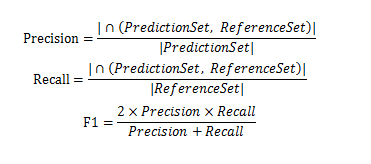
1、**[准确率（Precision）、召回率（Recall）以及综合评价指标（F1-Measure ）](https://www.cnblogs.com/bluepoint2009/archive/2012/09/18/precision-recall-f_measures.html)**

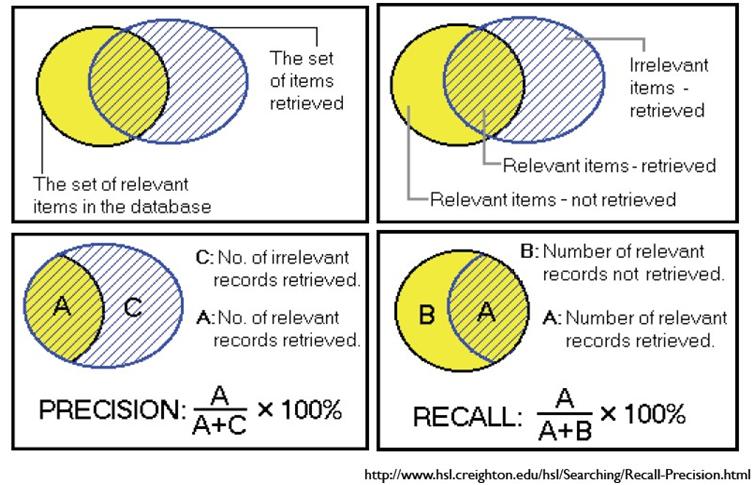
比赛采用经典的精确度(precision)、召回率(recall)和F1值作为评估指标。具体计算公式如下：



在信息检索和自然语言处理中经常会使用这些参数，下面简单介绍如下：

**准确率与召回率（Precision & Recall）**

我们先看下面这张图来加深对概念的理解，然后再具体分析。其中，用P代表Precision，R代表Recall



一般来说，Precision 就是检索出来的条目中（比如：文档、网页等）有多少是准确的，Recall就是所有准确的条目有多少被检索出来了。

下面这张表介绍了True Positive，False Negative等常见的概念，P和R也往往和它们联系起来。

|  |  |  |
| --- | --- | --- |
|  | **Relevant** | **NonRelevant** |
| **Retrieved** | true positives （tp） | false positives（fp） |
| **Not Retrieved** | false negatives（fn） | true negatives （tn） |

那么，

 P=\frac{tp} {tp+fp} \hfill (1) 

 R=\frac{tp} {tp+fn} \hfill (2) 

我们当然希望检索的结果P越高越好，R也越高越好，但事实上这两者在**某些情况下是矛盾的**。比如极端情况下，我们只搜出了一个结果，且是准确的，那么P就是100%，但是R就很低；而如果我们把所有结果都返回，那么必然R是100%，但是P很低。

因此在不同的场合中需要自己判断希望P比较高还是R比较高。如果是做实验研究，可以绘制Precision-Recall曲线来帮助分析。

**F1-Measure**

前面已经讲了，P和R指标有的时候是矛盾的，那么有没有办法综合考虑他们呢？我想方法肯定是有很多的，最常见的方法应该就是**F-Measure**了，有些地方也叫做F-Score，其实都是一样的。

F-Measure是Precision和Recall加权调和平均：

 F = \frac{(a^2+1)P*R} {a^2(P+R)} \hfill (3) 

当参数a=1时，就是最常见的F1了：

 F1 = \frac{2PR} {P+R} \hfill (4) 

很容易理解，F1综合了P和R的结果，当F1较高时则比较说明实验方法比较理想。