Measuring the Effectiveness of Retrieval Systems that Create Rankings

B. J. Briedis
School of Computer Science & Engineering
The University of New South Wales
Sydney NSW 2052 Australia

and

T. D. Gedeon
School of Computer Science & Engineering
The University of New South Wales
Sydney NSW 2052 Australia

ABSTRACT

The method used to measure the success of a retrieval system is of primary importance. The nature of retrieval engines has been evolving, with very large retrieval engines becoming common, and with the use of ranking becoming widespread. In particular many search engines on the World Wide Web return huge numbers of documents and it is left solely to the user to decide when to stop searching through a list of retrieved items. This paper presents a new measure suitable for testing search algorithms that provide documents in this fashion. The measure is a single figure that is easy to use when comparing different systems.

Keywords: Retrieval systems, ranked output, effectiveness measures.

1. INTRODUCTION

The foundations of information retrieval were laid by researchers in the 1950s and 1960s. These researchers not only devised retrieval methods that are commonly used today, they also formulated ways of testing the success of these retrieval methods. Although most modern retrieval systems rely largely on techniques devised during this time, modern retrieval systems have some important differences. Document collection sizes have grown, and the advent of the World Wide Web has brought about some huge collections. In addition the system of ranking retrieved items has become very popular. The choice of when to stop looking through a list of retrieved items is now generally left to the user. In the past, processing constraints often imposed an artificial limit on the number of documents returned and systems imposed their own cut-off points. Given these changes, different measures are now appropriate for the evaluation of retrieval systems.

A further reason for a new measure is that it is desirable to have a single figure to represent the effectiveness of a retrieval system. Much of the evaluation of retrieval systems currently relies on the use of a number of different statistics and graphs. By contrast, a single figure allows for the quick and simple evaluation of a retrieval system. This is particularly important when it is necessary to set parameters for a retrieval system or when the retrieval system incorporates a training algorithm. It is, for example, necessary with the LSI technique to decide on the number of dimensions the document representations are to have [4]. If a neural network is used, it may be necessary to decide when to stop training, when to add nodes or when to prune the network [2]. In the case of a genetic algorithm a single value of success is referred to continuously throughout evolution (e.g. [8]).

The traditional indicators of retrieval effectiveness are precision and recall. In order to calculate precision and recall it is necessary to decide which documents are relevant and which are not. In the case of recall, a relevance decision is required for every document in the collection for a number of standard queries. This is a very demanding requirement, and in some cases it may be better to use an alternative evaluation method, such as that proposed by Frei and Schäuble [5]. Nonetheless, having a complete set of relevancy judgments is very convenient: it allows for the consistent benchmarking of techniques and the benchmarking may be done quickly and repeatedly, without the need for an assessor to make subsequent relevance judgments, as is necessary with the

Frei and Schäuble method. Several standard test sets are already in existence.

2. LIMITATIONS OF PRECISION AND RECALL

The precision of a retrieval system is usually measured at a number of different recall values. A common way to represent this information is to plot precision values against recall values, as in Figure 1. It is well known that precision and recall are roughly inversely proportional. Retrieval algorithms are compared to one another by comparing their respective curves for the same document collection. The curve that is positioned further from the origin is deemed more effective.

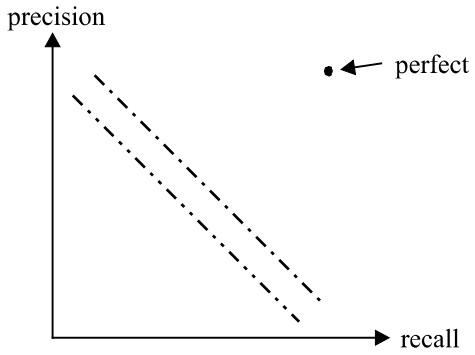

Figure 1: The precision-recall curve

The major problem with judging retrieval effectiveness in this way is that in order to decide which retrieval strategy works best it is necessary to somehow interpret the graph. It is difficult to automate this process, and it is not obvious how to interpret the graph when the curves cross. It is possible to pick one point on the graph as being the deciding statistic, but as different searches are of different lengths, there is no one level of recall that is a true reflection of usage.

While the recall axis appears to represent the rigour with which a user pursues a search, it does not do so directly. The number of documents that must be searched through in order to move a set distance along the recall axis varies not only between document collections, but even within the one collection. This complicates the process of interpreting the graph.

A further problem with the precision-recall curve as a measure of retrieval effectiveness is that users do not conduct searches at all levels of recall to the same degree. For instance, in very large retrieval systems such as WWW search engines it is often highly unrealistic to assume that there will be high recall, and consequently much of the curve is largely irrelevant. In other systems, such as various legal databases, a high level of recall may be sought [1, 13]. The reading of the graph must be varied according to the nature of the document collection.

In order to overcome some of the shortfalls of the recall-precision curve a number of statistics have been used, some of them seeking to summarize the curve itself. One commonly used measure is the mean of the precision taken at some number of evenly spread recall values (e.g. see [12]). It has also been suggested that the median of these values might give better results [11]. A problem with both these measures is that all recall values are treated as being of equal importance — an invalid assumption.

Other related measures that have been used are the harmonic mean of precision and recall [12], normalised recall, normalised precision, rank recall and log precision [9, 10]. These measures are all dependent on precision and/or recall. This strong reliance on precision and recall is surprising as they are themselves composite measures. The abundance of (quite different) measures based upon them suggests that they are not well suited to summarizing the effectiveness of retrieval systems. Sometimes several of these measures are used in the same paper (e.g. [9]), apparently because none of the measures in isolation is capable of adequately encapsulating all of the information required. A number of other evaluation strategies are summarized in van Rijsbergen [12].

It has occasionally been suggested that, rather than using precision and recall, an attempt should be made to model the situation the user is in when using a retrieval system [3, 5]. One simple move in this direction is to plot the number of relevant documents retrieved against the number of documents looked at by the user (e.g. see [13]). A problem with this approach is that different users view different numbers of documents. It is thus not possible to select any one point on the curve as being a good summary of retrieval performance.

3. A MEASURE OF EFFECTIVENESS

For the purpose of developing a new measure of effectiveness, it is assumed that there are a number of standard test queries available with matching relevancy judgments. The relevancy judgments may be binary (relevant or not relevant) or fuzzy (e.g. 70% relevant).

Another possible method for describing the relevance of documents is to have the assessor list those documents

that are relevant in the order of their relevance. This apparently results in more consistent assessments [5, 3]. In order for our measure to be used with relevance judgments expressed as a ranking it is necessary first to convert the ranking to fuzzy relevance judgments. This could be done by having the assessor group documents within the ranking and then assign to each group a fuzzy relevance. The process of assigning numbers to each group is somewhat problematic, but there are a few guidelines. Obviously 1 should be assigned to documents that are certainly relevant and 0 to those that are not. For the remainder, the proportionality of the judgments needs to be borne in mind. Retrieving one document of 0.8 should be equivalent to retrieving two documents of 0.4.

Instead of using relevance as a measure it may be better to have the assessor design the sort of ranking they would most like the retrieval system to produce. One likely effect of this would be for the documents to be rated according to *utility* rather than *relevance*, which has been suggested as being preferable [3]. It also allows an assessor to make judgments about the rankings themselves, not just the documents that they contain. An assessor might, for instance, decide that it is important to have a mix of different types of document in the top part of the ranking. This desire may be expressed using a ranking, and the ranking may then be converted into fuzzy judgments.

Search success

One assumption made in designing the following measure is that finding 1 relevant document is half as desirable as finding 2 relevant documents, one third as desirable as finding 3 relevant documents, and so on. Let F_q be the set of relevant documents that is found by a user after making a query (q). We define the *success* of a search (S_q) to be:

$$S_q = |F_q| \tag{1}$$

That is, S_q is the number of relevant documents found by the user. If fuzzy relevancy (or utility) judgments are used then,

$$S_q = \sum_{i \in F_q} J(i) \tag{2}$$

where $J(i)$ is the fuzzy relevance or utility judgment of document i and $0 < J(i) \leq 1$. For later purposes, let J be non-increasing.

In practice it is unknown how many documents a user will look at. A probability function may, however, be used to describe the likelihood of documents being studied. Let $P(x)$ be the estimated probability of retrieving document x and A_q be the set of all documents that are relevant, or partly relevant, to a query. The success of a search may now be redefined to be:

$$S_q = \sum_{i \in A_q} P(r_i) \times J(i) \tag{3}$$

where r_i is the ranking of document i and $J(i) = 1$ if binary judgments are used or $0 < J(i) \leq 1$ otherwise.

If high recall searches are rarely made by users, but it is nonetheless important that these searches when made are of reasonable quality, an additional cost function may be factored in as follows:

$$S_q = \sum_{i \in A_q} P(r_i) \times J(i) \times C(r_i) \tag{4}$$

where $C(r_i)$ is the cost of a relevant document appearing at position r_i in the ranking. Such a cost function would be increasing but would otherwise probably be quite arbitrary. For the rest of this paper it is assumed that $C(r_i) = 1$ for all i in A_q.

The success value may be normalized to give a value between 0 and 1. One method of performing the normalization is to divide by the best possible ranking:

$$N_q = S_q / S_q^* \tag{5}$$

where N_q is the normalised result and S_q^* is the maximum possible value of S_q. If binary relevance judgments are used:

$$S_q^* = P(1) + P(2) + \ldots + P(|A_q|) \tag{6}$$

For binary judgments it is possible to approximate S_q^* using $|A_q|$, the number of relevant documents. This provides a somewhat simpler equation but means the optimal result will usually be less than 1. Note that if $|A_q|$ is used as the denominator and the judgments are binary, the measure is the mean average of the probabilities of finding relevant documents. For fuzzy relevance judgments:

$$S_q^* = P(1) \times J(1) + P(2) \times J(2) + \ldots + P(n) * J(n) \tag{7}$$

where J is non-increasing and $n = |A_q|$.

The effectiveness of the retrieval system may be taken as the mean average of the set of normalised search success values,

$$M = \frac{1}{k} \sum_{i=1..k} N_i \qquad (8)$$

where k is number of queries, N_i is the normalised success value of query i, and M is the overall effectiveness of the retrieval system.

Distribution of cut-offs

One of the major decisions that must be made is that of which probability distribution of cut-offs to use. It is possible to estimate the actual probability curve by observing users making queries. This should not be difficult to do on retrieval systems that return items in blocks of 10 or so at a time. The probability curve would doubtless vary if the sample was restricted so as to cover only a specific search engine, user or type of search. Thus if it is convenient to do so, different probability curves may be obtained for use in different circumstances.

Despite the likely differences in probability distribution, there are several factors in common between searches on different systems. Users normally start to work their way through retrieved items at the top of the ranking list and abandon the list at some point. The major limiting factor is the amount of time the user wishes to spend searching. Thus the rate at which the probability drops is largely independent of the collection size. Relevant documents near the top of the ranking will almost certainly be found and thus have a probability of discovery of close to 1. Documents at the end of a long listing are almost certain to be missed, and thus have a probability of almost 0.

A number of standard distributions satisfy these requirements. One reasonable distribution to choose is the right half of the normal distribution which has been stretched vertically so that $P(0) = 1$. The equation of this curve is:

$$P(x \mid \sigma^2) = e^{\frac{-x^2}{2\sigma^2}} \qquad (9)$$

Having chosen this distribution, it is still necessary to set the variance. This may be done by sampling a given system, should the normal distribution be found to be a reasonable reflection of reality. A system in which high-recall searches are frequent would have a greater variance than a system which had few high-recall searches. An intuitive way of deciding on the variance is to set one point on the curve — say the ranking at which a relevant document has a 0.5 probability of being found. If x' is this ranking,

$$\sigma^2 = \frac{-x'^2}{2\ln(0.5)} \qquad (10)$$

It is possible in this way to estimate a reasonable value for a given system if testing is not feasible. Note that the curve is somewhat forgiving as there is a gradual transition from found to not found. It might be expected that x' is likely to carry over between systems of a similar nature — thus avoiding the need to run tests on every system.

Test collection results

The effectiveness of a simple retrieval algorithm has been calculated for a number of standard test collections and the results are given in Figure 2. The vector retrieval algorithm suggested by Salton [10] was used, with stemming being done using the Porter algorithm [7] and stop words being removed. These results may be useful to other researchers as a guide as to what constitutes reasonable scores on these test collections. If the case where one particular *retrieval system* is being evaluated then there exists an ideal value of x', and the retrieval effectiveness should be taken as close to this point as possible. If a *retrieval algorithm* is being evaluated, a range of x' values is significant.

As can be seen from Figure 2 the effectiveness score is dependent on the collection being tested. The test collection properties that are most likely to affect these scores are the collection size and the number of documents relevant to each query (see Table 1). The *cisi* collection, for example, has the largest number of documents relevant to each query and has the lowest effectiveness score whereas the *time* collection has the lowest number of relevant documents and one of the highest scores. The importance of the collection size may be seen in the figures for the *npl* and *adi* collections. The large *npl* collection has the second lowest set of scores while the very small *adi* collection has the highest score for many values of x'. The dependence of the measure on the collection size is not surprising as increasing the number of documents makes the task of discrimination harder and so causes the number of false hits to increase. The initial drop in effectiveness that occurs in the *cisi*

test set	collection size	Relevant documents per query (median)
adi	82	4
time	477	2
med	1033	22.5
cran	1400	7
cisi	1460	30.5
npl	11429	19

Table 1: Properties of the test collections

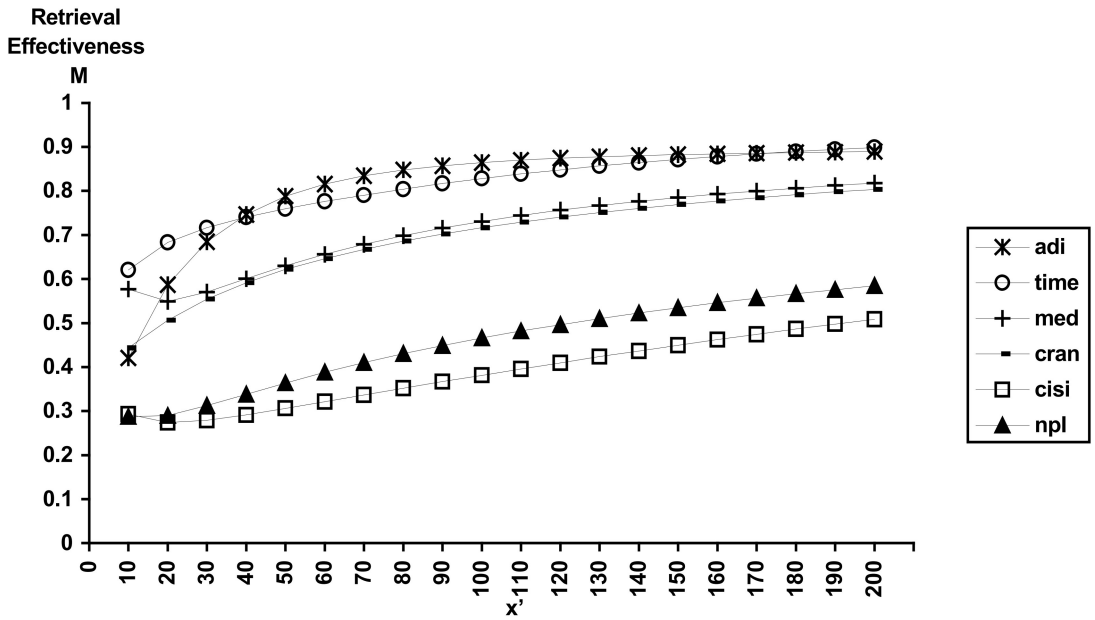

Figure 2: The retrieval effectiveness of various standard test sets[1]

and *med* collections is due to S_q^*, which increases quickly while x' is small when there are many relevant documents.

The differences in the values taken by the measure for different test collections may be reduced by adjusting the measure to take into account the effect of the most significant properties of the test collections. It is unlikely, however, that the measure could be altered to the point where a value of, say, 0.6 would indicate a successful retrieval algorithm, regardless of the collection being tested. It would still remain necessary to interpret the effectiveness score in light of the particular collection. While adjusting the measure has the appeal of making it easier to compare the performance of algorithms between collections, doing so distorts the measure so that it no longer truly represents the degree of success users experience. Properties of the collection, such as its size and the number of documents that are relevant to queries, affect the success users have and it is reasonable that this should be reflected in the operation of the measure.

4. CONCLUSION

The proposed measure yields a single value which allows for the easy evaluation of retrieval systems that return rankings of documents. This makes it particularly useful for setting parameters and for use with various learning algorithms.

Unlike most previous measures it takes into account the diminishing importance of items that are lower in the ranking. It does not assume that there is any one cut-off point, but rather that there is a mix of cut-off points which are described using a probability function. With an appropriate choice of probability function and/or halfway point, the measure should discriminate neither in favour of nor against recall. The measure can accommodate fuzzy relevance or utility judgments and also allows judgments to be expressed as rankings.

5. ACKNOWLEDGMENTS

This research has been funded in part by the Australian Research Council.

6. REFERENCES

[1] D. C. Blair and M. E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System", Communications of the ACM, Vol. 28, No. 3, pp. 289-299, 1985.

[2] B. J. Briedis and T. D. Gedeon, "Using the Grow-and-Prune Network to Solve Problems of Large Dimensionality", Proceedings of the 1998 Australian Conference on Neural Networks, Brisbane, 1998.

[3] C. W. Cleverdon, "User Evaluation of Information Retrieval Systems", Journal of Documentation, Vol. 30, No. 2, 1974, pp. 170-180.

[1] Some of the queries in the *cisi* collection have no documents that are relevant to them. These queries were excluded from testing.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, “Indexing by Latent Semantic Analysis”, Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391-407, 1990.

[5] H. P. Frei and P. Schäuble, “Determining the Effectiveness of Retrieval Algorithms”, Information Processing & Management, Vol. 27, No. 2/3, pp. 153-164, 1991.

[6] M. E. Lesk and G. Salton, “Relevance Assessments and Retrieval System Evaluation”, Information Storage and Retrieval, Vol. 4, pp. 343-359, 1969.

[7] M. F. Porter, “An Algorithm for Suffix Stripping”, Program, Vol. 14, No. 3, pp. 130-137, 1980.

[8] A. M. Robertson and P. Willett, “An Upperbound to the Performance of Ranked-Output Searching: Optimal Weighting of Query Terms using a Genetic Algorithm”, Journal of Documentation, Vol. 52, No. 4, pp. 405-420, 1996.

[9] G. Salton and M. E. Lesk, “Computer Evaluation of Indexing and Text Processing”, Journal of the Association for Computing Machinery, Vol. 15, No. 1, pp. 8-36, 1968.

[10] G. Salton, Introduction to Modern Information Retrieval, New York: McGraw-Hill, 1983.

[11] J. Savoy, “Statistical Inference in Retrieval Effectiveness Evaluation”, Information Processing & Management, Vol. 33, No. 4, pp. 495-512, 1997.

[12] C. J. van Rijsbergen, Information Retrieval, London: Butterworths, 1979.

[13] P. Wallis and J. A. Thom, “Relevance Judgments for Accessing Recall”, Information Processing & Management, Vol. 32, No. 3, pp. 273-286, 1996.