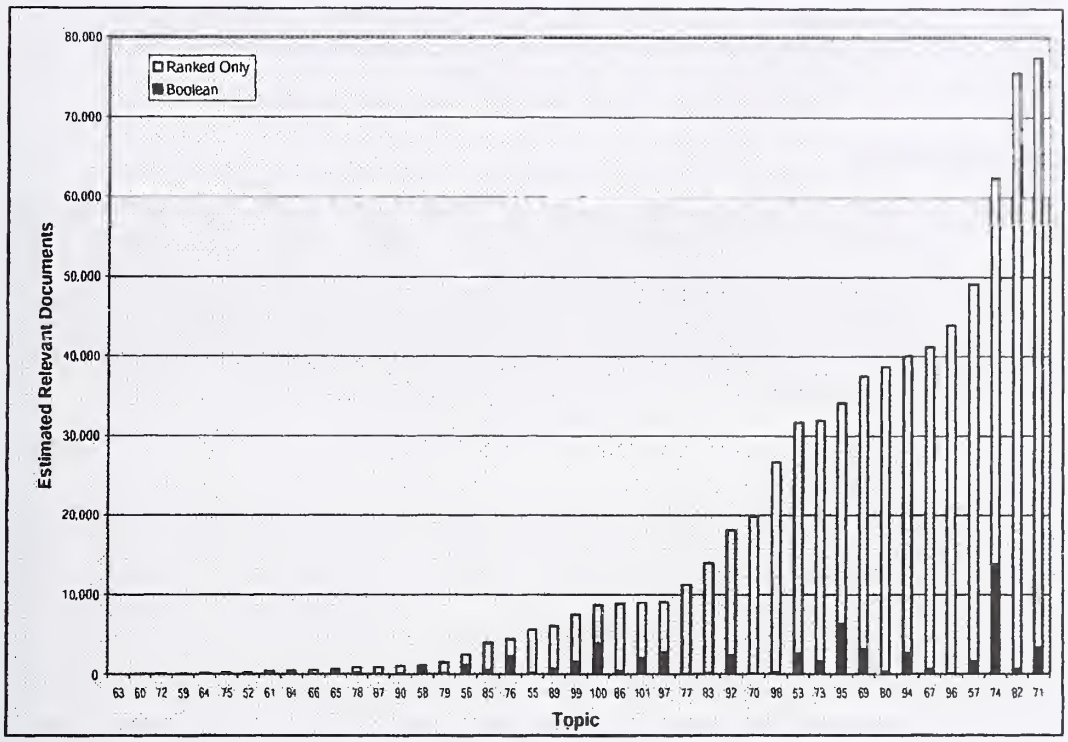


Figure 1: Estimated relevant documents found by the reference Boolean query (black) and found only by one or more ranked systems (white) for the 43 Ad Hoc topics.

teams from 3 sites took up the Interactive challenge. Some teams invested several hours per search topic, and most teams completed just 3 topics. It was found that there was substantial variation in the results of the participating teams, but most of them outscored the reference Boolean query in the task's utility measure for each of the 3 topics. This result is another encouraging one for expert searching, albeit one with many caveats. For instance, the participating teams in 2007 were limited to submitting 100 documents per topic (as was the sole expert searcher in 2006). We intend to remove this limit in 2008 so that the experts' ability to recall much larger numbers of relevant documents can be evaluated.

In the new Relevance Feedback task of 2007, 10 of the previous year's Ad Hoc topics were re-used. Participants were encouraged to use the previous year's document assessments as feedback to improve their results. 3 teams submitted a total of 8 runs, including 5 feedback runs. Residual evaluation was used, i.e., documents judged in the previous year were removed from the result sets before evaluating. The deep sampling approach of the Ad Hoc task was likewise applied to the Relevance Feedback task. An encouraging result from this pilot study was that the median of the 5 feedback runs found more relevant documents than the reference Boolean run by depth B_r for 7 of the 10 topics (where B_r is the number of documents matched by the reference Boolean query after removing documents judged the previous year), in contrast with the Ad Hoc task in which the median run outscored the reference Boolean run at depth B for just 8 of the 43 topics. We hope to run a larger study of Relevance Feedback (more test topics and more participants) in 2008.

The evaluation of e-discovery approaches remains a daunting challenge. The findings so far are very preliminary. Automated approaches can improve upon the recall of negotiated Boolean results, but typically at the expense of reviewing additional documents. Experts can improve upon automated approaches, but

they also vary a lot in performance. Feedback approaches are promising, but a larger study is needed. We are heartened that so many volunteers have contributed to the track's endeavours in 2006 and 2007 and look forward to working with everyone to make further advances in 2008.

Acknowledgments

This track would not have been possible without the efforts of a great many people. Our heartfelt thanks go to Ian Soboroff for creating the relevance assessment system; to the dedicated group of pro bono relevance assessors and the pro bono coordinators at the participating law schools; to Conor Crowley (DOAR Consulting), Joe Looby (FTI Consulting), Stephanie Mendelsohn (Genentech), and the team from H5 (Todd Elmer, Bruce Hedin, Jim Donahue and Michelle Luke) for their invaluable assistance with complaint drafting, topic formulation, and participating in Boolean negotiations; and to Richard Braman, Executive Director of The Sedona Conference®, for his continued support of the TREC Legal Track.

References

[1] TREC Legal Track Home Page. http://trec-legal.umiacs.umd.edu/.

[2] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Million query track 2007 overview. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, November 2007. http://trec.nist.gov.

[3] Jason R. Baron, David D. Lewis, and Douglas W. Oard. TREC-2006 legal track overview. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[4] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 619–620, 2006.

[5] Stefan Büttcher, Charles L. A. Clarke, and Ian Soboroff. The TREC 2006 terabyte track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[6] The Sedona Conference. The Sedona conference best practices commentary on the use of search and information retrieval methods in e-discovery. *The Sedona Conference Journal*, pages 189–223, 2007.

[7] Disability Rights Council of Greater Wash. v. Wash. Metro. Area Transit Auth. (D.D.C.). 2007 WL 1585452.

[8] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–666, 2006.

[9] George L. Paul and Jason R. Baron. Information inflation: Can the legal system adapt? *Richmond Journal of Law and Technology*, 13(3), 2007.

[10] H. Schmidt, K. Butter, and C. Rider. Building digital tobacco document libraries at the university of california, san francisco library/center for knowledge management. *D-Lib Magazine*, 8(2), 2002.

[11] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 705–706, 2006.

[12] Stephen Tomlinson. Experiments with the negotiated boolean queries of the TREC 2006 legal discovery track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, November 2006. http://trec.nist.gov.

[13] Williams v. Taser Intern, Inc. (N.D. Ga.). 2007 WL 1630875.

[14] Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th International Conference on Information and Knowledge Management (CIKM)*, pages 102–111, 2006.

[15] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, 1998.

Million Query Track 2007 Overview

James Allan*, Ben Carterette*, Javed A. Aslam+,
Virgil Pavlu+, Blagovest Dachev*, and Evangelos Kanoulas+

* Center for Intelligent Information Retrieval, Department of Computer Science
University of Massachusetts Amherst, Amherst, Massachusetts

+ College of Computer and Information Science, Northeastern University
Boston, Massachusetts

The Million Query (1MQ) track ran for the first time in TREC 2007. It was designed to serve two purposes. First, it was an exploration of ad-hoc retrieval on a large collection of documents. Second, it investigated questions of system evaluation, particularly whether it is better to evaluate using many shallow judgments or fewer thorough judgments.

Participants in this track were assigned two tasks: (1) run 10,000 queries against a 426Gb collection of documents at least once and (2) judge documents for relevance with respect to some number of queries.

Section 1 describes how the corpus and queries were selected, details the submission formats, and provides a brief description of all submitted runs. Section 2 provides an overview of the judging process, including a sketch of how it alternated between two methods for selecting the small set of documents to be judged. Sections 3 and 4 provide details of those two selection methods, developed at UMass and NEU, respectively. The sections also provide some analysis of the results.

In Section 6 we present some statistics about the judging process, such as the total number of queries judged, how many by each approach, and so on. We present some additional results and analysis of the overall track in Sections 7 and 8.

1 Phase I: Running Queries

The first phase of the track required that participating sites submit their retrieval runs.

1.1 Corpus

The 1MQ track used the so-called "terabyte" or "GOV2" collection of documents. This corpus is a collection of Web data crawled from Web sites in the .gov domain in early 2004. The collection is believed to include a large proportion of the .gov pages that were crawlable at that time, including HTML and text, plus the extracted text of PDF, Word, and PostScript files. Any document longer than 256Kb was truncated to that size at the time the collection was built. Binary files are not included as part of the collection, though were captured separately for use in judging.

The GOV2 collection includes 25 million documents in 426 gigabytes. The collection was made available by the University of Glasgow, distributed on a hard disk that was shipped to participants for an amount intended to cover the cost of preparing and shipping the data.

1.2 Queries

Topics for this task were drawn from a large collection of queries that were collected by a large Internet search engine. Each of the chosen queries is likely to have at least one relevant document in the GOV2 collection

because logs showed a clickthrough on one page captured by GOV2. Obviously there is no guarantee that the clicked page is relevant, but it increases the chance of the query being appropriate for the collection.

These topics are short, title-length (in TREC parlance) queries. In the judging phase, they were developed into full-blown TREC topics.

Ten thousand (10,000) queries were selected for the official run. The 10,000 queries included 150 queries that were judged in the context of the 2005 Terabyte Track [12] (though one of these had no relevant documents and was therefore excluded).

No quality control was imposed on the 10,000 selected queries. The hope was that most of them would be good quality queries, but it was recognized that some were likely to be partially or entirely non-English, to contain spelling errors, or even to be incomprehensible to anyone other than the person who originally created them.

The queries were distributed in a text file where each line has the format "N:query word or words". Here, N is the query number, is followed by a colon, and immediately followed by the query itself. For example, the line (from a training query) "32:barack obama internships" means that query number 32 is the 3-word query "barack obama internships". All queries were provided in lowercase and with no punctuation (it is not clear whether that formatting is a result of processing or because people use lowercase and do not use punctuation).

1.3 Submissions

Sites were permitted to provide up to five runs. Every submitted run was included in the judging pool and all were treated equally.

A run consisted of up to the top 1,000 documents for each of the 10,000 queries. The submission format was a standard TREC format of exactly six columns per line with at least one space between the columns. For example:

```
100 Q0 ZF08-175-870 1 9876 mysys1
100 Q0 ZF08-306-044 2 9875 mysys2
```

where:

1. The first column is the topic number.
2. The second column is unused but must always be the string "Q0" (letter Q, number zero).
3. The third column is the official document number of the retrieved document, found in the <DOCNO> field of the document.
4. The fourth column is the rank of that document for that query.
5. The fifth column is the score this system generated to rank this document.
6. The six column was a "run tag," a unique identifier for each group and run.

If a site would normally have returned no documents for a query, it instead returned the single document "GX000-00-0000000" at rank one. Doing so maintained consistent evaluation results (averages over the same number of queries) and did not break any evaluation tools being used.

1.4 Submitted runs

The following is a brief summary of some of the submitted runs. The summaries were provided by the sites themselves and are listed in alphabetical order. (When no full summary is available, the brief summary information from the submissions has been used.)

ARSC/University of Alaska Fairbanks The ARSC multisearch system is a heterogeneous distributed information retrieval simulation and demonstration implementation. The purpose of the simulation is to illustrate performance issues in Grid Information Retrieval applications by partitioning the GOV2 collection into a large number of hosts and searching each host independently of the others. Previous

TREC Terabyte Track experiments using the ARSC multisearch system have focused on the IR performance of multisearch result-set merging and the efficiency gains from truncating result-sets from a large collection of hosts before merging.

The primary task of the ARSC multisearch system in the 2007 TREC Million Query experiment is to estimate the number of hosts or subcollections of GOV2 that can be used to process 10,000 queries within the TREC Million Query Track time constraints. The secondary and ongoing task is to construct an effective strategy for picking a subsets of the GOV2 collections to search at query-time. The host-selection strategy used for this experiment was to restrict searches to hosts that returned the most relevant documents in previous TREC Terabyte Tracks.

Exegy Exegy's submission for the TREC 2007 million query track consisted of results obtained by running the queries against the raw data, i.e., the data was not indexed. The hardware-accelerated streaming engine used to perform the search is the Exegy Text Miner (XTM), developed at Exegy, inc. The search engine's architecture is novel: XTM is a hybrid system (heterogeneous compute platform) employing general purpose processors (GPPs) and field programmable gate arrays (FPGAs) in a hardware-software co-design architecture to perform the search. The GPPs are responsible for inputting the data to the FPGAs and reading and post-processing the search results that the FPGAs output. The FPGAs perform the actual search and due to the high degree of parallelism available (including pipelining) are able to do so much more efficiently than the GPP.

For the million query track the results for a particular query were obtained by searching for the exact query string within the corpus. This brute force approach, although naïve, returned relevant results for most of the queries. The mean-average precision for the results was 0.3106 and 0.0529 using the UMass and the NEU approaches, respectively. More importantly, XTM completed the search for the entire set of the 10,000 queries on the unindexed data in less than two and a half hours.

Heilongjiang Institute of Technology, China Used Lemur.

IBM Haifa This year, the experiments of IBM Haifa were focused on the scoring function of Lucene, an Apache open-source search engine. The main goal was to bring Lucene's ranking function to the same level as the state-of-the-art ranking formulas like those traditionally used by TREC participants. Lucene's scoring function was modified to include better document length normalization, and a better term-weight setting following to the SMART model.

Lucene then compared to Juru, the home-brewed search engine used by the group in previous TREC conferences. In order to examine the ranking function alone, both Lucene and Juru used the same HTML parser, the same anchor text, and the same query parsing process including stop-word removal, synonym expansion, and phrase expansion. Based on the 149 topics of the Terabyte tracks, the results of modified Lucene significantly outperform the original Lucene and are comparable to Juru's results.

In addition, a shallow query log analysis was conducted over the 10K query log. Based on the query log, a specific stop-list and a synonym-table were constructed to be used by both search engines.

Northeastern University We used several standard Lemur built in systems (tfidf_bm25, tfidf_log, kl_abs,kl_dir,inquery,cos, okapi) and combined their output (metasearch) using the hedge algorithm.

RMIT Zettair Dirichlet smoothed language model run.

SabIR Standard smart ltu.Lnu run.

University of Amsterdam The University of Amsterdam, in collaboration with the University of Twente, participated with the main aim to compare results of the earlier Terabyte tracks to the Million Query track. Specifically, what is the impact of shallow pooling methods on the (apparent) effectiveness of retrieval techniques? And what is the impact of substantially larger numbers of topics? We submitted a number of runs using different document representations (such as full-text, title-fields, or incoming anchor-texts) to increase pool diversity. The initial results show broad agreement in system rankings over various measures on topic sets judged at both Terabyte and Million Query tracks, with runs using the full-text index giving superior results on all measures. There are some noteworthy upsets: measures using the Million Query judged topics show stronger correlation with precision at early ranks.

University of Massachusetts Amherst The base UMass Amherst submissions were a simple query likelihood model and the dependence model approach fielded during the terabyte track last year. We also tried some simple automatic spelling correction on top of each baseline to deal with errors of that kind. All runs were done using the Indri retrieval system.

University of Melbourne Four types of runs were submitted:

1. A topic-only run using a similarity metric based on a language model with Dirichlet smoothing as describe by Zhai and Lafferty (2004).
2. Submit query to public web search engine, retrieve snippet information for top 5 documents, add unique terms from snippets to query, run expanded query using same similarity metric just described.
3. A standard impact-based ranking.
4. A merging of the language modeling and the impact runs.

2 Phase I: Relevance judgments and judging

After all runs were submitted, a subset of the topics were judged. The goal was to provide a small number of judgments for a large number of topics. For TREC 2007, over 1700 queries were judged, a large increase over the more typical 50 queries judged by other tracks in the past.

2.1 Judging overview

Judging was done by assessors at NIST and by participants in the track. Non-participants were welcome (encouraged!) to provide judgments, too, though very few such judgments occurred. Some of the judgments came from an Information Retrieval class project, and some were provided by hired assessors at UMass. The bulk of judgments, however, came from the NIST assessors.

The process looked roughly like this from the perspective of someone judging:

1. The assessment system presented 10 queries randomly selected from the evaluation set of 10,000 queries.
2. The assessor selected one of those ten queries to judge. The others were returned to the pool.
3. The assessor provided the description and narrative parts of the query, creating a full TREC topic. This information was used by the assessor to keep focus on what is relevant.
4. The system presented a GOV2 document (Web page) and asked whether it was relevant to the query. Judgments were on a three-way scale to mimic the Terabyte Track from years past: highly relevant, relevant, or not relevant. Consistent with past practice, the distinction between the first two was up to the assessor.
5. The assessor was required to continue judging until 40 documents has been judged. An assessor could optionally continue beyond the 40, but few did.

The system for carrying out those judgments was built at UMass on top of the Drupal content management platform[1]. The same system was used as the starting point for relevance judgments in the Enterprise track.

2.2 Selection of documents for judging

Two approaches to selecting documents were used:

Minimal Test Collection (MTC) method. In this method, documents are selected by how much they inform us about the difference in mean average precision given all the judgments that were made up to that point [10]. Because average precision is quadratic in relevance judgments, the amount each

[1] http://drupal.org

relevant document contributes is a function of the total number of judgments made and the ranks they appear at. Nonrelevant documents also contribute to our knowledge: if a document is nonrelevant, it tells us that certain terms cannot contribute anything to average precision. We quantify how much a document will contribute if it turns out to be relevant or nonrelevant, then select the one that we expect to contribute the most. This method is further described below in Section 3.

Statistical evaluation (statMAP) method. This method draws and judges a specific random sample of documents from the given ranked lists and produces unbiased, low-variance estimates of average precision, R-precision, and precision at standard cutoffs from these judged documents [1]. Additional (non-random) judged documents may also be included in the estimation process, further improving the quality of the estimates. This method is further described below in Section 4.

For each query, one of the following happened:

1. The pages to be judged for the query were selected by the "expected AP method." A minimum of 40 documents were judged, though the assessor was allowed to continue beyond 40 if so motivated.

2. The pages to be judged for the query were selected by the "statistical evaluation method." A minimum of 40 documents were judged, though the assessor was allowed continue beyond 40 if so motivated.

3. The pages to be judged were selected by alternating between the two methods until each has selected 20 pages. If a page was selected by more than one method, it was presented for judgment only once. The process continues until at least 40 pages have been judged (typically 20 per method), though the assessor was allowed continue beyond 40 if so motivated. (See Section 5.)

The assignments were made such that option (3) was selected half the time and the other two options each occurred 1/4 of the time. When completed, roughly half of the queries therefore had parallel judgments of 20 or more pages by each method, and the other half had 40 or more judgments by a single method.

In addition, a small pool of 50 queries were randomly selected for multiple judging. With a small random chance, the assessor's ten queries were drawn from that pool rather than the full pool. Whereas in the full pool no query was considered by more than one person, in the multiple judging pool, a query could be considered by any or even all assessors—though no assessor was shown the same query more than once.

3 UMass Method

The UMass algorithm is a greedy anytime algorithm. It iteratively orders documents according to how much information they provide about a difference in average precision, presents the top document to be judged, and, based on the judgment, re-weights and re-orders the documents.

Algorithm 3.1 shows the high-level pseudo-code for the algorithm, which we call MTC for *minimal test collection*.

Algorithm 3.1 MTC($\mathcal{S}, \mathcal{Q}$)

Require: a set of ranked lists $\mathcal{S}$, a set of qrels $\mathcal{Q}$ (possibly empty)

```
q = GET-QRELS(Q)
w = INIT-WEIGHTS(S, q)
loop
    i* = arg max_i w
    request judgment for document i*
    receive judgment j_i* for document i*
    w = UPDATE-WEIGHTS(i*, S)
    q_i* = j_i*
```

Here $\mathbf{q}$ is a vector of relevance judgments read in from a *qrels* file if one exists (for example if an assessor is resuming judging a topic that he had previously stopped). GET-QRELS simply translates (document, judgment) pairs into vector indexes such that $q_i = 1$ if the document has been judged relevant and 0

otherwise; if an assessor is just starting a topic, q_i will be 0 for all i. $\mathbf{w}$ is a vector of document weights (see below). We assume that there's a global ordering of documents, so that the relevance of document i can be found at index i in $\mathbf{q}$, and its weight at the same index in $\mathbf{w}$.

The INIT-WEIGHTS, SET-WEIGHTS, and UPDATE-WEIGHTS functions are where the real work happens. The pseudo-code below is rather complicated, so first some notational conventions: We shall use $i, j = 1 \cdots n$ to enumerate n documents and $s = 1 \cdots m$ to enumerate m systems. Capital bold letters are matrices. Column and row vectors for a matrix $\mathbf{M}$ are denoted $\mathbf{M}_{\cdot i}$ (for the ith column vector) or $\mathbf{M}_{i\cdot}$ (for the ith row vector). Matrix cells are referred to with nonbold subscripted letters, e.g. M_{ij}. Lowercase bold letters are vectors, and lowercase nonbold letters are scalars. Superscripts are never exponents, always some type of index.

Algorithm 3.2 INIT-WEIGHTS($\mathcal{S}, \mathbf{q}$)

Require: ranked lists $\mathcal{S}$, a vector of judgments $\mathbf{q}$

1: $\mathbf{V}^R = \mathbf{V}^N = [0]_{n \times m}$
2: **for all** $s \in \mathcal{S}$ **do**
3: $\quad \mathbf{C} = [0]_{n \times n}$
4: $\quad$ **for all** pairs of documents (i, j) **do**
5: $\quad\quad C_{ij} = 1/\max\{r_s(i), r_s(j)\}$
6: $\quad \mathbf{V}^R_{\cdot s} = \mathbf{Cq} + diag(\mathbf{C})$
7: $\quad \mathbf{V}^N_{\cdot s} = \mathbf{C}(1 - \mathbf{q})$
8: **return** SET-WEIGHTS()

Algorithm 3.3 SET-WEIGHTS()

Require: access to global weight matrices $\mathbf{V}^R, \mathbf{V}^N$

1: $\mathbf{w} = [0]_n$
2: **for all** unjudged documents i **do**
3: $\quad w_i^R = \max \mathbf{V}^R_{i\cdot} - \min \mathbf{V}^R_{i\cdot}$
4: $\quad w_i^N = \max \mathbf{V}^N_{i\cdot} - \min \mathbf{V}^N_{i\cdot}$
5: $\quad w_i = \max\{w_i^R, w_i^N\}$
6: **return** $\mathbf{w}$

Algorithm 3.2 initializes the weight vector. At line 1 we create two "global" weight matrices in which each element V_{is} is the effect a judgment will have on the average precision of system s (see below for more detail). We iterate over systems (line 2), for each run creating a coefficient matrix $\mathbf{C}$ (lines 3–5). Each pair of documents has an associated coefficient $1/\max\{r_s(i), r_s(j)\}$, where $r_s(i)$ is the rank of document i in system s (infinity if document i is unranked). In lines 6 and 7, we multiply the coefficient matrix by the qrels vector and assign the resulting vector to the corresponding system column of the weight matrix. At the end of this loop, the matrices $\mathbf{V}^R, \mathbf{V}^N$ contain the individual system weights for every document. Each column s contains the weights for system s and each row i the weights for document i.

The global weight of a document is the maximum difference between pairs of system weights. Global weights are set with the SET-WEIGHTS function, shown in Algorithm 3.3. For each row in the weight matrices, it finds the maximum and minimum weights in any system. The difference between these is the maximum pairwise difference. Then the maximum of w_i^R and w_i^N is the final weight of the document.

After each judgment, UPDATE-WEIGHTS (Algorithm 3.4) is called to update the global weight matrices and recomputes the document weights. $\mathbf{C}'$ is constructed by pulling the i^*th column from each of the m coefficient matrices $\mathbf{C}$ defined in set-weights. We construct it from scratch rather than keep all m $\mathbf{C}$ matrices in memory. Global weight matrices are updated simply by adding or subtracting $\mathbf{C}'$ depending on the judgment to i^*.

3.1 Running Time

MTC loops until the assessor quits or all documents have been judged. Within the loop, finding the maximum-weight document (line 4) is in $\mathcal{O}(n)$. UPDATE-WEIGHTS loops over systems and documents for a

Algorithm 3.4 UPDATE-WEIGHTS($i^*, \mathcal{S}$)

Require: the index of the most recent judgment i^*, a set of ranked lists $\mathcal{S}$
Require: access to global weight matrices $\mathbf{V}^R, \mathbf{V}^N$
1: $\mathbf{C}' = [0]_{n \times m}$
2: **for** $s \in \mathcal{S}$ **do**
3: **for all** documents i, $C'_{is} = 1/\max\{r_s(i^*), r_s(i)\}$
4: **if** i^* is relevant **then**
5: $\mathbf{V}^R = \mathbf{V}^R + \mathbf{C}'$
6: **else**
7: $\mathbf{V}^N = \mathbf{V}^N - \mathbf{C}'$
8: **return** SET-WEIGHTS()

runtime in $\mathcal{O}(m \cdot n)$. SET-WEIGHTS is also in $\mathcal{O}(n \cdot m)$: each max or min is over m elements, and four of them happen n times. Therefore the total runtime for each iteration is in $\mathcal{O}(m \cdot n)$.

INIT-WEIGHTS is in $\mathcal{O}(m \cdot n^2)$: we loop over m systems, each time performing $\mathcal{O}(n^2)$ operations to construct $\mathbf{C}$ and perform matrix-vector multiplication. Since MTC can iterate up to n times, the total runtime is in $\mathcal{O}(m \cdot n^2)$.

In practice, the algorithm was fast enough that assessors experienced no noticeable delay between submitting a judgment and receiving the next document, even though an entire $\mathcal{O}(m \cdot n)$ iteration takes place in between. INIT-WEIGHTS was slow enough to be noticed, but it ran only once, in the background while assessors defined a topic description and narrative.

3.2 Explanation

The pseudo-code is rather opaque, and it may not be immediately clear how it implements the algorithm described in our previous work. Here is the explanation.

In previous work we showed $AP_s \propto \sum_i \sum_j A_{ij} X_i X_j$, where X_i is a binary indicator of the relevance of document i and $A_{ij} = 1/\max\{r_s(i), r_s(j)\}$. See Section 3.3.1 for more details.

Define a lower bound for AP_s in which every unjudged document is assumed to be nonrelevant. An upper bound is similarly defined by assuming every unjudged document relevant. Denote the bounds $\lfloor AP_s \rfloor$ and $\lceil AP_s \rceil$ respectively.

Consider document i, ranked at $r_s(i)$ by system s. If we judge it relevant, $\lfloor AP_s \rfloor$ will increase by $\sum_{j|x_j=1} a_{ij} x_j$. If we judge it nonrelevant, $\lceil AP_s \rceil$ will decrease by $\sum_{j|x_j \neq 0} a_{ij} x_j$. These are matrix elements V^R_{is} and V^N_{is} respectively, computed at steps 4–7 in INIT-WEIGHTS and steps 2–7 in UPDATE-WEIGHTS.

Now suppose we have two systems s_1 and s_2. We want to judge the document that's going to have the greatest effect on $\Delta AP = AP_{s_1} - AP_{s_2}$. We can bound ΔAP as we did AP above, but the bounds are much hard to compute exactly. It turns out that that does not matter: it can be proven that the judgment that reduces the upper bound of ΔAP the most is a nonrelevant judgment to the document that maximizes $V^N_{is_1} - V^N_{is_2}$, and the judgment that increases the lower bound the most is a relevant judgment to the document that maximizes $V^R_{is_1} - V^R_{is_2}$. Since we of course do not know the judgment in advance, the final weight of document i is the maximum of these two quantities.

When we have more than two systems, we simply calculate the weight for each pair and take the maximum over all pairs as the document weight. Since the maximum over all pairs is simply the maximum weight for any system minus the minimum weight for any system, this can be calculated in linear time, as steps 3–5 of set-weights show.

3.3 UMass Evaluation

The evaluation tool `mtc-eval` takes as input one or more retrieval systems. It calculates $\mathcal{E}MAP$ (Eq. 1 below) for each system; these are used to rank the systems. Additionally, it computes $E[\Delta AP]$, $Var[\Delta AP]$, and $P(\Delta AP < 0)$ (Eqs. 2, 3, 4 respectively) for each topic and each pair of systems, and $\mathcal{E}\Delta MAP$, $\mathcal{V}\Delta MAP$, and $P(\Delta MAP < 0)$ (Eqs. 5, 6, 7 respectively) for each pair of systems. More details are provided below.

3.3.1 Expected Mean Average Precision

As we showed in Carterette et al., average precision can be written as a quadratic equation over Bernoulli trials X_i for the relevance of document i:

$$AP_s = \frac{1}{|R|} \sum_{i=1}^{n} \sum_{j \geq i} A_{ij} X_i X_j$$

where $A_{ij} = 1/\max\{r_s(i), r_s(j)\}$.

Let $p_i = p(X_i = 1)$. The expectation of AP_s is:

$$E[AP_s] \approx \frac{1}{\sum p_i} \sum_{i}^{n} \left(A_{ii} p_i + \sum_{j>i} A_{ij} p_i p_j \right)$$

We can likewise define the expected value of MAP, $\mathcal{E}MAP$, by summing over many topics:

$$\mathcal{E}MAP_s = \sum_{t \in \mathcal{T}} E[AP_{st}] \tag{1}$$

Systems submitted to the track were ranked by $\mathcal{E}MAP$. Probabilities p_i can be estimated in several different ways; Section 3 describes the method we used in detail.

3.3.2 ΔMAP and Confidence

In our previous work we have been more interested in the difference in MAP between two systems rather than the MAPs themselves. In this section we describe ΔMAP and the idea of confidence that an observed difference between systems is "real".

As in Section 3.2, suppose we have two retrieval systems s_1 and s_2. Define $\Delta AP = AP_{s_1} - AP_{s_2}$. We can write ΔAP in closed form as:

$$\Delta AP = \sum_{i=1}^{n} \sum_{j \geq i} C_{ij} X_i X_j$$

where $C_{ij} = 1/\max\{r_{s_1}(i), r_{s_1}(j)\} - 1/\max\{r_{s_2}(i), r_{s_2}(j)\}$.

ΔAP has a distribution over all possible assignments of relevance to the unjudged documents. Some assignments will result in $\Delta AP < 0$, some in $\Delta AP > 0$; if we believe that $\Delta AP < 0$ but there are many possible sets of judgments that could result in $\Delta AP > 0$, then we should say that we have low confidence in our belief.

As it turns out, ΔAP converges to a normal distribution. This makes it very easy to determine confidence: we simply calculate the expectation and variance of ΔAP and plug them into the normal cumulative density function provided by any statistics software package.

The expectation and variance of ΔAP are:

$$E[\Delta AP] = \frac{1}{\sum p_i} \sum_{i} \left(C_{ii} p_i + \sum_{j>i} C_{ij} p_i p_j \right) \tag{2}$$

$$\begin{aligned} Var[\Delta AP] = \frac{1}{(\sum p_i)^2} \Big(& \sum_{i} C_{ii}^2 p_i q_i + \sum_{j>i} C_{ij}^2 p_i p_j (1 - p_i p_j) \\ & + \sum_{i \neq j} 2 C_{ii} C_{ij} p_i p_j q_i + \sum_{k>j \neq i} 2 C_{ij} C_{ik} p_i p_j p_k q_i \Big) \end{aligned} \tag{3}$$

Confidence in a difference in average precision is then defined as

$$\text{confidence} = P(\Delta AP < 0) = \Phi \left(\frac{-E[\Delta AP]}{\sqrt{Var[\Delta AP]}} \right) \tag{4}$$

where Φ is the normal cumulative density function.

This can be very easily extended to determining our confidence in a difference in MAP. The expectation and variance of ΔMAP are:

$$\mathcal{E}\Delta MAP = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} E[\Delta AP_t] \tag{5}$$

$$\mathcal{V}\Delta MAP = \frac{1}{|\mathcal{T}|^2}\sum_{t\in\mathcal{T}} Var[\Delta AP_t] \tag{6}$$

and

$$\text{confidence} = P(\Delta MAP < 0) = \Phi\left(\frac{-\mathcal{E}\Delta MAP}{\sqrt{\mathcal{V}\Delta MAP}}\right) \tag{7}$$

3.3.3 Estimating Relevance

The formulas above require probabilities of relevance for unjudged documents. We used the "expert aggregation" model described in [9]. We will not present details here, but the goal is to estimate the relevance of unjudged documents based on the performance of systems over the judged documents. The model takes into account:

1. the relative frequency of relevant and nonrelevant documents for a topic;
2. the ability of a system to retrieve relevant documents;
3. the ability of a system to rank relevant documents highly;
4. the ability of a system to not retrieve nonrelevant documents;
5. variance over different systems using similar algorithms to rank.

Fitting the model is a three-step process: first, ranks are mapped to decreasing probabilities based on the number of judged relevant and judged nonrelevant documents identified for each topic. Second, these probabilities are calibrated to each system's ability to retrieve relevant documents at each rank. Finally, the systems' calibrated probabilities and the available judgments are used to train a logistic regression classifier for relevance. The model predicts probabilities of relevance for all unjudged documents.

4 NEU Evaluation Method

In this section, we describe the statistical sampling evaluation methodology, *statAP*, developed at Northeastern University and employed in the Million Query track. We begin with a simple example in order to provide intuition for the sampling strategy ultimately employed, and we then proceed to describe the specific application of this intuition to the general problem of retrieval evaluation.

4.1 Sampling Theory and Intuition

As a simple example, suppose that we are given a ranked list of documents $(d_1, d_2, \ldots)$, and we are interested in determining the precision-at-cutoff 1000, i.e., the fraction of the top 1000 documents that are relevant. Let $PC(1000)$ denote this value. One obvious solution is to examine each of the top 1000 documents and return the number of relevant documents seen divided by 1000. Such a solution requires 1000 relevance judgments and returns the *exact* value of $PC(1000)$ with *perfect certainty*. This is analogous to forecasting an election by polling each and every registered voter and asking how they intend to vote: In principle, one would determine, with certainty, the exact fraction of voters who would vote for a given candidate on that day. In practice, the cost associated with such "complete surveys" is prohibitively expensive. In election forecasting, market analysis, quality control, and a host of other problem domains, *random sampling* techniques are used instead [15].

In random sampling, one trades-off *exactitude* and *certainty* for *efficiency*. Returning to our $PC(1000)$ example, we could instead *estimate* $PC(1000)$ with some *confidence* by sampling in the obvious manner: Draw m documents uniformly at random from among the top 1000, judge those documents, and return the number of relevant documents seen divided by m — this is analogous to a random poll of registered voters in election forecasting. In statistical parlance, we have a *sample space* of documents indexed by $k \in \{1, \ldots, 1000\}$, we have a *sampling distribution* over those documents $p_k = 1/1000$ for all $1 \leq k \leq 1000$, and we have a *random variable* X corresponding to the relevance of documents,

$$x_k = rel(k) = \begin{cases} 0 & \text{if } d_k \text{ is non-relevant} \\ 1 & \text{if } d_k \text{ is relevant.} \end{cases}$$

One can easily verify that the *expected value* of a single random draw is $PC(1000)$

$$E[X] = \sum_{k=1}^{1000} p_k \cdot x_k = \frac{1}{1000} \sum_{k=1}^{1000} rel(k) = PC(1000),$$

and the Law of Large Numbers and the Central Limit Theorem dictate that the *average* of a set S of m such random draws

$$\widehat{PC}(1000) = \frac{1}{m} \sum_{k \in S} X_k = \frac{1}{m} \sum_{k \in S} rel(k)$$

will converge to its expectation, $PC(1000)$, quickly [13] — this is the essence of random sampling.

Random sampling gives rise to a number of natural questions: (1) How should the random sample be drawn? In *sampling with replacement*, each item is drawn independently and at random according to the distribution given (uniform in our example), and repetitions may occur; in *sampling without replacement*, a random subset of the items is drawn, and repetitions will not occur. While the former is much easier to analyze mathematically, the latter is often used in practice since one would not call the same registered voter twice (or ask an assessor to judge the same document twice) in a given survey. (2) How should the sampling distribution be formed? While $PC(1000)$ seems to dictate a uniform sampling distribution, we shall see that non-uniform sampling gives rise to much more efficient and accurate estimates. (3) How can one quantify the accuracy and confidence in a statistical estimate? As more samples are drawn, one expects the accuracy of the estimate to increase, but by how much and with what confidence? In the paragraphs that follow, we address each of these questions, in reverse order.

While statistical estimates are generally designed to be correct in expectation, they may be high or low in practice (especially for small sample sizes) due to the nature of random sampling. The variability of an estimate is measured by its *variance*, and by the Central Limit Theorem, one can ascribe 95% confidence intervals to a sampling estimate given its variance. Returning to our $PC(1000)$ example, suppose that (unknown to us) the actual $PC(1000)$ was 0.25; then one can show that the variance in our random variable X is 0.1875 and that the variance in our sampling estimate is $0.1875/m$, where m is the sample size. Note that the variance decreases as the sample size increases, as expected. Given this variance, one can derive 95% confidence intervals [13], i.e., an error range within which we are 95% confident that our estimate will lie.[2] For example, given a sample of size 50, our 95% confidence interval is $+/-0.12$, while for a sample of size 500, our 95% confidence interval is $+/-0.038$. This latter result states that with a sample of size 500, our estimate is likely to lie in the range $[0.212, 0.288]$. In order to increase the accuracy of our estimates, we must decrease the size of the confidence interval. In order to decrease the size of the confidence interval, we must decrease the variance in our estimate, $0.1875/m$. This can be accomplished by either (1) decreasing the variance of the underlying random variable X (the 0.1875 factor) or (2) increasing the sample size m. Since increasing m increases our judgment effort, we shall focus on decreasing the variance of our random variable instead.

While our $PC(1000)$ example seems to inherently dictate a uniform sampling distribution, one can reduce the variance of the underlying random variable X, and hence the sampling estimate, by employing *non-uniform* sampling. A maxim of sampling theory is that accurate estimates are obtained when one samples with *probability proportional to size* (PPS) [15]. Consider our election forecasting analogy: Suppose that

[2] For estimates obtained by averaging a random sample, the 95% confidence interval is roughly $+/-1.965$ *standard deviations*, where the standard deviation is the square root of the variance, i.e., $\sqrt{0.1875/m}$ in our example.

our hypothetical candidate is known to have strong support in rural areas, weaker support in the suburbs, and almost no support in major cities. Then to obtain an accurate estimate of the vote total (or fraction of total votes) this candidate is likely to obtain, it makes sense to spend your (sampling) effort "where the votes are." In other words, one should spend the greatest effort in rural areas to get very accurate counts there, somewhat less effort in the suburbs, and little effort in major cites where very few people are likely to vote for the candidate in question. However, one must now compensate for the fact that the sampling distribution is non-uniform — if one were to simply return the fraction of polled voters who intend to vote for our hypothetical candidate when the sample is highly skewed toward the candidate's areas of strength, then one would erroneously conclude that the candidate would win in a landslide. To compensate for non-uniform sampling, one must *under-count* where one *over-samples* and *over-count* where one *under-samples.*

Returning to our $PC(1000)$ example, employing a PPS strategy would dictate sampling "where the relevant documents are." Analogous to the election forecasting problem, we do have a prior belief about where the relevant documents are likely to reside — in the context of ranked retrieval, relevant documents are generally more likely to appear toward the top of the list. We can make use of this fact to reduce our sampling estimate's variance, so long as our assumption holds. Consider the non-uniform sampling distribution shown in Figure 1 where

$$p_k = \begin{cases} 1.5/1000 & 1 \le k \le 500 \\ 0.5/1000 & 501 \le k \le 1000. \end{cases}$$

Here we have *increased* our probability of sampling the top half (where more relevant documents are likely to reside) and *decreased* our probability of sampling the bottom half (where fewer relevant documents are likely to reside).

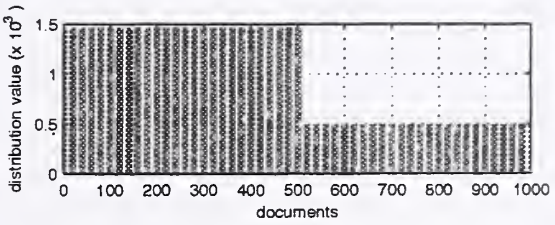


Figure 1: Non-uniform sampling distribution.

In order to obtain the correct estimate, we must now "under-count" where we "over-sample" and "over-count" where we "under-sample." This is accomplished by modifying our random variable X as follows:

$$x_k = \begin{cases} rel(k)/1.5 & 1 \le k \le 500 \\ rel(k)/0.5 & 501 \le k \le 1000. \end{cases}$$

Note that we over/under-count by precisely the factor that we under/over-sample; this ensures that the expectation is correct:

$$\begin{aligned} E[X] &= \sum_{k=1}^{1000} p_k \cdot x_k = \sum_{k=1}^{500} \frac{1.5}{1000} \cdot \frac{rel(k)}{1.5} + \sum_{k=1}^{500} \frac{0.5}{1000} \cdot \frac{rel(k)}{0.5} \\ &= \frac{1}{1000} \sum_{k=1}^{1000} rel(k) = PC(1000). \end{aligned}$$

For a given sample S of size m, our estimator is then a weighted average

$$\begin{aligned} \widehat{PC}(1000) &= \frac{1}{m} \sum_{k \in S} X_k \\ &= \frac{1}{m} \left(\sum_{k \in S\,:\,k \le 500} \frac{rel(k)}{1.5} + \sum_{k \in S\,:\,k > 500} \frac{rel(k)}{0.5} \right) \end{aligned}$$

where we over/under-count appropriately.

Note that our expectation and estimator are correct, *independent of whether our assumption about the location of the relevant documents actually holds!* However, if our assumption holds, then the variance of our

random variable (and sampling estimate) will be reduced (and vice versa). Suppose that all of the relevant documents were located where we over-sample. Our expectation would be correct, and one can show that the variance of our random variable is reduced from 0.1875 to 0.1042 — we have sampled where the relevant documents are and obtained a more accurate count as a result. This reduction in variance yields a reduction in the 95% confidence interval for a sample of size 500 from $+/-0.038$ to $+/-0.028$, a 26% improvement. Conversely, if the relevant documents were located in the bottom half, the confidence interval would increase.

One could extend this idea to three (or more) strata, as in Figure 2. For each document k, let α_k be the factor by which it is over/under-sampled with respect to the uniform distribution; for example, in Figure 1, α_k is 1.5 or 0.5 for the appropriate ranges of k, while in Figure 2, α_k is 1.5, 1, or 0.5 for appropriate ranges of k. For a sample S of size m drawn according to the distribution in question, the sampling estimator would be

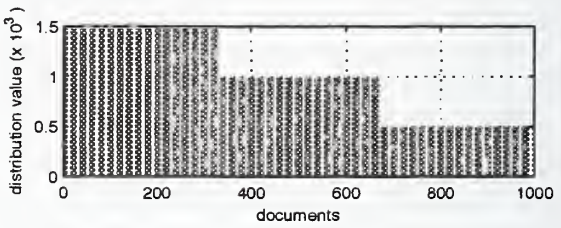


Figure 2: Non-uniform distrib. with three strata.

$$\widehat{PC}(1000) = \frac{1}{m} \sum_{k \in S} \frac{rel(k)}{\alpha_k}.$$

In summary, one can sample with respect to any distribution, and so long as one over/under-counts appropriately, the estimator will be correct. Furthermore, if the sampling distribution places higher weight on the items of interest (e.g., relevant documents), then the variance of the estimator will be reduced, yielding higher accuracy. Finally, we note that sampling is often performed *without replacement* [15]. In this setting, the estimator changes somewhat, though the principles remain the same: sample where you think the relevant documents are in order to reduce variance and increase accuracy. The α_k factors are replaced by *inclusion probabilities* π_k, and the estimator must be normalized by the size of the sample space:

$$\widehat{PC}(1000) = \frac{1}{1000} \sum_{k \in S} \frac{rel(k)}{\pi_k}.$$

Modularity. The evaluation and sampling modules are completely independent: the sampling module produces the sample in a specific format but does not impose or assume a particular evaluation being used; conversely, the evaluation module uses the given sample, with no knowledge of or assumptions about the sampling strategy strategy empolyed (a strong improvement over method presented in [5]). In fact, the sampling technique used is known to work with many other estimators, while the estimator used is known to work with other sampling strategies [8]. This flexibility is particularly important if one has reason to believe that a different sampling strategy might work better for a given evaluation.

4.2 The sample

The sample is the set of documents selected for judging together with all information required for evaluation: in our case, that means (1) the documents ids, (2) the relevance assessments, and (3) the inclusion probability for each document.

The inclusion probability π_k is simply the probability that the document k would be included in any sample of size m. In without-replacement sampling, $\pi_k = p_k$ when $m = 1$ and π_k approaches 1 as the sample size grows. For most common without-replacement sampling approaches, these inclusion probabilities are notoriously difficult to compute, especially for large sample sizes [8].

Additional judged documents, obtained deterministically, can be added to the existing sample with associated inclusion probability of 1. This is a useful feature as often in practice separate judgments are available; it matches perfectly the design of the Million Query Track pooling strategy, where for more than 800 topics a mixed pool of documents was created (half randomly sampled, half deterministically chosen).

Additionally, deterministic judgments may arise in at least two other natural ways: First when large-scale judging is done by assessors, it might be desirable to deterministically judge a given depth-pool (say the top 50 documents of every list to be evaluated) and then invoke the sampling strategy to judge additional

documents. (This strategy was employed in Terabyte 06 Track). Second, if it is determined that additional samples are required (say for a new run with many unjudged documents), one can judge either hand-picked documents and/or sampled documents and combine them with the original sample. Any collisions (where a document is sampled and separately deterministically judged) are handled by setting inclusion probability to 1.

4.3 Evaluation

Given a sample S of judged documents along with inclusion probabilities, we discuss here how to estimate quantities of interest (AP, R-precision, precision-at-cutoff).

For AP estimates, which we view as mean of a population of precision values, we adapt the generalized ratio estimator for unequal probability designs (very popular on polls, election strategies, market research etc.), as described in [15]:

$$\widehat{X} = \frac{\sum_{k \in S} v_k/\pi_k}{\sum_{k \in S} 1/\pi_k}$$

where v_k is the *value* associated with item k (e.g., the relevance of a document, a vote for a candidate, the size of a potential donation, etc.). For us, the "values" we wish to average are the precisions at relevant documents, and the ratio estimator for AP is thus

$$\widehat{AP} = \frac{\sum\limits_{rel(k)=1} \widehat{PC}(rank(k))/\pi_k}{\sum\limits_{rel(k)=1} 1/\pi_k} \tag{8}$$

where $\widehat{PC}(r) = \frac{1}{r} \sum\limits_{rank(k) \le r} \frac{rel(k)}{\pi_k}$ estimates precision at rank r and $k \in S$ iterates through *sampled documents only*).

Note that $\widehat{AP}$ mimics the well known formula $AP = \frac{sum\_of\_precisions\_at\_rel\_docs}{number\_of\_rel\_docs}$ because the numerator is an unbiased estimator for the sum of precision values at relevant ranks, while the denominator is an unbiased estimator of the number of relevant documents in the collection: $\widehat{R} = \sum_{rel(k)=1} \frac{1}{\pi_k}$. Combining the estimates for R and for precision at rank, $PC(r)$, we obtain also an estimate for R-precision:

$$\widehat{RP} = \widehat{PC}(\widehat{R}) = \frac{1}{\widehat{R}} \sum_{rank(k) \le \widehat{R}} \frac{rel(k)}{\pi_k}. \tag{9}$$

Finally, we note that the variance in our estimates can be estimated as well, and from this, one can determine confidence intervals in all estimates produced. Details may be found our companion paper [1].

4.4 Sampling strategy

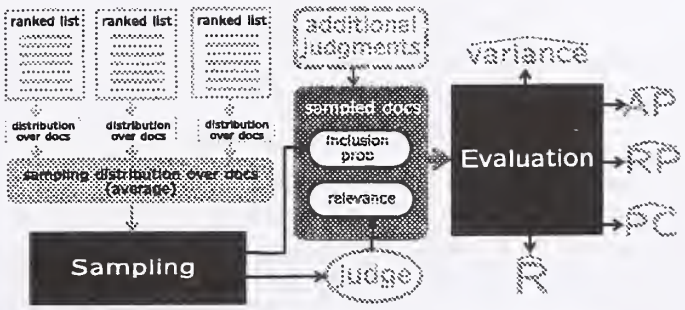


Figure 3: statAP: Sampling and evaluation design.

There are many ways one can imagine sampling from a given distribution [8]. Essentially, sampling consists of a *sampling distribution* over documents (that should be dictated by the ranks of documents in the ranked lists and therefore naturally biased towards relevant documents) and a *sampling strategy* (sometimes called "selection") that produces inclusion probabilities roughly proportional to the sampling distribution. Following are our proposed choices for both the distribution and the selection algorithms; many others could work just as well. In the Million Query Track, due to unexpected server behavior, both the

sampling distribution and the selection strategy were altered, yielding suboptimal chosen samples; nevertheless, we were able to compute the inclusion probability for each selected document and run the evaluation, though at a reduced accuracy and efficiency.

Sampling distribution (prior). It has been shown that average precision induces a good relevance prior over the ranked documents of a list. The AP-prior has been used with sampling techniques[5]; in metasearch (data fusion) [4]; in automatic assessment of query difficulty [2]; and in on-line application to pooling[3]. It has also been shown that this prior can be averaged over multiple lists to obtain a global prior over documents[5]. An accurate description together with motivation and intuition can be found in [5].

For a given ranked list of documents, let Z be the size of the list. Then the prior distribution weight associated with any rank r, $1 \leq r \leq Z$, is given by

$$W(r) = \frac{1}{2Z}\left(1 + \frac{1}{r} + \frac{1}{r+1} + \cdots + \frac{1}{Z}\right) \approx \frac{1}{2Z}\log\frac{Z}{r}. \tag{10}$$

We used for experimentation the above described prior, averaged per document over all run lists; Note that the our sampling strategy works with any prior over documents.

Stratified sampling strategy. The most important considerations are: handle *non-uniform* sampling distribution; *without replacement* so we can easily add other judged documents; *probabilities proportional with size (pps)* minimizes variance by obtaining inclusion probabilities π_k roughly proportional with precision values $PC_{rank(d)}$; and *computability of inclusion probabilities* for documents (π_k) and for pairs of documents (π_{kl}). We adopt a method developed by Stevens [8, 14], sometimes referred to as *stratified sampling*, that has all of the features enumerated above and it is very straight forward for our application. The details of our proposed sampling strategy can be found in [1]. Figure 3 provides an overall view of the statAP sampling and evaluation methodology.

5 Alternation of methods

Half of the queries were served by alternating between the UMass method MTC and the NEU method statMAP. The alternation was kept on separate "tracks", so that a judgment on a document served by statMAP would not affect the document weights for MTC. If, say, statMAP selected a document that MTC had already served (and therefore that had already been judged), the judgment was recorded for statMAP without showing the document to the user.

6 Statistics

The following statistics reflect the status of judgments as of October 16, 2007. Those are not the same judgments that were used by the track participants for their notebook papers, though the differences are small.

1,755 queries were judged
A total of 22 of those queries were judged by more than one person.
10 were judged by two people
5 were judged by 3 people
4 were judged by 4 people
3 were judged by 5 or more people

The actual assignment of topics to judging method was done in advance based on topic number. The following was the distribution of topics to methods:

443	of those used the MTC (UMass-only) method
471	used the statMAP (NEU-only) method
432	alternated, starting with MTC
409	alternated, starting with statMAP

Since assessors were shown a set of queries and could choose from them, we wondered whether there was an order effect. That is, did people tend to select the first query or not. Here is the number of times someone selected a query for judging based on where in the list of 10 it was.

	149 Terabyte		1MQ		
run name	unjudg	MAP	unjudg	$\mathcal{E}MAP$	statMAP
UAms.AnLM	64.72	0.0278‡	90.75	0.0281	0.0650
UAms.TiLM	61.43	0.0392‡	89.40	0.0205	0.0938
exegyexact	8.81	0.0752‡	13.67	0.0184	0.0517
umelbexp	61.17	0.1251	91.85	0.0567*†	0.1436†
ffind07c	22.91	0.1272‡	77.94	0.0440	0.1531
ffind07d	24.07	0.1360	82.11	0.0458	0.1612
sabmq07a1	21.69	0.1376	86.51	0.0494	0.1519
UAms.Sum6	32.74	0.1398‡	81.37	0.0555	0.1816
UAms.Sum8	24.40	0.1621	79.92	0.0580	0.1995
UAms.TeVS	21.11	0.1654	81.35	0.0503	0.1805
hedge0	16.90	0.1708‡	80.44	0.0647	0.2175
umelbimp	15.40	0.2499	80.83	0.0870	0.2568
umelbstd	11.48	0.2532‡	82.17	0.0877	0.2583
umelbsim	10.38	0.2641‡	80.17	0.1008*†	0.2891†
hitir	9.06	0.2873	80.25	0.0888	0.2768
rmitbase	8.32	0.2936	79.28	0.0945	0.2950
indriQLSC	7.34	0.2939	79.18	0.0969	0.3040
LucSynEx	13.02	0.2939	78.23	0.1032*	0.3184*
LucSpel0	13.08	0.2940	78.27	0.1031	0.3194*
LucSyn0	13.08	0.2940	78.27	0.1031	0.3194*
indriQL	7.12	0.2960‡	78.80	0.0979*	0.3086
JuruSynE	8.86	0.3135	78.36	0.1080	0.3117
indriDMCSC	9.79	0.3197	80.36	0.0962*	0.2981*
indriDM	8.67	0.3238	79.51	0.0981*	0.3060*

Table 1: Performance on 149 Terabyte topics, 1692 partially-judged topics per $\mathcal{E}MAP$, and 1084 partially-judged queries per statMAP, along with the number of unjudged documents in the top 100 for both sets.

Rank	1	2	3	4	5	6	7	8	9	10
Count	213	157	144	148	169	141	118	145	156	139
Percent	13.9%	10.3%	9.4%	9.7%	11.0%	9.2%	7.7%	9.5%	10.2%	9.1%

(The numbers add up to 1530 rather than 1755 because this logging was included partway through the judging process.)

Judgments came from the following sources:

1,478	NIST assessors
97	CIIR hired annotators
47	IR class project

The remaining judgments came from different sites, some (though not all) of which were participants. The number of judged queries ranked from 1 to 37 per site (other than those listed above).

7 Results

The 24 runs were evaluated over the TB set using `trec_eval` and over the 1MQ set using $\mathcal{E}MAP$ and statMAP. If TB is representative, we should see that $\mathcal{E}MAP$ and statMAP agree with each other as well as TB about the relative ordering of systems. Our expectation is that statMAP will present better estimates of MAP while $\mathcal{E}MAP$ is more likely to present a correct ranking of systems.

The left side of Table 1 shows the MAP for our 24 systems over the 149 Terabyte queries, ranked from lowest to highest. The average number of unjudged documents in the top 100 retrieved is also shown. Since some of these systems did not contribute to the Terabyte judgments, they ranked quite a few unjudged documents.

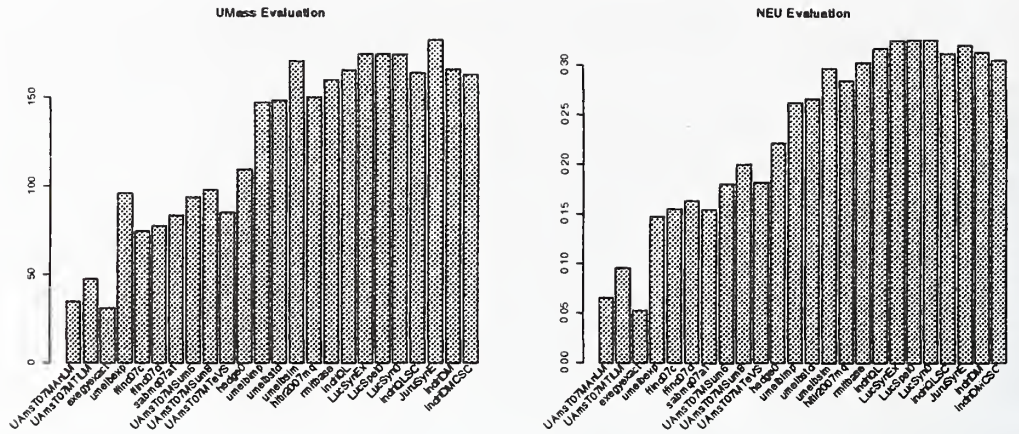


Figure 4: MTC and statMAP evaluation results sorted by evaluation over 149 Terabyte topics.

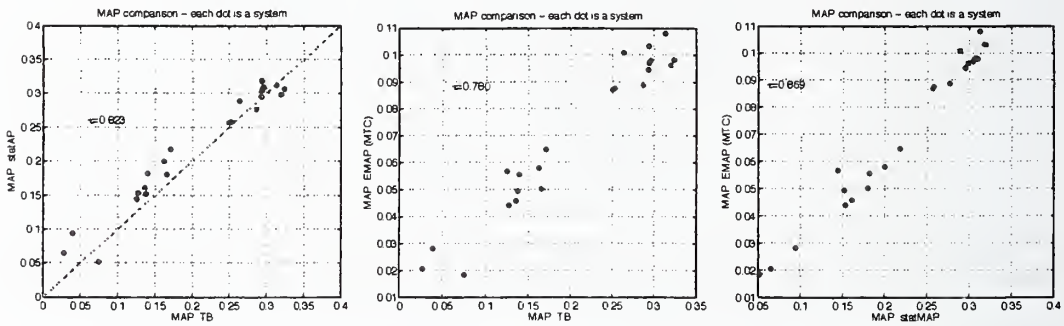


Figure 5: From left, evaluation over Terabyte queries versus statMAP, evaluation over Terabyte queries versus $\mathcal{EMAP}$, and statMAP evaluation versus $\mathcal{EMAP}$ evaluation.

The right side shows $\mathcal{EMAP}$ and statMAP over the queries judged for our experiment, in order of increasing MAP over Terabyte queries. It also shows the number of unjudged documents in the top 100. $\mathcal{EMAP}$ and statMAP are evaluated over somewhat different sets of queries; statMAP excludes queries judged by MTC and queries for which no relevant documents were found, while $\mathcal{EMAP}$ includes all queries, with those that have no relevant documents having some probability that a relevant document may yet be found.

Overall, the rankings by $\mathcal{EMAP}$ and statMAP are fairly similar, and both are similar to the "gold standard". Figure 4 shows a graphical representation of the two rankings compared to the ranking by Terabyte systems. Figure 5 shows how statMAP, $\mathcal{EMAP}$, and MAP over TB queries correlate. All three methods have identified the same three clusters of systems, separated in Table 1 by horizontal lines; within those clusters there is some variation in the rankings between methods. For statMAP estimates (Figure 5, left plot), besides the ranking correlation, we note the accuracy in terms of absolute difference with the TB MAP values by the line corresponding to the main diagonal.

Some of the bigger differences between the methods are noted in Table 1 by a * indicating that the run moved four or more ranks from its position in the TB ranking, or a † indicating a difference of four or more ranks between $\mathcal{EMAP}$ and statMAP. Both methods presented about the same number of such disagreements, though not on the same systems. The biggest disagreements between $\mathcal{EMAP}$ and statMAP were on umelbexp and umelbsim, both of which $\mathcal{EMAP}$ ranked five places higher than statMAP. Each method settled on a different "winner": indriDM for the TB queries, JuruSynE for $\mathcal{EMAP}$, and LucSpel0 and LucSyn0 tied by statMAP. However, these systems are all quite close in performance by all three methods.

An obvious concern about the gold standard is the correlation between the number of unjudged documents and MAP: the tau correlation is $-.517$, or $-.608$ when exegyexact (which often retrieved only one document) is excluded. This correlation persists for the number unjudged in the top 10. To ensure that we were not inadvertently ranking systems by the number of judged documents, we selected some of the top-retrieved documents in sparsely-judged systems for additional judgments. A total of 533 additional judgments only discovered 7 new relevant documents for the UAms systems, 4 new relevant documents for the ffind systems, but 58 for umelbexp. The new relevant judgments caused umelbexp to move up one rank. This suggests that while the ranking is fair for most systems, it is likely underestimating umelbexp's performance.

It is interesting that the three evaluations disagree as much as they do in light of work such as Zobel's [16]. There are at least three possible reasons for the disagreement: (1) the gold standard queries represent a different sample space than the rest; (2) the gold standard queries are incompletely judged; and (3) the assessors did not pick queries truly randomly. The fact that $\mathcal{E}MAP$ and statMAP agree with each other more than either agrees with the gold standard suggests to us that the gold standard is most useful as a loose guide to the relative differences between systems, but does not meaningfully reflect "truth" over the larger query sample. But the possibility of biased sampling affects the validity of the other two sets as well: as described above, assessors were allowed to choose from 10 different queries, and it is possible they chose queries that they could decide on clear intents for rather than queries that were unclear. It is difficult to determine how random query selection was. We might hypothesize that, due to order effects, if selection was entirely random we would expect to see the top most query selected most, followed by the second-ranked query, followed by the third, and so on, roughly conforming to a log-normal distribution. This in fact is *not* what happened; as the click rates in Section 6 show, assessors chose the top-ranked query slightly more often than the others (13.9% of all clicks), but the rest were roughly equal (around 10%). But this would only disprove random selection if we could guarantee that presentation bias holds in this situation. Nevertheless, it does lend weight to the idea that query selection was not random.

8 Additional Analysis

In this section we present some additional statistics and analysis over the collected data. For more detailed analysis, in particular on the stability of rankings, tradeoffs between the numbers of queries and judgments, and reusability, we refer the reader to our companion work [11].

8.1 Assessments

Assessors made a total of 33,077 judgments for the 801 alternating queries. Of these, 15,028 (45%) were chosen by both methods. 12,489 (38%) were chosen only by MTC, and 5,560 (17%) were chosen only by statMAP.

Forty-two of the 149 Terabyte topics ended up being selected by 1MQ assessors to be rejudged. For these 42 queries, there were 2,011 documents judged for both the 2007 1MQ track and the 2005 Terabyte track. Agreement on the relevance of these documents was 75%.

Looking at the difference in system rankings produced by the NIST assessors only versus those produced by the non-NIST assessors may provide a sort of "upper bound" Kendall's τ correlation, the best that can be expected given disagreement between assessors. Though $\tau = 0.9$ is the usual standard, our observed correlation is 0.802. Nearly all of this is due to swaps in the top-ranked systems, which are very similar in quality.

8.2 Agreement on Statistical Significance

We evaluated statistical significance over the TB queries by a one-sided paired t-test at $\alpha = 0.05$. A run denoted by a ‡ in Table 1 has a MAP significantly less than the next run in the ranking. (Considering the number of unjudged documents, some of these results should be taken with a grain of salt.) Significance is not transitive, so a significant difference between two adjacent runs does not always imply a significant difference between other runs. Both $\mathcal{E}MAP$ and statMAP swapped some significant pairs, though they agreed with each other for nearly all such swaps.

pair	confidence
exegyexact & UAmsT07MAnLM	0.9577
sabmq07a1 & UAmsT07MTeVS	0.7113
UAmsT07MSum6 & umelbexp	0.6639
umelbexp & UAmsT07MSum8	0.6920
umelbimp & umelbstd	0.6095
umelbimp & hitir2007mq	0.7810
umelbstd & hitir2007mq	0.6909
rmitbase & indriDMCSC	0.8412
indriDMCSC & indriQLSC	0.6552
indriDMCSC & indriQL	0.8480
indriQLSC & indriDM	0.7748
indriQL & indriDM	0.5551
LucSyn0 & LucSpel0	0.5842
LucSyn0 & LucSynEx	0.6951
LucSpel0 & LucSynEx	0.6809

Table 2: Probability that a difference in MAP is less than zero for selected pairs of systems.

Overall, MTC agreed with 92.4% of the significant differences between systems as measured over the 149 Terabyte topics. NEU agreed with 93.7% of the significant differences. This difference is not significant by a one-sample proportion test ($p = 0.54$).

8.3 Confidence

As we described in Section 4, MTC is more interested in differences between pairs than in the value of $\mathcal{EMAP}$. For nearly all the pairs the confidence was 1, meaning that we predict that additional judgments will not change the relative ordering of pairs. Table 2 shows the confidence in the difference in $\mathcal{EMAP}$ for selected pairs that had less than 100% confidence. Note that many of the high-ranked systems (the indri set and the Luc set) were difficult to differentiate.

8.4 ANOVA and Generalizability Theory

As extensively discussed in previous sections, 24 different runs were submitted to the Million Query track, where each run output a ranked list of documents for each one of 10,000 queries. The ranked lists produced by all systems for a subset of 1,755 queries were judged and their quality was assessed employing two different methodologies, MTC and NEU. Each of the two methodologies evaluated the quality of the ranked lists for the 1,755 queries by the means of some estimate of average precision (AP) and the overall quality of each system by some estimate of mean average precision (MAP), resulting into two test collections.

There are two questions that naturally arise: (1) How reliable are the given performance comparisons, and (2) how good are the test collections? We answer these two questions by employing Generalizability Theory (GT) [6, 7].

In particular, GT provides the appropriate tools to answers the question: To what extent does the variance of the observed average precision (AP) values reflect real performance difference between the systems as opposed to other sources of variance? During the first step of GT (the G-study), the AP value for a ranked list of documents produced by a single system ran over a single query can be decomposed into a number of uncorrelated effects (sources of variance),

$$AP_{aq} = \mu + v_a + v_q + v_{aq,err}$$

where μ is the grand mean over all AP values, v_a is the system effect, v_q is the query effect and $v_{aq,err}$ is the system-query interaction effect along with any other effect not being considered. Apart from the grand mean that is a constant, each of the other effects is modeled as a random variable and therefore it has mean and variance. In the same manner as the AP value decomposition, the variance of the observed AP value is

Effects	Variance	% of total variance
System Effect	0.0008	10%
Query Effect	0.0054	69%
S-Q Interaction Effect	0.0016	21%

Table 3: Variance components analysis based on 429 queries employing the MTC methodology.

Effects	Variance	% of total variance
System Effect	0.0069	11%
Query Effect	0.0247	39%
S-Q Interaction Effect	0.0310	50%

Table 4: Variance components analysis based on 459 queries employing the NEU methodology.

decomposed into the corresponding variance components,

$$\sigma^2(AP_{aq}) = \sigma^2(a) + \sigma^2(q) + \sigma^2(aq, err)$$

Table 3 and Table 4 provide estimates of those variance components when the MTC and the NEU methodology is employed, respectively. The figures in Table 3 are based on 429 queries selected by MTC, while the figures in Table 4 are based on 459 selected by NEU. Note that each variance component reported in the two tables along with the corresponding percentage is calculated on a per query basis. Therefore, 65.42% (or 78.64%) would be the percentage of the total variance that corresponds to the system-query interaction if a system runs on a single query when the MTC (or NEU) methodology is employed.

While the G-study copes with the decomposition of variance of a single AP value into variance components due to a single system and a single query, the next step of GT (the D-study) considers the decomposition of the variance of the average of the AP values over all queries (MAP) into variance components due to a single system and a set of N_q queries. The variance components in the D-study can be easily computed by using the variance components computed during G-study as follows,

$$\sigma^2(Q) = \sigma^2(q)/N_q, \sigma^2(aQ) = \sigma^2(aq)/N_q$$

while the variance due to the system effect remains the same.

Table 5 and Table 6 provide the percentage of the variance of the MAP values that is due to the system effect, i.e. $\sigma^2(a)/(\sigma^2(a) + \sigma^2(q)/N_q + \sigma^2(aq)/N_q)$ for different number of queries (N_q) for the two methodologies. As can be observed, for both MTC and NEU methodologies, the variance in the MAP scores, in a test design of 450 queries (i.e. approximately the design used in Million Query track) reflect the real performance difference between the systems, since the percentage of the total MAP variance that is due to the system effect is 98% for both methodologies. Therefore, any disagreement in the overall ranking of the systems by the two methodologies are due to the different estimators used by the two methodologies for computing AP and MAP values.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023, in part by NSF grants IIS-0534482 and IIS-0533625, and in part by Microsoft Live Labs. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

[1] J. A. Aslam and V. Pavlu. A practical sampling strategy for efficient retrieval evaluation. Working draft available at the following URL: http://www.ccs.neu.edu/home/jaa/papers/drafts/statAP.html, May 2007.

Number of Queries (N_q)	50	100	200	450
% of total variance due to system	85%	92%	95%	98%

Table 5: % of total variance due to system employing the MTC methodology.

Number of Queries (N_q)	50	100	200	450
% of total variance due to system	86%	93%	96%	98%

Table 6: % of total variance due to system employing the NEU methodology.

[2] J. A. Aslam and V. Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of ECIR*, pages 198–209, 2007.

[3] J. A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of CIKM*, pages 484–491, 2003.

[4] J. A. Aslam, V. Pavlu, and E. Yilmaz. Measure-based metasearch. In *Proceedings of SIGIR*, pages 571–572, 2005.

[5] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of SIGIR*, pages 541–548, 2006.

[6] D. Bodoff and P. Li. Test theory for assessing ir test collection. In *Proceedings of SIGIR*, pages 367–374, 2007.

[7] R. L. Brennan. *Generalizability Theory.* Springer-Verlag, New York, 2001.

[8] K. R. W. Brewer and M. Hanif. *Sampling With Unequal Probabilities.* Springer, New York, 1983.

[9] B. Carterette. Robust test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 55–62, 2007.

[10] B. Carterette, J. Allan, and R. K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 268–275, 2006.

[11] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proceedings of SIGIR*, 2008.

[12] C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 terabyte track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[13] J. A. Rice. *Mathematical Statistics and Data Analysis.* Duxbury Press, second edition, 1995.

[14] W. L. Stevens. Sampling without replacement with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 20, No. 2. (1958), pp. 393-397.*

[15] S. K. Thompson. *Sampling.* Wiley-Interscience, second edition, 2002.

[16] J. Zobel. How reliable are the results of large-scale retrieval experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.

Overview of the TREC 2007 Question Answering Track

Hoa Trang Dang[1], Diane Kelly[2], and Jimmy Lin[3]

[1]National Institute of Standards and Technology
Gaithersburg, MD 20899
hoa.dang@nist.gov

[2]University of North Carolina
Chapel Hill, NC 27599
dianek@email.unc.edu

[3]University of Maryland
College Park, MD 20742
jimmylin@umd.edu

Abstract

The TREC 2007 question answering (QA) track contained two tasks: the main task consisting of series of factoid, list, and "Other" questions organized around a set of targets, and the complex, interactive question answering (ciQA) task. The main task differed from previous years in that the document collection comprised blogs in addition to newswire documents, requiring systems to process diverse genres of unstructured text. The evaluation of factoid and list responses distinguished between answers that were globally correct (with respect to the document collection), and those that were only locally correct (with respect to the supporting document but not to the overall document collection). The ciQA task provided a framework for participants to investigate interaction in the context of complex information needs. Standing in for surrogate users, assessors interacted with QA systems live over the Web; this setup allowed participants to experiment with more complex interfaces but also revealed limitations in the ciQA design for evaluation of interactive systems.

1 Introduction

The goal of the TREC question answering (QA) track is to foster research on systems that directly return answers, rather than documents containing answers, in response to a natural language question. Since its inception in TREC-8 (1999), the track has steadily expanded both the type and difficulty of the questions asked. The first several editions of the track focused on *factoid* questions. A factoid question is a fact-based, short answer question such as *How many calories are there in a Big Mac?* The task in the TREC 2003 QA track contained list and definition questions in addition to factoid questions (Voorhees, 2004). A list question asks for different answer instances that satisfy the information need, such as *List the names of chewing gums*. Answering such questions requires a system to assemble a response from information located in multiple documents. A definition question asks for interesting information about a particular person or thing such as *Who is Vlad the Impaler?* or *What is a golden parachute?* Definition questions also require systems to

locate information in multiple documents, but in this case the information of interest is much less crisply delineated.

Since TREC 2004 (Voorhees, 2005a), factoid and list questions have been grouped into different series, where each series is associated with a target and the questions in the series ask for some information about the target. In addition, the final question in each series is an explicit "Other" question, which is to be interpreted as "Tell me other interesting things about this target I don't know enough to ask directly". This last question is roughly equivalent to the definition questions in the TREC 2003 task. The series format supports the evaluation of different types of questions (factoid, list and Other) while providing an abstraction of a real user session with a QA system.

In TREC 2004, the target for a series could be a person, organization, or thing. Events were added as possible targets in TREC 2005, requiring that answers must be temporally correct with respect to the time-frame defined by the series. In TREC 2006, that requirement for sensitivity to temporal dependencies was made explicit in the distinction between locally and globally correct answers, so that answers for questions phrased in the present tense must not only be supported by the supporting document (locally correct), but must also be the most up-to-date answer in the document collection (globally correct).

The main task in the TREC 2007 QA track repeated the question series format, but with a significant change in the genre of the document collection. Instead of just newswire, the document collection contained both newswire and blogs. Mining blogs for answers introduced significant new challenges in at least two aspects that are very important for real-world QA systems: 1) being able to handle language that is not well-formed, and 2) dealing with discourse structures that are more informal and less reliable than newswire. Based on its successful application in TREC 2006 (Dang and Lin, 2007), the nugget pyramid evaluation method became the official evaluation method for the Other questions in TREC 2007.

In addition to the main task, the TREC 2007 QA track repeated the complex, interactive QA (ciQA) task of TREC 2006. At the TREC 2006 workshop, participants indicated that they wanted to have longer, more complex interactions in the ciQA task rather than short interactions via cached interaction forms. Participants proposed trying "live interactions" for 2007. Under this setup, the interactive QA system was located at a URL (Uniform Resource Locator) on the participant's machine, and NIST assessors simply navigated to the URL. The advantage was that participants were able to explore more complex interactions and interfaces. However, this setup placed the burden on participants to have their systems accessible during the entire interaction period and to record all desired data during the interaction.

The remainder of this paper describes each of the two tasks in the TREC 2007 QA track in more detail. Section 2 describes the questions, evaluation methods, and results for the main task, while Section 3 discusses the ciQA task.

2 Main Task

The scenario for the main task in the TREC 2007 QA track was that an adult, native speaker of English is looking for information about a target of interest. The target could be a person, organization, thing, or event. The user was assumed to be an "average" reader of U.S. newspapers. Serving as surrogate users, NIST assessors developed the questions and judged the system responses.

The main task required systems to provide answers to a series of related questions. A question series, which focused on a target, consisted of several factoid questions, one or two list questions, and exactly one Other question. The order of questions in the series and the type of each question (factoid, list, or Other) were all explicitly encoded in the test set. Example series are shown in Figure 1. The final test set contained 70 series; the targets of these series are given in Table 1. Of the 70 targets, 19 were PERSONs, 17 were

219	Target: Iraqi defector Curveball		
	219.1	FACTOID	What year did Curveball defect?
	219.2	FACTOID	What was Curveball's profession?
	219.3	FACTOID	What is Curveball's real name?
	219.4	FACTOID	Which intelligence service employed Curveball?
	219.5	LIST	Which US government officials accepted his claims regarding Iraqi weapons labs?
	219.6	FACTOID	Where does Curveball now live?
	219.7	OTHER	
254	Target: House of Chanel		
	254.1	FACTOID	Who founded the House of Chanel?
	254.2	FACTOID	In what year was the company founded?
	254.3	FACTOID	Who is the president of the House of Chanel?
	254.4	FACTOID	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	FACTOID	What Chanel creation is the top-selling fragrance in the world?
	254.8	OTHER	
269	Target: Pakistan earthquakes of October 2005		
	269.1	FACTOID	On what date did this earthquake strike?
	269.2	LIST	What countries were affected by this earthquake?
	269.3	FACTOID	What was the final death toll from this earthquake?
	269.4	FACTOID	What was the strength of this earthquake?
	269.5	FACTOID	Where was the epicenter (latitude and longitude)?
	269.6	LIST	What countries supplied aid?
	269.7	OTHER	

Figure 1: Sample question series from the test set. Series 219 has a PERSON as the target, series 254 has an ORGANIZATION as the target, and series 269 has an EVENT as the target.

ORGANIZATIONs, 15 were EVENTs, and 19 were THINGs. The series contained a total of 360 factoid questions, 85 list questions, and 70 Other questions. Each series contained 6–10 questions (counting the Other question), with most series containing 7 questions.

Answers were to be drawn from a document collection comprising the Blog06 corpus (Macdonald and Ounis, 2006) and the AQUAINT-2 Corpus of English News Text. The AQUAINT-2 collection contains approximately 2.5 GB of text (about 907K documents) spanning the time period of October 2004 - March 2006; articles are in English and come from a variety of sources including Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, and The Associated Press. Blog06 documents were collected by polling 100,649 RSS and Atom feeds over an 11 week period (December 6, 2005 - February 21, 2006). A blog document is defined to be a blog post plus its follow-up comments (a permalink). As a convenience for track participants, NIST made available document rankings of the top 1000 documents per target for each of two corpora, as produced using the PRISE document retrieval system, with the target as the query.

Participants were allowed two weeks to download the test data and submit their results. All processing of the questions was required to be strictly automatic. Systems were required to process series independently

216	Paul Krugman	251	Lyme disease
217	Jay-Z	252	American Girl dolls
218	impressionist Darrell Hammond	253	Kurt Weill
219	Iraqi defector Curveball	254	House of Chanel
220	International Management Group (IMG)	255	British American Tobacco (BAT)
221	U.S. Mint	256	Buffalo Soldiers
222	3M	257	2005 DARPA Grand Challenge
223	Merrill Lynch & Co.	258	2005 presidential election in Egypt
224	WWE	259	2005 World Snooker Championships
225	Sago Mine disaster	260	Teenage Mutant Ninja Turtles (TMNT)
226	Harriet Miers withdraws nomination to Supreme Court	261	marsupials
227	Robert Blake criminal trial	262	kumquat
228	March Madness 2006	263	Ayn Rand
229	first partial face transplant	264	Alan Greenspan
230	AMT	265	Mahmud (or Mahmood, Mahmoud) Ahmadinejad
231	USS Abraham Lincoln	266	Rafik Hariri, former Lebanese Prime Minister
232	Dulles Airport	267	FISA Court
233	comic strip Blondie	268	Israel evacuation of the Gaza Strip
234	Irving Berlin	269	Pakistan earthquakes of October 2005
235	Susan Butcher	270	The Mars rovers, Spirit and Opportunity
236	Boston Pops	271	Jon Bon Jovi
237	Cunard Cruise Lines	272	Barack Obama
238	2004 Baseball World Series	273	Rush Limbaugh
239	game show Jeopardy	274	Exxon Mobile Corp
240	Harry Potter and the Goblet of Fire	275	Dixie Chicks
241	Jasper Fforde	276	B-17 bomber
242	Guinness Brewery	277	Boeing 777 aircraft
243	2005 London terror bombing attacks	278	St. Peter's Basilica
244	Rubik's Cube Competitions	279	Australian wine
245	hybrid cars	280	Angkor Wat temples
246	Michael Brown	281	Joseph Steffen
247	Ella Fitzgerald	282	Orhan Pamuk
248	CSPI	283	Habitat for Humanity
249	Fulbright Program	284	CAFTA approval by U.S. Congress
250	publication of Danish cartoons of Mohammed	285	Yeti

Table 1: Targets of the 70 question series.

from one another, and to process an individual series in question order. That is, systems were allowed to use questions and answers from earlier questions in a series to answer later questions in the same series, but could not "look ahead" and use later questions to help answer earlier questions. Thus, question series can be viewed as an abstraction of an information-seeking dialogue between the user and the system; cf. (Kato et al., 2004). In total, 51 runs from 21 participants were submitted to the main task.

The evaluation of a single run can be decomposed into component evaluations for each of the question types and a final per-series score. Each of the three question types has its own response format and evaluation method. The individual component evaluations in 2007 were identical to those used in the TREC 2006 QA track, except that the official scores for Other questions were computed using multiple assessors' judgments of the importance of information nuggets, and assessors were not restricted in the criteria they could use in distinguishing between locally correct and globally correct answers for factoid and list questions. An aggregate score was computed for each series in a run using a simple average of the component scores of questions in that series, and the final score for the run was computed as the average of its per-series scores.

2.1 Factoid questions

The system response to a factoid question was either exactly one [*doc-id*, *answer-string*] pair or the literal string 'NIL'. Since there was no guarantee that a factoid question had an answer in the document collection, NIL was returned by the system when it believed there was no answer. Otherwise, *answer-string* was a string containing precisely an answer to the question, and *doc-id* was the id of a document in the collection that supported *answer-string* as an answer.

Each response was independently judged by two human assessors. When the two assessors disagreed in their judgments, a third adjudicator made the final determination. Each response was assigned exactly one of the following five judgments:

incorrect: the answer string does not contain a correct answer or the answer is not responsive;

not supported: the answer string contains a correct answer but the document returned does not support that answer;

not exact: the answer string contains a correct answer and the document supports that answer, but the string contains more than just the answer or is missing bits of the answer;

locally correct: the answer string consists of exactly a correct answer that is supported by the document returned, but the document collection contains a contradictory answer that the assessor believes is better;

globally correct: the answer string consists of exactly the correct answer, that answer is supported by the document returned, and the document collection does not contain a contradictory answer that the assessor believes is better.

To be responsive, an answer string was required to contain appropriate units and to refer to the correct "famous" entity (e.g., the Taj Mahal casino is not responsive if the question asks about "the Taj Mahal"). Questions also had to be interpreted in the time frame implied by the question series. For example, if the target was the event "France wins World Cup in soccer" and the question was *Who was the coach of the French team?* then the correct answer must be "Aime Jacquet", the name of the coach of the French team in 1998 when France won the World Cup, and not just the name of any past or current coach of the French

Run Tag	Submitter	Accuracy	NIL Prec	NIL Recall
LymbaPA07	Lymba Corporation	0.706	0.000	0.000
LCCFerret	Language Computer Corporation	0.494	0.000	0.000
lsv2007c	Saarland University	0.289	–	0.000
UofL	University of Lethbridge	0.258	0.052	0.500
QASCU1	Concordia University	0.256	0.000	0.000
FDUQAT16A	Fudan University	0.236	0.053	0.312
pronto07run3	Universita di Roma "La Sapienza"	0.222	0.000	0.000
ILQUA1	State University of New York (SUNY) at Albany	0.222	0.000	0.000
Ephyra3	Carnegie Mellon University and Universitaet Karlsruhe	0.208	0.048	0.062
QUANTA	Tsinghua University (State Key Lab)	0.206	0.091	0.062

Table 2: Evaluation scores for the factoid component. Scores are shown for the best run from the top 10 groups.

team. NIL responses were correct only if there was no known answer to the question in the collection. NIL was correct for 16 of the 360 factoid questions in the test set. For 26 questions, no system returned the correct answer, although those questions did have a correct answer found by the assessors.

It may be the case (especially with the inclusion of blogs) that different documents support contradictory answers as being correct. An exact answer-string that is supported in its associated document is assumed to be globally correct unless there is a *better, contradictory* answer supported elsewhere in the document collection. The assessor was allowed to use any number of criteria in determining that one answer was better than another, including recency of the supporting document, the amount of support provided by each supporting document, the number of distinct sources that support the answer as being correct, and the credibility or authoritativeness of the source. The assessor marked as globally correct one or more of the most credible of the known locally correct answers. "Global" correctness was defined with respect to the document collection, and not necessarily with respect to the real world.

The main evaluation metric for the factoid component was *accuracy*, the fraction of questions judged to be globally correct. Table 2 shows the most accurate run for the factoid component for each of the top 10 groups. Also reported are the recall and precision of recognizing when no answer exists in the document collection. NIL precision is the ratio of the number of times NIL was returned and correct to the number of times it was returned; NIL recall is the ratio of the number of times NIL was returned and correct to the number of times it was correct in the entire test set (16). If NIL was never returned, NIL precision is undefined and NIL recall is zero.

2.2 List questions

A list question asks for different instances of a particular type. The correct answer for a list question is the set of all such distinct instances in the document collection. A system's response to a list question consists of an unordered set of [*doc-id*, *answer-string*] pairs such that each *answer-string* represents a correct answer instance.

Each instance was evaluated in the same manner as the factoid questions, i.e., assigned one of the following judgments: incorrect, not supported, not exact, locally correct, and globally correct. Instances that were judged to be globally correct were then manually grouped into equivalence classes, where each

Run Tag	Submitter	F
LymbaPA07	Lymba Corporation	0.479
LCCFerret	Language Computer Corporation	0.324
ILQUA1	State University of New York (SUNY) at Albany	0.147
QASCU3	Concordia University	0.145
Ephyra3	Carnegie Mellon University and Universitaet Karlsruhe	0.144
UofL	University of Lethbridge	0.132
FDUQAT16B	Fudan University	0.131
IITDIBM2007T	Indian Institute Of Technology, Delhi	0.125
FDUQAT16A	Fudan University	0.107
pronto07run3	Universita di Roma "La Sapienza"	0.103

Table 3: Average F-scores for the list question component. Scores are shown for the best run from the top 10 groups.

equivalence class was considered a distinct answer. Thus, systems were not rewarded (and were in fact penalized) for returning equivalent answers multiple times.

The final set of known globally correct answers for a list question was compiled from the union of distinct globally correct answers across all runs plus additional distinct answers the assessor found during question development. For the 85 list questions in the test set, the median number of known distinct globally correct answers per question was 7, with a minimum of 2 and a maximum of 64. A system's response to a list question was scored using instance precision (IP) and instance recall (IR) based on the complete list of known distinct globally correct answers. Let S be the number of such answers, D be the number of distinct globally correct answers returned by the system, and N be the total number of instances returned by the system. Then $IP = D/N$ and $IR = D/S$. Precision and recall were then combined to produce an F-score with equal weight given to recall and precision:

$$F = \frac{2 \times IP \times IR}{IP + IR}$$

The score for the list component of a run was the average F-score over the 85 questions. Table 3 gives the average F-score of the run with the best list component score for each of the top 10 groups.

2.3 Other questions

The Other questions were evaluated using the methodology originally developed for the TREC 2003 definition questions. A system's response for an Other question consisted of an unordered set of [*doc-id*, *answer-string*] pairs. The answer strings were presumed to contain interesting "nuggets" about the series target that had not yet been covered by earlier questions in the series. The requirement to not repeat information already covered by earlier questions in the series made answering Other questions more difficult than answering TREC 2003 definition questions.

Judging the quality of the systems' responses was performed in two steps. In the first step, all of the answer strings from all of the systems were presented to an assessor in a single list. Using all the answer strings and searches performed during question development, the assessor created a list of information nuggets about the target. An information nugget in the context of an Other question is defined as an atomic

Run Tag	Submitter	F($\beta = 3$)
FDUQAT16B	Fudan University	0.329
lsv2007c	Saarland University	0.299
QASCU2	Concordia University	0.281
LymbaPA07	Lymba Corporation	0.281
LCCFerret	Language Computer Corporation	0.261
ILQUA1	State University of New York (SUNY) at Albany	0.242
csail3	Massachusetts Institute of Technology (MIT)	0.235
uams07main	University of Amsterdam	0.209
IITDIBM2007S	Indian Institute Of Technology, Delhi	0.209
QUANTA	Tsinghua University (State Key Lab)	0.194

Table 4: Average F-scores ($\beta = 3$) for the Other questions. Scores are shown for the best run from the top 10 groups.

piece of information about the target that is interesting (in the assessor's opinion) and is not part of an earlier question in the series or an answer to an earlier question in the series. An information nugget is considered atomic if the assessor could make a binary decision as to whether the nugget appears in a response. Once the nugget list was created for a target, the assessor decided which were vital, meaning that the information must be returned for a response to be good. Non-vital ("okay") nuggets acted as "don't care" conditions in that the assessor believed the information in the nugget to be interesting enough that returning the information was acceptable in, but not necessary for, a good response.

In the second step of the evaluation process, the assessor went through each system's output in turn and marked which nuggets appeared in the response. An answer string contained a nugget if there was a *conceptual* match between the answer string and the nugget; that is, the match was independent of the particular wording used in either the nugget or the system output. A nugget match was marked at most once per response—if the system output contained more than one match for a nugget, an arbitrary match was marked and the remainder were left unmarked. A single [*doc-id, answer-string*] pair in a system response could match 0, 1, or multiple nuggets.

To address some of the weaknesses of using vital/okay judgments from a single assessor (Hildebrandt et al., 2004), Lin and Demner-Fushman (2006) proposed an extension called "nugget pyramids", in which multiple assessors provide judgments of whether a nugget was vital or simply okay. Dang and Lin (2007) subsequently verified the efficacy of this method, and thus NIST adopted the pyramid extension for computing F-scores for Other responses. Nine different sets of vital/okay judgments were solicited from eight unique assessors (the primary assessor who originally created the nuggets later assigned vital/okay labels again). Each assessor was given all the questions for the series, as well as the nuggets created by the primary assessor. Using the pyramid procedure, a weight was assigned to each nugget based on the number of assessors who marked it as vital.

Given the nugget list and the set of nuggets matched in a system's response, nugget recall was computed as the ratio of the sum of weights of matched nuggets to the sum of weights of all nuggets in the list. Nugget precision was much more difficult to compute since there was no effective way of enumerating all the concepts contained in a particular answer string. Instead, a measure based on length (in non-whitespace characters) was used as an approximation to nugget precision. The length-based measure granted an allowance of 100 characters for each nugget matched. If the total system output was less than this number of

characters, the value of nugget precision was 1.0. Otherwise, the measure's value decreased as the length increased according to the following formula:

$$1 - \frac{length - allowance}{length}.$$

The final score for an Other question was an F-score, with nugget recall weighted more heavily than nugget precision:

$$F(\beta) = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}.$$

The score for the Other questions component was the average F-score (β=3) over the 70 Other questions. Table 4 gives the F-score for the best scoring Other question component for each of the top 10 groups.

2.4 Per-series Combined Scores

The three component scores measure a system's ability to process each type of question, but may not reflect the system's overall usefulness to a user. Since each individual series is an abstraction of a single user's interaction with the system, taking the individual series as the basic unit of evaluation should provide a more accurate representation of the effectiveness of the system from an individual user's perspective. Since each series is a mixture of different question types, we can compute a weighted average of the scores of the three question types on a per-series basis, and take the average of the per-series weighted scores as the final score for the run (Voorhees, 2005b). In 2007, the weighted score for an individual series was computed as:

$$\text{WeightedScore} = \frac{1}{3} \times \text{Factoid} + \frac{1}{3} \times \text{List} + \frac{1}{3} \times \text{Other}.$$

To compute the weighted score for an individual series, only the scores for questions belonging to that series were included in the computation. Since each of the component scores ranges between 0 and 1, the weighted score is also in that range. The final per-series score of each run is simply the average of individual per-series weighted scores.

We fit a two-way analysis of variance (ANOVA) model with the target type and the best run from each group as factors, and the per-series score as the dependent variable; we found significant differences between runs (p essentially equal to 0). To determine which runs were significantly different from each other, we

RunID	Submitter	Score					
LymbaPA07	Lymba Corporation	0.4839	A				
LCCFerret	Language Computer Corporation	0.3575		B			
FDUQAT16B	Fudan University	0.2310			C		
lsv2007c	Saarland University	0.2296			C		
QASCU1	Concordia University	0.2216			C	D	
ILQUA1	State University of New York (SUNY) at Albany	0.2023			C	D	E
Ephyra1	Carnegie Mellon University and Universitaet Karlsruhe	0.1804			C	D	E	F
IITDIBM2007T	Indian Institute Of Technology, Delhi	0.1735				D	E	F
QUANTA	Tsinghua University (State Key Lab)	0.1592					E	F
csail3	Massachusetts Institute of Technology (MIT)	0.1415						F

Table 5: Multiple comparison of the best run from the top 10 groups, based on ANOVA of per-series score.

performed a multiple comparison using Tukey's honestly significant difference criterion and controlling for the experiment-wise Type I error so that the probability of declaring a difference between two runs to be significant, when it is actually not, is at most 5%. Table 5 shows the results of the multiple comparison for the 10 groups with the highest final per-series score; runs sharing a common letter are not significantly different.

2.5 Discussion

Despite the inclusion of the blog corpus, which was expected to make the QA task more difficult, the best component scores in the main task were higher in 2007, after having generally declined each year since TREC 2004.

For each series, an attempt had been made during question development to include at least one question whose answer was found in the Blog06 corpus but not in the AQUAINT-2 corpus. This could be the answer to a factoid question, one of the items answering a list question, or (in rare cases) a nugget for the Other question. NIST assessors varied in their ability to locate blog-specific information that was suitable for the series. In some cases, the assessor could not find an answer in the AQUAINT-2 corpus during topic development, but the answer was later found in AQUAINT-2 during the assessment of system responses. In the end, only 15.0% (54/360) of the factoid questions had an answer that could be found only in the Blog06 corpus; 24.8% (235/946) of the distinct items answering a list question could be found only in the Blog06 corpus; and at most 6.1% (45/735) of the distinct nuggets answering an Other question could be found only in the Blog06 corpus.

The positive contribution of answers from blog documents to the various component scores was likely depressed due to the nature of the questions asked. Because factoid and list questions generally requested factual information, it is not surprising that most of their answers would come from newswire rather than blogs. In addition, assessors tend to place more credibility on newswire documents than blog posts, so if a blog answer contradicted a newswire answer, the newswire answer would be judged as the globally correct one, and the blog answer would at best be judged as locally correct; the effect would be more pronounced for factoid questions (which generally have only one globally correct answer) than for list questions (which allow multiple answers). Finally, assessors were most interested in factual information about their targets, and consequently found very little interesting Other information nuggets in the blog documents.

3 Complex Interactive QA (ciQA) Task

In TREC 2007, the goals of the complex, interactive question answering (ciQA) task remained unchanged from the previous year—to push the frontiers of question answering in two directions:

- A move away from "factoid" questions towards more complex information needs that exist within richer user contexts.
- A move away from the one-shot interaction model implicit in previous evaluations towards a model based on interactions with users.

The ciQA task in TREC 2007 represented the second iteration of the evaluation, which started in 2006. The TREC 2006 ciQA task (Dang et al., 2007), in turn, descended from the TREC 2005 HARD track, which focused on single-iteration clarification dialogues (Allan, 2006). However, there were substantial changes in the evaluation methodology: in TREC 2006, participants "encapsulated" their interactions in HTML forms

that were sent to NIST. This year, the task moved to completely "live" systems where assessors accessed individual QA systems, running at the participants' sites, over the Web.

3.1 Task Definition

3.1.1 Corpus

The ciQA task used the newswire portion of the corpus used by the main QA task (excluding the blog data). Participants were provided with the top 100 documents as retrieved by the PRISE system, using the question template verbatim as the query.

3.2 Complex "Relationship" Questions

The complex information needs explored by ciQA remained unchanged from last year; they represent extensions and refinements of so-called "relationship" questions piloted in TREC 2005 (Voorhees and Dang, 2006).

The concept of a "relationship" is defined as the ability of one entity to influence another, including both the means to influence and the motivation for doing so. Eight "spheres of influence" were noted in a previous pilot study funded by the AQUAINT research program: financial, movement of goods, family ties, communication pathways, organizational ties, co-location, common interests, and temporal. Evidence for both the existence or absence of ties is relevant. The particular relationships of interest naturally depend on the context.

A relationship question in the ciQA task, referred to as a topic, is composed of two parts. Consider an example:

> **Template:** What evidence is there for transport of [drugs] from [Mexico] to [the U.S.]?
> **Narrative:** The analyst would like to know of efforts to curtail the transport of drugs from Mexico to the U.S. Specifically, the analyst would like to know of the success of the efforts by local or international authorities.

The question template is a stylized information need that has a fixed structure and free slots whose instantiation varies across different topics. The narrative is free-form natural language text that elaborates on the information need, providing, for example, user context, a more articulated statement of interest, focus on particular topical aspects, etc.

The ciQA task employed the following templates, which were the same as those used in TREC 2006:

- What evidence is there for transport of [goods] from [entity] to [entity]?
- What [relationship] exist between [entity] and [entity]? (where [relationship] is an element of {"financial relationships", "organizational ties", "familial ties", "common interests"})
- What influence/effect do(es) [entity] have on/in [entity]?
- What is the position of [entity] with respect to [issue]?
- Is there evidence to support the involvement of [entity] in [event/entity]?

Thirty topics were developed for this year's task, but they were not distributed evenly across the five templates. In addition, one "throw-away" topic was included for training purposes.

Assessor	Topics
1	57, 69, 83
2	56, 63, 64, 74
3	65, 75, 76, 82
4	61, 68, 80, 85
5	58, 66, 70, 77
6	60, 72, 79, 84
7	62, 73, 81
8	59, 67, 71, 78

Table 6: Mapping between each NIST assessor and the topics they were responsible for.

3.2.1 Interaction Design

The purpose of the interactive aspect of ciQA was to provide a framework for participants to investigate interaction in the QA context. Unlike in TREC 2006, participants were able to deploy full-fledged Web-based QA systems with which the assessors engaged for five minutes per topic. There were no restrictions on the nature of the interaction or the system, except that it had to be accessible from a Web browser. Anything ranging from mixed-initiative dialogues to graphical interfaces was allowed.

To initiate interactions, assessors were directed to URLs provided by the participants. Assessors interacted with each system for five minutes per topic. The interaction length included time spent loading/rendering the page, as well as any delay caused by network traffic. It was the participant's responsibility to ensure that the QA system was Web-accessible during the period of time the assessors were scheduled to interact with submitted systems (a three-day period). If assessors were unable to access the participant's QA system, they skipped that interaction and did not return to it later.

The "throw-away" topic described earlier was used to familiarize assessors with systems before they completed actual test topics. Like other topics, the training period lasted five minutes, and could consist of anything that the participants wanted (e.g., a structured tutorial to introduce system features).

The interactions were completed at NIST using a Redhat Enterprise Linux 4 workstation with a 20-inch LCD monitor with 1600×1200 resolution and true color display (millions of colors), and a Firefox Web browser, v2.0.0.6. In addition, Flash, Acroread, and RealPlayer were enabled.

3.2.2 Experimental Protocol

The basic setup for the task was as follows: Participants first submitted initial runs and URL files to NIST. The URL files provided pointers to the participants' Web-based QA system (one for each topic). Included in the URL files were also pointers to screenshots of the interface, supplied by the participants for archival purposes. NIST assessors interacted with the Web-based QA systems during a three-day period. Results of those interactions were available immediately to participants, since they hosted their own systems. It was each participant's own responsibility to instrument their system to collect whatever data was necessary; NIST did not keep track of the interactions. Eight assessors participated in the task. Most assessors completed four topics; the mapping between assessors and topics is shown in Table 6.

Approximately two weeks following the interaction period, participants submitted final runs based on the results of the interactions to NIST. Assessors evaluated both initial and final runs.

Each participant was allowed to submit a maximum of 2 initial runs, 2 URL files, and 2 final runs. Manual runs were accepted, but had to be marked as such in the run submission interface. The interactive part of ciQA was optional; groups that did not wish to participate in the interactive aspect were asked to simply not submit URL files (however, every team engaged in the interactions). For each final run, participants were asked to supply the run tag of its corresponding initial run—this provided pairs of corresponding initial–final runs that isolated the effects of the interaction.

3.2.3 Evaluation Methodology

System responses were evaluated using the "nugget pyramid" extension of the nugget-based methodology used in previous TREC QA tasks (Lin and Demner-Fushman, 2006). Nine different sets of vital/okay judgments were solicited from eight unique assessors (the assessor who originally created the nuggets later assigned vital/okay labels again). Additional analyses included recall by length plots, as described in (Lin, 2007). A recall plot quantifies pyramid recall as a function of response length, which provides a rough model of how quickly a user can learn about the topic by reading system responses in sequential order. For more information on how this is computed, please refer to (Dang et al., 2007).

In addition to runs submitted by participants, we separately prepared a sentence retrieval baseline, similar to the one prepared last year. This provided a task-wide baseline to serve as a point of comparison. For each topic, the verbatim question template was used as a query to Lucene, which returned the top 20 documents. These documents were then tokenized into individual sentences. Sentences that contained at least one non-stopword from the question were retained and returned as the baseline run (up to a quota of 5,000 characters). Sentence order within each document and across the ranked list was preserved. The interaction associated with this run asked the assessor for relevance judgments on each of the sentences. Three options were given: "relevant", "not relevant", and "no opinion". The final run was prepared by simply removing those sentences judged not relevant—this had the effect of pulling more sentences from documents lower in the ranked list.

After assessors finished their interactions, they completed an online exit questionnaire which asked them to evaluate the various interactions. Assessors evaluated interactions according to several dimensions related to ease of use, usefulness, and effectiveness using 5-point scales. Assessors were also able to provide qualitative feedback about each interaction. Small screenshots of each system were displayed to remind assessors of each interaction. The order of these screenshots (and the order in which assessors evaluated each interaction) was random. A portion of the exit questionnaire, displaying the ciQA baseline interaction (described above), can be seen in Figure 2. At the end of the exit questionnaire, assessors were presented with four open-ended questions that asked them about their overall experiences. These questions were:

1. Of all interactions, which was your favorite and why?
2. What annoyed you about the interactions and why?
3. How different did you find the various interactions from one another and why?
4. Anything else?

3.3 Results

The ciQA task drew participation from seven groups. NIST received twelve initial runs and twelve final runs. A total of fourteen URL files were submitted. For the purposes of evaluation, the sentence retrieval baseline was treated like any other submission. In total, there were twelve initial–final pairs (and the sentence retrieval baseline).

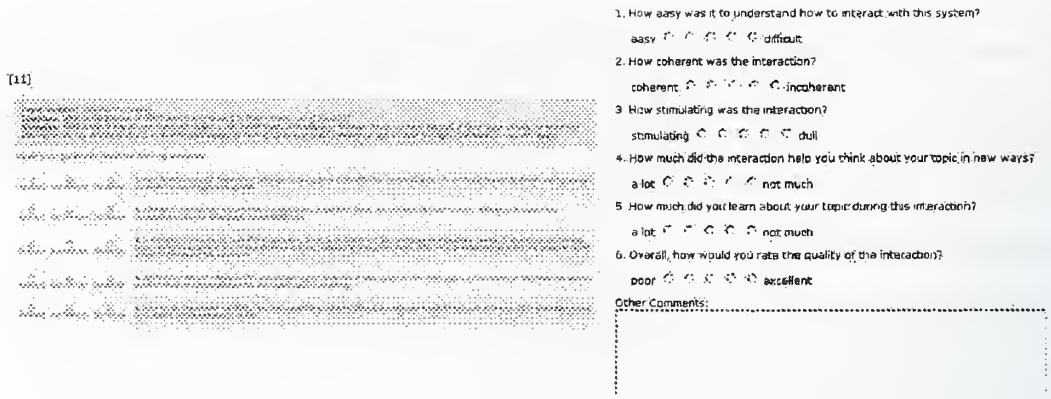


Figure 2: Portion of exit questionnaire for the baseline interaction. On the left the assessor sees a screenshot of the system (not meant to be readable, but simply as a reminder); questions are shown on the right.

3.3.1 System Effectiveness

The pyramid F-scores of the initial–final run pairs are shown in Table 7. By comparing the score of the corresponding runs, we can quantify the effect of the interaction on system performance. The scatter plot in Figure 3 presents a different view of the results—the initial score is plotted on the x axis, and the final score is plotted on the y axis. Points above the reference line $y = x$ represent cases where interaction improved performance.

We note two striking observations: First, unlike last year (Dang et al., 2007), most systems outperformed the baseline.[1] This is encouraging for the development of the field as a whole. Second, many interactions were detrimental, i.e., the pyramid F-score of the final run was higher than that of the initial run. Once again, this was different from last year, where interactions generally yielded small gains. We believe this effect to be caused by a combination of factors: problems with the task setup (more below); technical issues in deploying live Web-based QA systems; and the broadening of the design space that truly allows for effective and non-effective interactions.

3.3.2 Assessors Feedback about Interactions

The majority of interactions submitted by participants involved eliciting some type of relevance feedback from assessors. Items presented to assessors for feedback varied and included terms, sentences, articles from Wikipedia, and entire answer sets. A couple of systems asked assessors to interactively construct answers to their questions using sentences and documents. One interaction technique asked assessors to respond to open-ended questions modeled after a reference exchange, while another technique asked assessors to indicate their preferences for answer types. While most of the interactions went smoothly, at least two sites had network difficulties which impacted the interactions assessors had with their systems.

Figure 4 presents the mean quantitative ratings provided by subjects for three questions:

1. How easy was it to understand how to interact with this system?

[1]There were indexing issues with UNC's initial submission, which readily explains one of the two below-baseline performers. The other run, from the University of Maryland, experimented with *abstractive* techniques for question answering—i.e., the runs contained responses that were not found in any source document.

Organization	Type	Run tags		Pyramid F-Score	
		Initial	Final	Initial	Final
Michigan State U.	automatic	MSUciQAiHeu	MSUciQAfCol	0.359	0.361
Michigan State U.	automatic	MSUciQAiHeu	MSUciQAfInt	0.359	0.370
RMIT	automatic	rmitrun2	rmitrun5	0.361	0.343
RMIT	automatic	rmitrun2	rmitrun6	0.361	0.333
U. Mass	automatic	UMassBaseAut	UMassIntA	0.318	0.347
U. Mass	manual	UMassBaseAut	UMassIntM	0.318	0.503
U. Maryland	automatic	UMD07iMASCa	UMD07iMASCb	0.182	0.156
U. Maryland	automatic	UMD07MMRa	UMD07MMRb	0.333	0.334
U. NC and Yahoo!	automatic	UNCYABL30	UNCYAEX2	0.062	0.374
U. Strathclyde	manual	sicka	sicka2	0.410	0.394
U. Waterloo	manual	UWinitWIKI	UWfinalMAN	0.388	0.386
U. Waterloo	automatic	UWinitWIKI	UWfinalWIKI	0.388	0.380
baseline	automatic	baseA	baseB	0.327	0.327

Table 7: Performance of the twelve initial–final run pairs submitted to the TREC 2007 ciQA task. The sentence retrieval baseline is provided as a reference.

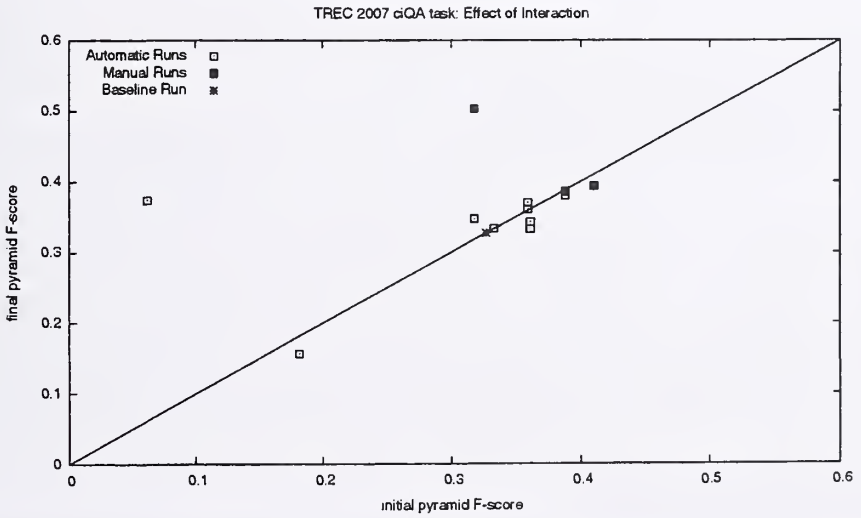


Figure 3: Scatter plot showing initial and final pyramid F-scores for each run pair submitted to the TREC 2007 ciQA task. Points above the line $y = x$ represent interactions that increased answer quality. Note that most systems outperformed the sentence retrieval baseline

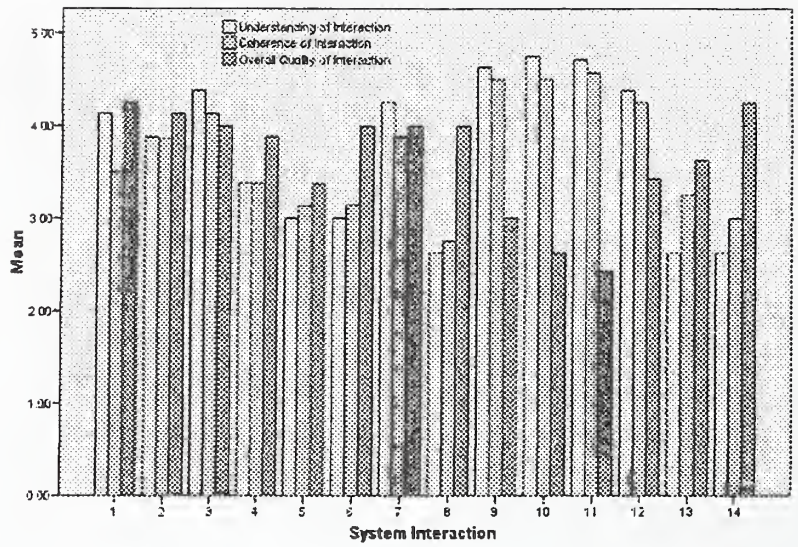


Figure 4: Mean assessors' ratings for each interaction along three dimensions: comprehensibility of interaction, coherence of interaction, and overall quality.

2. How coherent was this interaction?

3. Overall, how would you rate the quality of the interaction?

In all cases, higher scores are more positive. It is important to note that this is a small, unrepresentative, and unusual sample, so these results should be viewed cautiously. These results are by no means definitive and/or generalizable beyond this evaluation context.

Interactions that were rated the lowest with respect to understanding (interactions 8, 13 and 14) and coherence (interactions 8 and 14) had information-dense interfaces and often required multiple steps (one of these interactions was answer construction). Interactions rated most positively for these two attributes were traditional relevance feedback interfaces. Interestingly, understanding and coherence were positively correlated with one another ($r = 0.949$, $p < 0.01$), but were negatively correlated with assessors' overall quality ratings ($r = -0.533$, $p < 0.05$ and $r = -0.674, p < 0.01$, respectively). The interaction that received the lowest quality assessment scored fairly high on understanding and coherence. This interaction was the ciQA baseline interaction, which elicited sentence-level relevance feedback. The interaction that received some of the lowest assessments for understanding and coherence received one of the highest overall quality scores (interaction 14). This interaction consisted of building answers and may have received a higher quality score because of its novelty. It also engaged assessors in the most interaction which may be why scores on these three measures differ.

The qualitative feedback from the final set of questions asking assessors about their entire experiences showed that assessors preferred the traditional relevance feedback interactions, felt considerable time pressure, and did not like the complicated interactions. One assessor indicated a preference for one of the answer construction interactions, while another did not like this interaction. At least two assessors were puzzled about the use of Wikipedia and were displeased with this interaction.

Data from the exit questionnaire should be viewed cautiously for several reasons. Some interactions

were less than perfect because of network problems, so assessors' evaluations, in part, reflect this. Assessors' comments indicated that they felt huge time pressures, which may by why such an overwhelming preference was indicated for simple, easily understood and executed interactions such as those that employed relevance feedback.

One of the most interesting results of this evaluation was that it revealed several limitations of this style of evaluation in the context of TREC. Many of the limitations stem from the fact that assessors already know a great deal about their topics before they engage in interactions. The approach in TREC has traditionally been to have the same person develop the topic and assess its answers, since the assessor is supposed to act as a surrogate user with his/her own particular information needs. However, in developing the topic for ciQA, this "user" researches the topic (to make sure that it is a suitable topic for the particular document collection) and consequently knows more about the topic than a naive user issuing the query.[2] NIST assessors are unusual "users" and it is unrealistic to expect them to assume dual roles as assessors (during topic development and answer evaluation) and naive users (during the interactions).

Helping users learn more about their topics and helping systems learn more about users are central goals of interactive systems. The exit questionnaire reveals that interactive techniques for addressing these goals cannot be evaluated using the ciQA experimental framework. Additionally, not all ciQA participants understood that assessors already knew the answers to the questions they were asking so there may also have been a mismatch between participants' and assessors' expectations of the interactions.

4 Future of the QA Track

TREC 2007 revealed limitations in the ciQA design for evaluating interactive systems. These limitations could not be reconciled within the NIST evaluation framework, and hence it was decided not to attempt another interactive QA task in 2008.

The primary goal of the TREC 2007 main task (and what distinguished it from previous TREC QA tasks) was the introduction of blog text to encourage research in NLP techniques that would handle ill-formed language and discourse structures that are more informal and less reliable than newswire. Questions were asked over a combined newswire (AQUAINT-2) and blog (Blog06) corpus, rather than only a blog corpus, in order to ease participants' transition from newswire. However, because most of the TREC 2007 questions requested factual information, they did not specifically test systems' ability to process blog text, as answers still came predominantly from the AQUAINT-2 corpus.

This mismatch between the corpus and the information need expressed in the questions naturally suggests that in order to move away from traditional newswire towards blogs, the QA task should be changed so that the questions are more targeted towards characteristics that are particular to blogs. Because blogs naturally contain a large amount of opinions, it was decided that the QA task for 2008 should focus on questions that ask about people's *opinions*. Questions would still be grouped into series focused by a particular target (person, organization, etc.), but there would be no factoid questions.[3] Rather, each series would comprise *rigid* list questions (e.g., "What people have good opinions of Sean Hannity?") which would be evaluated in the same manner as TREC 2007 list questions, and *squishy* list questions (e.g., "What reasons do people give for liking Sean Hannity?") which would be evaluated with the nugget pyramid method used for TREC 2007 Other questions.

[2]Results of questions 3, 4, and 5 from the exit questionnaire, which asked assessors to indicate how much they learned about their topics through the interaction (see Figure 2 for specific questions) are not presented because some assessors indicated that these values were low because they already knew about their topics.

[3]It was pointed out that asking factoid type questions about opinions seemed inappropriate, and after nine years of factoid questions (starting in TREC 1999), it was time to retire factoids from the QA track in any case.

References

James Allan. 2006. HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Hoa Dang and Jimmy Lin. 2007. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 768–775, Prague, Czech Republic.

Hoa Trang Dang, Jimmy Lin, and Diane Kelly. 2007. Overview of the TREC 2006 Question Answering Track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, Maryland.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*, pages 49–56.

Tsuneaki Kato, Jun'ichi Fukumoto, Fumito Masui, and Noriko Kando. 2004. Handling information access dialogue through QA technologies—A novel challenge for open-domain question answering. In *Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics of Question Answering*, pages 70–77, May.

Jimmy Lin and Dina Demner-Fushman. 2006. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 383–390, New York, New York.

Jimmy Lin. 2007. Is question answering better than information retrieval? A task-based evaluation framework for question series. In *Proceedings of the 2007 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2007)*, pages 212–219, Rochester, New York.

Craig Macdonald and Iadh Ounis. 2006. The TREC blog06 collection: Creating and analysing a blog test collection. Technical Report DCS Technical Report TR-2006-224, Department of Computing Science, University of Glasgow.

Ellen M. Voorhees and Hoa T. Dang. 2006. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Ellen M. Voorhees. 2004. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68.

Ellen M. Voorhees. 2005a. Overview of the TREC 2004 Question Answering track. In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2004)*, pages 52–62.

Ellen M. Voorhees. 2005b. Using question series to evaluate question answering system effectiveness. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 299–306.

TREC 2007 Spam Track Overview

Gordon V. Cormack
University of Waterloo
Waterloo, Ontario, Canada

1 Introduction

TREC's *Spam Track* uses a standard testing framework that presents a set of chronologically ordered email messages a spam filter for classification. In the filtering task, the messages are presented one at at time to the filter, which yields a binary judgment (*spam* or *ham* [i.e. non-spam]) which is compared to a human-adjudicated *gold standard*. The filter also yields a *spamminess* score, intended to reflect the likelihood that the classified message is spam, which is the subject of post-hoc ROC (Receiver Operating Characteristic) analysis. Four different forms of user feedback are modeled: with *immediate feedback* the gold standard for each message is communicated to the filter immediately following classification; with *delayed feedback* the gold standard is communicated to the filter sometime later (or potentially never), so as to model a user reading email from time to time and perhaps not diligently reporting the filter's errors; with *partial feedback* the gold standard for only a subset of email recipients is transmitted to the filter, so as to model the case of some users never reporting filter errors; with *active on-line learning* (suggested by D. Sculley from Tufts University [11]) the filter is allowed to request immediate feedback for a certain quota of messages which is considerably smaller than the total number. Two test corpora – email messages plus gold standard judgments – were used to evaluate subject filters. One *public* corpus (trec07p) was distributed to participants, who ran their filters on the corpora using a track-supplied toolkit implementing the framework and the four kinds of feedback. One private *corpus* (MrX 3) was not distributed to participants; rather, participants submitted filter implementations that were run, using the toolkit, on the private data. Twelve groups participated in the track, each submitting up to four filters for evaluation in each of the four feedback modes (immediate; delayed; partial; active).

Task guidelines and tools may be found on the web at http://plg.uwaterloo.ca/~gvcormac/spam/ .

1.1 Filtering – Immediate Feedback

The immediate feedback filtering task is identical to the TREC 2005 and TREC 2006 (immediate) tasks [3, 5]. A chronological sequence of messages is presented to the filter using a standard interface. The the filter classifies each message in turn as either *spam* or *ham,* also computes a *spamminess score* indicating its confidence that the message is spam. The test setup simulates an ideal user who communicates the correct (gold standard) classification to the filter for each message immediately after the filter classifies it.

Participants were supplied with tools, sample filters, and sample corpora (including the TREC 2005 and TREC 2006 public corpora) for training and development. Filters were evaluated on the two new corpora developed for TREC 2007.

1.2 Filtering – Delayed Feedback

Real user's don't immediately report the correct classification to filters. They read their email, typically in batches, some time after it is classified. Last year (TREC 2006) the delayed learning task sought to simulate user behavior by withholding feedback for some random number of messages after which feedback was given; this delay followed by feedback was repeated in several cycles. This year (TREC 2007) the track seeks instead to measure the effect of delay. To this end, immediate feedback is given for the first several thousand messages (10,000 for trec07p; 20,000 for MrX 3) after which no feedback at all is given. Thus, the majority of the corpus is classified with no feedback and the cumulative effect of delay may be evaluated by examining the learning curve.

Participants trained on the TREC 2006 corpus. While the 2007 guidelines specified that feedback might never be given, they did not specify the exact nature of the task. It was anticipated that the delayed feedback task would be more difficult for the filters, and that filter performance would degrade during the interval for which no feedback was given. It was anticipated that participants might be able to harness information from unlabeled messages (the ones for which feedback was not given) to improve performance.

1.3 Partial Feedback

Partial feedback is a variant on delayed feedback effected with exactly the same tools. As for "delayed feedback" the feedback was in fact either given immediately or not at all. In this case, however, the messages for which feedback was given were those sent to a subset of the recipients in the corpus; that is, the filter was trained on some users' messages but asked to classify every users' messages. Partial feedback was used only for the trec07p corpus, as it contained email addressed to many recipients. It was not applicable to MrX 3, being a single-user corpus.

1.4 The On-line Active Learning Task

For the on-line task, filters were passed an additional parameter – the quota of messages for which feedback could be requested – and were expected to return an additional result – to request or decline feedback for each message classified. Filters that were unaware of these parameters were assumed to request feedback for each message classified until the quota was exhausted; thus the default behavior was identical to the delayed feedback task. However, filters were able to decline feedback for some messages (presumably those whose classification the filter found unimportant) in order to preserve quota so as to be able to request feedback for later messages.

A naive solution to this problem would be to have the filter make a label request for every message. This would request labels and train normally for the first N messages, where N is the initial quota, and then would not update for the remainder of the run. The testing jig is backward compatible with filters from prior years by making the naive approach the default method if no label request is specified. This allows prior filters to run on this task without modification.

2 Evaluation Measures

We used the same evaluation measures developed for TREC 2005. The tables and figures in this overview report Receiver Operating Characteristic (ROC) Curves, as well as $1 - ROCA(\%)$ – the area above the ROC curve, indicating the probability that a random spam message will receive a lower spamminess score than a random ham message.

The appendix contains detailed summary reports for each participant run, including ROC curves, 1-ROCA%, and the following statistics. The *ham misclassification percentage* (*hm%*) is the fraction of all ham classified as spam; the *spam misclassification percentage* (*sm%*) is the fraction of all spam classified as ham.

There is a natural tension between ham and spam misclassification percentages. A filter may improve one at the expense of the other. Most filters, either internally or externally, compute a spamminess score that reflects the filter's estimate of the likelihood that a message is spam. This score is compared against some fixed threshold t to determine the ham/spam classification. Increasing t reduces *hm%* while increasing *sm%* and vice versa. Given the score for each message, it is possible to compute *sm%* as a function of *hm%* (that is, *sm%* when t is adjusted to as to achieve a specific *hm%*) or vice versa. The graphical representation of this function is a Receiver Operating Characteristic (ROC) curve; alternatively a recall-fallout curve. The area under the ROC curve is a cumulative measure of the effectiveness of the filter over all possible values. ROC area also has a probabilistic interpretation: the probability that a random ham will receive a lower score than a random spam. For consistency with *hm%* and *sm%*, which measure failure rather than effectiveness, spam track reports the area *above* the ROC curve, as a percentage ($(1 - ROCA)\%$). The appendix further reports *sm%* when the threshold is adjusted to achieve several specific levels of *hm%,* and vice versa.

A single quality measure, based only on the filter's binary ham/spam classifications, is nonetheless desirable. To this end, the appendix reports *logistic average misclassification percentage* (*lam%*) defined as $lam\% = logit^{-1}(\frac{logit(hm\%)+logit(sm\%)}{2})$ where $logit(x) = log(\frac{x}{100\%-x})$. That is, *lam%* is the geometric mean of the

odds of ham and spam misclassification, converted back to a proportion[1]. This measure imposes no a priori relative importance on ham or spam misclassification, and rewards equally a fixed-factor improvement in the odds of either.

For each measure and each corpus, the appendix reports 95% confidence limits computed using a bootstrap method [4] under the assumption that the test corpus was randomly selected from some source population with the same characteristics.

3 Spam Filter Evaluation Tool Kit

All filter evaluations were performed using the *TREC Spam Filter Evaluation Toolkit*, developed for this purpose. The toolkit is free software and is readily portable.

Participants were required to provide filter implementations for Linux or Windows implementing five command-line operations mandated by the toolkit:

- **initialize** – creates any files or servers necessary for the operation of the filter
- **classify** *message [quota]* – returns ham/spam classification and spamminess score for *message*. *[quota]* is used only in active learning feedback.
- **train ham** *message* – informs filter of correct (ham) classification for previously classified *message*
- **train spam** *message* – informs filter of correct (spam) classification for previously classified *message*
- **finalize** – removes any files or servers created by the filter.

Track guidelines prohibited filters from using network resources, and constrained temporary disk storage (1 GB), RAM (1 GB), and run-time (2 sec/message, amortized). These limits were enforced incrementally, so that individual long-running filters were granted more than 2 seconds provided the overall average time was less than 2 second per query plus one minute to facilitate start-up.

The toolkit takes as input a test corpus consisting of a set of email messages, one per file, and an index file indicating the chronological sequence and gold-standard judgments for the messages. It calls on the filter to classify each message in turn, records the result, and at some time later (perhaps immediately, perhaps never, and perhaps only on request of the filter) communicates the gold standard judgment to the filter.

The recorded results are post-processed by an evaluation component supplied with the toolkit. This component computes statistics, confidence intervals, and graphs summarizing the filter's performance.

4 Test Corpora

	Ham	Spam	Total
trec07p	25220	50199	75419
MrX3	8082	153893	161975
Total	33302	204092	237394

Table 1: Corpus Statistics

For TREC 2007, we used one public corpus and one private corpus with a total of 237,394 messages (see table 1).

4.1 Public Corpus – trec07p

The public corpus contains all the messages delivered to a particular server from April 8 through July 6, 2007. The server contains many accounts that have fallen into disuse but continue to receive a lot of spam. To these accounts were added a number of "honeypot" accounts published on the web and used to sign up for

[1] For small values, odds and proportion are essentially equal. Therefore *lam%* shares much with the geometric mean average precision used in the robust track.

a number of services – some legitimate and some not. Several services were canceled and several "opt-out" links from spam messages were clicked. All messages were adjudicated using the methodology developed for previous spam tracks. [6] This corpus is the first TREC public corpus that contains exclusively ham and spam sent to the same server within the same time period. The messages were unaltered except for a few systematic substitutions of names.

4.2 Private Corpus – MrX3

The MrX3 corpus was derived from the same source as the MrX and MrX2 corpora used for TREC 2006 and TREC 2006 respectively. All of X's email from December 2006 through July 11, 2007 was used. The proportion of spam has grown substantially since 2005[2]; Ham volume was insubstantially different.

5 Spam Track Participation

Group	Filter Prefix
Beijing University of Posts and Telecommunications	kid
Fudan University-WIM Lab	fdw
Heilongjiang Institute of Technology	hit
Indiana University	iub
International Institute of Information Technology	III
Jozef Stefan Institute	ijs
Mitsubishi Electric Research Labs	crm
National University of Defense Technology	ndt
Shanghai Jiao Tong University	sjt
South China University of Technology	scu
Tufts University	tft
University of Waterloo	wat

Table 2: Participant filters

Corpus / Task	Filter Suffix
trec07p / immediate feedback	pf
trec07p / delayed feedback	pd
trec07p / partial feedback	pp
trec07p / active feedback	p1000
MrX3 / immediate feedback	x3f
MrX3 /delayed feedback	x3d

Table 3: Run-id suffixes

Twelve groups participated in the TREC 2007 spam track. Each participant submitted up to four filter implementations for evaluation on the private corpora; in addition, each participant ran the same filters on the public corpora, which were made available following filter submission. All test runs are labeled with an identifier whose prefix indicates the group and filter priority and whose suffix indicates the corpus to which the filter is applied. Table 2 shows the identifier prefix for each submitted filter. All test runs have a suffix indicating the corpus and task, detailed in figure 3 .

6 Results

Figures 2 through 6 show the results of the best seven systems for each type of feedback with respect to each corpus. The left panel of each figure shows the ROC curve, while the right panel shows the learning curve: cumulative 1-ROCA% as a function of the number of messages processed. Only the best run for each

[2]Note that the MrX and MrX3 corpora include all email delivered during a particular time period, MrX2 was sampled so as to yield the same ham:spam ratio as MrX.

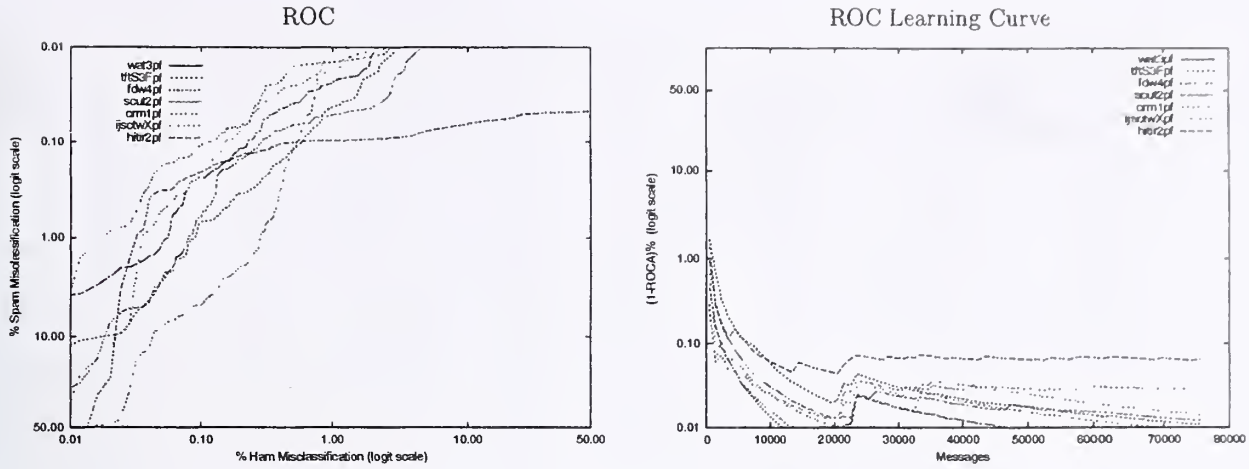


Figure 1: trec07p Public Corpus – Immediate Feedback

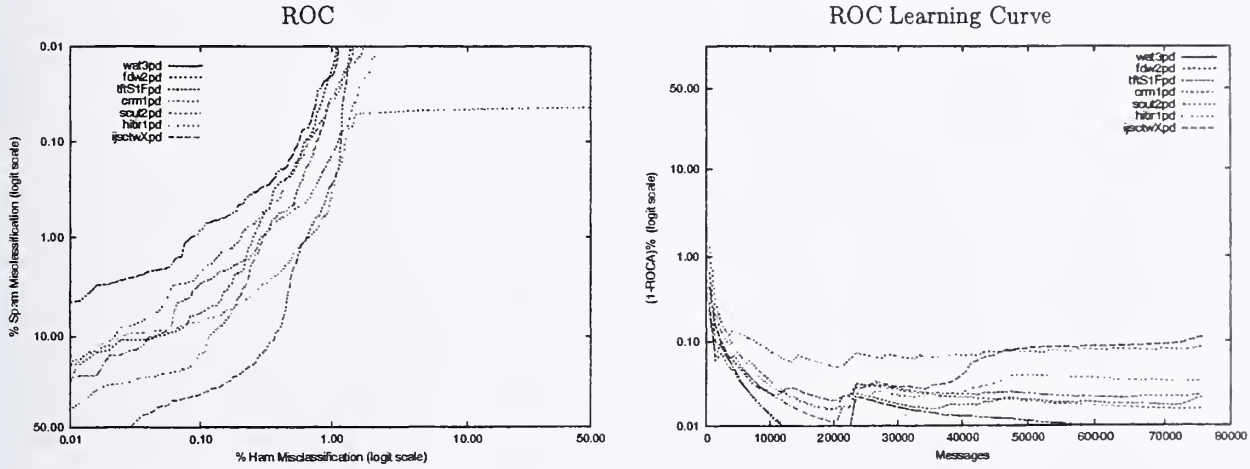


Figure 2: trec07p Public Corpus – Delayed Feedback

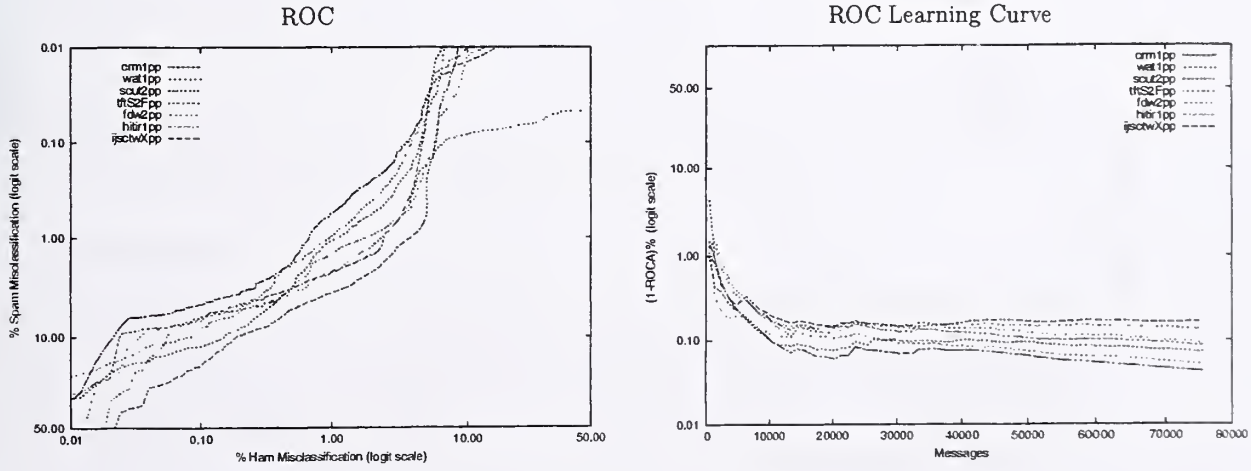


Figure 3: trec07p Public Corpus – Partial Feedback

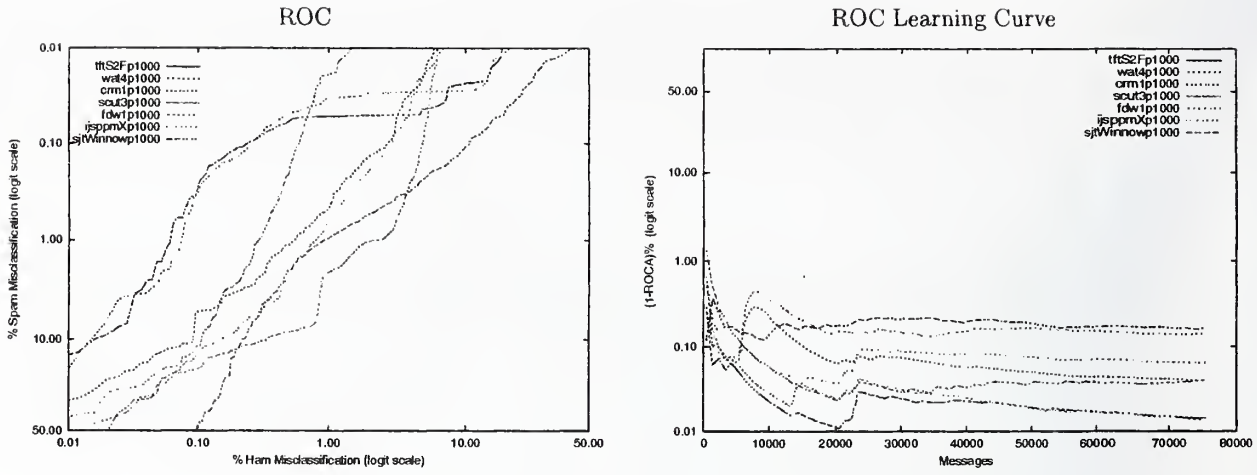


Figure 4: trec07p Public Corpus – Active Learning (quota 1000)

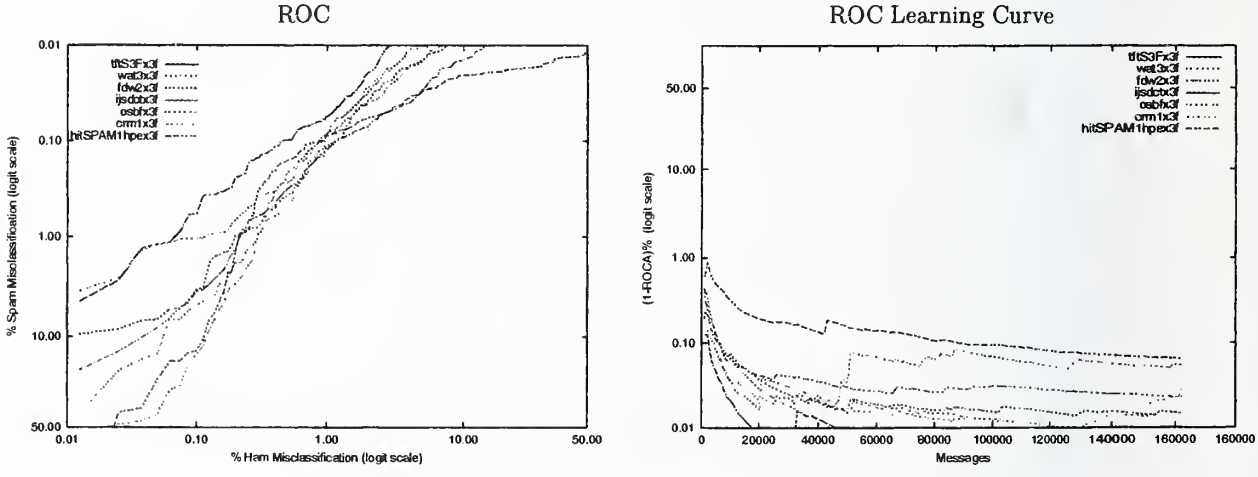


Figure 5: MrX3 Corpus – Immediate Feedback

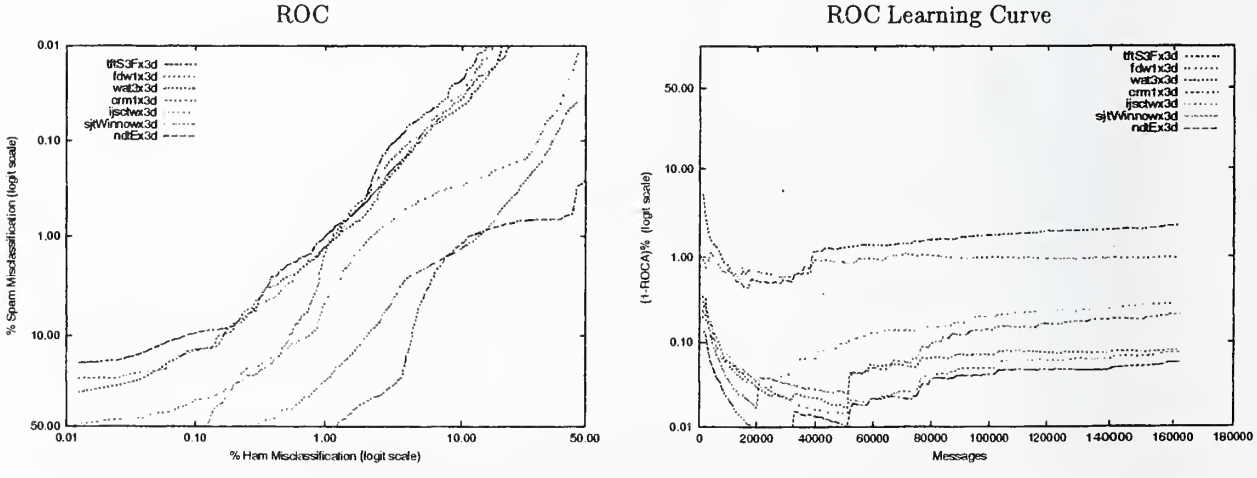


Figure 6: MrX3 Corpus – Delayed Feedback

	Immediate feed.		Delayed feed.		Partial feed.		Active learning	
Rank	run tag	1-ROCA(%)	run tag	1-ROCA(%)	run tag	1-ROCA(%)	run tag	1-ROCA(%)
1	wat3pf	0.0055	wat3pd	0.0086	crm1pp	0.0425	tftS2Fp1000	0.0144
2	wat1pf	0.0057	wat1pd	0.0105	wat1pp	0.0514	wat4p1000	0.0145
3	wat4pf	0.0057	wat4pd	0.0105	wat4pp	0.0514	crm1p1000	0.0401
4	wat2pf	0.0077	fdw2pd	0.0159	wat3pp	0.0516	scut3p1000	0.0406
5	tftS3Fpf	0.0093	wat2pd	0.0207	scut2pp	0.0719	tftS1Fp1000	0.0413
6	tftS1Fpf	0.0099	tftS1Fpd	0.0214	tftS2Fpp	0.0858	tftS3Fp1000	0.0475
7	tftS2Fpf	0.0103	fdw1pd	0.0223	tftS1Fpp	0.0878	scut2p1000	0.0533
8	fdw4pf	0.0109	tftS2Fpd	0.0225	tftS3Fpp	0.0919	fdw1p1000	0.0641
9	scut2pf	0.0121	tftS3Fpd	0.0226	fdw2pp	0.0921	fdw2p1000	0.0881
10	crm1pf	0.0142	crm1pd	0.0229	fdw1pp	0.1066	wat1p1000	0.1193
11	fdw3pf	0.0157	fdw3pd	0.0229	wat2pp	0.1087	wat2p1000	0.1193
12	fdw2pf	0.0195	fdw4pd	0.0229	fdw3pp	0.1109	wat3p1000	0.1215
13	fdw1pf	0.0198	scut2pd	0.0342	fdw4pp	0.1151	ijsppmXp1000	0.1417
14	ijsctwXpf	0.0297	scut3pd	0.0516	hitir1pp	0.1351	ijsctwXp1000	0.1473
15	ijsppmXpf	0.0299	hitir1pd	0.0855	hitir2pp	0.1356	sjtWinnowp1000	0.1626
16	scut1pf	0.0348	hitir2pd	0.0876	scut1pp	0.1534	fdw3p1000	0.1629
17	ijsdcwXpf	0.0371	ijsctwXpd	0.1111	ijsctwXpp	0.1656	scut1p1000	0.1939
18	ijsdctXpf	0.0382	ijsppmXpd	0.1148	ijsppmXpp	0.1724	fdw4p1000	0.2029
19	scut3pf	0.0406	sjtWinnowpd	0.2813	crm4pp	0.1866	hitir2p1000	0.2800
20	crm4pf	0.0457	crm2pd	0.3186	scut3pp	0.1898	crm2p1000	0.3244
21	hitir2pf	0.0644	scut1pd	0.3251	ijsdctXpp	0.1962	hitir1p1000	0.3246
22	hitir1pf	0.0652	crm4pd	0.3354	ijsdcwXpp	0.2477	ndtAp1000	0.7507
23	sjtMulti1pf	0.0709	sjtMulti1pd	0.4250	crm2pp	0.3882	ndtBp1000	1.3037
24	sjtMulti2pf	0.0732	ndtApd	0.4359	sjtMulti1pp	0.4250	sjtMulti1p1000	1.3102
25	IIITHpf	0.1041	ndtBpd	0.5842	sjtMulti2pp	0.4830	ndtCp1000	1.3932
26	crm2pf	0.1289	ndtCpd	0.6547	crm3pp	0.6743	kidult2p1000	1.5239
27	ndtApf	0.1662	crm3pd	0.8844	sjtBayespp	0.6910	kidult3p1000	1.5895
28	ndtBpf	0.1931	kidult3pd	0.9006	ndtApp	0.7910	kidult1p1000	1.6267
29	ndtCpf	0.2164	kidult0pd	1.1703	ndtBpp	0.9366	kidult0p1000	1.9030
30	sjtWinnowpf	0.2209	kidult2pd	1.4355	sjtWinnowpp	1.0133	ndtDp1000	2.3704
31	crm3pf	0.2364	kidult1pd	1.4959	ndtCpp	1.0191	sjtMulti2p1000	2.6864
32	sjtBayespf	0.3155	iube5c5pd	1.5241	kidult3pp	3.1509	sjtBayesp1000	4.0136
33	kidult0pf	0.3599	iube2c3pd	1.5911	kidult1pp	3.1711	iube2c3p1000	10.3933
34	kidult3pf	0.4515	iube2c6pd	1.9411	kidult2pp	3.1940	iube5c5p1000	10.3933
35	kidult2pf	0.4532	ndtDpd	1.9486	kidult0pp	3.5517	iube2c6p1000	12.5153
36	kidult1pf	0.4579	sjtMulti2pd	17.2297	iube5c5pp	4.0446	crm4p1000	50.3043

Table 4: Summary 1-ROCA (%) – trec07p Public Corpus

participant is shown in the figures; tables 4 and 5 show 1-ROCA% for all feedback regimens on trec07p and MrX3 respectively. Full details for all runs are given in the notebook appendix.

7 Conclusions

Once again, the general performance of filters has improved over previous techniques. Support vector machines [12, 9] and logistic regression [7], specifically engineered for spam filtering, show exceptionally strong performance. Delayed and partial feedback degrade filter performance; at the time of writing we are unaware of any special methods used by participants mitigate this degradation [10]. The learning curves do not show substantial de-learning as delay increases.

The best-performing techniques for active learning use techniques akin to "uncertainty scheduling" [12] in which feedback is requested only for those messages whose score is near the filter's threshold.

	Immediate feed.		Delayed feed.	
Rank	run tag	1-ROCA(%)	run tag	1-ROCA(%)
1	tftS3Fx3f	0.0042	tftS3Fx3d	0.0568
2	tftS2Fx3f	0.0054	tftS2Fx3d	0.0683
3	wat3x3f	0.0076	tftS1Fx3d	0.0685
4	wat1x3f	0.0096	fdw1x3d	0.0747
5	wat4x3f	0.0096	fdw2x3d	0.0751
6	fdw2x3f	0.0147	wat3x3d	0.0787
7	fdw3x3f	0.0154	wat1x3d	0.0896
8	fdw1x3f	0.0155	fdw3x3d	0.1062
9	tftS1Fx3f	0.0166	fdw4x3d	0.1258
10	wat2x3f	0.0219	crm1x3d	0.2079
11	ijsdctx3f	0.0229	wat2x3d	0.2512
12	fdw4x3f	0.0255	ijsctwx3d	0.2830
13	ijsdcwx3f	0.0281	ijsppmx3d	0.3055
14	osbfx3f	0.0281	crm2x3d	0.3811
15	ijsctwx3f	0.0392	ijsdcwx3d	0.5036
16	ijsppmx3f	0.0397	ijsdctx3d	0.5288
17	crm1x3f	0.0543	crm4x3d	0.7589
18	hitSPAM1hpex3f	0.0650	sjtWinnowx3d	0.9674
19	hitSPAM2chix3f	0.1032	ndtEx3d	2.2840
20	crm4x3f	0.1145	crm3x3d	2.5169
21	crm2x3f	0.1296	kid0x3d	2.5383
22	sjtWinnowx3f	0.1666	ndtDx3d	4.6920
23	sjtMulti1x3f	0.3413	sjtMulti1x3d	5.0656
24	crm3x3f	0.9476	ndtAx3d	5.3401
25	IIITx3f	1.0234	sjtBayesx3d	28.7693
26	kidult0x3f	1.0313	IIITx3d	49.9682
27	ndtDx3f	1.3985	-	-
28	sjtBayesx3f	2.0811	-	-
29	ndtAx3f	2.4078	-	-
30	scut2x3f	4.7596	-	-
31	iube5c6x3f	19.0336	-	-
32	hitSPAM3bayx3f	49.9682	-	-

Table 5: Summary 1-ROCA (%) – MrX3 Private Corpus

8 Epilogue

In each of the three years that TREC has hosted the spam track, new techniques have dominated the previous state of the art. In TREC 2005, sequential compression models showed outstanding performance [2] – much better than that achieved by commonly deployed "Bayesian" filters. In TREC 2006, OSBF-Lua achieved dominance through Orthogonal Sparse Bigrams and iterative training [1]. This year, SVM and logistic regression methods – based on character features – were for the first time shown to be superior for spam.

CEAS 2008, the Conference on Email and Anti-Spam (www.ceas.cc) will host a laboratory evaluation modeled after the spam track. In addition, CEAS will run the Live Challenge – a real-time version of the task using a live email feed rather than an archival corpus. Other evaluation efforts – and their results – are compared and contrasted with the spam track in a recent survey [8].

9 Acknowledgments

The author thanks D. Sculley for suggesting the active feedback task and making the necessary modifications to the spam filter evaluation toolkit.

References

[1] Assis, F. OSBF-Lua - A Text Classification Module for Lua The Importance of the Training Method. In *Fifteenth Text REtrieval Conference (TREC-2006)* (Gaithersburg, MD, 2006), NIST.

[2] Bratko, A., Cormack, G. V., Filipič, B., Lynam, T. R., and Zupan, B. Spam filtering using statistical data compression models. *Journal of Machine Learning Research 7*, Dec (2006), 2673–2698.

[3] Cormack, G. Trec 2005 spam track overview. In *Proceedings of TREC 2005* (Gaithersburg, MD, 2005).

[4] Cormack, G. Statistical precision of information retrieval evaluation. In *Proceedings of SIGIR 2006* (Seattle, WA, 2006).

[5] Cormack, G. Trec 2006 spam track overview. In *Proceedings of TREC 2006* (Gaithersburg, MD, 2006).

[6] Cormack, G., and Lynam, T. Spam Corpus Creation for TREC. In *Proceedings of the Second Conference on Email and Anti-Spam (CEAS-2005)* (Mountain View, CA, 2005).

[7] Cormack, G. V. Waterloo participation in the TREC 2007 spam track. In *Proceedings of TREC 2007* (Gaithersburg, MD, 2007).

[8] Cormack, G. V. *Email Spam Filtering: A systematic review*, vol. 1. 2008.

[9] Kato, M., Langeway, J., Wu, Y., and Yerazunis, W. S. Three non-Bayesian methods of spam filtration: CRM114 at TREC 2007. In *Proceedings of TREC 2007* (Gaithersburg, MD, 2007).

[10] Mojdeh, M., and Cormack, G. V. Semi-supervised spam filtering: does it work? In *SIGIR 2008* (Singapore, 2008).

[11] Sculley, D. Online active learning methods for fast label-efficient spam filtering. In *CEAS 2007: Fourth Conference on Email and AntiSpam* (August 2007).

[12] Sculley, D., and Wachman, G. M. Relaxed online SVMs in the TREC spam filtering track. In *Proceedings of TREC 2007* (Gaithersburg, MD, 2007).


Text REtrieval Conference (TREC)
...to encourage research in information retrieval
from large text collections
SP 500-274
NIST
Trec 2007
Conference
Proceedings

NIST Technical Publications

Periodical

Journal of Research of the National Institute of Standards and Technology—Reports NIST research and development in metrology and related fields of physical science, engineering, applied mathematics, statistics, biotechnology, and information technology. Papers cover a broad range of subjects, with major emphasis on measurement methodology and the basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Institute's technical and scientific programs. Issued six times a year.

Nonperiodicals

Monographs—Major contributions to the technical literature on various subjects related to the Institute's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NIST, NIST annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NIST under the authority of the National Standard Data Act (Public Law 90-396). NOTE:The Journal of Physical and Chemical Reference Data (JPCRD) is published bimonthly for NIST by the American Institute of Physics (AIP). Subscription orders and renewals are available from AIP, P.O. Box 503284, St. Louis, MO63150-3284.

National Construction Safety Team Act Reports—This series comprises the reports of investigations carried out under Public Law 107-231, the technical cause(s) of the building failure investigated; any technical recommendations for changes to or the establishment of evacuation and emergency response procedures; any recommended specific improvements to building standards, codes, and practices; and recomendations for research and other approprate actions to help prevent future building failures.

Building Science Series—Disseminates technical information developed at the Institute on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NIST under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NIST administers this program in support of the efforts of private-sector standardizing organizations.

Order the following NIST publications—FIPS and NISTIRs—from the National Technical Information Service, Springfield, VA 22161.

Federal Information Processing Standards Publications (FIPS PUB)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NIST pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NIST Interagency or Internal Reports (NISTIR)—The series includes interim or final reports on work performed by NIST for outside sponsors (both government and nongovernment). In general, initial distribution is handled by the sponsor; public distribution is handled by sales through the National Technical Information Service, Springfield, VA 22161, in hard copy, electronic media, or microfiche form. NISTIR's may also report results of NIST projects of transitory or limited interest, including those that will be published subsequently in more comprehensive form.