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.

The standard approach to information retrieval system evaluation revolves

RELEVANCE 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

GOLD STANDARD the gold standard or ground truth judgment of relevance. The test document

GROUND TRUTH 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 reduc-

ing 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 be-

cause 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.

But, nevertheless, an information need is present. If a user types python into a

web search engine, they might be wanting to know where they can purchase

a pet python. Or they might be wanting information on the programming

language Python. From a one word query, it is very difficult for a system to

know what the information need is. But, nevertheless, the user has one, and

can judge the returned results on the basis of their relevance to it. To evalu-

ate a system, we require an overt expression of an information need, which

can be used for judging returned documents as relevant or nonrelevant. At

this point, we make a simplification: relevance can reasonably be thought

of as a scale, with some documents highly relevant and others marginally

so. But for the moment, we will use just a binary decision of relevance. We

Online edition (c)

2009 Cambridge UP

8.2 Standard test collections 153

discuss the reasons for using binary relevance judgments and alternatives in

Section 8.5.1.

Many systems contain various weights (often known as parameters) that

can be adjusted to tune system performance. It is wrong to report results on

a test collection which were obtained by tuning these parameters to maxi-

mize performance on that collection. That is because such tuning overstates

the expected performance of the system, because the weights will be set to

maximize performance on one particular set of queries rather than for a ran-

dom sample of queries. In such cases, the correct procedure is to have one

DEVELOPMENT TEST or more development test collections, and to tune the parameters on the devel-

COLLECTION opment test collection. The tester then runs the system with those weights

on the test collection and reports the results on that collection as an unbiased

estimate of performance.