

Estimating the Number of Relevant Documents in Enormous Collections

Paul Kantor, Myung-Ho Kim, Ulukbek Ibraev, Koray Atasoy

Rutgers University

4 Huntington St, New Brunswick, NJ 08903

kantor@scils.rutgers.edu <http://aplab.rutgers.edu/ant/>

Abstract

In assessing information retrieval systems, it is important to know not only the precision of the retrieved set, but also to compare the number of retrieved relevant items to the total number of relevant items. For large collections, such as the TREC test collections, or the World Wide Web, it is not possible to enumerate the entire set of relevant documents. If the retrieved documents are evaluated, a variant of the statistical “capture-recapture” method can be used to estimate the total number of relevant documents, providing the several retrieval systems used are sufficiently independent. We show that the underlying signal detection model supporting such an analysis can be extended in two ways. First, assuming that there are two distinct performance characteristics (corresponding to the chance of retrieving a relevant, and retrieving a given non-relevant document), we show that if there are three or more independent systems available it is possible to estimate the number of relevant documents without actually having to decide whether each individual document is relevant. We report applications of this 3-system method to the TREC data, leading to the conclusion that the independence assumptions are not satisfied. We then extend the model to a multi-system, multi-problem model, and show that it is possible to include statistical dependencies of all orders in the model, and determine the number of relevant documents for each of the problems in the set. Application to the TREC setting will be presented.

INTRODUCTION

It has long been recognized (Salton and McGill, 1983) that when estimating the effectiveness of an Information Retrieval System, one ought to take into account some indication of how easy the retrieval problem is. The standard measure of retrieval performance today is a curve which plots the precision of a retrieved set as a function of the recall of that retrieved set. This measure applies easily to modern retrieval systems which rank documents for retrieval. For any length of the ranked set, r , we can compute the number of relevant documents that have been retrieved, $g(r)$. The precision is defined as $g(r)/r$, and the recall is defined as $g(r)/\mathbf{G}$. In this equation \mathbf{G} represents the total number of relevant documents for the given problem in the entire collection available for retrieval. It serves as an indication of the ease or difficulty of the problem in the sense that large \mathbf{G} corresponds to a problem that is in some sense easier, and low \mathbf{G} represents a problem that is in some sense difficult. If a problem is easy achieving high precision, particularly at the early ranks in a retrieval list, is not particularly impressive, as there are a great many relevant documents and presumably some of them are quite easy to find.

In the most complete and authoritative evaluation framework existing today, the TREC Conferences (Harman 1997), the problem of estimating \mathbf{G} is finessed by using a so-called pooled estimate of the number of relevant documents. In particular each of some 20 to 70 competing schemes presents a list of 1000 documents in

decreasing order of their estimated relevance to the particular problem or topic at hand. The top 100 documents from each list are placed in an evaluation pool, and each unique document in that pool (many documents of course appear on multiple lists) is evaluated by a single expert for its relevance to the topic. The total number of documents judged to be relevant in this process is used as the estimate of the true number of relevant documents \mathbf{G}_{true} . In what follows, to be more precise, we will refer to it as $\mathbf{G}_{\text{judged}}$. It is obvious that (barring some miraculous accident) the true number of relevant documents which we represent by \mathbf{G}_{true} , is necessarily larger than this number. This paper is a progress report on a variety of approaches that are being explored to see whether we can make some estimate of \mathbf{G}_{true} , based on the fact that we have information from a large number of more or less distinct retrieval systems all dealing with the same problem.

We can explain the reasons for optimism by considering the so-called “capture-recapture” method that is used in environmental management. To estimate the number of wolves in a forest, one captures a certain number N_1 of wolves (presumably at random) and tags them. Sometime later (but not too long afterward) one captures another random set of N_2 wolves and determines how many of them, N_{12} , have tags. The estimate of the total population is then given $N_1 N_2 / N_{12}$. This follows immediately from the observation that the probability that a wolf is captured twice is given by the product $d_1 d_2$, of the individual probabilities that the wolf will be captured in each of the two trapping sessions.

A method something like this was apparently used in an effort to estimate the size of the web (Lawrence and Giles, 1998), although the results of our work indicate that one should be very cautious in believing such an estimate.

Before entering on the technical details we consider a number of ways in which the proposed assumption of independence might fail. Consider first the method of trapping used. Suppose for example the trap is a cage with a door of a certain size which closes automatically when an animal enters it. Even if all wolves are equally likely to be attracted by the cage, the cage will only trap those wolves that are small enough to get into it. Thus there may be some fraction of the wolf population which is not seen in any of our sampling runs and which is completely impervious to estimation by the method proposed.

A second difficulty may be described as follows. Suppose we are catching not wolves, but raccoons, which are known to be quite clever with their hands. And suppose the trap has a door which can be opened by a raccoon of above average intelligence but not by an ordinary raccoon. In that case when we come round to check the traps we will have found a certain fraction of the population and not have found another. If as is more likely, the ability to escape from the trap somehow increases monotonically with raccoon intelligence, we would have to admit that our method is more heavily biased towards an estimate of the number of dull-witted raccoons, than to an estimate of the total number of raccoons.

The analogy in the information retrieval situation is as follows. Documents corresponding to the wolves that don't fit into the trap are documents whose description in some way carries none of the clues that any of the retrieval systems now in use rely upon for the identification of relevant documents. As all current systems in the TREC setting base their retrieval on the texts of the documents, an obvious example (in fact, excluded from the TREC competition) would be a news article whose accompanying photograph contained useful information about a problem. Since the photograph is not itself represented to any of the systems, it cannot possibly form the basis for the retrieval. The meaning of the second notion, the dull-witted raccoons, is a little bit harder to articulate. It corresponds to the notion that in some sense some of the relevant documents are much easier to retrieve than others. We do not have a precise model for this, but it might, for example, occur when a document contains many synonyms for each of its important concepts. Thus, no matter which of those synonyms was used in a query formulated by a particular system, the document would have a chance of being retrieved. In what follows we will represent this notion instead empirically, considering that a document which is relevant is easy to retrieve if it is retrieved by a large number of the systems in a particular TREC setting.

A SURPRISING MATHEMATICAL RESULT

It would seem, from the preceding discussion, that estimate of the total number of relevant documents in a collection, even if the systems are independent in precisely the way required by the analysis, would nonetheless require an examination of retrieved documents to see whether they are relevant. It's rather the same as having to look in the cage and see whether what we've caught is a wolf or some other kind of an animal.

However, in the TREC setting (this is of course not true of a setting such as the World Wide Web) we actually know the number of animals in the forest to begin with. If we assume that every animal either is a wolf or is not, that is that every document either is relevant or is not, then given a minimum of three retrieval systems, all of which trap relevant documents independently from each other, we are able to develop a set of coupled equations that permit determination of the total number of relevant documents \mathbf{G}_{true} . What is remarkable is that these equations, shown below, do not require us to determine whether specific retrieved documents are relevant or not, but only require us to count the number of documents retrieved by each system, and to count the overlaps of the retrieved sets. What results is a set of equations with eight unknowns: the number of relevant documents, the number of documents that are not relevant, and two performance parameters characterizing each of the three systems.

The first performance parameter, called the detection power, is interpreted as the probability that a random relevant document will be in the set retrieved by that system. The second parameter characterizing each system, called the false alarm rate, is the probability that a document which is not relevant will be in the retrieved set given by that system. In the particular case of the TREC setting we have sets of size 1000 retrieved by each of the three systems, and we know the total size of the collection, so the only new quantities to be determined are the pair wise and three-fold overlap of the retrieved sets. Together, these provide eight equations for eight unknowns, which are highly non-linear.

FEASIBILITY OF SOLVING THE NON-LINEAR EQUATIONS

In a set of simulation runs, we have explored the solvability of these equations using the solver tool built into the Excel spreadsheet. To do this we calculated the total squared deviation between the left hand sides of the equations and the right hand side and sought to minimize this as a function of the eight parameters. In a wide range of trial situations, where the right hand sides of the equations were generated by direct calculation from assumed values of the system parameters, we were able to reproduce with high precision the input values. In some additional studies, we were able to show that these are fairly robust to the addition of noise in the right hand side of the equation at the level of a few percent.

Thus, although the equations are not solvable analytically, there should be no difficulty in solving them in real situations.

APPLICATION OF THE EQUATIONS TO THE TREC-6 ROUTING DATA

We then turned to the TREC-6 data, and computed the actual overlaps of the retrieved sets and the corresponding sets of equations. There were 31 systems which participated in this particular TREC task, and we broke them into 10 groups of 3 systems each. For any particular topic, if the model is correct, the computations of \mathbf{G} resulting from each such triple should be close to each other and should form an estimate of the real value of the number of relevant documents. This work produced estimates of \mathbf{G} that were obviously incorrect, in that they fall substantially below $\mathbf{G}_{\text{judged}}$, while the true value must fall above. Some

further investigation, using the intersection of the retrieved sets themselves, showed that these intersections are much too large to be the result of stochastically independent retrieval systems. This result set the stage for our further analysis.

OVERLAPS OF SETS OF JUDGED RELEVANT DOCUMENTS

We turned our attention next to the sets of retrieved and judged relevant documents produced by the 31 systems in the TREC-6 routing task. Our goal was to explore more precisely the relations among them to see if we could bring their lack of independence under sufficient mathematical control. In doing this analysis we are able to deal with any pair of systems without having to consider three systems at a time. If two systems retrieve relevant documents in a stochastically independent way, then the size of the overlap between documents retrieved by System i and those retrieved by System j , which we denote by G_{ij} , is given by $G_{ij} = d_i d_j \mathbf{G}$. Since the retrieved sets of relevant documents themselves are given by $G_i = d_i \mathbf{G}$, we can immediately produce, by the capture/recapture argument, an estimate of \mathbf{G} which we represent as G^{ij} . As might be expected, these estimates are also all too small.

These results, while disappointing, are not surprising, because we have good reason to expect that the retrieval judgments made by different systems in the TREC setting have some degree of positive correlation. In this case, the number of documents retrieved by both systems will be larger than the value predicted by the assumption of stochastic independence, and correspondingly, the estimates G^{ij} will be too low. Thus, while the mathematical result involving three independent systems holds out a tantalizing prospect, the conditions necessary for its validity are simply not met in the experimental data.

EXTRAPOLATION MODELS

As another way of attacking the problem of estimating \mathbf{G} , we try to develop a model for the accumulation of relevant documents, as one moves down the list of documents retrieved by a particular system, for a particular topic.

Let $g_i(t, r)$ be the number of relevant documents retrieved by system i , and $g_{ij}(t, r)$ the number of relevant documents retrieved by both system i and j , up to rank r , for a topic t . We define $g^{ij}(t, r)$ by

$$g^{ij}(t, r) = g_i(t, r)g_j(t, r)/g_{ij}(t, r)$$

and let $\hat{g}(t, r)$ be the average of $g_{ij}(t, r)$ over all pairs of systems (i, j) .

We have found, for almost all topics, that we can approximately fit $\hat{g}(t, r)$ by an expression of the form $K(t)(1 - e^{-\lambda(t)r})$, where $K(t)$ and $\lambda(t)$ are certain constants depending on the topic t . In a variant calculation, we used SAS non-linear regression to fit several $g_i(r)$ simultaneously, using a single value of K , presumed to be the asymptotic “true” value of \mathbf{G} , together with topic-dependent values $\lambda(t)$. To avoid excessive computation (and increase the chance that we would approach a reasonable value) we restricted this calculation to a set of 6 systems, by choosing one from each of the highest 6 pairs of systems at TREC6.

Unfortunately, both attacks are thwarted, as the average of the estimates again falls short of even the number of judged relevant documents, and thus can not possibly measure the even larger number of documents that are truly relevant. In retrospect it is obvious that this must fail. The true number of relevant documents is larger than the number retrieved by even the best system, and *ipso facto* larger than the average among systems that are both good and bad.

Figure 1 shows the low values of the asymptotic estimate for \mathbf{G} that result from averaging many systems.

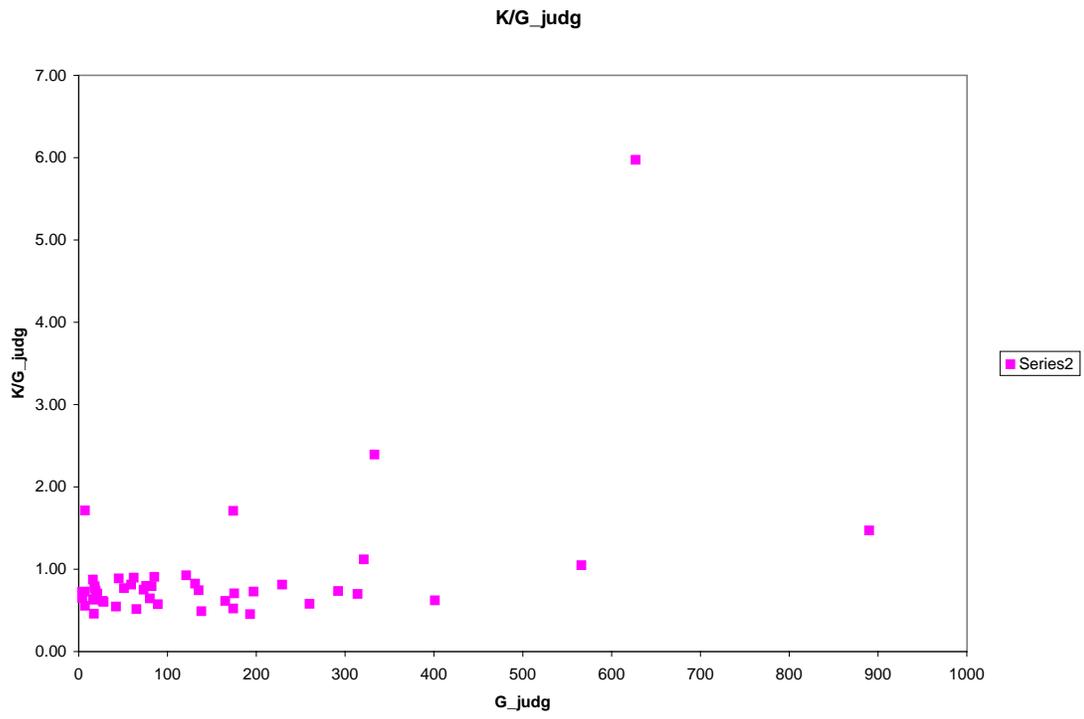


Figure 1: The ratio of K_{ave} to G_{judg} . While there is some scatter, nearly all values are less than 1.

THE UNION SYSTEM

Our final attempt to estimate the number of relevant documents, making use of all the available judgments on documents retrieved by 31 systems participating in the TREC6 routing task, involves forming a hypothetical super-system called the *union* system. The cumulated relevant document function for this system is defined by the following procedure. For each rank r , increase $g_{\text{union}}(r)$ by the number of relevant documents appearing at rank r in any of the lists produced by the 31 retrieval systems, and not having already appeared previously on any of the lists. In contrast to the averaging methods used before, this at least ensures that the curve computed will do better than any specific system and, if TREC judging were to strictly follow the published rules, $g_{\text{union}}(100) = G_{\text{judged}}$. In fact, there are slight discrepancies. However, this method of extrapolation does tend to give estimates that are higher than the number already known to be relevant through human judging.

In Table 1, the Topic ID refers to the TREC classification of topics. $\hat{G}(100)$ is the average of g^{ij} over all pairs of systems, evaluated when each system has retrieved 100 documents. g_{union} is defined in the text. G_{judg} is the total number of documents judged relevant for each topic, as reported in the TREC `qrels` files. K_{ave} is the asymptote of the non-linear model, applied to the function $\hat{G}(r)$, which is the average of the estimates produced by all possible pairs of systems. The fitting is done over the range $r = 1, \dots, 100$. The ratio of K_{ave} to G_{judg} is denoted by K/G . K_{union} is the coefficient of the non-linear regression model, applied now to the cumulated function $g_{\text{union}}(r)$, which includes all of the documents judged relevant by the TREC analysts. The table has been sorted in decreasing order of the ratio of K_{ave} to G_{judg} . If our model were achieving realistic estimates of the true number of relevant documents, \mathbf{G} , this ratio should always be greater than 1. In fact, we see that it is greater than 1 in only 13 cases, and falls as low as 61%.

DISCUSSION

We are in possession of a mathematical tool, the coupled equations for estimating the number of relevant documents, which is very tantalizing, but cannot quite fulfill its promise. The assumptions of independence are simply not met in the available data. This of course raises the possibility that other estimates, such as the estimate of the size of the WWW, are also unreasonably low.

We can see several paths to further exploration. One is to develop alternative models for the dependence of $g(r)$ on r . The exponential model, which is appropriate for infinite collections, may not be the right one to use. An attractive alternative is found in logarithmic models, in which $g(r) = a \ln(r)$, which provides an extrapolation based on the true size of the collection, which is a large number (175,000), but not infinite.

Another path for further exploration is based on attempts to model the lack of stochastic independence among systems. If systems are “similar”, perhaps that similarity is more or less independent of the specific topic. This can be expressed, for example, in a log-linear model, which sets $\ln G_{ij} = \ln \mathbf{G} + \ln d_i + \ln d_j + \chi_{ij}$, permitting d_i to depend on t , while χ_{ij} is independent of t .

ACKNOWLEDGMENTS

This work is supported by the Defense Advanced Research Projects Agency (DARPA) Contract Number N66001-97-C-8537.

Topic ID	$\hat{G}(100)$	$g_{\text{union}}(100)$	G_{judg}	K_{ave}	K/G	K_{union}	$K_{\text{union}}/G_{\text{judg}}$
180	6.7	17	17	7.8	0.46	27	1.588235
202	410.9	534	627	3746.4	5.98	762.7	1.216427
10002	185.6	267	321	359.8	1.12	390	1.214953
111	368.8	480	566	594.6	1.05	674.5	1.191696
161	93.1	118	121	112.1	0.93	131.2	1.084298
100	115.2	179	197	143.7	0.73	210	1.06599
148	151.9	241	260	150.8	0.58	270.6	1.040769
189	646.3	667	890	1309.4	1.47	926.1	1.040562
94	84.9	174	193	87.7	0.45	196	1.015544
58	13.9	18	18	13.9	0.77	18.2	1.011111
125	18.7	27	27	16.6	0.61	27.2	1.007407
62	246.5	368	401	249.8	0.62	402.2	1.002993
44	3.4	4	4	2.9	0.73	4	1
6	89.5	146	165	101.4	0.61	164.9	0.999394
108	201.9	293	314	219.8	0.70	312.1	0.993949
78	41.9	45	45	40	0.89	44.6	0.991111
123	52.6	62	62	55.7	0.90	61.4	0.990323
1	40.3	51	51	39.3	0.77	50.3	0.986275
77	14.2	16	16	14	0.88	15.7	0.98125
10003	55	72	73	55	0.75	71.6	0.980822
173	11.1	16	16	10.1	0.63	15.6	0.975
194	2.6	4	4	2.6	0.65	3.9	0.975
12	179.4	270	292	214.7	0.74	283.5	0.97089
5	5.8	7	7	5.1	0.73	6.7	0.957143
192	6.9	7	7	12	1.71	6.7	0.957143
119	50.2	76	85	77.2	0.91	81.3	0.956471
185	15.1	18	18	14.3	0.79	17.2	0.955556
10001	93.1	127	135	100.7	0.75	128.5	0.951852
82	65.7	80	82	65.2	0.80	77.6	0.946341
126	15.5	19	19	14	0.74	17.9	0.942105
128	362.2	281	333	796.9	2.39	312.5	0.938438
154	132.4	168	175	123.7	0.71	163.8	0.936
114	48	57	59	48	0.81	55.1	0.933898
142	150.6	200	229	186.3	0.81	212.7	0.928821
282	17.8	28	28	16.9	0.60	25.2	0.9
3	61.7	70	76	60.5	0.80	68.2	0.897368
95	63.5	123	138	67.7	0.49	123.8	0.897101
118	55.8	83	89	51.2	0.58	79.2	0.889888
10004	12.5	17	18	11.4	0.63	16	0.888889
11	88.4	150	174	91.2	0.52	153.2	0.88046
240	82.9	121	131	108	0.82	115	0.877863
54	173.4	164	174	297.6	1.71	151.9	0.872989
187	12.7	19	21	14.8	0.70	18.1	0.861905
4	52.2	70	80	51.7	0.65	68.5	0.85625
228	36.8	58	65	33.6	0.52	53.2	0.818462
24	22.7	34	42	23	0.55	30.5	0.72619
23	4	6	7	3.9	0.56	4.3	0.614286

Table 1:

REFERENCES

Lawrence, S. and Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280(5360):98.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.