Precision and Recall

CS5154/6054

Yizong Cheng

9/13/2022

8 *Evaluation in information retrieval*

8.1 Information retrieval system evaluation

To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

1. A document collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either *relevant* or *nonrelevant* for each query-document pair.

RELEVANCE

GOLD STANDARD
GROUND TRUTH

The standard approach to information retrieval system evaluation revolves around the notion of *relevant* and *nonrelevant* documents. With respect to a user information need, a document in the test collection is given a binary classification as either relevant or nonrelevant. This decision is referred to as the *gold standard* or *ground truth* judgment of relevance. The test document collection and suite of information needs have to be of a reasonable size: you need to average performance over fairly large test sets, as results are highly variable over different documents and information needs. As a rule of thumb, 50 information needs has usually been found to be a sufficient minimum.

INFORMATION NEED

Relevance is assessed relative to an information need, *not* a query. For example, an information need might be:

> Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.

This might be translated into a query such as:

> wine AND red AND white AND heart AND attack AND effective

A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query. This distinction is often misunderstood in practice, because the information need is not overt.

8.2 Standard test collections

Here is a list of the most standard test collections and evaluation series. We focus particularly on test collections for ad hoc information retrieval system evaluation, but also mention a couple of similar test collections for text classification.

CRANFIELD The *Cranfield* collection. This was the pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness, but is nowadays too small for anything but the most elementary pilot experiments. Collected in the United Kingdom starting in the late 1950s, it contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgments of all (query, document) pairs.

TREC

Text Retrieval Conference (TREC). The U.S. National Institute of Standards and Technology (NIST) has run a large IR test bed evaluation series since 1992. Within this framework, there have been many tracks over a range of different test collections, but the best known test collections are the ones used for the TREC Ad Hoc track during the first 8 TREC evaluations between 1992 and 1999. In total, these test collections comprise 6 CDs containing 1.89 million documents (mainly, but not exclusively, newswire articles) and relevance judgments for 450 information needs, which are called *topics* and specified in detailed text passages. Individual test collections are defined over different subsets of this data. The early TRECs each consisted of 50 information needs, evaluated over different but overlapping sets of documents. TRECs 6–8 provide 150 information needs over about 528,000 newswire and Foreign Broadcast Information Service articles. This is probably the best subcollection to use in future work, because it is the largest and the topics are more consistent. Because the test document collections are so large, there are no exhaustive relevance judgments. Rather, NIST assessors' relevance judgments are available only for the documents that were among the top k returned for some system which was entered in the TREC evaluation for which the information need was developed.

File Edit View History Bookmarks Tools Help

Text REtrieval Conference (TREC) Ho ×

https://trec.nist.gov

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

Overview

Other Evaluations

Publications

Information for Active Participants

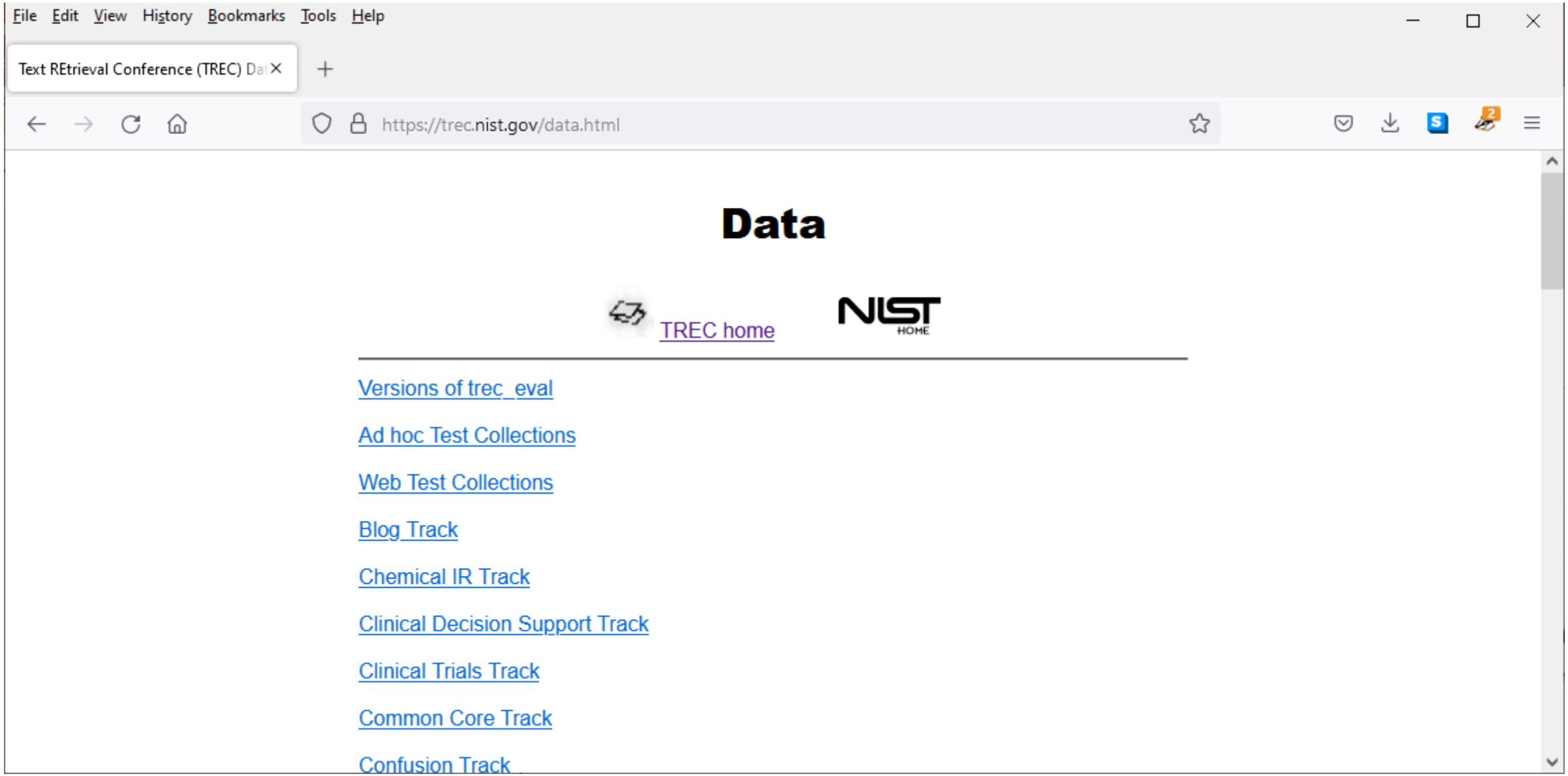
Frequently Asked Questions

Tracks

Data

Past TREC Results

Contact Information


File Edit View History Bookmarks Tools Help
Text REtrieval Conference (TREC) Dat
https://trec.nist.gov/data.html
Data
TREC home
NIST HOME
Versions of trec_eval
Ad hoc Test Collections
Web Test Collections
Blog Track
Chemical IR Track
Clinical Decision Support Track
Clinical Trials Track
Common Core Track
Confusion Track

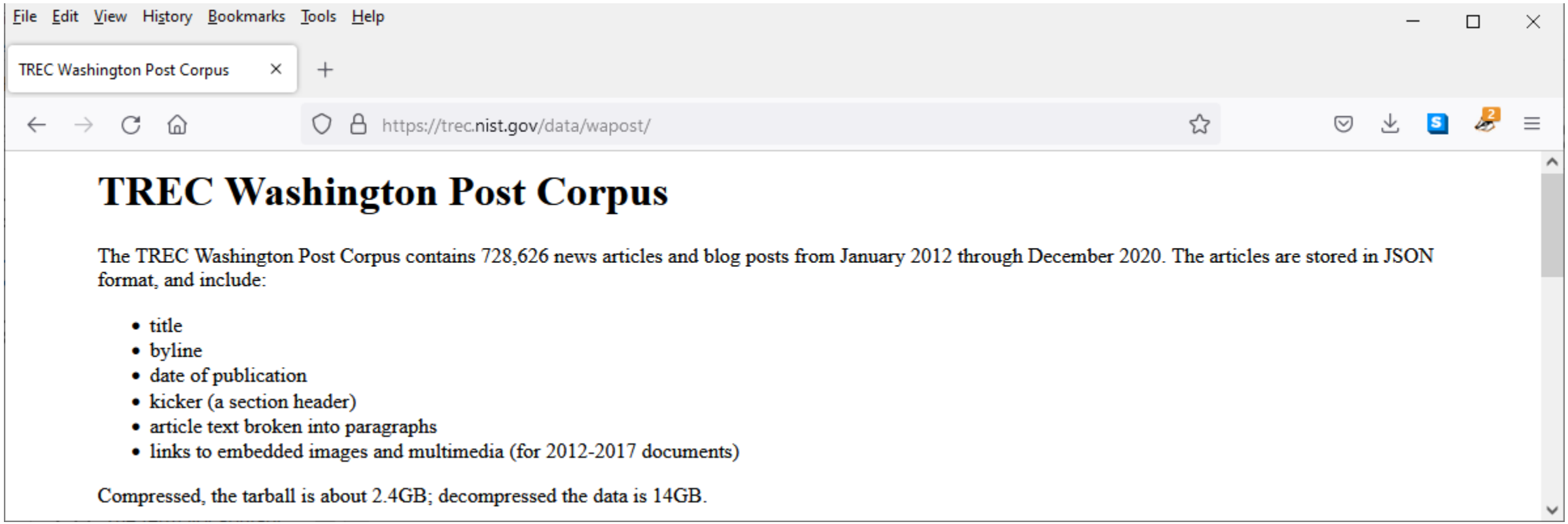
Data - *English Documents Introduction*

TREC home Data home NIST HOME

The following table shows the english documents used for testing in the various TRECs, along with the appropriate topic numbers used in each of the two main tasks.

Document and Topic Sets for TRECs			
TREC	**Task**	**Documents**	**Topics**
TREC-1	ad hoc routing	disks 1 & 2 disk 2	51 - 100 1 - 50
TREC-2	ad hoc routing	disks 1 & 2 disk 3	101 - 150 51 - 100
TREC-3	ad hoc routing	disks 1 & 2 disk 3*	151 - 200 101 - 150
TREC-4	ad hoc routing	disks 2 & 3 CS+FR	201 - 250 assorted
TREC-5	ad hoc routing	disks 2 & 4 FBIS-1	251 - 300 assorted
TREC-6	ad hoc routing	disks 4 & 5 FBIS-2	301 - 350 assorted
TREC-7	adhoc	disks 4 & 5 (no CR)	351 - 400
TREC-8	adhoc	disks 4 & 5 (no CR)	401 - 450
** re-use of disk3 forced by lack of new data*			

Last updated: Monday, 15-Apr-2019 08:22:44 MDT
Date created: Tuesday, 01-Aug-00
trec@nist.gov


File Edit View History Bookmarks Tools Help
TREC Washington Post Corpus
https://trec.nist.gov/data/wapost/
TREC Washington Post Corpus
The TREC Washington Post Corpus contains 728,626 news articles and blog posts from January 2012 through December 2020. The articles are stored in JSON format, and include:
title
byline
date of publication
kicker (a section header)
article text broken into paragraphs
links to embedded images and multimedia (for 2012-2017 documents)
Compressed, the tarball is about 2.4GB; decompressed the data is 14GB.

REUTERS

Reuters-21578 and Reuters-RCV1. For text classification, the most used test collection has been the Reuters-21578 collection of 21578 newswire articles; see Chapter 13, page 279. More recently, Reuters released the much larger Reuters Corpus Volume 1 (RCV1), consisting of 806,791 documents; see Chapter 4, page 69. Its scale and rich annotation makes it a better basis for future research.

20 NEWSGROUPS

20 Newsgroups. This is another widely used text classification collection, collected by Ken Lang. It consists of 1000 articles from each of 20 Usenet newsgroups (the newsgroup name being regarded as the category). After the removal of duplicate articles, as it is usually used, it contains 18941 articles.

Reuters-21578 Text Categorization Collection Data Set

Download: Data Folder, Data Set Description

Abstract: This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories.

Data Set Characteristics:	Text	**Number of Instances:**	21578	**Area:**	N/A
Attribute Characteristics:	Categorical	**Number of Attributes:**	5	**Date Donated**	1997-09-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	231953

Source:

David D. Lewis
AT&T Labs - Research
lewis '@' research.att.com

Documents came from Reuters newswire in 1987.

20 Newsgroups

The 20 Newsgroups data set

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his *Newsweeder: Learning to filter netnews* paper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Organization

The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. **comp.sys.ibm.pc.hardware / comp.sys.mac.hardware**), while others are highly unrelated (e.g **misc.forsale / soc.religion.christian**). Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Data

The data available here are in .tar.gz bundles. You will need tar and gunzip to open them. Each subdirectory in the bundle represents a newsgroup; each file in a subdirectory is the text of some newsgroup document that was posted to that newsgroup.

8.3 Evaluation of unranked retrieval sets

PRECISION *Precision* (P) is the fraction of retrieved documents that are relevant

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved}) \tag{8.1}$$

RECALL *Recall* (R) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant}) \tag{8.2}$$

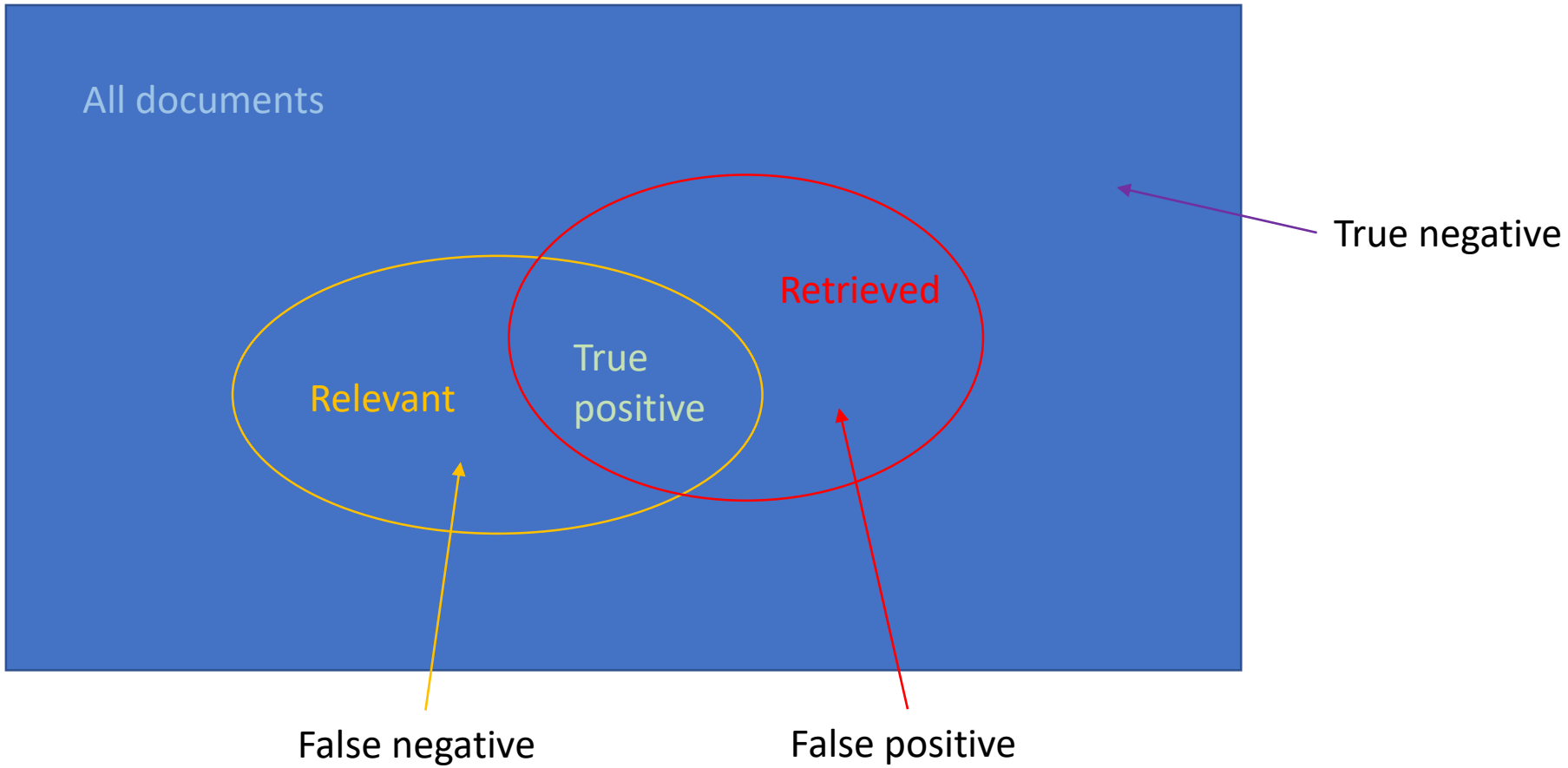
(8.3)

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Then:

(8.4)
$$\begin{aligned} P &= tp/(tp+fp) \\ R &= tp/(tp+fn) \end{aligned}$$

Two Sets of Documents (again!)



ACCURACY

An obvious alternative that may occur to the reader is to judge an information retrieval system by its *accuracy*, that is, the fraction of its classifications that are correct. In terms of the contingency table above, accuracy $= (tp + tn)/(tp + fp + fn + tn)$. This seems plausible, since there are two actual classes, relevant and nonrelevant, and an information retrieval system can be thought of as a two-class classifier which attempts to label them as such (it retrieves the subset of documents which it believes to be relevant). This is precisely the effectiveness measure often used for evaluating machine learning classification problems.

F MEASURE

A single measure that trades off precision versus recall is the *F measure*, which is the weighted harmonic mean of precision and recall:

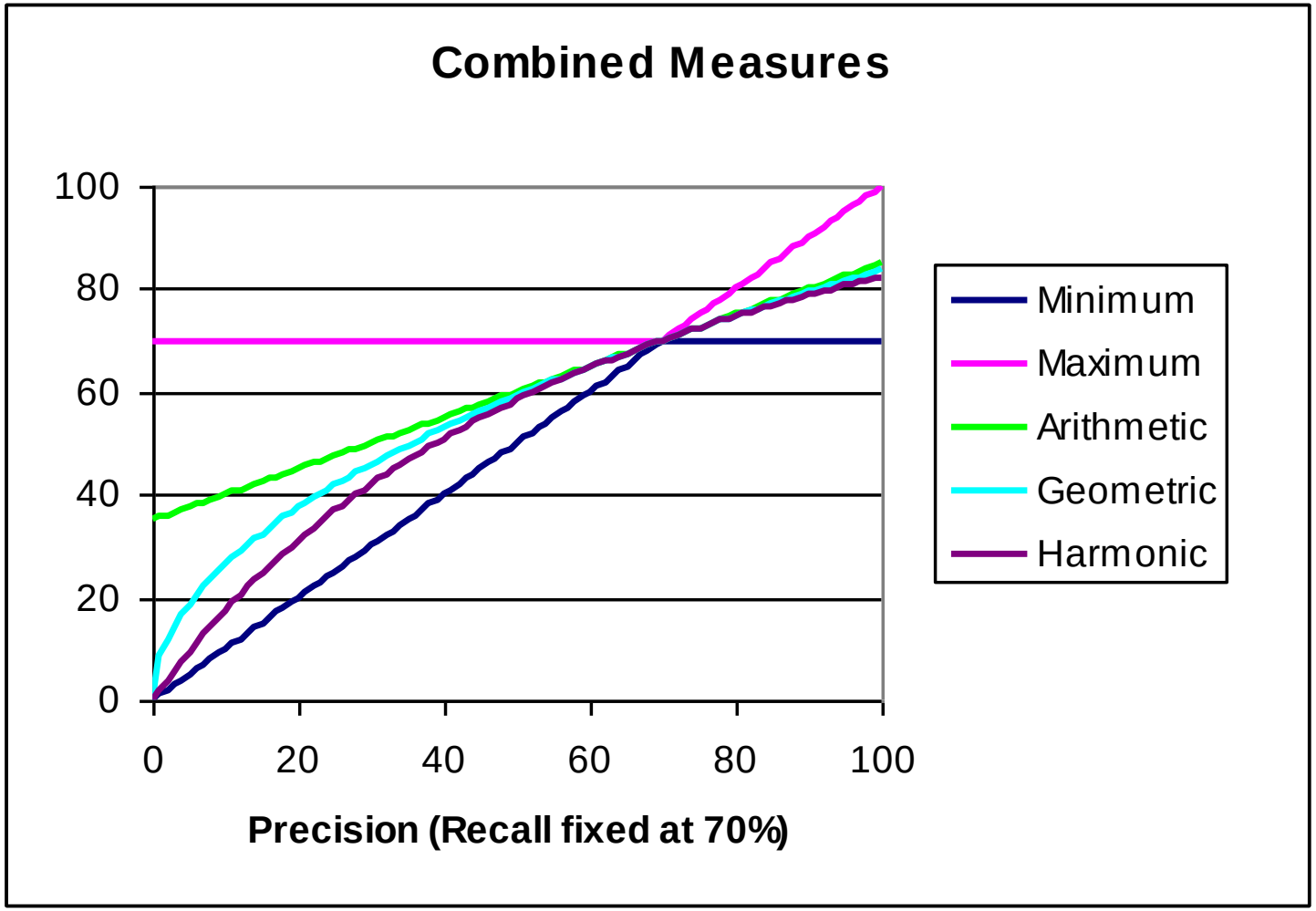
$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha} \tag{8.5}$$

where $\alpha \in [0, 1]$ and thus $\beta^2 \in [0, \infty]$. The default *balanced F measure* equally weights precision and recall, which means making $\alpha = 1/2$ or $\beta = 1$. It is commonly written as F_1, which is short for $F_{\beta=1}$, even though the formulation in terms of α more transparently exhibits the F measure as a weighted harmonic mean. When using $\beta = 1$, the formula on the right simplifies to:

$$F_{\beta=1} = \frac{2PR}{P+R} \tag{8.6}$$

However, using an even weighting is not the only choice. Values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall. For example, a

Why do we use a harmonic mean rather than the simpler average (arithmetic mean)? Recall that we can always get 100% recall by just returning all documents, and therefore we can always get a 50% arithmetic mean by the same process. This strongly suggests that the arithmetic mean is an unsuitable measure to use. In contrast, if we assume that 1 document in 10,000 is relevant to the query, the harmonic mean score of this strategy is 0.02%. The harmonic mean is always less than or equal to the arithmetic mean and the geometric mean. When the values of two numbers differ greatly, the harmonic mean is closer to their minimum than to their arithmetic mean; see Figure 8.1.



► **Figure 8.1** Graph comparing the harmonic mean to other means. The graph shows a slice through the calculation of various means of precision and recall for the fixed recall value of 70%. The harmonic mean is always less than either the arithmetic or geometric mean, and often quite close to the minimum of the two numbers. When the precision is also 70%, all the measures coincide.

8.5 Assessing relevance

To properly evaluate a system, your test information needs must be germane to the documents in the test document collection, and appropriate for predicted usage of the system. These information needs are best designed by domain experts. Using random combinations of query terms as an information need is generally not a good idea because typically they will not resemble the actual distribution of information needs.

Given information needs and documents, you need to collect relevance assessments. This is a time-consuming and expensive process involving human beings. For tiny collections like Cranfield, exhaustive judgments of relevance for each query and document pair were obtained. For large modern collections, it is usual for relevance to be assessed only for a subset of the documents for each query. The most standard approach is *pooling*, where relevance is assessed over a subset of the collection that is formed from the top k documents returned by a number of different IR systems (usually the ones

POOLING

KAPPA STATISTIC

Nevertheless, it is interesting to consider and measure how much agreement between judges there is on relevance judgments. In the social sciences, a common measure for agreement between judges is the *kappa statistic*. It is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement.

(8.10)
$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of the times the judges agreed, and $P(E)$ is the proportion of the times they would be expected to agree by chance. There are choices in how the latter is estimated: if we simply say we are making a two-class decision and assume nothing more, then the expected chance agreement rate is 0.5. However, normally the class distribution assigned is skewed, and it is usual to use *marginal* statistics to calculate expected agreement.[2] There are still two ways to do it depending on whether one pools the marginal distribution across judges or uses the marginals for each judge separately; both forms have been used, but we present the pooled version because it is more conservative in the presence of systematic differences in assessments across judges. The calculations are shown in Table 8.2. The kappa value will be 1 if two judges always agree, 0 if they agree only at the rate given by chance, and negative if they are worse than random. If there are

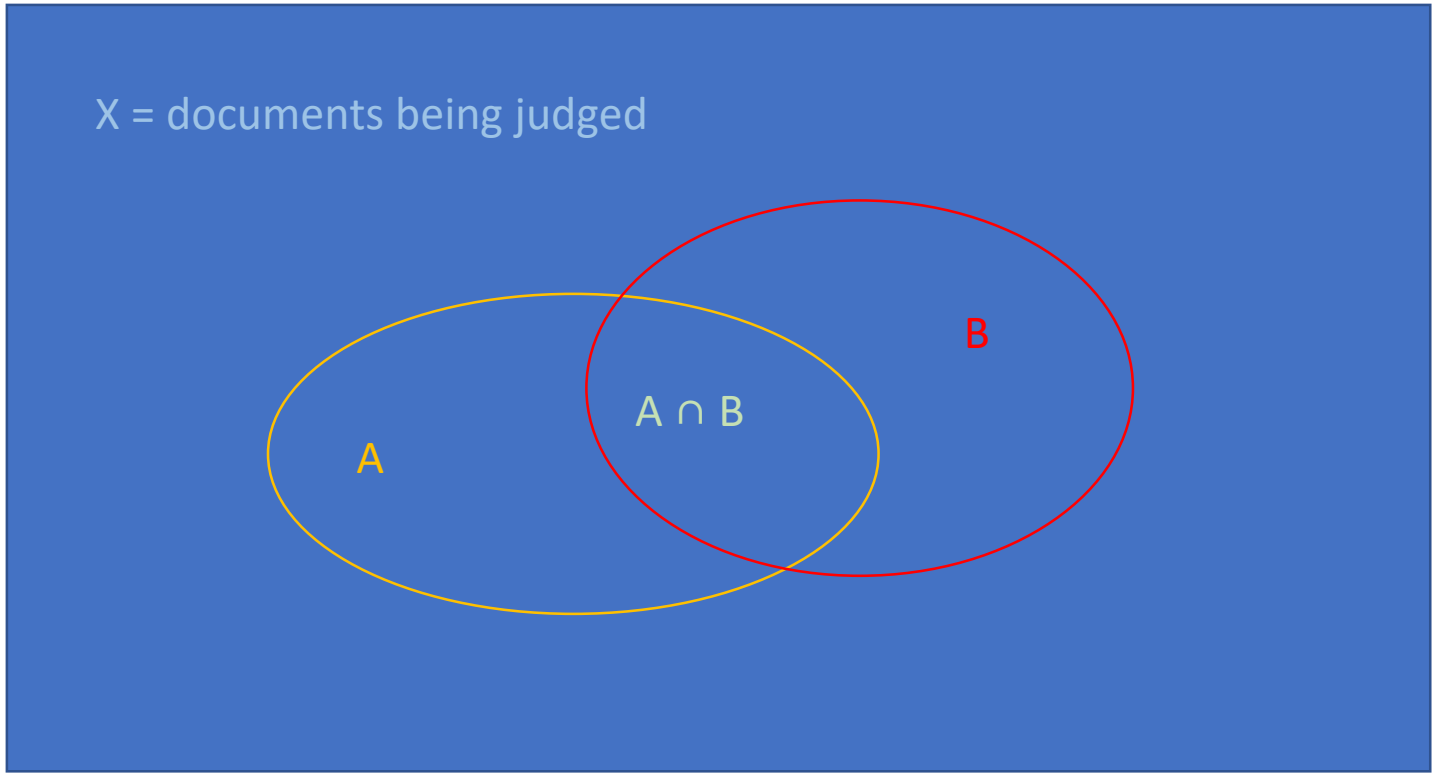
MARGINAL

		Judge 2 Relevance		
		Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

Observed proportion of the times the judges agreed
$P(A) = (300 + 70)/400 = 370/400 = 0.925$
Pooled marginals
$P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$
$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$
Probability that the two judges agreed by chance
$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$
Kappa statistic
$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$

► **Table 8.2** Calculating the kappa statistic.

Two Sets of Relevant Docs By Two Judges



Four Intersections and Their Counts

	B	not B	
A	$\|A \cap B\|$	$\|A\| - \|A \cap B\|$	$\|A\|$
not A	$\|B\| - \|A \cap B\|$	$\|X\| - \|A\| - \|B\| + \|A \cap B\|$	$\|X\| - \|A\|$
	$\|B\|$	$\|X\| - \|B\|$	$\|X\|$

	yes2	no2	
yes1	N11	N10	N1x
no1	N01	N00	N0x
	Nx1	Nx0	N

Kappa Statistic Calculation

- Observed proportion of the times the judges agreed:
  - $P(A) = (N11 + N00)/N$
- Pooled marginals
  - $P(\text{nonrelevant}) = (Nx0/N + N0x/N)/2 = (Nx0 + N0x)/(N + N)$
  - $P(\text{relevant}) = (Nx1/N + N1x/N)/2 = (Nx1 + N1x)/(N + N)$
- Probability that the two judges agreed by chance
  - $P(E) = P(\text{nonrelevant})^2 + P(\text{relevant})^2$