

Information Retrieval

CS 6900

Lecture 10

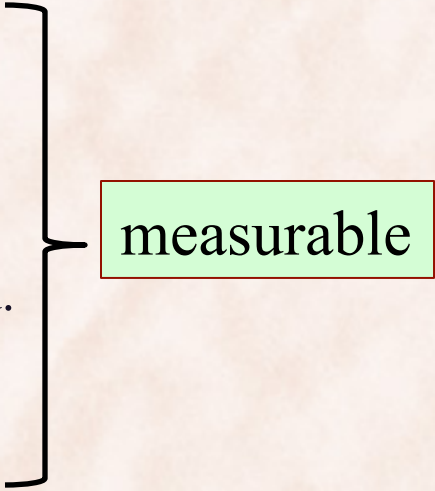
Razvan C. Bunescu

School of Electrical Engineering and Computer Science

bunescu@ohio.edu

IR Evaluation Measures

- 1) **How fast does it index?**
 - Number of bytes per second.
- 2) **How fast does it search?**
 - Latency as a function of queries per second.
- 3) **What is the cost per query?**
 - \$/query.
- 4) **What is the level of **user happiness**?**
 - How can we quantify user happiness?



measurable

User Happiness

- Who is the user we are trying to make happy?
 - Web search engine: searcher. Success: Searcher finds what she was looking for. Measure: rate of return to this search engine.
 - Web search engine: advertiser. Success: Searcher clicks on ad. Measure: clickthrough rate.
 - Ecommerce: buyer. Success: Buyer buys something. Measures: time to purchase, fraction of “conversions” of searchers to buyers.
 - Ecommerce: seller. Success: Seller sells something. Measure: profit per item sold.
 - Enterprise: CEO. Success: Employees are more productive (because of effective search). Measure: profit of the company.

Relevance as Proxy for User Happiness

- User **happiness** \approx the **relevance** of search results.
- Relevance is assessed relative to the **user need**, *not* the **query**.
 - Note: **user need** is translated into a **query**.
 - Information need: *I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
 - Query: red wine white wine heart attack
 - Assess whether the retrieved document addresses the underlying need, not whether it has these words.
 - Binary Assessments: **Relevant** or **Nonrelevant**.

Standard Methodology for Measuring Relevance in IR

- To measure relevance effectiveness of ad-hoc IR, we need:
 1. A **document collection**.
 2. A suite of information needs, expressible as **queries**.
 - Must be representative of actual user needs.
 - Sample from query logs, if available.
 3. **Binary assessments** of either Relevant or Nonrelevant for each query and each document.
 - Can be more nuanced: *numerical* (0, 1, 2, 3, ...) or *ordinal*.
 - Use *pooling*, when it is unfeasible to assess every (q, d) pair.

Unranked Retrieval Measures

- **Precision:** fraction of retrieved docs that are relevant = $P(\text{relevant} \mid \text{retrieved})$
- **Recall:** fraction of relevant docs that are retrieved = $P(\text{retrieved} \mid \text{relevant})$

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

$$\text{Precision } P = tp / (tp + fp)$$

$$\text{Recall } R = tp / (tp + fn)$$

Precision and Recall

- **Precision** reflects the ability to retrieve top-ranked documents that are mostly relevant.
- **Recall** reflects the ability of the search to find *all* of the relevant items in the corpus.
 - Difficult to estimate, since total number of relevant documents may not be available.
 - **Pooling**: Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items in the top k results from each algorithm is taken as the total relevant set.

F-measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P + R} = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.
- Instantiation of more general F_β , for $\beta=1$:

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R} = \frac{1 + \beta^2}{\frac{\beta^2}{R} + \frac{1}{P}}$$

Ranked Retrieval Measures

- Binary relevance:
 - R-precision.
 - Precision@K (P@K) and Recall@K (R@K).
 - 11-point Interpolated Average Precision.
 - Mean Average Precision (MAP).
 - Mean Reciprocal Rank (MRR).
- Multiple levels of relevance:
 - Normalized Discounted Cumulative Gain (NDCG)

R-precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = $4/6 = 0.67$

Precision@K

1. Set a rank threshold K.
2. Compute % of documents relevant in top K.
 - Ignores documents ranked lower than K.

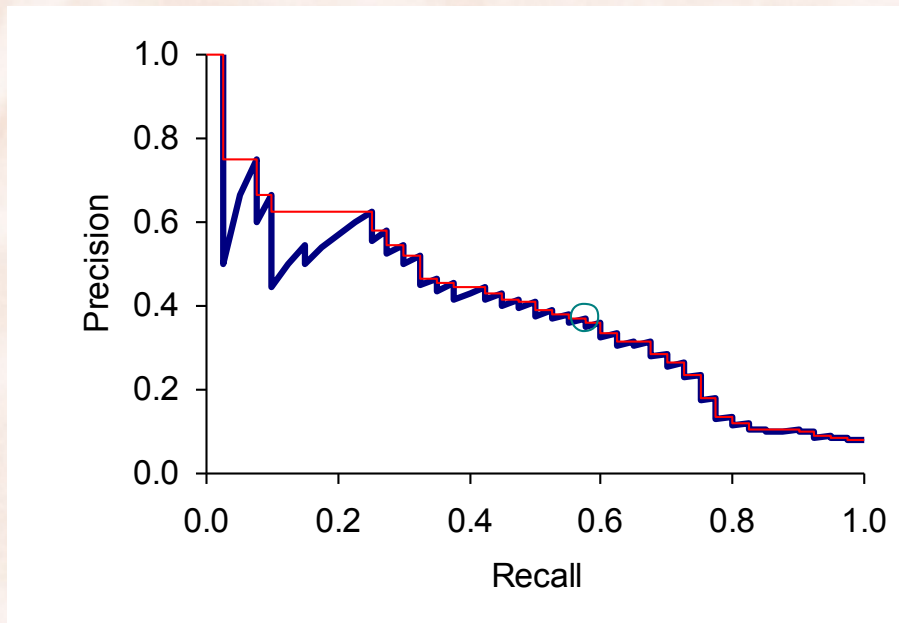
- Example:

- Prec@3 of 2/3
- Prec@4 of 2/4
- Prec@5 of 3/5



- In a similar fashion we have Recall@K

Precision vs. Recall Curves

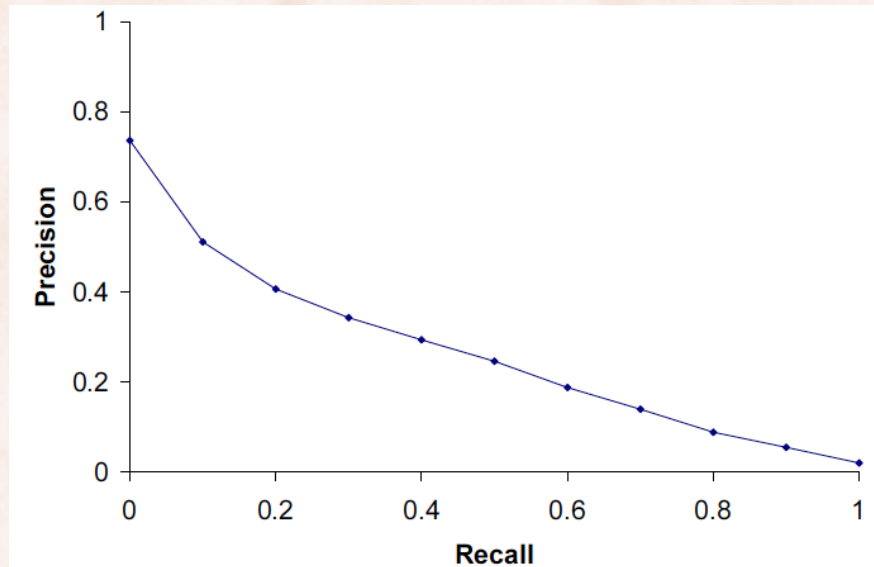


- Each point corresponds to a result for top k hits ($k = 1, 2, 3, \dots$).
- Interpolation (in red): Take maximum of all future points.
 - Rationale: The user is willing to look at more stuff if both precision and recall get better.

11-point Interpolated Average Precision

Recall	Interp. Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

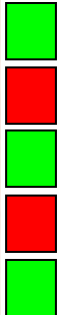
- 11-point interpolated average precision is about 0.425.
- Used in first 8 TREC evaluations.



► Figure 8.3 Averaged 11-point precision/recall graph across 50 queries for a representative TREC system. The Mean Average Precision for this system is 0.2553.

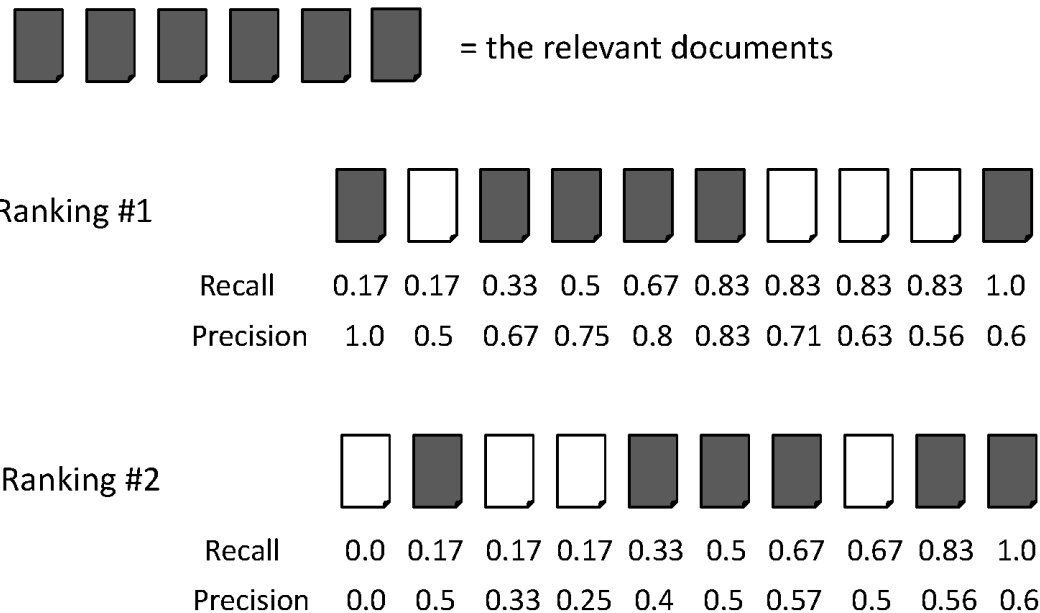
Mean Average Precision (MAP)

1. Consider rank position of each of the R relevant docs:
 - K_1, K_2, \dots, K_R
2. Compute Precision@K for each K_1, K_2, \dots, K_R .
3. Average precision = average of P@K.

Example:  has AvgPrec of $\frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$

- MAP is Average Precision across multiple queries.

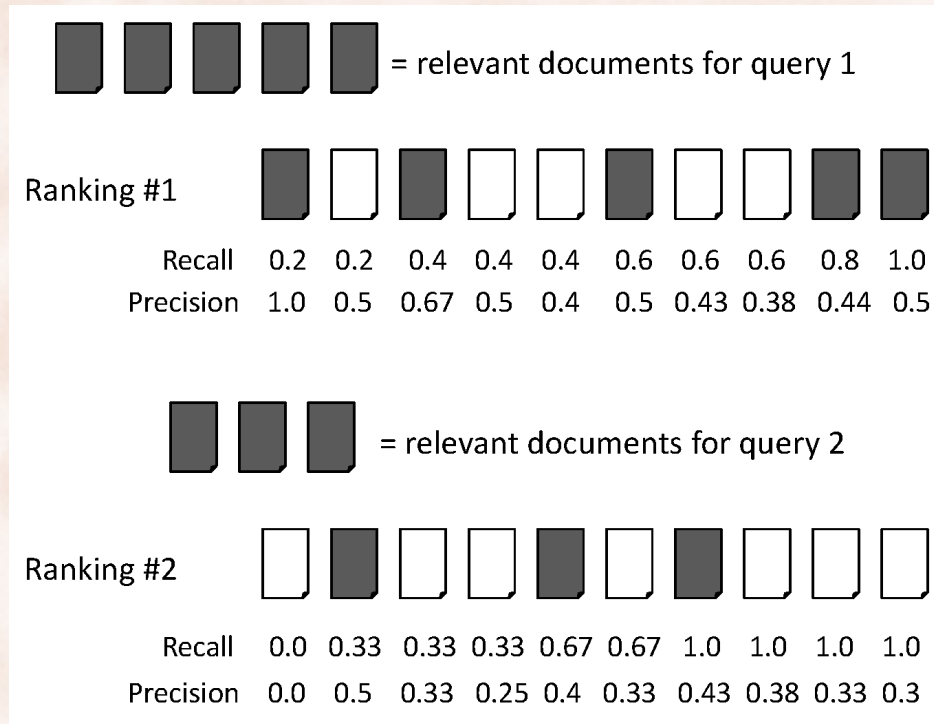
Average Precision



$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

Mean Average Precision (MAP)



average precision query 1 = $(1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$

average precision query 2 = $(0.5 + 0.4 + 0.43)/3 = 0.44$

mean average precision = $(0.62 + 0.44)/2 = 0.53$

Mean Average Precision (MAP)

- If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc. to be zero.
- MAP is macro-averaging: each query counts equally.
- Now perhaps most commonly used measure in research papers.
- Good for web search?
 - MAP assumes user is interested in finding many relevant documents for each query
 - MAP requires many relevance judgments in a text collection.

Multiple Levels of Relevance

- Documents are rarely entirely relevant or non-relevant to a query.
- Many sources of *graded relevance judgments*:
 - Relevance judgments on a 5-point scale.
 - Averaging among multiple judges.

Search Pad

SearchScan - On

108,000,000 results for Toyota safety:

Show All

Toyota

Motor Trend

CarsDirect

Shopping Sites

Also try: [toyota safety ratings](#), [toyota safety recall](#), [More...](#)

Toyota Recall

Toyota Takes Care of its Customers. Read the FAQs at **Toyota.com**.
www.Toyota.com/Recall

Sponsored Results

Toyota Safety

& Latest Prices. Free Info. **Toyota** Research, Reviews.
www.Toyota.Edmunds.com

TOYOTA | Car Safety Innovation and Technology

Toyota home page for car **safety** and car technology Prius model.
www.safetytoyota.com - [Cached](#)

Toyota home page for car safety and car technology ...

We are presenting **Toyota's safety** technologies for cars. We clearly explain about car **safety** and car technology using movies and more.
www.safetytoyota.com/en-gb - [Cached](#)

Toyota Safety Ratings - Toyota Safety Features - Motor Trend ...

MotorTrend offers **Toyota safety** ratings, comprehensive auto **safety** reports, and more. View a all of the standard **Toyota safety** features. ...
motortrend.com/new_cars/07/toyota/safety_ratings/index.html - 149k - [Cached](#)

Toyota Motor Europe Corporate Site Safety

Our approach. **Toyota** believes that all stakeholders in the road **safety** equation share a responsibility to reduce the frequency of road accidents. ...
www.toyota.eu/Safety - [Cached](#)

pdf European Safety Brochure 2005

4047k - Adobe PDF - [View as html](#)
not guarantee that all accidents or injuries will be avoided when driving a **Toyota** and/or Lexus brand motor vehicle equipped with the **safety** systems ...
www.toyota.no/Images/Safety_Brochure_tcm308-344461.pdf

Toyota - Star Safety System

Star **Safety** System ... **Toyota** Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. **Toyota** Newsroom. sign up for info ...
www.toyota.com/vehicles/demos/star-safety.html - 58k - [Cached](#)

Toyota Prius Safety Ratings - CarsDirect

Get overall **safety** ratings and NHTSA crash test results for the **Toyota** Prius at CarsDirect.

Safety for a Toyota

Research **Safety** Ratings and Reviews For New Car at Kelley Blue Book.
www.kbb.com

Sponsored Results

Toyota Safety

Find **Toyota Safety** dealers, new cars, prices, and photos.
www.NewCars.org

Toyota Safety

Toyota safety Discount Prices Save Money Shopping Online Today.
www.smarter.com

Safety Toyoto

Explore 5,000+ Pro Sports Choices. Save On Safety Toyota.
BaseballGear.Shopzilla.com

[See your message here...](#)

Fair

Fair

Good

Cummulative Gain

- With graded relevance judgments, we can compute the *gain* at each rank.
- **Cumulative Gain** at rank n :

$$CG_n = \sum_{i=1}^n rel_i$$

- Where rel_i is the graded relevance of the document at position i .

n	doc #	relevance (gain)	CG _n
1	588	1.0	1.0
2	589	0.6	1.6
3	576	0.0	1.6
4	590	0.8	2.4
5	986	0.0	2.4
6	592	1.0	3.4
7	984	0.0	3.4
8	988	0.0	3.4
9	578	0.0	3.4
10	985	0.0	3.4
11	103	0.0	3.4
12	591	0.0	3.4
13	772	0.2	3.6
14	990	0.0	3.6

Discounted Cumulative Gain

- Users care more about high-ranked documents, so we **discount** results by $1/\log_2(rank)$
- Popular measures for evaluating web search and related tasks.
- **Discounted Cumulative Gain:**

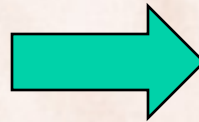
$$DCG_n = rel_1 + \sum_{i=2}^n \frac{rel_i}{\log_2 i}$$

n	doc #	rel (gain)	CG _n	log _n	DCG _n
1	588	1.0	1.0	-	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44

Normalized Discounted Cumulative Gain (NDCG)

- To compare DCGs, normalize values so that an *ideal ranking* would have a **Normalized DCG** of 1.0.
- Ideal ranking:

n	doc #	rel (gain)	CG _n	log _n	DCG _n
1	588	1.0	1.0	0.00	1.00
2	589	0.6	1.6	1.00	1.60
3	576	0.0	1.6	1.58	1.60
4	590	0.8	2.4	2.00	2.00
5	986	0.0	2.4	2.32	2.00
6	592	1.0	3.4	2.58	2.39
7	984	0.0	3.4	2.81	2.39
8	988	0.0	3.4	3.00	2.39
9	578	0.0	3.4	3.17	2.39
10	985	0.0	3.4	3.32	2.39
11	103	0.0	3.4	3.46	2.39
12	591	0.0	3.4	3.58	2.39
13	772	0.2	3.6	3.70	2.44
14	990	0.0	3.6	3.81	2.44



n	doc #	rel (gain)	CG _n	log _n	IDCG _n
1	588	1.0	1.0	0.00	1.00
2	592	1.0	2.0	1.00	2.00
3	590	0.8	2.8	1.58	2.50
4	589	0.6	3.4	2.00	2.80
5	772	0.2	3.6	2.32	2.89
6	576	0.0	3.6	2.58	2.89
7	986	0.0	3.6	2.81	2.89
8	984	0.0	3.6	3.00	2.89
9	988	0.0	3.6	3.17	2.89
10	578	0.0	3.6	3.32	2.89
11	985	0.0	3.6	3.46	2.89
12	103	0.0	3.6	3.58	2.89
13	591	0.0	3.6	3.70	2.89
14	990	0.0	3.6	3.81	2.89

Normalized Discounted Cumulative Gain (NDCG)

- Normalize by DCG of the ideal ranking:

$$NDCG_n = \frac{DCG_n}{IDCG_n}$$

- $NDCG \leq 1$ at all ranks.
- $NDCG$ is now comparable across different queries:
 - Useful for contrasting queries with varying numbers of relevant results.
 - Quite popular for Web search.

n	doc #	rel (gain)	DCG _n	IDCG _n	NDCG _n
1	588	1.0	1.00	1.00	1.00
2	589	0.6	1.60	2.00	0.80
3	576	0.0	1.60	2.50	0.64
4	590	0.8	2.00	2.80	0.71
5	986	0.0	2.00	2.89	0.69
6	592	1.0	2.39	2.89	0.83
7	984	0.0	2.39	2.89	0.83
8	988	0.0	2.39	2.89	0.83
9	578	0.0	2.39	2.89	0.83
10	985	0.0	2.39	2.89	0.83
11	103	0.0	2.39	2.89	0.83
12	591	0.0	2.39	2.89	0.83
13	772	0.2	2.44	2.89	0.84
14	990	0.0	2.44	2.89	0.84

Issues with Relevance

- **Marginal Relevance:** Do later documents in the ranking add new information beyond what is already given in higher documents.
 - Choice of retrieved set should encourage **diversity** and **novelty**.
- **Coverage Ratio:** The proportion of relevant items retrieved out of the total relevant documents *known* to a user prior to the search.
 - Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000).

A/B Testing at Web Search Engines

- Can exploit an existing user base to provide useful feedback on a **single innovation**.
- Randomly send a small fraction (1–10%) of incoming users to a variant of the system that includes a single change.
 - Have most users use the old system.
- Judge effectiveness by measuring change in ***clickthrough***: the percentage of users that click on the top result (or any result on the first page).
- Probably the evaluation methodology that large search engines trust the most.

Standard Methodology for Measuring Relevance in IR

- To measure relevance effectiveness of ad-hoc IR, we need:
 1. A **document collection**.
 2. A suite of information needs, expressible as **queries**.
 - Must be representative of actual user needs.
 - Sample from query logs, if available.
 3. **Binary assessments** of either Relevant or Nonrelevant for each query and each document.
 - Can be more nuanced: *numerical* (0, 1, 2, 3, ...) or *ordinal*.
 - Use *pooling*, when it is unfeasible to assess every (q, d) pair.

Early Test Collections

- Previous experiments were based on the SMART collection which is fairly small. (<ftp://ftp.cs.cornell.edu/pub/smart>)

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

- Different researchers used different test collections and evaluation techniques.

The TREC Benchmark

- TREC: **T**ext **RE**trieval **C**onference (<http://trec.nist.gov/>)
 - Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
 - Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.
 - Participants are given parts of a standard set of documents and **TOPICS** (from which queries have to be derived) in different stages for training and testing.
 - Participants submit the P/R values for the final document and query corpus and present their results at the conference.

TREC Objectives

- Provide a common ground for comparing different IR techniques.
 - Same set of documents and queries, and same evaluation method.
- Sharing of resources and experiences in developing the benchmark.
 - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
 - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

TREC Advantages

- Large scale (compared to a few MB in the SMART Collection).
- Relevance judgments provided.
- Under continuous development with support from the U.S. Government.
- Wide participation:
 - TREC 1: 28 papers 360 pages.
 - TREC 4: 37 papers 560 pages.
 - TREC 7: 61 papers 600 pages.
 - TREC 8: 74 papers.

TREC Tasks

- **Ad hoc:** New questions are being asked on a static set of data.
- **Routing:** Same questions are being asked, but new information is being searched. (news clipping, library profiling).
- New tasks added after TREC 5:
 - Interactive, multilingual, natural language, multiple database merging, filtering, very large corpus (20 GB, 7.5 million documents), question answering.

The TREC Collection

- Both long and short documents (from a few hundred to over one thousand unique terms in a document).
 - Both SGML documents and SGML queries contain many different kinds of information (fields).
 - Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system.
 - A system may be very good at ranking, but if it generates poor queries from the topic, its final P/R would be poor.
- Test documents consist of:

WSJ	Wall Street Journal articles (1986-1992)	550 M
AP	Associate Press Newswire (1989)	514 M
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 M
FR	Federal Register	469 M
DOE	Abstracts from Department of Energy reports	190 M

Sample SGML Document

<DOC>

<DOCNO> WSJ870324-0001 </DOCNO>

<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>

<DD> 03/24/87</DD>

<SO> WALL STREET JOURNAL (J) </SO>

<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)
MARKETING, ADVERTISING (MKT) TELECOMMUNICATIONS,
BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>

<DATELINE> NEW YORK </DATELINE>

<TEXT>

John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said. Industry sources put the value of the proposed acquisition at more than \$100 million. ...

</TEXT>

</DOC>

Sample SGML Query

```
<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language processing
        technology which is being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution
        developing or marketing a natural language processing technology, identify the
        technology, and identify one of more features of the company's product.
<con> Concept(s): 1. natural language processing ;2. translation, language,
        dictionary
<fac> Factor(s):
<nat> Nationality: U.S.</nat>
</fac>
<def> Definitions(s):
</top>
```

TREC Evaluation

- **Summary table statistics:** Number of topics, number of documents retrieved, number of relevant documents.
- **11-point Interpolated Precision:** Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- **Document level average:** Average precision when 5, 10, ..., 100, ... 1000 documents are retrieved.
- **Average precision histogram:** Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

GOV2 Web Corpus

- Recent web-based gold-standard corpus assembled by NIST.
 - The largest web collection easily available for research.
 - Still more than 2 orders of magnitude smaller than collections indexed by large web search companies.
- 25 million web pages in the .gov domain
 - High proportion of .gov pages in 2004.
- Total of 426 GB of text.
- Set of 50 relevance-judged queries.

Cystic Fibrosis (CS) Collection

- 1,239 abstracts of medical journal articles on CF.
- 100 information requests (queries) in the form of complete English questions.
- Relevant documents determined and rated by 4 separate medical experts on 0 to 2 scale:
 - 0: Not relevant.
 - 1: Marginally relevant.
 - 2: Highly relevant.

CF Document Fields

- MEDLINE access number
- Author
- Title
- Source
- Major subjects
- Minor subjects
- Abstract (or extract)
- References to other documents
- Citations to this document

Sample CF Document

AN 74154352

AU Burnell-R-H. Robertson-E-F.

TI Cystic fibrosis in a patient with Kartagener syndrome.

SO Am-J-Dis-Child. 1974 May. 127(5). P 746-7.

MJ CYSTIC-FIBROSIS: co. KARTAGENER-TRIAD: co.

MN CASE-REPORT. CHLORIDES: an. HUMAN. INFANT. LUNG: ra. MALE.

SITUS-INVERSUS: co, ra. SODIUM: an. SWEAT: an.

AB A patient exhibited the features of both Kartagener syndrome and cystic fibrosis. At most, to the authors' knowledge, this represents the third such report of the combination. Cystic fibrosis should be excluded before a diagnosis of Kartagener syndrome is made.

RF 001 KARTAGENER M BEITR KLIN TUBERK 83 489 933

002 SCHWARZ V ARCH DIS CHILD 43 695 968

003 MACE JW CLIN PEDIATR 10 285 971

...

CT 1 BOCHKOVA DN GENETIKA (SOVIET GENETICS) 11 154 975

2 WOOD RE AM REV RESPIR DIS 113 833 976

3 MOSSBERG B MT SINAI J MED 44 837 977

...

Sample CF Queries

QN 00002

QU Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respiratory tract in CF?

NR 00007

RD 169 1000 434 1001 454 0100 498 1000 499 1000 592 0002 875 1011

QN 00004

QU What is the lipid composition of CF respiratory secretions?

NR 00009

RD 503 0001 538 0100 539 0100 540 0100 553 0001 604 2222 669 1010
711 2122 876 2222

- NR: Number of Relevant documents
- RD: Relevant Documents
- Ratings code: Four 0-2 ratings, one from each expert