Content-based retrieval of visual information

Oerlemans, A.A.J.

Chapter 4

Performance Evaluation

For testing and comparing the effectiveness of retrieval and classification methods, ways of evaluating the performance are required. This chapter discusses several of these methods, such as precision, recall, precision-recall graphs and average precision. Note that we use the term 'documents' in the descriptions because most of these methods were originally designed for evaluating text search engines, but in evaluating the performance of CBIR systems, 'documents' can be directly translated to 'images'.

4.1 Precision

In a retrieval task, precision is defined as the number of relevant documents retrieved as a fraction of the total number of documents retrieved:

$$precision = \frac{\#_{retrieved\ and\ relevant}}{\#_{retrieved}} \tag{4.1}$$

Precision values can be between 0.0 and 1.0. As an example, suppose a query returns 10 search results and 4 of these results are relevant for the user, then the precision of this result set is said to be 0.4. Note that a precision value always needs the number of documents in the result set to be meaningful in comparisons. A precision of 0.5 when returning just two documents gives a different perspective than a precision of 0.5 when the retrieval system returns 100 documents.

For a classification task, the precision is defined as:

$$precision = \frac{\#_{true\ positives}}{\#_{true\ positives} + \#_{false\ positives}} \tag{4.2}$$

This states that the precision of a classifier with respect to positive classification is the fraction of correctly classified positives in the total number of positively classified documents.

4.2 Recall

Recall is another measure of the effectiveness of a retrieval method. Recall is defined as the number of relevant documents retrieved as a fraction of the total number of relevant documents that are in the database:

$$recall = \frac{\#_{retrieved\ and\ relevant}}{\#_{relevant\ in\ database}} \tag{4.3}$$

Recall values have the same range as precision values, between 0.0 and 1.0. As an example, if a retrieval result set with 10 documents contains 4 relevant documents and there are 40 relevant documents in the database, the recall value for this result set is 0.1. Note that the recall value also needs to be accompanied by the number of documents that were retrieved. Also note that the recall value will always be 1.0 if the entire database is returned as a result set in response to a query.

In a classification task, the definition of recall is given by:

$$recall = \frac{\#_{true\ positives}}{\#_{true\ positives} + \#_{false\ negatives}} \tag{4.4}$$

This formula tells us the fraction of positively classified documents with respect to all positives that are in the data set.

4.3 Precision-Recall graphs

To give a graphical impression of the performance of a retrieval method, precision-recall graphs can be created. (Some researchers refer to them as recall-precision graphs, probably a more correct naming, but the term precision-recall is most widely used.) To generate such a graph, a single query is repeatedly executed and the number of returned results is varied. For each of these results sets, the precision and recall are determined and both these values are plotted as a single coordinate in the graph. The shape of the resulting graph gives a quick indication of the performance of a retrieval method. Note that for very low recall values, it is often not very elegant to plot the corresponding precision values, since the result sets used are probably very small and the values can fluctuate rapidly. It is common to start the graph at a recall value of 0.1.

First, the two extreme shapes of a precision-recall graph will be discussed. The ideal shape of a precision-recall graph would be the situation where all returned

documents are always relevant. For each recall value, the precision would always be 1.0. Only if no more relevant documents can be found, the search results will contain irrelevant documents. Note that the graph has only been plotted up to the first recall value of 1.0.

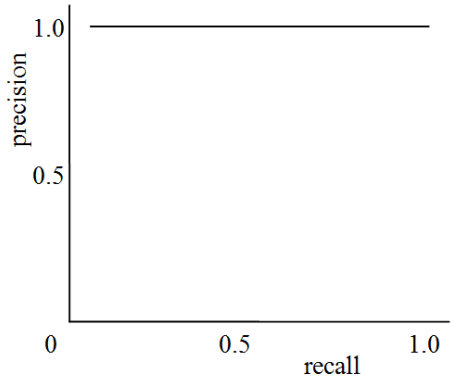

Figure 4.1: The optimal precision-recall graph, with every precision value at 1.0.

The worst case scenario would be when all relevant results only show up after all irrelevant documents have been returned. In this case, when recall values increase from 0.0 to 1.0, the precision would increase slowly from 0.0 to a value specific for the database. (Note that we can plot the coordinate 0.0, 0.0 now.) Assuming that the database contains 100 documents and 10 relevant documents, the final precision would be 0.1 for a recall value of 1.0.

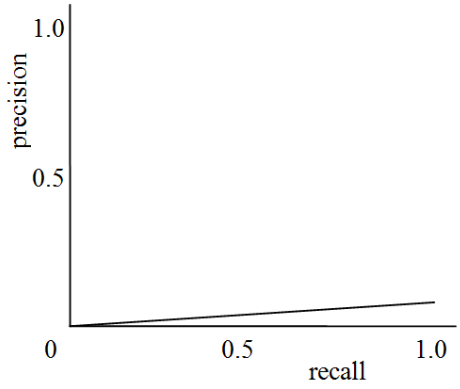

Figure 4.2: The worst case scenario for a precision-recall graph: all relevant documents are ranked lowest.

Some examples of precision-recall graphs are given below, together with an explanation on how to read these graphs.

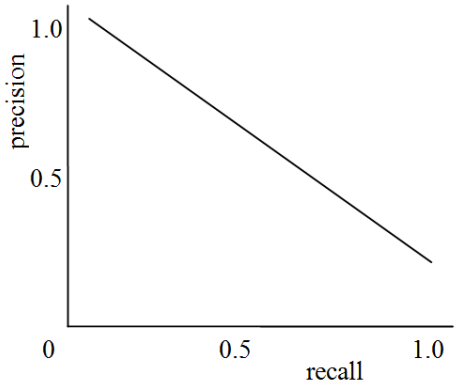

Figure 4.3: A linear relation between recall and precision.

This graph shows a linear relation between precision and recall. For this graph, with increasing recall, precision decreases until the point where all relevant documents are retrieved. This graph tells us that for any given number of results, the percentage of relevant documents has an inverse linear relation with the number of retrieved documents.

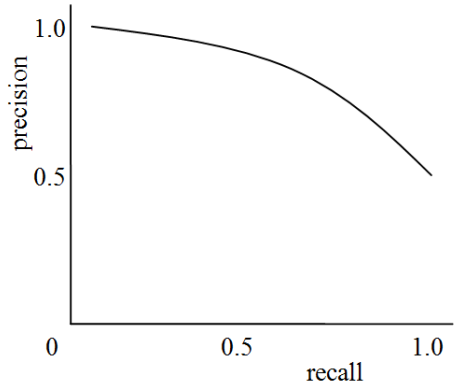

Figure 4.4: A precision-recall graph that indicates high retrieval precision.

This graph shows higher precision values for lower recall values, compared to the first graph. This tells us that for short result sets, the precision is very high, but as the number of retrieved documents increases, the precision decreases. In other words, this probably means that most of the relevant are results are returned in

the top of the result set, but some of the relevant documents are not detected by the retrieval method and are mixed with the rest of the results.

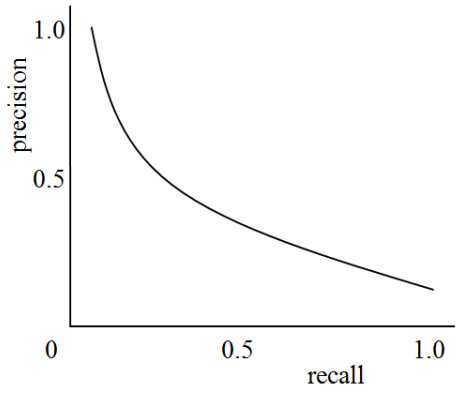

Figure 4.5: A precision-recall graph that indicates a rapidly decreasing precision with increasing result set sizes.

This graph shows a lower precision at low recall levels than the first graph. This translates to the notion that the percentage of relevant documents in the search results will decrease sharply when the number of search results is increased. In other words, many irrelevant documents show up in the top results.

4.4 Average precision

Another method for measuring the performance of a retrieval method is the average precision. This is a value that does not need a fixed length of the result set to be usable in comparisons. The average precision is calculated by averaging the precision values at each relevant document in the result set, usually up to the point where recall is 1.0.

Assume the last relevant document is retrieved at position N in the result set and that the function *relevant* returns 1 when a document is relevant and that *precision* returns the precision of the result set up to a certain point. Then the formula for the average precision can be given as:

$$average\ precision = \frac{\sum_{i=1}^{N} relevant(i) precision(i)}{\sum_{i=1}^{N} relevant(i)} \tag{4.5}$$

With this value, the overall performance of a retrieval method can be assessed with one number, without the need for a graph or a fixed number of returned documents.

4.5 Accuracy

In classification tasks, another measure is available to determine the performance of the classifier: the accuracy. It is defined as

$$accuracy = \frac{\#_{true\ positives} + \#_{true\ negatives}}{\#_{true\ positives} + \#_{true\ negatives} + \#_{false\ positives} + \#_{false\ negatives}} \tag{4.6}$$

This yields the fraction of correctly classified documents with respect to the total number of documents. It can be seen as the probability that a classification, either positive or negative, is correct.