Text REtrieval Conference (TREC)

Rekha Ramesh (114388002)
Abhishek Mallik (113050019)
Aliabbas Petiwala (114380003)
Motivation

• You are experimenting in Information Retrieval area
• You have built a retrieval system
• You want to test it with realistically sized test data
• You want to compare results across other similar systems
• Where to go?
• The answer is TREC ……
What is TREC?

- A workshop series that provides the infrastructure for large-scale testing of (text) retrieval technology
  - Realistic test collections
  - Uniform, appropriate scoring procedures
  - A forum for the exchange of research methodology
TREC Philosophy

• TREC is a modern example of the Cranfield tradition
  – system evaluation based on test collection

• Emphasis on advancing the state of the art from evaluation results
  – TREC’s primary purpose is not competitive benchmarking
  – Experimental workshop
Yearly Conference Cycle

- Call for Participation
- Task Definition
- Document Procurement
- Topic Development
- IR Experiments
- Relevance Assessments
- Results Evaluation
- Results Analysis
- TREC Conference
- Proceedings Publication
A Brief History of TREC

• 1992: first TREC conference
  – Started by Donna Harman and Charles Wayne as 1 of 3 evaluations in DARPA’s TIPSTER program
  – First 3 CDS of documents from this era
  – Open to IR not funded by DARPA
    • 25 groups submitted runs
  – Two tasks: ad hoc retrieval, routing
    • 2GB of text, 50 topics
    • Primarily an exercise in scaling up systems
Adhoc and Routing Task

• **Routing task**
  – Same questions are always being asked, but that new data is being searched. For e.g. News clipping services, library profiling systems.

• **Adhoc task**
  – New questions are being asked against a static set of data. For e.g. Researcher using a library
Goals of TREC

- To encourage research in text retrieval 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

- To increase the availability of appropriate evaluation techniques for use by industry and academia
TREC Tracks

- TREC-1(1992) and TREC-2(1993) concentrated on ad hoc task and the routing task

- Starting in TREC-3, a variety of other “tracks,” were introduced (shown in table)

- The tracks invigorate TREC by focusing research on new areas or particular aspects of text retrieval

- The tracks represent the majority of the work done in TREC
TREC Tracks

• Set of tracks in a particular TREC depends on:
  – Interest of participants
  – Appropriateness of task to TREC
  – Needs of sponsors
  – resource constraints

• Need to submit proposal for new track in writing to NIST
Tasks Performed in TREC

- **Personal documents**
  - Blog, Microblog
  - Spam

- **Retrieval in a domain**
  - Chemical IR
  - Genomics, Medical Records

- **Answers, not documents**
  - Novelty
  - QA, Entity

- **Searching corporate repositories**
  - Legal Enterprise

- **Size, efficiency, & web search**
  - Terabyte, Million Query
  - Web
  - VLC

- **Beyond text**
  - Video
  - Speech
  - OCR

- **Beyond just English**
  - Cross-language
  - Chinese
  - Spanish

- **Human-in-the-loop**
  - HARD, Feedback
  - Interactive, Session

- **Streamed text**
  - Filtering
  - Routing

- **Static text**
  - Ad Hoc, Robust

**Years:**
- 1992
- 1994
- 1995
- 1996
- 1997
- 1998
- 1999
- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009
- 2010
- 2011
Use of only static text

• Million Query Track:
  – to explore ad hoc retrieval on a large document set.

• Robust Track:
  – to focus research on improving the consistency of retrieval technology
    by concentrating on poorly performing topics.

Use of streamed text

• Filtering:
  – for each document in a document stream, decide whether to
    retrieve it in response to standing query
  – Routing, batch filtering, adaptive filtering
Human in the LOOP

• **Interactive**: Investigate searching as an interactive task by examining the process as well as the outcome
  – Observational study of subjects using live web to perform search task
  – Controlled laboratory experiment of hypothesis suggested by observations

• **HARD**: High Accuracy Retrieval from Documents
  – to improve retrieval performance by targeting retrieval results to the specific user
Beyond just English

• **CLIR**: Cross Lingual Information Retrieval
  – documents are in one language and topics are in another language

• **Multilingual (Spanish & Chinese):**
  – Investigates the retrieval performance when both documents and topics is in a language other than English
Beyond text

• **Confusion Track**: investigates how retrieval performance is affected through noisy and confused data
  - Data processed through OCR

• **SDR Track**: retrieval methodology for spoken documents

• **The Video Track**: to promote progress in content based retrieval task
Web Searching

- **Web Track**: build a test system that more closely mimics the retrieval environment of WWW
  - Documents were a collection of web pages

- **Terabyte Track**: develop an evaluation methodology for terabyte-scale document collections.
  - This collection contains a large proportion of the crawlable pages in .gov, including html and text, plus extracted text of pdf, word and postscript files

- **Enterprise Track**: 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,
Answers, not Docs

• **Novelty Track**: investigate systems’ abilities to locate relevant and new (non redundant) information within an ordered set of documents.

• **Q&A Track**: systems that return actual answers, as opposed to ranked lists of documents, in response to a question.
  – the test set of questions consisted of factoid, list, and definition questions.

  DEFINITION What is a comet?
  FACTOID How often does it approach the earth?
  LIST In what countries was the comet visible on its last return?
  OTHER
Retrieval in Domains

• **Legal Track**: to focus specifically on the problem of e-discovery, the effective production of digital or digitized documents as evidence in litigation.

• **Genome Track**: a forum for evaluation of information access systems in the genomics domain – document collections were a set of full-text articles from several biomedical journals.
Personal Documents

• **Blogs Track:** explore information seeking behavior in the blogosphere
  – to discover the similarities and differences between blog search and other types of search

• **Spams Track:** to evaluate how well systems are able to separate spam and ham (non-spam) when given an email sequence
  – The task involved classifying email messages as ham or spam, differing in the amount and frequency of the feedback the system received.
Participation in TREC

More than 250 distinct groups have participated in at least one TREC.
TREC Impacts

• **Test collections**
  – Papers in general literature use TREC collections

• **Incubator for new research areas**
  – PhD theses resulting from CLIR, SDR, QA

• **Common evaluation methodology and improved measures for text retrieval**
  – Document best practices in IR research methodology for new researchers
TREC Impacts

• Open forum for exchange of research
  – TREC papers figure prominently in IR syllabi on the web
  – Publication of all results prevents unsuccessful research from being duplicated

• Technology transfer
  – Impact is far greater than just those who actually participate
TREC Approach

Assessors create topics at NIST

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

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

Systems are evaluated using relevance judgments
Creating a test collection for Adhoc task

- **topic statements**
  - Automatic: no manual intervention
  - Manual: everything else, including interactive feedback

- **queries**

- **representative document set**

- **ranked list**
Documents

• **Must be representative of real task of interest**
  - Genre
  - Diversity (subject, style and vocabulary)
  - Amount
  - Full text vs. abstract

• **TREC**
  - Generally newswire/news paper
  - General interest topics
  - Full text
Topics

• **Distinguish between statement of user need (topic) & system data structure (query)**
  – Topics gives criteria for relevance
  – Allows for different query construction techniques

• **TREC topics are NOT all created equal**
  – 1-150: very detailed, rich content
  – 151-250: method of topic creation resulted in focused and easy topics
  – 201-250: single sentence only
  – 301-450: title is set of hand-picked keywords

• **A topic generally consists of four sections:**
  – an identifier,
  – a title (consist of up to three words that best describe the topic),
  – a description (one sentence description of the topic area), and
  – a narrative (a concise description of what makes a document relevant)
An Example Topic

<num> Number: 951
<title> Mutual Funds
<desc> Description: Blogs about mutual funds performance and trends.
<narr> Narrative: Ratings from other known sources (Morningstar) correlative 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.)
Relevance Judgments (1)

- The relevance judgments turn a set of documents and topics into a test collection.

- Retrieve all of the relevant documents and none of the irrelevant documents.

- TREC usually uses binary relevance judgments.
Relevance Judgments (2)

• In test collections, judgments are usually binary, static, and assumed to be complete.

• But...
  – Relevance is idiosyncratic.
  – Relevance does not entail utility.
  – Documents have different degrees of relevance.
  – Relevance can change over time with the same user.
  – For realistic collections judgments cannot be complete.
Relevance Judgments (3)

- **Consistency**
  - Idiosyncratic nature of relevance judgments does not effect comparative results

- **Incompleteness**
  - Complete judgments should be unbiased
  - TREC pooling has been adequate to produce unbiased judgments
Creating relevance judgments
Evaluation at TREC

• In TREC, ad hoc tasks are evaluated using the trec-eval package

• This package reports about 85 different numbers for a run

• Including recall and precision at various cut-off levels

• And single valued summary measures that are derived from recall and precision.
Evaluation Measure Criteria

• Related to user satisfaction
• Interpretable
• Able to average or collect
• Have high discrimination power
• Able to be analyzed
Ranked Retrieval Chart
Evaluation Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Relevant</th>
<th>Non Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieved</td>
<td>r</td>
<td>n-r</td>
</tr>
<tr>
<td>Non Retrieved</td>
<td>R-r</td>
<td>N-n-R+r</td>
</tr>
</tbody>
</table>

\[ \begin{align*}
N & - \text{ Number of docs in collection} \\
n & - \text{ Number of docs retrieved} \\
R & - \text{ Number of docs relevant} \\
r & - \text{ Number of relevant docs retrieved}
\end{align*} \]
Recall Precision Graph

![Graph showing recall versus precision for different runs]

- run1
- run2
- run3
Uninterpolated R-P curve for Single Topic
Interpolated R-P curve for Single Topic
Single Number Summary Score

- **Precision** (n) : $r / n$
- **Recall** (n) : $r / R$
- **F-measure** : $2PR/(P+R)$
- **Average** : $\text{Avg}_{rd}(\text{Prec (rank of rd)})$
- **Precision at K** : Precision at fixed low levels of retrieved result
- **R-Precision** : $r/|R|$ where $r$ is no. of relevant documents from the top $|R|$ documents of a system.
- **Rank of first relevant** (expected search length)
Average Precision

- **Advantage**
  - Sensitive to entire ranking: changing a rank will change final score
  - Stable: a small change in ranking makes relatively small change in scores
  - Has both precision and recall oriented factors
    - Ranks closest to 1 receives largest weight
    - computed over all relevant documents

- **Disadvantage**
  - Less easily interpretable
Summary

• TREC emphasizes individual experiments evaluated against benchmark task
  – Leverages modest government investments into substantially more R&D than could funded directly
  – Improves state-of-the-art
  – Accelerates technology transfer
References (1)


- All subsequent 19 Text REtrieval Conference (TREC) publications

References (2)

